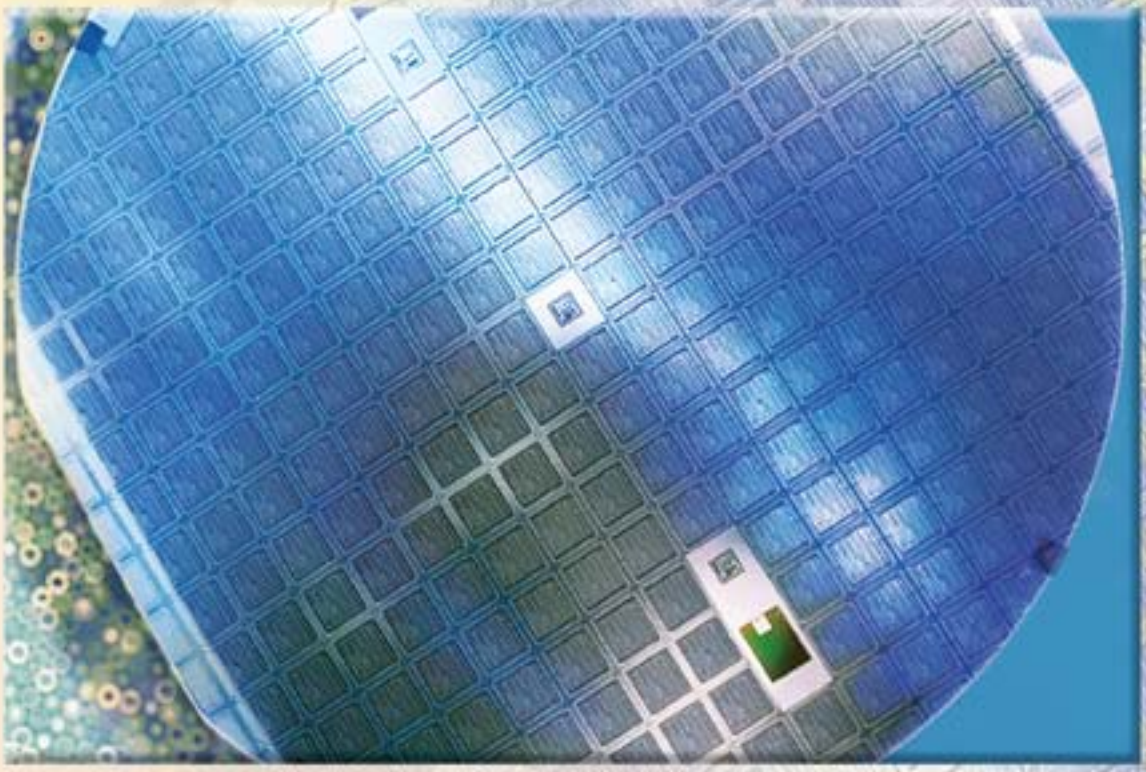


Semiconductor Physics and Devices

Basic Principles

Donald A. Neamen

Fourth Edition

Semiconductor Physics and Devices

Basic Principles

Fourth Edition

Donald A. Neamen
University of New Mexico

The McGraw-Hill Companies

SEMICONDUCTOR PHYSICS & DEVICES: BASIC PRINCIPLES, FOURTH EDITION

This book is printed on acid-free paper.

1 2 3 4 5 6 7 8 9 0 DOC/DOC 1 0 9 8 7 6 5 4 3 2 1

ISBN 978-0-07-352958-5
MHID 0-07-352958-3

Vice President & Editor-in-Chief: *Marty Lange*
Vice President EDP/Central Publishing Services: *Kimberly Meriwether David*
Publisher: *Raghu Srinivasan*
Sponsoring Editor: *Peter E. Massar*
Marketing Manager: *Curt Reynolds*
Development Editor: *Lora Neyens*
Project Manager: *Melissa M. Leick*
Design Coordinator: *Brenda A. Rolwes*
Cover Designer: *Studio Montage, St. Louis, Missouri*
Cover Image: © *Getty Images RF*
Buyer: *Sherry L. Kane*
Media Project Manager: *Balaji Sundararaman*
Compositor: *MPS Limited, a Macmillan Company*
Typeface: *10/12 Times Roman*
Printer: *RR Donnelley, Crawfordsville*

Library of Congress Cataloging-in-Publication Data

Neamen, Donald A.
Semiconductor physics and devices : basic principles / Donald A. Neamen.
— 4th ed.
p. cm.
Includes index.
ISBN 978-0-07-352958-5 (alk. paper)
1. Semiconductors. I. Title.
QC611.N39 2011
537.6'22—dc22

2010045765

www.mhhe.com

ABOUT THE AUTHOR

Donald A. Neamen is a professor emeritus in the Department of Electrical and Computer Engineering at the University of New Mexico where he taught for more than 25 years. He received his Ph.D. from the University of New Mexico and then became an electronics engineer at the Solid State Sciences Laboratory at Hanscom Air Force Base. In 1976, he joined the faculty in the ECE department at the University of New Mexico, where he specialized in teaching semiconductor physics and devices courses and electronic circuits courses. He is still a part-time instructor in the department. He also recently taught for a semester at the University of Michigan-Shanghai Jiao Tong University (UM-SJTU) Joint Institute in Shanghai, China.

In 1980, Professor Neamen received the Outstanding Teacher Award for the University of New Mexico. In 1983 and 1985, he was recognized as Outstanding Teacher in the College of Engineering by Tau Beta Pi. In 1990, and each year from 1994 through 2001, he received the Faculty Recognition Award, presented by graduating ECE students. He was also honored with the Teaching Excellence Award in the College of Engineering in 1994.

In addition to his teaching, Professor Neamen served as Associate Chair of the ECE department for several years and has also worked in industry with Martin Marietta, Sandia National Laboratories, and Raytheon Company. He has published many papers and is the author of *Microelectronics Circuit Analysis and Design,* 4th edition, and *An Introduction to Semiconductor Devices*.

CONTENTS

PREFACE

PHILOSOPHY AND GOALS

The purpose of the fourth edition of this book is to provide a basis for understanding the characteristics, operation, and limitations of semiconductor devices. In order to gain this understanding, it is essential to have a thorough knowledge of the physics of the semiconductor material. The goal of this book is to bring together quantum mechanics, the quantum theory of solids, semiconductor material physics, and semiconductor device physics. All of these components are vital to the understanding of both the operation of present-day devices and any future development in the field.

The amount of physics presented in this text is greater than what is covered in many introductory semiconductor device books. Although this coverage is more extensive, the author has found that once the basic introductory and material physics have been thoroughly covered, the physics of the semiconductor device follows quite naturally and can be covered fairly quickly and efficiently. The emphasis on the underlying physics will also be a benefit in understanding and perhaps in developing new semiconductor devices.

Since the objective of this text is to provide an introduction to the theory of semiconductor devices, there is a great deal of advanced theory that is not considered. In addition, fabrication processes are not described in detail. There are a few references and general discussions about processing techniques such as diffusion and ion implantation, but only where the results of this processing have direct impact on device characteristics.

PREREQUISITES

This text is intended for junior and senior undergraduates majoring in electrical engineering. The prerequisites for understanding the material are college mathematics, up to and including differential equations, basic college physics, and an introduction to electromagnetics. An introduction to modern physics would be helpful, but is not required. Also, a prior completion of an introductory course in electronic circuits is helpful, but not essential.

ORGANIZATION

The text is divided into three parts—Part I covers the introductory quantum physics and then moves on to the semiconductor material physics; Part II presents the physics of the fundamental semiconductor devices; and Part III deals with specialized semiconductor devices including optical, microwave, and power devices.

Part I consists of Chapters 1 through 6. Chapter 1 presents an introduction to the crystal structure of solids leading to the ideal single-crystal semiconductor material.

Chapters 2 and 3 introduce quantum mechanics and the quantum theory of solids, which together provide the necessary basic physics. Chapters 4 through 6 cover the semiconductor material physics. Chapter 4 considers the physics of the semiconductor in thermal equilibrium, Chapter 5 treats the transport phenomena of the charge carriers in a semiconductor, and the nonequilibrium excess carrier characteristics are developed in Chapter 6. Understanding the behavior of excess carriers in a semiconductor is vital to the goal of understanding the device physics.

Part II consists of Chapters 7 through 13. Chapter 7 treats the electrostatics of the basic pn junction and Chapter 8 covers the current–voltage, including the dc and small-signal, characteristics of the pn junction diode. Metal–semiconductor junctions, both rectifying and ohmic, and semiconductor heterojunctions are considered in Chapter 9. The basic physics of the metal–oxide–semiconductor field-effect transistor (MOSFET) is developed in Chapters 10 with additional concepts presented in Chapter 11. Chapter 12 develops the theory of the bipolar transistor and Chapter 13 covers the junction field-effect transistor (JFET). Once the physics of the pn junction is developed, the chapters dealing with the three basic transistors may be covered in any order—these chapters are written so as not to depend on one another.

Part III consists of Chapters 14 and 15. Chapter 14 considers optical devices, such as the solar cell and light emitting diode. Finally, semiconductor microwave devices and semiconductor power devices are presented in Chapter 15.

Eight appendices are included at the end of the book. Appendix A contains a selected list of symbols. Notation may sometimes become confusing, so this appendix may aid in keeping track of all the symbols. Appendix B contains the system of units, conversion factors, and general constants used throughout the text. Appendix H lists answers to selected problems. Most students will find this appendix helpful.

USE OF THE BOOK

The text is intended for a one-semester course at the junior or senior level. As with most textbooks, there is more material than can be conveniently covered in one semester; this allows each instructor some flexibility in designing the course to his or her own specific needs. Two possible orders of presentation are discussed later in a separate section in this preface. However, the text is not an encyclopedia. Sections in each chapter that can be skipped without loss of continuity are identified by an asterisk in both the table of contents and in the chapter itself. These sections, although important to the development of semiconductor device physics, can be postponed to a later time.

The material in the text has been used extensively in a course that is required for junior-level electrical engineering students at the University of New Mexico. Slightly less than half of the semester is devoted to the first six chapters; the remainder of the semester is devoted to the pn junction, the metal–oxide–semiconductor field-effect transistor, and the bipolar transistor. A few other special topics may be briefly considered near the end of the semester.

As mentioned, although the MOS transistor is discussed prior to the bipolar transistor or junction field-effect transistor, each chapter dealing with the basic types of transistors is written to stand alone. Any one of the transistor types may be covered first.

NOTES TO THE READER

This book introduces the physics of semiconductor materials and devices. Although many electrical engineering students are more comfortable building electronic circuits or writing computer programs than studying the underlying principles of semiconductor devices, the material presented here is vital to an understanding of the limitations of electronic devices, such as the microprocessor.

Mathematics is used extensively throughout the book. This may at times seem tedious, but the end result is an understanding that will not otherwise occur. Although some of the mathematical models used to describe physical processes may seem abstract, they have withstood the test of time in their ability to describe and predict these physical processes.

The reader is encouraged to continually refer to the preview sections at the beginning of each chapter so that the objective of the chapter and the purpose of each topic can be kept in mind. This constant review is especially important in the first six chapters, dealing with the basic physics.

The reader must keep in mind that, although some sections may be skipped without loss of continuity, many instructors will choose to cover these topics. The fact that sections are marked with an asterisk does not minimize the importance of these subjects.

It is also important that the reader keep in mind that there may be questions still unanswered at the end of a course. Although the author dislikes the phrase, "it can be shown that . . . ," there are some concepts used here that rely on derivations beyond the scope of the text. This book is intended as an introduction to the subject. Those questions remaining unanswered at the end of the course, the reader is encouraged to keep "in a desk drawer." Then, during the next course in this area of concentration, the reader can take out these questions and search for the answers.

ORDER OF PRESENTATION

Each instructor has a personal preference for the order in which the course material is presented. Listed below are two possible scenarios. The first case, called the MOSFET approach, covers the MOS transistor before the bipolar transistor. It may be noted that the MOSFET in Chapters 10 and 11 may be covered before the pn junction diode.

The second method of presentation listed, called the bipolar approach, is the classical approach. Covering the bipolar transistor immediately after discussing the pn junction diode is the traditional order of presentation. However, because the MOSFET is left until the end of the semester, time constraints may shortchange the amount of class time devoted to this important topic.

Unfortunately, because of time constraints, every topic in each chapter cannot be covered in a one-semester course. The remaining topics must be left for a second-semester course or for further study by the reader.

MOSFET approach	
Chapter 1	Crystal structure
Chapters 2, 3	Selected topics from quantum mechanics and theory of solids
Chapter 4	Semiconductor physics
Chapter 5	Transport phenomena
Chapter 6	Selected topics from nonequilibrium characteristics
Chapter 7	The pn junction
Chapters 10, 11	The MOS transistor
Chapter 8	The pn junction diode
Chapter 9	A brief discussion of the Schottky diode
Chapter 12	The bipolar transistor
	Other selected topics

Bipolar approach	
Chapter 1	Crystal structure
Chapters 2, 3	Selected topics from quantum mechanics and theory of solids
Chapter 4	Semiconductor physics
Chapter 5	Transport phenomena
Chapter 6	Selected topics from nonequilibrium characteristics
Chapters 7, 8	The pn junction and pn junction diode
Chapter 9	A brief discussion of the Schottky diode
Chapter 12	The bipolar transistor
Chapters 10, 11	The MOS transistor
	Other selected topics

NEW TO THE FOURTH EDITION

Order of Presentation: The two chapters dealing with MOSFETs were moved ahead of the chapter on bipolar transistors. This change emphasizes the importance of the MOS transistor.

Semiconductor Microwave Devices: A short section was added in Chapter 15 covering three specialized semiconductor microwave devices.

New Appendix: A new Appendix F has been added dealing with effective mass concepts. Two effective masses are used in various calculations in the text. This appendix develops the theory behind each effective mass and discusses when to use each effective mass in a particular calculation.

Preview Sections: Each chapter begins with a brief introduction, which then leads to a preview section given in bullet form. Each preview item presents a particular objective for the chapter.

Exercise Problems: Over 100 new Exercise Problems have been added. An Exercise Problem now follows each example. The exercise is very similar to the worked example so that readers can immediately test their understanding of the material just covered. Answers are given to each exercise problem.

Test Your Understanding: Approximately 40 percent new Test Your Understanding problems are included at the end of many of the major sections of the chapter. These exercise problems are, in general, more comprehensive than those presented at the end of each example. These problems will also reinforce readers' grasp of the material before they move on to the next section.

End-of-Chapter Problems: There are 330 new end-of-chapter problems, which means that approximately 48 percent of the problems are new to this edition.

RETAINED FEATURES OF THE TEXT

- *Mathematical Rigor:* The mathematical rigor necessary to more clearly understand the basic semiconductor material and device physics has been maintained.
- *Examples:* An extensive number of worked examples are used throughout the text to reinforce the theoretical concepts being developed. These examples contain all the details of the analysis or design, so the reader does not have to fill in missing steps.
- *Summary section:* A summary section, in bullet form, follows the text of each chapter. This section summarizes the overall results derived in the chapter and reviews the basic concepts developed.
- *Glossary of important terms:* A glossary of important terms follows the Summary section of each chapter. This section defines and summarizes the most important terms discussed in the chapter.
- *Checkpoint:* A checkpoint section follows the Glossary section. This section states the goals that should have been met and the abilities the reader should have gained. The Checkpoints will help assess progress before moving on to the next chapter.
- *Review questions:* A list of review questions is included at the end of each chapter. These questions serve as a self-test to help the reader determine how well the concepts developed in the chapter have been mastered.
- *End-of-chapter problems:* A large number of problems are given at the end of each chapter, organized according to the subject of each section in the chapter.
- *Summary and Review Problems:* A few problems, in a Summary and Review section, are open-ended design problems and are given at the end of most chapters.
- *Reading list:* A reading list finishes up each chapter. The references, which are at an advanced level compared with that of this text, are indicated by an asterisk.
- *Answers to selected problems:* Answers to selected problems are given in the last appendix. Knowing the answer to a problem is an aid and a reinforcement in problem solving.

ONLINE RESOURCES

A website to accompany this text is available at www.mhhe.com/neamen. The site includes the solutions manual as well as an image library for instructors. Instructors can also obtain access to C.O.S.M.O.S. for the fourth edition. C.O.S.M.O.S. is a Complete Online Solutions Manual Organization System instructors can use to create exams and assignments, create custom content, and edit supplied problems and solutions.

ACKNOWLEDGMENTS

I am indebted to the many students I have had over the years who have helped in the evolution of this fourth edition as well as to the previous editions of this text. I am grateful for their enthusiasm and constructive criticism.

I want to thank the many people at McGraw-Hill for their tremendous support. To Peter Massar, sponsoring editor, and Lora Neyens, development editor, I am grateful for their encouragement, support, and attention to the many details of this project. I also appreciate the efforts of project managers who guided this work through its final phase toward publication. This effort included gently, but firmly, pushing me through proofreading.

Let me express my continued appreciation to those reviewers who read the manuscripts of the first three editions in its various forms and gave constructive criticism. I also appreciate the efforts of accuracy checkers who worked through the new problem solutions in order to minimize any errors I may have introduced. Finally, my thanks go out to those individuals who have reviewed the book prior to this new edition being published. Their contributions and suggestions for continued improvement are very valuable.

REVIEWERS FOR THE FOURTH EDITION

The following reviewers deserve thanks for their constructive criticism and suggestions for the fourth edition of this book.

Sandra Selmic, *Louisiana Tech University*
Terence Brown, *Michigan State University*
Timothy Wilson, *Oklahoma State University*
Lili He, *San Jose State University*
Jiun Liou, *University of Central Florida*
Michael Stroscio, *University of Illinois-Chicago*
Andrei Sazonov, *University of Waterloo*

McGRAW-HILL CREATE™

Craft your teaching resources to match the way you teach! With McGraw-Hill Create™, www.mcgrawhillcreate.com, you can easily rearrange chapters, combine material from other content sources, and quickly upload content you have written like your course syllabus or teaching notes. Find the content you need in Create by searching through thousands of leading McGraw-Hill textbooks. Arrange your book to fit your teaching style. Create even allows you to personalize your book's appearance by selecting the cover and adding your name, school, and course information. Order a Create book and you'll receive a complimentary print review copy in 3–5 business days or a complimentary electronic review copy (eComp) via email in minutes. Go to www.mcgrawhillcreate.com today and register to experience how McGraw-Hill Create™ empowers you to teach your students *your* way.

MCGRAW-HILL HIGHER EDUCATION AND BLACKBOARD HAVE TEAMED UP.

Blackboard, the Web-based course-management system, has partnered with McGraw-Hill to better allow students and faculty to use online materials and activities to complement face-to-face teaching. Blackboard features exciting social learning and teaching tools that foster more logical, visually impactful and active learning opportunities for students. You'll transform your closed-door classrooms into communities where students remain connected to their educational experience 24 hours a day.

This partnership allows you and your students access to McGraw-Hill's Create™ right from within your Blackboard course—all with one single sign-on. McGraw-Hill and Blackboard can now offer you easy access to industry leading technology and content, whether your campus hosts it, or we do. Be sure to ask your local McGraw-Hill representative for details.

ELECTRONIC TEXTBOOK OPTIONS

E-textbooks are an innovative way for students to save money and create a greener environment at the same time. An e-book can save students about half the cost of a traditional textbook and offers unique features like a powerful search engine, highlighting, and the ability to share notes with classmates using e-books.

McGraw-Hill offers this text as an e-book. To talk about the e-book options, contact your McGraw-Hill sales rep or visit the site www.coursesmart.com to learn more.

Semiconductors and the Integrated Circuit

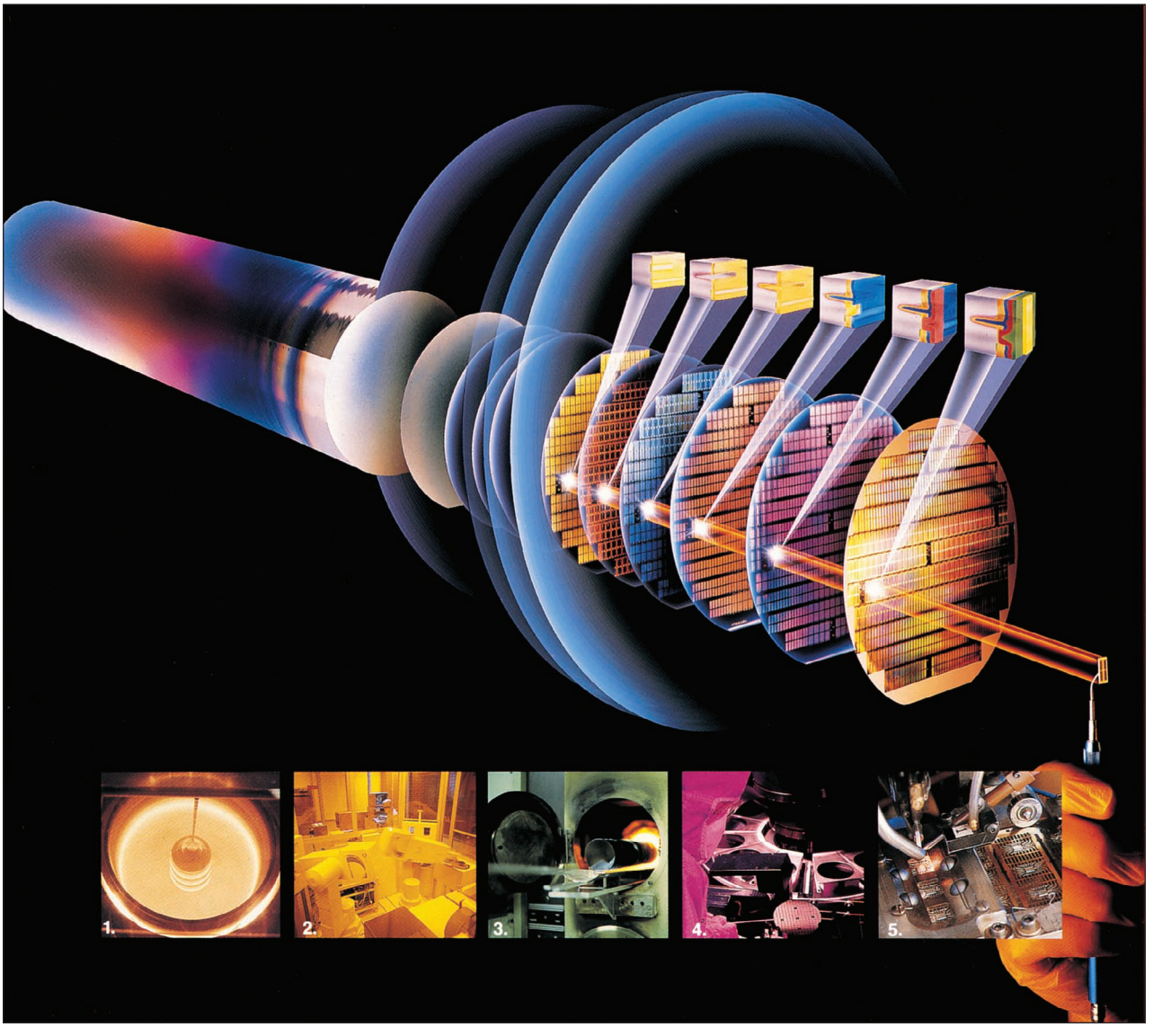
PREVIEW

We often hear that we are living in the information age. Large amounts of information can be obtained via the Internet, for example, and can be obtained very quickly over long distances via satellite communications systems. The information technologies are based upon digital and analog electronic systems, with the transistor and integrated circuit (IC) being the foundation of these remarkable capabilities. Wireless communication systems, including printers, faxes, laptop computers, ipods, and of course the cell phones are big users of today's IC products. The cell phone is not just a telephone any longer, but includes e-mail services and video cameras, for example. Today, a relatively small laptop computer has more computing capability than the equipment used to send a man to the moon a few decades ago. The semiconductor electronics field continues to be a fast-changing one, with thousands of technical papers published and many new electronic devices developed each year. ■

HISTORY

The semiconductor device has a fairly long history, although the greatest explosion of IC technology has occured during the last two or three decades.[1] The metal–semiconductor contact dates back to the early work of Braun in 1874, who discovered the asymmetric nature of electrical conduction between metal contacts and semiconductors, such as copper, iron, and lead sulfide. These devices were used as detectors in early experiments on radio. In 1906, Pickard took out a patent for a point contact

[1]This brief introduction is intended to give a flavor of the history of the semiconductor device and integrated circuit. Thousands of engineers and scientists have made significant contributions to the development of semiconductor electronics—the few events and names mentioned here are not meant to imply that these are the only significant events or people involved in the semiconductor history.



Compliments of Texas Instruments Incorporated

detector using silicon and, in 1907, Pierce published rectification characteristics of diodes made by sputtering metals onto a variety of semiconductors.

By 1935, selenium rectifiers and silicon point contact diodes were available for use as radio detectors. A significant advance in our understanding of the metal–semiconductor contact was aided by developments in semiconductor physics. In 1942, Bethe developed the thermionic-emission theory, according to which the current is determined by the process of emission of electrons into the metal rather than by drift or diffusion. With the development of radar, the need for better and more reliable detector diodes and mixers increased. Methods of achieving high-purity silicon and germanium were developed during this time and germanium diodes became a key component in radar systems during the Second World War.

Another big breakthrough came in December 1947 when the first transistor was constructed and tested at Bell Telephone Laboratories by William Shockley, John Bardeen, and Walter Brattain. This first transistor was a point contact device and used polycrystalline germanium. The transistor effect was soon demonstrated in silicon as well. A significant improvement occurred at the end of 1949 when single-crystal material was used rather than the polycrystalline material. The single crystal yields uniform and improved properties throughout the whole semiconductor material.

The next significant step in the development of the transistor was the use of the diffusion process to form the necessary junctions. This process allowed better control of the transistor characteristics and yielded higher-frequency devices. The diffused mesa transistor was commercially available in germanium in 1957 and in silicon in 1958. The diffusion process also allowed many transistors to be fabricated on a single silicon slice, so the cost of these devices decreased.

THE INTEGRATED CIRCUIT (IC)

The transistor led to a revolution in electronics since it is smaller and more reliable than vacuum tubes used previously. The circuits at that time were discrete in that each element had to be individually connected by wires to form the circuit. The integrated circuit has led to a new revolution in electronics that was not possible with discrete devices. Integration means that complex circuits, consisting of millions of devices, can be fabricated on a single chip of semiconductor material.

The first IC was fabricated in February of 1959 by Jack Kilby of Texas Instruments. In July 1959, a planar version of the IC was independently developed by Robert Noyce of Fairchild. The first integrated circuits incorporated bipolar transistors. Practical MOS transistors were then developed in the mid-1960s and 1970s. The MOS technologies, especially CMOS, have become a major focus for IC design and development. Silicon is the main semiconductor material, while gallium arsenide and other compound semiconductor materials are used for optical devices and for special applications requiring very high frequency devices.

Since the first IC, very sophisticated and complex circuits have been designed and fabricated. A single silicon chip may be on the order of 1 square centimeter and some ICs may have more than a hundred terminals. An IC can contain the arithmetic, logic, and memory functions on a single chip—the primary example of this type of IC

is the microprocessor. Integration means that circuits can be miniaturized for use in satellites and laptop computers where size, weight, and power are critical parameters.

An important advantage of ICs is the result of devices being fabricated very close to each other. The time delay of signals between devices is short so that high-frequency and high-speed circuits are now possible with ICs that were not practical with discrete circuits. In high-speed computers, for example, the logic and memory circuits can be placed very close to each other to minimize time delays. In addition, parasitic capacitance and inductance between devices are reduced which also provides improvement in the speed of the system.

Intense research on silicon processing and increased automation in design and manufacturing have led to lower costs, higher fabrication yields, and greater reliability of integrated circuits.

FABRICATION

The integrated circuit is a direct result of the development of various processing techniques needed to fabricate the transistor and interconnect lines on the single chip. The total collection of these processes for making an IC is called a *technology*. The following few paragraphs provide an introduction to a few of these processes. This introduction is intended to provide the reader with some of the basic terminology used in processing.

Thermal Oxidation A major reason for the success of silicon ICs is the fact that an excellent native oxide, SiO_2, can be formed on the surface of silicon. This oxide is used as a gate insulator in the MOSFET and is also used as an insulator, known as the field oxide, between devices. Metal interconnect lines that connect various devices can be placed on top of the field oxide. Most other semiconductors do not form native oxides that are of sufficient quality to be used in device fabrication.

Silicon will oxidize at room temperature in air forming a thin native oxide of approximately 25 Å thick. However, most oxidations are done at elevated temperatures since the basic process requires that oxygen diffuse through the existing oxide to the silicon surface where a reaction can occur. A schematic of the oxidation process is shown in Figure 0.1. Oxygen diffuses across a stagnant gas layer directly adjacent

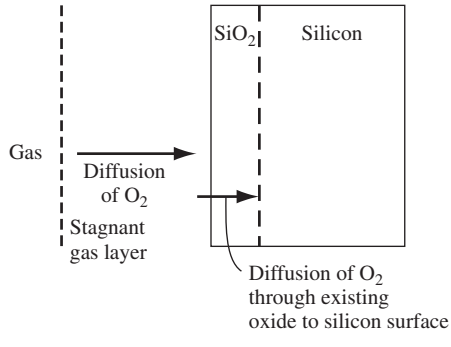


Figure 0.1 | Schematic of the oxidation process.

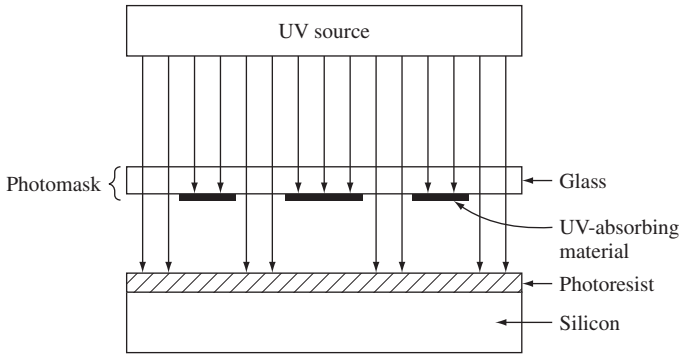


Figure 0.2 | Schematic showing the use of a photomask.

to the oxide surface and then diffuses through the existing oxide layer to the silicon surface where the reaction between O_2 and Si forms SiO_2. Because of this reaction, silicon is actually consumed from the surface of the silicon. The amount of silicon consumed is approximately 44 percent of the thickness of the final oxide.

Photomasks and Photolithography The actual circuitry on each chip is created through the use of photomasks and photolithography. The photomask is a physical representation of a device or a portion of a device. Opaque regions on the mask are made of an ultraviolet-light-absorbing material. A photosensitive layer, called photoresist, is first spread over the surface of the semiconductor. The photoresist is an organic polymer that undergoes chemical change when exposed to ultraviolet light. The photoresist is exposed to ultraviolet light through the photomask as indicated in Figure 0.2. The photoresist is then developed in a chemical solution. The developer is used to remove the unwanted portions of the photoresist and generate the appropriate patterns on the silicon. The photomasks and photolithography process is critical in that it determines how small the devices can be made. Instead of using ultraviolet light, electrons and x-rays can also be used to expose the photoresist.

Etching After the photoresist pattern is formed, the remaining photoresist can be used as a mask, so that the material not covered by the photoresist can be etched. Plasma etching is now the standard process used in IC fabrication. Typically, an etch gas such as chlorofluorocarbons is injected into a low-pressure chamber. A plasma is created by applying a radio-frequency voltage between cathode and anode terminals. The silicon wafer is placed on the cathode. Positively charged ions in the plasma are accelerated toward the cathode and bombard the wafer normal to the surface. The actual chemical and physical reaction at the surface is complex, but the net result is that silicon can be etched anisotropically in very selected regions of the wafer. If photoresist is applied on the surface of silicon dioxide, then the silicon dioxide can also be etched in a similar way.

Diffusion A thermal process that is used extensively in IC fabrication is diffusion. Diffusion is the process by which specific types of "impurity" atoms can be introduced into the silicon material. This doping process changes the conductivity type of the silicon so that pn junctions can be formed. (The pn junction is a basic building block of semiconductor devices.) Silicon wafers are oxidized to form a layer of

silicon dioxide, and windows are opened in the oxide in selected areas using photolithography and etching as just described.

The wafers are then placed in a high-temperature furnace (about 1100°C) and dopant atoms such as boron or phosphorus are introduced. The dopant atoms gradually diffuse or move into the silicon due to a density gradient. Since the diffusion process requires a gradient in the concentration of atoms, the final concentration of diffused atoms is nonlinear, as shown in Figure 0.3. When the wafer is removed from the furnace and the wafer temperature returns to room temperature, the diffusion coefficient of the dopant atoms is essentially zero so that the dopant atoms are then fixed in the silicon material.

Ion Implantation A fabrication process that is an alternative to high-temperature diffusion is ion implantation. A beam of dopant ions is accelerated to a high energy and is directed at the surface of a semiconductor. As the ions enter the silicon, they collide with silicon atoms and lose energy and finally come to rest at some depth within the crystal. Since the collision process is statistical in nature, there is a distribution in the depth of penetration of the dopant ions. Figure 0.4 shows such an example of the implantation of boron into silicon at a particular energy.

Two advantages of the ion implantation process compared to diffusion are (1) the ion implantation process is a low-temperature process and (2) very well

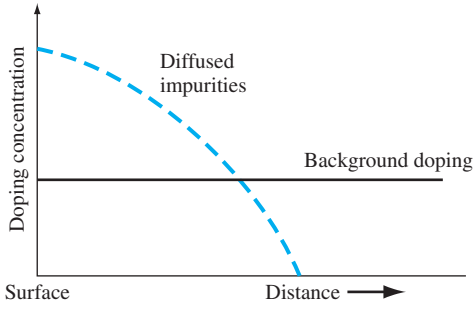


Figure 0.3 | Final concentration of diffused impurities into the surface of a semiconductor.

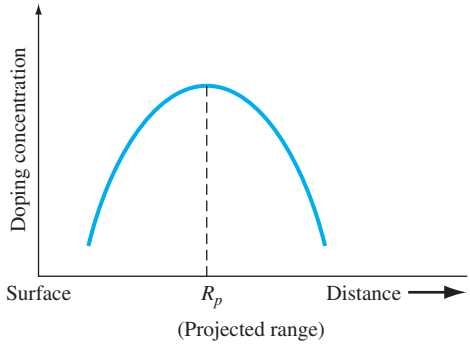


Figure 0.4 | Final concentration of ion-implanted boron into silicon.

defined doping layers can be achieved. Photoresist layers or layers of oxide can be used to block the penetration of dopant atoms so that ion implantation can occur in very selected regions of the silicon.

One disadvantage of ion implantation is that the silicon crystal is damaged by the penetrating dopant atoms because of collisions between the incident dopant atoms and the host silicon atoms. However, most of the damage can be removed by thermal annealing the silicon at an elevated temperature. The thermal annealing temperature, however, is normally much less that the diffusion process temperature.

Metallization, Bonding, and Packaging After the semiconductor devices have been fabricated by the processing steps discussed, they need to be connected to each other to form the circuit. Metal films are generally deposited by a vapor deposition technique, and the actual interconnect lines are formed using photolithography and etching. In general, a protective layer of silicon nitride is finally deposited over the entire chip.

The individual integrated circuit chips are separated by scribing and breaking the wafer. The integrated circuit chip is then mounted in a package. Lead bonders are finally used to attach gold or aluminum wires between the chip and package terminals.

Summary: Simplified Fabrication of a pn Junction Figure 0.5 shows the basic steps in forming a pn junction. These steps involve some of the processing described in the previous paragraphs.

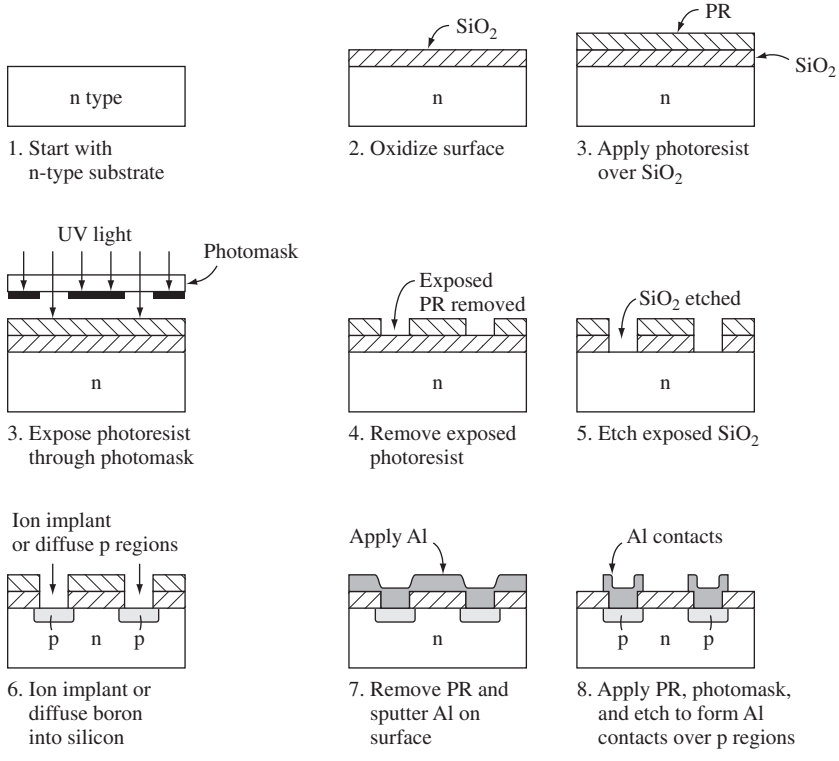


Figure 0.5 | The basic steps in forming a pn junction.

READING LIST

1. Campbell, S. A. *The Science and Engineering of Microelectronic Fabrication*. 2nd ed. New York: Oxford University Press, 2001.
2. Ghandhi, S. K. *VLSI Fabrication Principles: Silicon and Gallium Arsenide.* New York: John Wiley and Sons, 1983.
3. Rhoderick, E. H. *Metal-Semiconductor Contacts.* Oxford: Clarendon Press, 1978.
4. Runyan, W. R., and K. E. Bean. *Semiconductor Integrated Circuit Processing Technology.* Reading, MA: Addison-Wesley, 1990.
5. Torrey, H. C., and C. A. Whitmer. *Crystal Rectifiers.* New York: McGraw-Hill, 1948.
6. Wolf, S., and R. N. Tauber. *Silicon Processing for the VLSI Era,* 2nd ed. Sunset Beach, CA: Lattice Press, 2000.

CHAPTER 1

The Crystal Structure of Solids

This text deals with the electrical properties and characteristics of semiconductor materials and devices. The electrical properties of solids are therefore of primary interest. The semiconductor is in general a single-crystal material. The electrical properties of a single-crystal material are determined not only by the chemical composition but also by the arrangement of atoms in the solid; this being true, a brief study of the crystal structure of solids is warranted. The formation, or growth, of the single-crystal material is an important part of semiconductor technology. A short discussion of several growth techniques is included in this chapter to provide the reader with some of the terminology that describes semiconductor device structures. ■

1.0 | PREVIEW

In this chapter, we will:

- Describe three classifications of solids—amorphous, polycrystalline, and single crystal.
- Discuss the concept of a unit cell.
- Describe three simple crystal structures and determine the volume and surface density of atoms in each structure.
- Describe the diamond crystal structure.
- Briefly discuss several methods of forming single-crystal semiconductor materials.

1.1 | SEMICONDUCTOR MATERIALS

Semiconductors are a group of materials having conductivities between those of metals and insulators. Two general classifications of semiconductors are the elemental semiconductor materials, found in group IV of the periodic table, and the compound semiconductor materials, most of which are formed from special combinations of group III and group V elements. Table 1.1 shows a portion of the periodic table in

Table 1.1 | A portion of the periodic table

III	IV	V
5 **B** Boron	6 **C** Carbon	
13 **Al** Aluminum	14 **Si** Silicon	15 **P** Phosphorus
31 **Ga** Gallium	32 **Ge** Germanium	33 **As** Arsenic
49 **In** Indium		51 **Sb** Antimony

Table 1.2 | A list of some semiconductor materials

Elemental semiconductors	
Si	Silicon
Ge	Germanium
Compound semiconductors	
AlP	Aluminum phosphide
AlAs	Aluminum arsenide
GaP	Gallium phosphide
GaAs	Gallium arsenide
InP	Indium phosphide

which the more common semiconductors are found and Table 1.2 lists a few of the semiconductor materials. (Semiconductors can also be formed from combinations of group II and group VI elements, but in general these will not be considered in this text.)

The elemental materials, those that are composed of single species of atoms, are silicon and germanium. Silicon is by far the most common semiconductor used in integrated circuits and will be emphasized to a great extent.

The two-element, or *binary,* compounds such as gallium arsenide or gallium phosphide are formed by combining one group III and one group V element. Gallium arsenide is one of the more common of the compound semiconductors. Its good optical properties make it useful in optical devices. GaAs is also used in specialized applications in which, for example, high speed is required.

We can also form a three-element, or *ternary,* compound semiconductor. An example is $Al_xGa_{1-x}As$, in which the subscript x indicates the fraction of the lower atomic number element component. More complex semiconductors can also be formed that provide flexibility when choosing material properties.

1.2 | TYPES OF SOLIDS

Amorphous, polycrystalline, and single crystals are the three general types of solids. Each type is characterized by the size of an ordered region within the material. An ordered region is a spatial volume in which atoms or molecules have a regular geometric arrangement or periodicity. Amorphous materials have order only within a few atomic or molecular dimensions, while polycrystalline materials have a high degree of order over many atomic or molecular dimensions. These ordered regions, or single-crystal regions, vary in size and orientation with respect to one another. The single-crystal regions are called grains and are separated from one another by grain boundaries. Single-crystal materials, ideally, have a high degree of order, or regular geometric periodicity, throughout the entire volume of the material. The advantage of a single-crystal material is that, in general, its electrical properties are superior

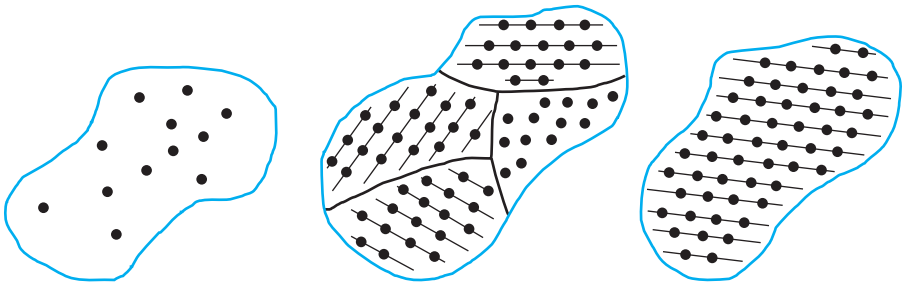


Figure 1.1 | Schematics of three general types of crystals: (a) amorphous, (b) polycrystalline, (c) single.

to those of a nonsingle-crystal material, since grain boundaries tend to degrade the electrical characteristics. Two-dimensional representations of amorphous, polycrystalline, and single-crystal materials are shown in Figure 1.1.

1.3 | SPACE LATTICES

Our primary emphasis in this text will be on the single-crystal material with its regular geometric periodicity in the atomic arrangement. A representative unit, or a group of atoms, is repeated at regular intervals in each of the three dimensions to form the single crystal. The periodic arrangement of atoms in the crystal is called the *lattice*.

1.3.1 Primitive and Unit Cell

We can represent a particular atomic array by a dot that is called a *lattice point*. Figure 1.2 shows an infinite two-dimensional array of lattice points. The simplest means of repeating an atomic array is by translation. Each lattice point in Figure 1.2 can be translated a distance a_1 in one direction and a distance b_1 in a second noncolinear direction to generate the two-dimensional lattice. A third noncolinear translation will produce the three-dimensional lattice. The translation directions need not be perpendicular.

Since the three-dimensional lattice is a periodic repetition of a group of atoms, we do not need to consider the entire lattice, but only a fundamental unit that is being repeated. A *unit cell* is a small volume of the crystal that can be used to reproduce the entire crystal. A unit cell is not a unique entity. Figure 1.3 shows several possible unit cells in a two-dimensional lattice.

The unit cell A can be translated in directions a_2 and b_2, the unit cell B can be translated in directions a_3 and b_3, and the entire two-dimensional lattice can be constructed by the translations of either of these unit cells. The unit cells C and D in Figure 1.3 can also be used to construct the entire lattice by using the appropriate translations. This discussion of two-dimensional unit cells can easily be extended to three dimensions to describe a real single-crystal material.

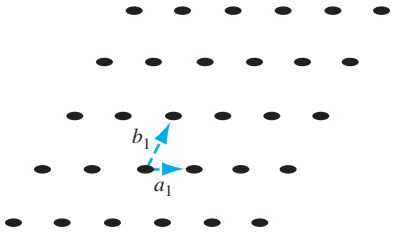


Figure 1.2 | Two-dimensional representation of a single-crystal lattice.

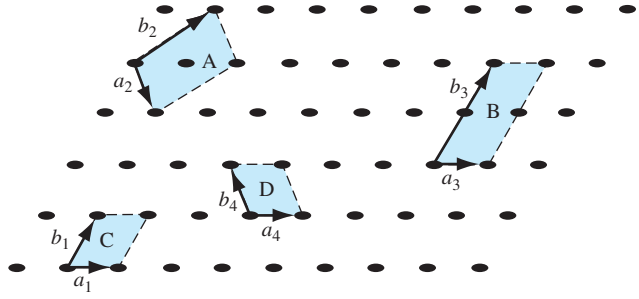


Figure 1.3 | Two-dimensional representation of a single-crystal lattice showing various possible unit cells.

A *primitive cell* is the smallest unit cell that can be repeated to form the lattice. In many cases, it is more convenient to use a unit cell that is not a primitive cell. Unit cells may be chosen that have orthogonal sides, for example, whereas the sides of a primitive cell may be nonorthogonal.

A generalized three-dimensional unit cell is shown in Figure 1.4. The relationship between this cell and the lattice is characterized by three vectors $\bar{a}$, $\bar{b}$, and $\bar{c}$, which need not be perpendicular and which may or may not be equal in length. Every equivalent lattice point in the three-dimensional crystal can be found using the vector

$$\bar{r} = p\bar{a} + q\bar{b} + s\bar{c} \tag{1.1}$$

where p, q, and s are integers. Since the location of the origin is arbitrary, we will let p, q, and s be positive integers for simplicity. The magnitudes of the vectors $\bar{a}$, $\bar{b}$, and $\bar{c}$ are the lattice constants of the unit cell.

1.3.2 Basic Crystal Structures

Before we discuss the semiconductor crystal, let us consider three crystal structures and determine some of the basic characteristics of these crystals. Figure 1.5 shows the simple cubic, body-centered cubic, and face-centered cubic structures. For these simple structures, we may choose unit cells such that the general vectors $\bar{a}$, $\bar{b}$, and $\bar{c}$

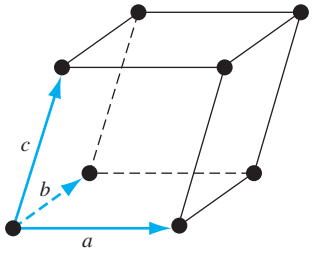


Figure 1.4 | A generalized primitive unit cell.

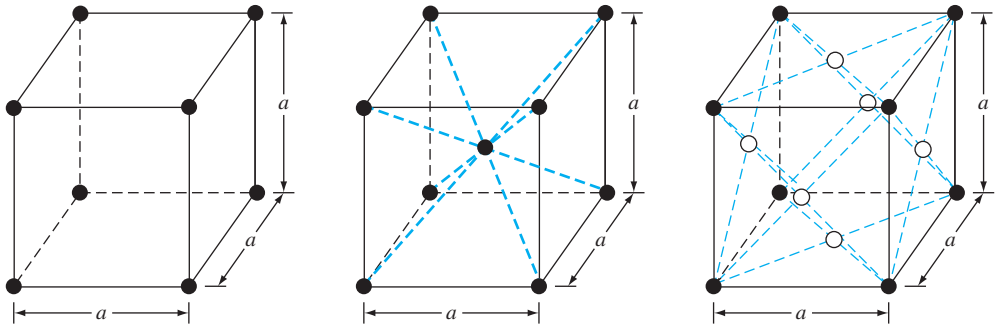


Figure 1.5 | Three lattice types: (a) simple cubic, (b) body-centered cubic, (c) face-centered cubic.

are perpendicular to each other and the lengths are equal. The lattice constant of each unit cell in Figure 1.5 is designated as "*a*." The *simple cubic* (sc) structure has an atom located at each corner; the *body-centered cubic* (bcc) structure has an additional atom at the center of the cube; and the *face-centered cubic* (fcc) structure has additional atoms on each face plane.

By knowing the crystal structure of a material and its lattice dimensions, we can determine several characteristics of the crystal. For example, we can determine the volume density of atoms.

EXAMPLE 1.1

Objective: Find the volume density of atoms in a crystal.

Consider a single-crystal material that is a body-centered cubic, as shown in Figure 1.5b, with a lattice constant $a = 5$ Å $= 5 \times 10^{-8}$ cm. A corner atom is shared by eight unit cells that meet at each corner so that each corner atom effectively contributes one-eighth of its volume to each unit cell. The eight corner atoms then contribute an equivalent of one atom to the unit cell. If we add the body-centered atom to the corner atoms, each unit cell contains an equivalent of two atoms.

■ Solution

The number of atoms per unit cell is $\frac{1}{8} \times 8 + 1 = 2$

The volume density of atoms is then found as

$$\text{Volume Density} = \frac{\#\ atoms\ per\ unit\ cell}{volume\ of\ unit\ cell}$$

So

$$\text{Volume Density} = \frac{2}{a^3} = \frac{2}{(5 \times 10^{-8})^3} = 1.6 \times 10^{22}\ \text{atoms/cm}^3$$

■ EXERCISE PROBLEM

Ex 1.1 The lattice constant of a face-centered cubic lattice is 4.25 Å. Determine the (a) effective number of atoms per unit cell and (b) volume density of atoms.

[Ans. (a) 4; (b) 5.21×10^{22} cm^{-3}]

1.3.3 Crystal Planes and Miller Indices

Since real crystals are not infinitely large, they eventually terminate at a surface. Semiconductor devices are fabricated at or near a surface, so the surface properties may influence the device characteristics. We would like to be able to describe these surfaces in terms of the lattice. Surfaces, or planes through the crystal, can be described by first considering the intercepts of the plane along the $\bar{a}$, $\bar{b}$, and $\bar{c}$ axes used to describe the lattice.

EXAMPLE 1.2

Objective: Describe the plane shown in Figure 1.6. (The lattice points in Figure 1.6 are shown along the $\bar{a}$, $\bar{b}$, and $\bar{c}$ axes only.)

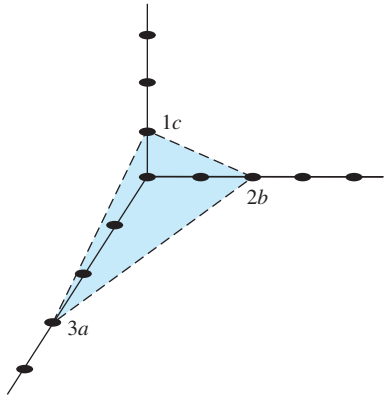


Figure 1.6 | A representative crystal-lattice plane.

■ Solution

From Equation (1.1), the intercepts of the plane correspond to $p = 3$, $q = 2$, and $s = 1$. Now write the reciprocals of the intercepts, which gives

$$\left(\frac{1}{3}, \frac{1}{2}, \frac{1}{1}\right)$$

Multiply by the lowest common denominator, which in this case is 6, to obtain (2, 3, 6). The plane in Figure 1.6 is then referred to as the (236) plane. The integers are referred to as the Miller indices. We will refer to a general plane as the (*hkl*) plane.

■ Comment

We can show that the same three Miller indices are obtained for any plane that is parallel to the one shown in Figure 1.6. Any parallel plane is entirely equivalent to any other.

■ EXERCISE PROBLEM

Ex 1.2 Describe the lattice plane shown in Figure 1.7. [Ans. (211) plane]

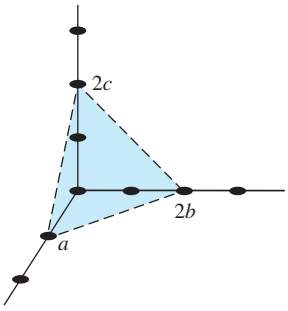


Figure 1.7 | Figure for Exercise Problem Ex 1.2.

Three planes that are commonly considered in a cubic crystal are shown in Figure 1.8. The plane in Figure 1.8a is parallel to the $\bar{b}$ and $\bar{c}$ axes so the intercepts are given as $p = 1$, $q = \infty$, and $s = \infty$. Taking the reciprocal, we obtain the Miller indices as (1, 0, 0), so the plane shown in Figure 1.8a is referred to as the (100) plane. Again, any plane parallel to the one shown in Figure 1.8a and separated by an integral number of lattice constants is equivalent and is referred to as the (100) plane. One advantage to taking the reciprocal of the intercepts to obtain the Miller indices is that the use of infinity is avoided when describing a plane that is parallel to an axis. If we were to describe a plane passing through the origin of our system, we would obtain infinity as one or more of the Miller indices after taking the reciprocal of the intercepts. However, the location of the origin of our system is entirely arbitrary and so, by translating the origin to another equivalent lattice point, we can avoid the use of infinity in the set of Miller indices.

For the simple cubic structure, the body-centered cubic, and the face-centered cubic, there is a high degree of symmetry. The axes can be rotated by 90° in each

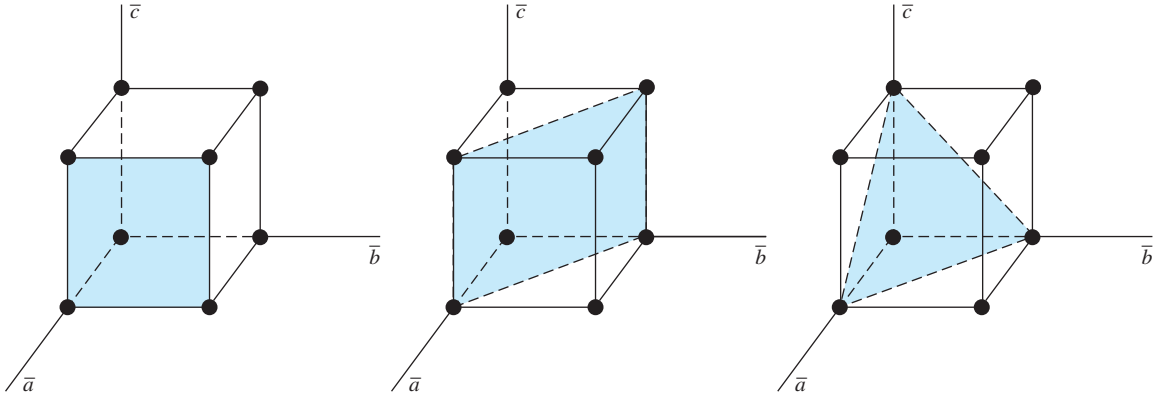


Figure 1.8 | Three lattice planes: (a) (100) plane, (b) (110) plane, (c) (111) plane.

of the three dimensions and each lattice point can again be described by Equation (1.1) as

$$\bar{r} = p\bar{a} + q\bar{b} + s\bar{c} \tag{1.1}$$

Each face plane of the cubic structure shown in Figure 1.8a is entirely equivalent. These planes are grouped together and are referred to as the {100} set of planes.

We may also consider the planes shown in Figures 1.8b and 1.8c. The intercepts of the plane shown in Figure 1.8b are $p = 1$, $q = 1$, and $s = \infty$. The Miller indices are found by taking the reciprocal of these intercepts and, as a result, this plane is referred to as the (110) plane. In a similar way, the plane shown in Figure 1.8c is referred to as the (111) plane.

One characteristic of a crystal that can be determined is the distance between nearest equivalent parallel planes. Another characteristic is the surface concentration of atoms, number per square centimeter (#/cm^2), that are cut by a particular plane. Again, a single-crystal semiconductor is not infinitely large and must terminate at some surface. The surface density of atoms may be important, for example, in determining how another material, such as an insulator, will "fit" on the surface of a semiconductor material.

EXAMPLE 1.3

Objective: Calculate the surface density of atoms on a particular plane in a crystal.

Consider the body-centered cubic structure and the (110) plane shown in Figure 1.9a. Assume the atoms can be represented as hard spheres with the closest atoms touching each other. Assume the lattice constant is $a_1 = 5$ Å. Figure 1.9b shows how the atoms are cut by the (110) plane.

The atom at each corner is shared by four similar equivalent lattice planes, so each corner atom effectively contributes one-fourth of its area to this lattice plane as indicated in the figure. The four corner atoms then effectively contribute one atom to this lattice plane. The atom in the center is completely enclosed in the lattice plane. There is no other equivalent plane that

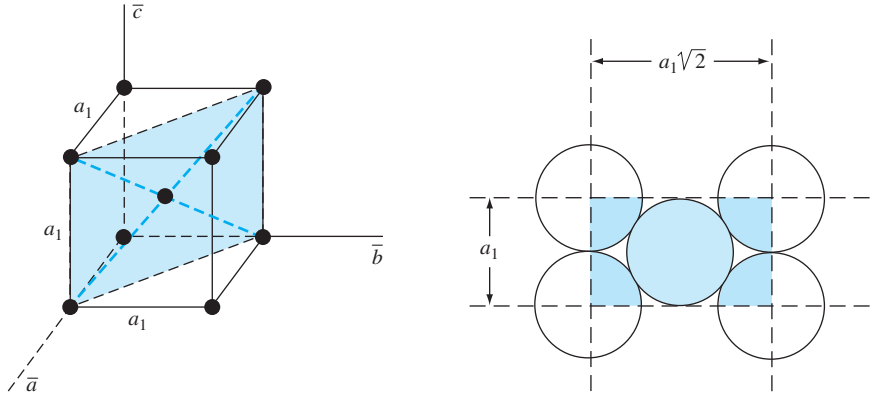


Figure 1.9 | (a) The (110) plane in a body-centered cubic and (b) the atoms cut by the (110) plane in a body-centered cubic.

cuts the center atom and the corner atoms, so the entire center atom is included in the number of atoms in the crystal plane. The lattice plane in Figure 1.9b, then, contains two atoms.

■ Solution

The number of atoms per lattice plane is $\frac{1}{4} \times 4 + 1 = 2$

The surface density of atoms is then found as

$$\text{Surface Density} = \frac{\#\ of\ atoms\ per\ lattice\ plane}{area\ of\ lattice\ plane}$$

So

$$\text{Surface Density} = \frac{2}{(a_1)(a_1\sqrt{2})} = \frac{2}{(5 \times 10^{-8})^2\sqrt{2}}$$

$$= 5.66 \times 10^{14}\ \text{atoms/cm}^2$$

■ Comment

The surface density of atoms is a function of the particular crystal plane in the lattice and generally varies from one crystal plane to another.

■ EXERCISE PROBLEM

Ex 1.3 The lattice constant of a face-centered-cubic structure is 4.25 Å. Calculate the surface density of atoms for a (a) (100) plane and (b) (110) plane.

[Ans. (a) $1.11 \times 10^{15}\ \text{cm}^{-2}$; (b) $7.83 \times 10^{14}\ \text{cm}^{-2}$]

1.3.4 Directions in Crystals

In addition to describing crystal planes in a lattice, we may want to describe a particular direction in the crystal. The direction can be expressed as a set of three integers that are the components of a vector in that direction. For example, the body diagonal in a simple cubic lattice is composed of vector components 1, 1, 1. The body diagonal is then described as the [111] direction. The brackets are used to designate direction as distinct from the parentheses used for the crystal planes. The three basic directions and the associated crystal planes for the simple cubic structure are shown in Figure 1.10. Note that in the simple cubic lattices, the [*hkl*] direction is perpendicular to the (*hkl*) plane. This perpendicularity may not be true in noncubic lattices.

TEST YOUR UNDERSTANDING

TYU 1.1 The volume density of atoms for a simple cubic lattice is $4 \times 10^{22}\ \text{cm}^{-3}$. Assume that the atoms are hard spheres with each atom touching its nearest neighbor. Determine the lattice constant and the radius of the atom. (Ans. $a = 2.92$ Å, $r = 1.46$ Å)

TYU 1.2 Consider a simple cubic structure with a lattice constant of $a = 4.65$ Å. Determine the surface density of atoms in the (a) (100) plane, (b) (110) plane, and (c) (111) plane. [Ans. (a) $4.62 \times 10^{14}\ \text{cm}^{-2}$; (b) $3.27 \times 10^{14}\ \text{cm}^{-2}$; (c) $2.67 \times 10^{14}\ \text{cm}^{-2}$]

TYU 1.3 (a) Determine the distance between nearest (100) planes in a simple cubic lattice with a lattice constant of $a = 4.83$ Å. (b) Repeat part (a) for the (110) plane.

[Ans. (a) 4.83 Å; (b) 3.42 Å]

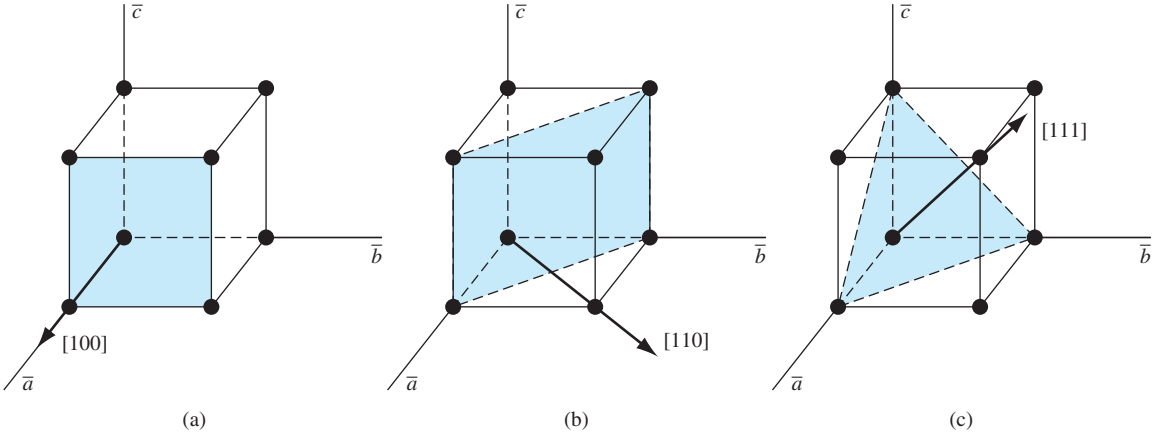


Figure 1.10 | Three lattice directions and planes: (a) (100) plane and [100] direction, (b) (110) plane and [110] direction, (c) (111) plane and [111] direction.

1.4 | THE DIAMOND STRUCTURE

As already stated, silicon is the most common semiconductor material. Silicon is referred to as a group IV element and has a diamond crystal structure. Germanium is also a group IV element and has the same diamond structure. A unit cell of the diamond structure, shown in Figure 1.11, is more complicated than the simple cubic structures that we have considered up to this point.

We may begin to understand the diamond lattice by considering the tetrahedral structure shown in Figure 1.12. This structure is basically a body-centered cubic with four of the corner atoms missing. Every atom in the tetrahedral structure has four nearest neighbors and it is this structure that is the basic building block of the diamond lattice.

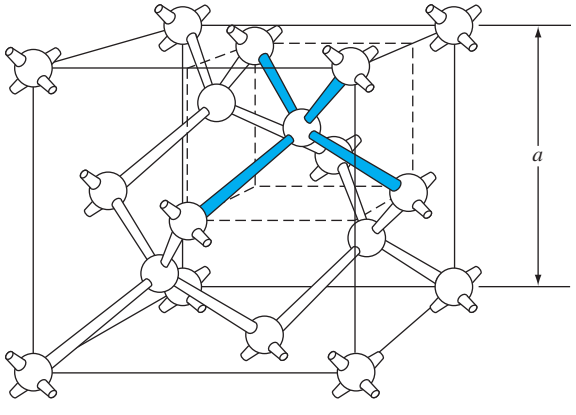


Figure 1.11 | The diamond structure.

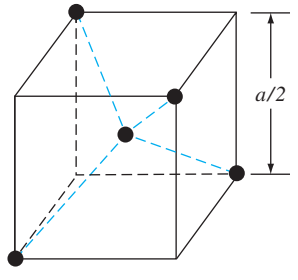


Figure 1.12 | The tetrahedral structure of closest neighbors in the diamond lattice.

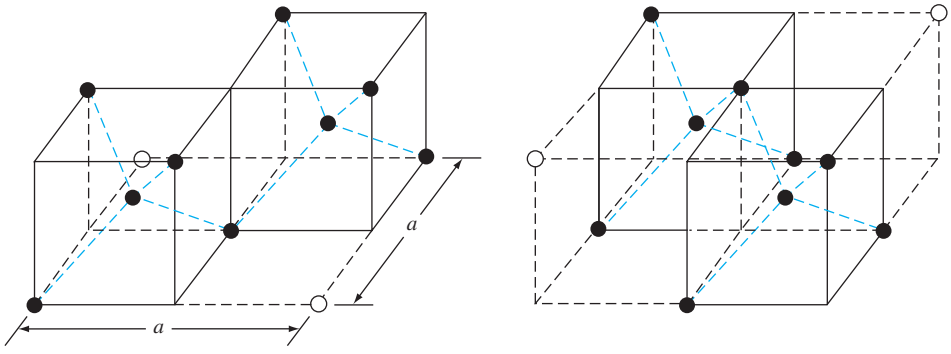


Figure 1.13 | Portions of the diamond lattice: (a) bottom half and (b) top half.

There are several ways to visualize the diamond structure. One way to gain a further understanding of the diamond lattice is by considering Figure 1.13. Figure 1.13a shows two body-centered cubic, or tetrahedral, structures diagonally adjacent to each other. The open circles represent atoms in the lattice that are generated when the structure is translated to the right or left, one lattice constant, *a*. Figure 1.13b represents the top half of the diamond structure. The top half again consists of two tetrahedral structures joined diagonally, but which are at 90° with respect to the bottom-half diagonal. An important characteristic of the diamond lattice is that any atom within the diamond structure will have four nearest neighboring atoms. We will note this characteristic again in our discussion of atomic bonding in the next section.

The diamond structure refers to the particular lattice in which all atoms are of the same species, such as silicon or germanium. The zincblende (sphalerite) structure differs from the diamond structure only in that there are two different types of atoms in the lattice. Compound semiconductors, such as gallium arsenide, have the zincblende structure shown in Figure 1.14. The important feature of both the diamond and the zincblende structures is that the atoms are joined together to form a tetrahedron. Figure 1.15 shows the basic tetrahedral structure of GaAs in which each Ga atom has four nearest As neighbors and each As atom has four nearest Ga neighbors. This figure also begins to show the interpenetration of two sublattices that can be used to generate the diamond or zincblende lattice.

TEST YOUR UNDERSTANDING

TYU 1.4 Consider the diamond unit cell shown in Figure 1.11. Determine the (a) number of corner atoms, (b) number of face-centered atoms, and (c) number of atoms totally enclosed in the unit cell. [Ans. (a) 8; (b) 6; (c) 4]

TYU 1.5 The lattice constant of silicon is 5.43 Å. Calculate the volume density of silicon atoms. (Ans. $5 \times 10^{22}\,\text{cm}^{-3}$)

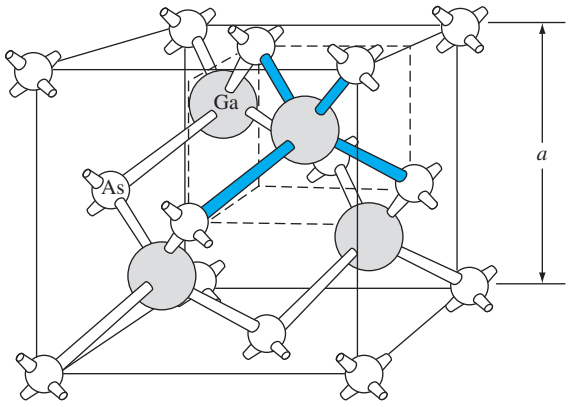


Figure 1.14 | The zincblende (sphalerite) lattice of GaAs.

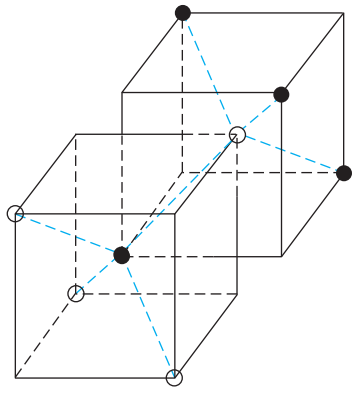


Figure 1.15 | The tetrahedral structure of closest neighbors in the zincblende lattice.

1.5 | ATOMIC BONDING

We have been considering various single-crystal structures. The question arises as to why one particular crystal structure is favored over another for a particular assembly of atoms. A fundamental law of nature is that the total energy of a system in thermal equilibrium tends to reach a minimum value. The interaction that occurs between atoms to form a solid and to reach the minimum total energy depends on the type of atom or atoms involved. The type of bond, or interaction, between atoms, then, depends on the particular atom or atoms in the crystal. If there is not a strong bond between atoms, they will not "stick together" to create a solid.

The interaction between atoms can be described by quantum mechanics. Although an introduction to quantum mechanics is presented in the next chapter, the quantum-mechanical description of the atomic bonding interaction is still beyond the scope of this text. We can nevertheless obtain a qualitative understanding of how various atoms interact by considering the valence, or outermost, electrons of an atom.

The atoms at the two extremes of the periodic table (excepting the inert elements) tend to lose or gain valence electrons, thus forming ions. These ions then essentially have complete outer energy shells. The elements in group I of the periodic table tend to lose their one electron and become positively charged, while the elements in group VII tend to gain an electron and become negatively charged. These oppositely charged ions then experience a coulomb attraction and form a bond referred to as an *ionic bond.* If the ions were to get too close, a repulsive force would become dominant, so an equilibrium distance results between these two ions. In a crystal, negatively charged ions tend to be surrounded by positively charged ions and positively charged ions tend to be surrounded by negatively charged ions, so a periodic array of the atoms is formed to create the lattice. A classic example of ionic bonding is sodium chloride.

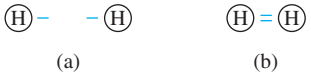


Figure 1.16 | Representation of (a) hydrogen valence electrons and (b) covalent bonding in a hydrogen molecule.

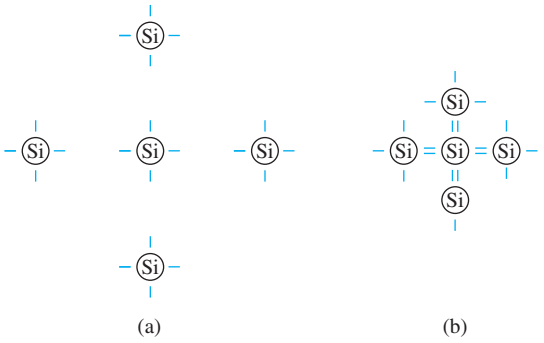


Figure 1.17 | Representation of (a) silicon valence electrons and (b) covalent bonding in the silicon crystal.

The interaction of atoms tends to form closed valence shells such as we see in ionic bonding. Another atomic bond that tends to achieve closed-valence energy shells is *covalent bonding,* an example of which is found in the hydrogen molecule. A hydrogen atom has one electron and needs one more electron to complete the lowest energy shell. A schematic of two noninteracting hydrogen atoms, and the hydrogen molecule with the covalent bonding, is shown in Figure 1.16. Covalent bonding results in electrons being shared between atoms, so that in effect the valence energy shell of each atom is full.

Atoms in group IV of the periodic table, such as silicon and germanium, also tend to form covalent bonds. Each of these elements has four valence electrons and needs four more electrons to complete the valence energy shell. If a silicon atom, for example, has four nearest neighbors, with each neighbor atom contributing one valence electron to be shared, then the center atom will in effect have eight electrons in its outer shell. Figure 1.17a schematically shows five noninteracting silicon atoms with the four valence electrons around each atom. A two-dimensional representation of the covalent bonding in silicon is shown in Figure 1.17b. The center atom has eight shared valence electrons.

A significant difference between the covalent bonding of hydrogen and of silicon is that, when the hydrogen molecule is formed, it has no additional electrons to form additional covalent bonds, while the outer silicon atoms always have valence electrons available for additional covalent bonding. The silicon array may then be formed into an infinite crystal, with each silicon atom having four nearest neighbors and eight shared electrons. The four nearest neighbors in silicon forming the covalent bond correspond to the tetrahedral structure and the diamond lattice, which were shown in Figures 1.12 and 1.11 respectively. Atomic bonding and crystal structure are obviously directly related.

The third major atomic bonding scheme is referred to as *metallic bonding*. Group I elements have one valence electron. If two sodium atoms ($Z = 11$), for example, are brought into close proximity, the valence electrons interact in a way similar to that in covalent bonding. When a third sodium atom is brought into close proximity with the

first two, the valence electrons can also interact and continue to form a bond. Solid sodium has a body-centered cubic structure, so each atom has eight nearest neighbors with each atom sharing many valence electrons. We may think of the positive metallic ions as being surrounded by a sea of negative electrons, the solid being held together by the electrostatic forces. This description gives a qualitative picture of the metallic bond.

A fourth type of atomic bond, called the *Van der Waals* bond, is the weakest of the chemical bonds. A hydrogen fluoride (HF) molecule, for example, is formed by an ionic bond. The effective center of the positive charge of the molecule is not the same as the effective center of the negative charge. This nonsymmetry in the charge distribution results in a small electric dipole that can interact with the dipoles of other HF molecules. With these weak interactions, solids formed by the Van der Waals bonds have a relatively low melting temperature—in fact, most of these materials are in gaseous form at room temperature.

*1.6 | IMPERFECTIONS AND IMPURITIES IN SOLIDS

Up to this point, we have been considering an ideal single-crystal structure. In a real crystal, the lattice is not perfect, but contains imperfections or defects; that is, the perfect geometric periodicity is disrupted in some manner. Imperfections tend to alter the electrical properties of a material and, in some cases, electrical parameters can be dominated by these defects or impurities.

1.6.1 Imperfections in Solids

One type of imperfection that all crystals have in common is atomic thermal vibration. A perfect single crystal contains atoms at particular lattice sites, the atoms separated from each other by a distance we have assumed to be constant. The atoms in a crystal, however, have a certain thermal energy, which is a function of temperature. The thermal energy causes the atoms to vibrate in a random manner about an equilibrium lattice point. This random thermal motion causes the distance between atoms to randomly fluctuate, slightly disrupting the perfect geometric arrangement of atoms. This imperfection, called *lattice vibrations,* affects some electrical parameters, as we will see later in our discussion of semiconductor material characteristics.

Another type of defect is called a *point defect.* There are several of this type that we need to consider. Again, in an ideal single-crystal lattice, the atoms are arranged in a perfect periodic arrangement. However, in a real crystal, an atom may be missing from a particular lattice site. This defect is referred to as a *vacancy;* it is schematically shown in Figure 1.18a. In another situation, an atom may be located between lattice sites. This defect is referred to as an *interstitial* and is schematically shown in Figure 1.18b. In the case of vacancy and interstitial defects, not only is the perfect

*Indicates sections that will aid in the total summation of understanding of semiconductor devices, but may be skipped the first time through the text without loss of continuity.

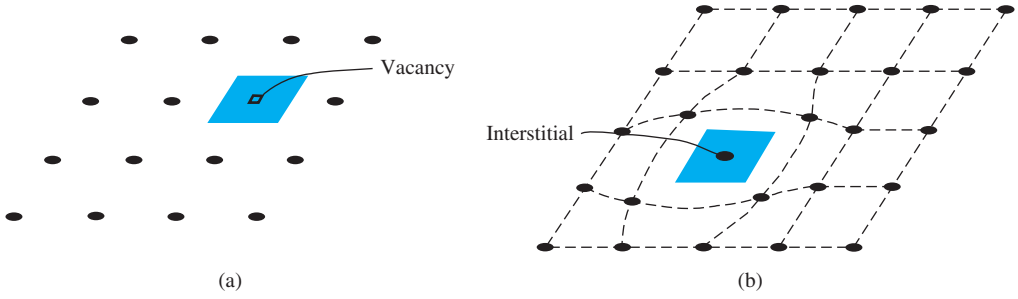


Figure 1.18 | Two-dimensional representation of a single-crystal lattice showing (a) a vacancy defect and (b) an interstitial defect.

geometric arrangement of atoms broken but also the ideal chemical bonding between atoms is disrupted, which tends to change the electrical properties of the material. A vacancy and interstitial may be in close enough proximity to exhibit an interaction between the two point defects. This vacancy–interstitial defect, also known as a *Frenkel defect,* produces different effects than the simple vacancy or interstitial.

The point defects involve single atoms or single-atom locations. In forming single-crystal materials, more complex defects may occur. A line defect, for example, occurs when an entire row of atoms is missing from its normal lattice site. This defect is referred to as a *line dislocation* and is shown in Figure 1.19. As with a point defect, a line dislocation disrupts both the normal geometric periodicity of the lattice and the ideal atomic bonds in the crystal. This dislocation can also alter the electrical properties of the material, usually in a more unpredictable manner than the simple point defects.

Other complex dislocations can also occur in a crystal lattice. However, this introductory discussion is intended only to present a few of the basic types of defect, and to show that a real crystal is not necessarily a perfect lattice structure. The effect of these imperfections on the electrical properties of a semiconductor will be considered in later chapters.

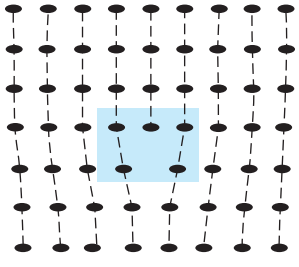


Figure 1.19 | A two-dimensional representation of a line dislocation.

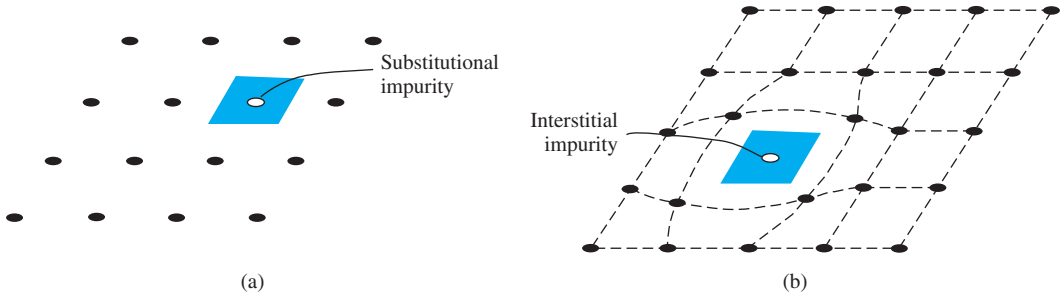


Figure 1.20 | Two-dimensional representation of a single-crystal lattice showing (a) a substitutional impurity and (b) an intersititial impurity.

1.6.2 Impurities in Solids

Foreign atoms, or impurity atoms, may be present in a crystal lattice. Impurity atoms may be located at normal lattice sites, in which case they are called *substitutional* impurities. Impurity atoms may also be located between normal sites, in which case they are called *interstitial* impurities. Both these impurities are lattice defects and are schematically shown in Figure 1.20. Some impurities, such as oxygen in silicon, tend to be essentially inert; however, other impurities, such as gold or phosphorus in silicon, can drastically alter the electrical properties of the material.

In Chapter 4 we will see that, by adding controlled amounts of particular impurity atoms, the electrical characteristics of a semiconductor material can be favorably altered. The technique of adding impurity atoms to a semiconductor material in order to change its conductivity is called *doping*. There are two general methods of doping: impurity diffusion and ion implantation.

The actual diffusion process depends to some extent on the material but, in general, impurity diffusion occurs when a semiconductor crystal is placed in a high-temperature ($\approx$1000°C) gaseous atmosphere containing the desired impurity atom. At this high temperature, many of the crystal atoms can randomly move in and out of their single-crystal lattice sites. Vacancies may be created by this random motion so that impurity atoms can move through the lattice by hopping from one vacancy to another. Impurity diffusion is the process by which impurity particles move from a region of high concentration near the surface to a region of lower concentration within the crystal. When the temperature decreases, the impurity atoms become permanently frozen into the substitutional lattice sites. Diffusion of various impurities into selected regions of a semiconductor allows us to fabricate complex electronic circuits in a single semiconductor crystal.

Ion implantation generally takes place at a lower temperature than diffusion. A beam of impurity ions is accelerated to kinetic energies in the range of 50 keV or greater and then directed to the surface of the semiconductor. The high-energy impurity ions enter the crystal and come to rest at some average depth from the surface. One advantage of ion implantation is that controlled numbers of impurity atoms can be introduced into specific regions of the crystal. A disadvantage of this technique is that the incident impurity atoms collide with the crystal atoms, causing

lattice-displacement damage. However, most of the lattice damage can be removed by thermal annealing, in which the temperature of the crystal is raised for a short time. Thermal annealing is a required step after implantation.

*1.7 | GROWTH OF SEMICONDUCTOR MATERIALS

The success in fabricating very large scale integrated (VLSI) circuits is a result, to a large extent, of the development of and improvement in the formation or growth of pure single-crystal semiconductor materials. Semiconductors are some of the purest materials. Silicon, for example, has concentrations of most impurities of less than 1 part in 10^{10} atoms. The high purity requirement means that extreme care is necessary in the growth and the treatment of the material at each step of the fabrication process. The mechanics and kinetics of crystal growth are extremely complex and will be described in only very general terms in this text. However, a general knowledge of the growth techniques and terminology is valuable.

1.7.1 Growth from a Melt

A common technique for growing single-crystal materials is called the *Czochralski method.* In this technique, a small piece of single-crystal material, known as a *seed,* is brought into contact with the surface of the same material in liquid phase, and then slowly pulled from the melt. As the seed is slowly pulled, solidification occurs along the plane between the solid–liquid interface. Usually the crystal is also rotated slowly as it is being pulled, to provide a slight stirring action to the melt, resulting in a more uniform temperature. Controlled amounts of specific impurity atoms, such as boron or phosphorus, may be added to the melt so that the grown semiconductor crystal is intentionally doped with the impurity atom. Figure 1.21a shows a schematic of the Czochralski growth process and a silicon ingot or boule grown by this process.

Some impurities may be present in the ingot that are undesirable. Zone refining is a common technique for purifying material. A high-temperature coil, or r-f induction coil, is slowly passed along the length of the boule. The temperature induced by the coil is high enough so that a thin layer of liquid is formed. At the solid–liquid interface, there is a distribution of impurities between the two phases. The parameter that describes this distribution is called the *segregation coefficient:* the ratio of the concentration of impurities in the solid to the concentration in the liquid. If the segregation coefficient is 0.1, for example, the concentration of impurities in the liquid is a factor of 10 greater than that in the solid. As the liquid zone moves through the material, the impurities are driven along with the liquid. After several passes of the r-f coil, most impurities are at the end of the bar, which can then be cut off. The moving molten zone, or the zone-refining technique, can result in considerable purification.

After the semiconductor is grown, the boule is mechanically trimmed to the proper diameter and a flat is ground over the entire length of the boule to denote

*Indicates sections that will aid in the total summation of understanding of semiconductor devices, but may be skipped the first time through the text without loss of continuity.

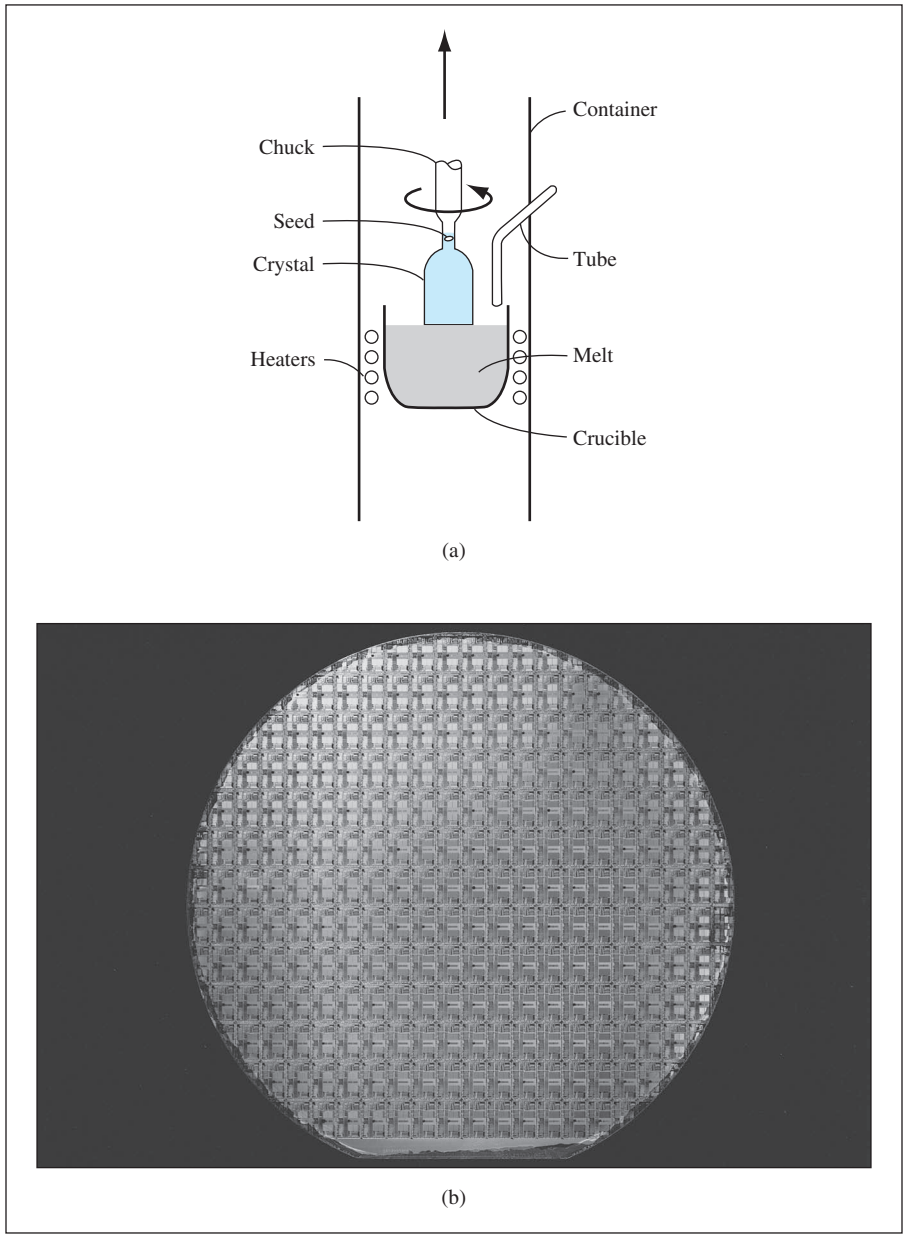


Figure 1.21 | (a) Model of a crystal puller and (b) photograph of a silicon wafer with an array of integrated circuits. The circuits are tested on the wafer then sawed apart into chips that are mounted into packages. (Photo courtesy of Intel Corporation.)

the crystal orientation. The flat is perpendicular to the [110] direction or indicates the (110) plane. (See Figure 1.21b.) This then allows the individual chips to be fabricated along given crystal planes so that the chips can be sawed apart more easily. The boule is then sliced into wafers. The wafer must be thick enough to mechanically support itself. A mechanical two-sided lapping operation produces a flat wafer of uniform thickness. Since the lapping procedure can leave a surface damaged and contaminated by the mechanical operation, the surface must be removed by chemical etching. The final step is polishing. This provides a smooth surface on which devices may be fabricated or further growth processes may be carried out. This final semiconductor wafer is called the substrate material.

1.7.2 Epitaxial Growth

A common and versatile growth technique that is used extensively in device and integrated circuit fabrication is epitaxial growth. *Epitaxial growth* is a process whereby a thin, single-crystal layer of material is grown on the surface of a single-crystal substrate. In the epitaxial process, the single-crystal substrate acts as the seed, although the process takes place far below the melting temperature. When an epitaxial layer is grown on a substrate of the same material, the process is termed *homoepitaxy.* Growing silicon on a silicon substrate is one example of a homoepitaxy process. At present, a great deal of work is being done with *heteroepitaxy.* In a heteroepitaxy process, although the substrate and epitaxial materials are not the same, the two crystal structures should be very similar if single-crystal growth is to be obtained and if a large number of defects are to be avoided at the epitaxial–substrate interface. Growing epitaxial layers of the ternary alloy AlGaAs on a GaAs substrate is one example of a heteroepitaxy process.

One epitaxial growth technique that has been used extensively is called *chemical vapor-phase deposition* (CVD). Silicon epitaxial layers, for example, are grown on silicon substrates by the controlled deposition of silicon atoms onto the surface from a chemical vapor containing silicon. In one method, silicon tetrachloride reacts with hydrogen at the surface of a heated substrate. The silicon atoms are released in the reaction and can be deposited onto the substrate, while the other chemical reactant, HCl, is in gaseous form and is swept out of the reactor. A sharp demarcation between the impurity doping in the substrate and in the epitaxial layer can be achieved using the CVD process. This technique allows great flexibility in the fabrication of semiconductor devices.

Liquid-phase epitaxy is another epitaxial growth technique. A compound of the semiconductor with another element may have a melting temperature lower than that of the semiconductor itself. The semiconductor substrate is held in the liquid compound and, since the temperature of the melt is lower than the melting temperature of the substrate, the substrate does not melt. As the solution is slowly cooled, a single-crystal semiconductor layer grows on the seed crystal. This technique, which occurs at a lower temperature than the Czochralski method, is useful in growing group III–V compound semiconductors.

A versatile technique for growing epitaxial layers is the *molecular beam epitaxy* (MBE) process. A substrate is held in vacuum at a temperature normally in the range

of 400 to 800°C, a relatively low temperature compared with many semiconductor-processing steps. Semiconductor and dopant atoms are then evaporated onto the surface of the substrate. In this technique, the doping can be precisely controlled resulting in very complex doping profiles. Complex ternary compounds, such as AlGaAs, can be grown on substrates, such as GaAs, where abrupt changes in the crystal composition are desired. Many layers of various types of epitaxial compositions can be grown on a substrate in this manner. These structures are extremely beneficial in optical devices such as laser diodes.

1.8 | SUMMARY

- A few of the most common semiconductor materials were listed. Silicon is the most common semiconductor material and appears in column IV of the periodic table.
- The properties of semiconductors and other materials are determined to a large extent by the single-crystal lattice structure. The unit cell is a small volume of the crystal that is used to reproduce the entire crystal. Three basic unit cells are the simple cubic, body-centered cubic, and face-centered cubic.
- Silicon has the diamond crystal structure. Atoms are formed in a tetrahedral configuration with four nearest neighbor atoms. The binary semiconductors have a zincblende lattice that is basically the same as the diamond lattice.
- Miller indices are used to describe planes in a crystal lattice. These planes may be used to describe the surface of a semiconductor material. The Miller indices are also used to describe directions in a crystal.
- Imperfections do exist in semiconductor materials. A few of these imperfections are vacancies, substitutional impurities, and interstitial impurities. Small amounts of controlled substitutional impurities can favorably alter semiconductor properties as we will see in later chapters.
- A brief description of semiconductor growth methods was given. Bulk growth, such as the Czochralski method, produces the starting semiconductor material or substrate. Epitaxial growth can be used to control the surface properties of a semiconductor. Most semiconductor devices are fabricated in the epitaxial layer.

GLOSSARY OF IMPORTANT TERMS

binary semiconductor A two-element compound semiconductor, such as gallium arsenide (GaAs).

covalent bonding The bonding between atoms in which valence electrons are shared.

diamond lattice The atomic crystal structure of silicon, for example, in which each atom has four nearest neighbors in a tetrahedral configuration.

doping The process of adding specific types of atoms to a semiconductor to favorably alter the electrical characteristics.

elemental semiconductor A semiconductor composed of a single species of atom, such as silicon or germanium.

epitaxial layer A thin, single-crystal layer of material formed on the surface of a substrate.

ion implantation One particular process of doping a semiconductor.

lattice The periodic arrangement of atoms in a crystal.

Miller indices The set of integers used to describe a crystal plane.

primitive cell The smallest unit cell that can be repeated to form a lattice.

substrate A semiconductor wafer or other material used as the starting material for further semiconductor processing, such as epitaxial growth or diffusion.

ternary semiconductor A three-element compound semiconductor, such as aluminum gallium arsenide (AlGaAs).

unit cell A small volume of a crystal that can be used to reproduce the entire crystal.

zincblende lattice A lattice structure identical to the diamond lattice except that there are two types of atoms instead of one.

CHECKPOINT

After studying this chapter, the reader should have the ability to:

- List the most common elemental semiconductor material.
- Describe the concept of a unit cell.
- Determine the volume density of atoms for various lattice structures.
- Determine the Miller indices of a crystal-lattice plane.
- Sketch a lattice plane given the Miller indices.
- Determine the surface density of atoms on a given crystal-lattice plane.
- Describe the tetrahedral configuration of silicon atoms.
- Understand and describe various defects in a single-crystal lattice.

REVIEW QUESTIONS

1. List two elemental semiconductor materials and two compound semiconductor materials.

2. Sketch three lattice structures: (*a*) simple cubic, (*b*) body-centered cubic, and (*c*) face-centered cubic.

3. Describe the procedure for finding the volume density of atoms in a crystal.

4. Describe the procedure for obtaining the Miller indices that describe a plane in a crystal.

5. Describe the procedure for finding the surface density of atoms on a particular lattice plane.

6. Describe why a unit cell, that is not a primitive unit cell, might be preferable to a primitive unit cell.

7. Describe covalent bonding in silicon.

8. What is meant by a substitutional impurity in a crystal? What is meant by an interstitial impurity?

PROBLEMS

Section 1.3 Space Lattices

1.1 Determine the number of atoms per unit cell in a (*a*) face-centered cubic, (*b*) body-centered cubic, and (*c*) diamond lattice.

1.2 Assume that each atom is a hard sphere with the surface of each atom in contact with the surface of its nearest neighbor. Determine the percentage of total unit cell volume

that is occupied in (*a*) a simple cubic lattice, (*b*) a face-centered cubic lattice, (*c*) a body-centered cubic lattice, and (*d*) a diamond lattice.

1.3 If the lattice constant of silicon is 5.43 Å, calculate (*a*) the distance from the center of one silicon atom to the center of its nearest neighbor, (*b*) the number density of silicon atoms (#/cm^3), and (*c*) the mass density (g/cm^3) of silicon.

1.4 (*a*) The lattice constant of GaAs is 5.65 Å. Determine the number of Ga atoms and As atoms per cm^3. (*b*) Determine the volume density of germanium atoms in a germanium semiconductor. The lattice constant of germanium is 5.65 Å.

1.5 The lattice constant of GaAs is $a = 5.65$ Å. Calculate (*a*) the distance between the centers of the nearest Ga and As atoms, and (*b*) the distance between the centers of the nearest As atoms.

1.6 Calculate the angle between any pair of bonds in the tetrahedral structure.

1.7 Assume the radius of an atom, which can be represented as a hard sphere, is $r = 1.95$ Å. The atom is placed in a (*a*) simple cubic, (*b*) fcc, (*c*) bcc, and (*d*) diamond lattice. Assuming that nearest atoms are touching each other, what is the lattice constant of each lattice?

1.8 A crystal is composed of two elements, A and B. The basic crystal structure is a face-centered cubic with element A at each of the corners and element B in the center of each face. The effective radius of element A is $r_A = 1.035$ Å. Assume that the elements are hard spheres with the surface of each A-type atom in contact with the surface of its nearest A-type neighbor. Calculate (*a*) the maximum radius of the B-type element that will fit into this structure, (*b*) the lattice constant, and (*c*) the volume density (#/cm^3) of both the A-type atoms and the B-type atoms.

1.9 (*a*) A crystal with a simple cubic lattice structure is composed of atoms with an effective radius of $r = 2.25$ Å and has an atomic weight of 12.5. Determine the mass density assuming the atoms are hard spheres and nearest neighbors are touching each other. (*b*) Repeat part (*a*) for a body-centered cubic structure.

1.10 A material, with a volume of 1 cm^3, is composed of an fcc lattice with a lattice constant of 2.5 mm. The "atoms" in this material are actually coffee beans. Assume the coffee beans are hard spheres with each bean touching its nearest neighbor. Determine the volume of coffee after the coffee beans have been ground. (Assume 100% packing density of the ground coffee.)

1.11 The crystal structure of sodium chloride (NaCl) is a simple cubic with the Na and Cl atoms alternating positions. Each Na atom is then surrounded by six Cl atoms and likewise each Cl atom is surrounded by six Na atoms. (*a*) Sketch the atoms in a (100) plane. (*b*) Assume the atoms are hard spheres with nearest neighbors touching. The effective radius of Na is 1.0 Å and the effective radius of Cl is 1.8 Å. Determine the lattice constant. (*c*) Calculate the volume density of Na and Cl atoms. (*d*) Calculate the mass density of NaCl.

1.12 (*a*) A material is composed of two types of atoms. Atom A has an effective radius of 2.2 Å and atom B has an effective radius of 1.8 Å. The lattice is a bcc with atoms A at the corners and atom B in the center. Determine the lattice constant and the volume densities of A atoms and B atoms. (*b*) Repeat part (*a*) with atoms B at the corners and atom A in the center. (*c*) What comparison can be made of the materials in parts (*a*) and (*b*)?

1.13 (*a*) Consider the materials described in Problem 1.12(*a*) and 1.12(*b*). For each case, calculate the surface density of A atoms and B atoms in the (100) plane. What comparison can be made of the two materials? (*b*) Repeat part (*a*) for the (110) plane.

1.14 (*a*) The crystal structure of a particular material consists of a single atom in the center of a cube. The lattice constant is a_0 and the diameter of the atom is a_0. Determine the volume density of atoms and the surface density of atoms in the (110) plane. (*b*) Compare the results of part (*a*) to the results for the case of the simple cubic structure shown in Figure 1.5a with the same lattice constant.

1.15 The lattice constant of a simple cubic lattice is a_0. (*a*) Sketch the following planes: (i) (110), (ii) (111), (iii) (220), and (iv) (321). (*b*) Sketch the following directions: (i) [110], (ii) [111], (iii) [220], and (iv) [321].

1.16 For a simple cubic lattice, determine the Miller indices for the planes shown in Figure P1.16.

1.17 A body-centered cubic lattice has a lattice constant of 4.83 Å. A plane cutting the lattice has intercepts of 9.66 Å, 19.32 Å, and 14.49 Å along the three cartesian coordinates. What are the Miller indices of the plane?

1.18 The lattice constant of a simple cubic primitive cell is 5.28 Å. Determine the distance between the nearest parallel (*a*) (100), (*b*) (110), and (*c*) (111) planes.

1.19 The lattice constant of a single crystal is 4.73 Å. Calculate the surface density (#/cm^2) of atoms on the (i) (100), (ii) (110), and (iii) (111) plane for a (*a*) simple cubic, (*b*) body-centered cubic, and (*c*) face-centered cubic lattice.

1.20 Determine the surface density of atoms for silicon on the (*a*) (100) plane, (*b*) (110) plane, and (*c*) (111) plane.

1.21 Consider a face-centered cubic lattice. Assume the atoms are hard spheres with the surfaces of the nearest neighbors touching. Assume the effective radius of the atom is 2.37 Å. (*a*) Determine the volume density of atoms in the crystal. (*b*) Calculate the surface density of atoms in the (110) plane. (*c*) Determine the distance between nearest (110) planes. (*d*) Repeat parts (*b*) and (*c*) for the (111) plane.

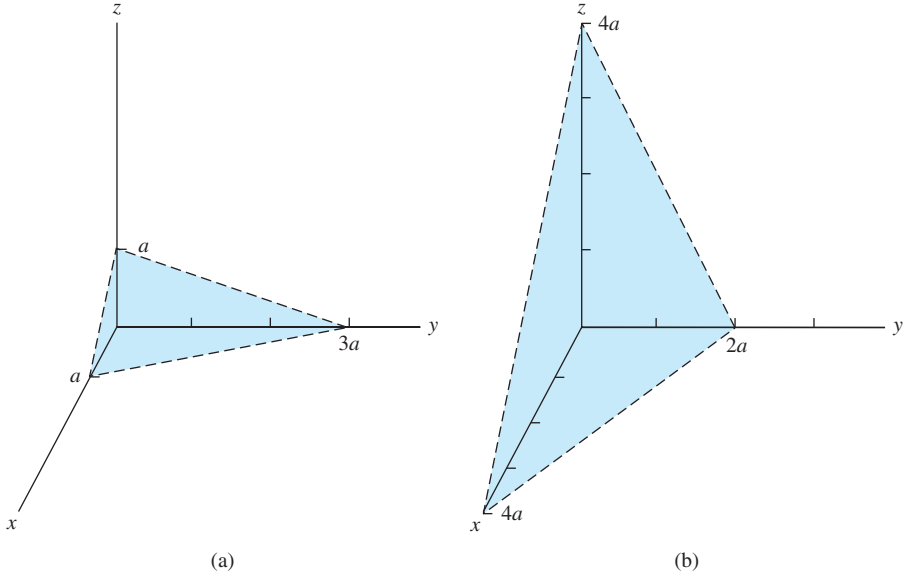


Figure P1.16 | Figure for Problem 1.16.

Section 1.5 Atomic Bonding

1.22 Calculate the density of valence electrons in silicon.

1.23 The structure of GaAs is the zincblende lattice. The lattice constant is 5.65 Å. Calculate the density of valence electrons in GaAs.

Section 1.6 Imperfections and Impurities in Solids

1.24 (*a*) If 5×10^{17} phosphorus atoms per cm^3 are add to silicon as a substitutional impurity, determine the percentage of silicon atoms per unit volume that are displaced in the single crystal lattice. (*b*) Repeat part (a) for 2×10^{15} boron atoms per cm^3 added to silicon.

1.25 (*a*) Assume that 2×10^{16} cm^{-3} of boron atoms are distributed homogeneously throughout single crystal silicon. What is the fraction by weight of boron in the crystal? (*b*) If phosphorus atoms, at a concentration of 10^{18} cm^{-3}, are added to the material in part (*a*), determine the fraction by weight of phosphorus.

1.26 If 2×10^{16} cm^{-3} boron atoms are added to silicon as a substitutional impurity and are distributed uniformly throughout the semiconductor, determine the distance between boron atoms in terms of the silicon lattice constant. (Assume the boron atoms are distributed in a rectangular or cubic array.)

1.27 Repeat Problem 1.26 for 4×10^{15} cm^{-3} phosphorus atoms being added to silicon.

READING LIST

1. Azaroff, L.V., and J. J. Brophy. *Electronic Processes in Materials.* New York: McGraw-Hill, 1963.

2. Campbell, S. A. *The Science and Engineering of Microelectronic Fabrication.* New York: Oxford University Press, 1996.

3. Dimitrijev, S. *Principles of Semiconductor Devices*. New York: Oxford University Press, 2006.

4. Kittel, C. *Introduction to Solid State Physics,* 7th ed. Berlin: Springer-Verlag, 1993.

***5.** Li, S. S. *Semiconductor Physical Electronics.* New York: Plenum Press, 1993.

6. McKelvey, J. P. *Solid State Physics for Engineering and Materials Science.* Malabar, FL: Krieger, 1993.

7. Pierret, R. F. *Semiconductor Device Fundamentals.* Reading, MA: Addison-Wesley, 1996.

8. Runyan, W. R., and K. E. Bean. *Semiconductor Integrated Circuit Processing and Technology.* Reading, MA: Addison-Wesley, 1990.

9. Singh, J. *Semiconductor Devices: Basic Principles.* New York: John Wiley and Sons, 2001.

10. Streetman, B. G., and S. K. Banerjee. *Solid State Electronic Devices,* 6th ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2006.

11. Sze, S. M. *VLSI Technology.* New York: McGraw-Hill, 1983.

***12.** Wolfe, C. M., N. Holonyak, Jr., and G. E. Stillman. *Physical Properties of Semiconductors.* Englewood Cliffs, NJ: Prentice Hall, 1989.

*Indicates references that are at an advanced level compared to this text.

CHAPTER 2

Introduction to Quantum Mechanics

The goal of this text is to help readers understand the operation and characteristics of semiconductor devices. Ideally, we would like to begin discussing these devices immediately. However, in order to understand the current–voltage characteristics, we need some knowledge of electron behavior in a semiconductor when the electron is subjected to various potential functions.

The motion of large objects, such as planets and satellites, can be predicted to a high degree of accuracy using classical theoretical physics based on Newton's laws of motion. But certain experimental results, involving electrons and light waves, appear to be inconsistent with classical physics. However, these experimental results can be predicted using the principles of quantum mechanics. The behavior and characteristics of electrons in a semiconductor can be described by the formulation of quantum mechanics called wave mechanics. The essential elements of this wave mechanics, using Schrodinger's wave equation, are presented in this chapter.

The goal of this chapter is to provide a brief introduction to quantum mechanics so that readers gain an understanding of and become comfortable with the analysis techniques. This introductory material forms the basis of semiconductor physics. ■

2.0 | PREVIEW

In this chapter, we will:

- Discuss a few basic principles of quantum mechanics that apply to semiconductor device physics.
- State Schrodinger's wave equation and discuss the physical meaning of the wave function.
- Consider the application of Schrodinger's wave equation to various potential functions to determine some of the fundamental properties of electron behavior in a crystal.

- Apply Schrodinger's wave equation to the one-electron atom. The result of this analysis yields the four basic quantum numbers, the concept of discrete energy bands, and the initial buildup of the periodic table.

2.1 | PRINCIPLES OF QUANTUM MECHANICS

Before we delve into the mathematics of quantum mechanics, there are three principles we need to consider: the principle of energy quanta, the wave–particle duality principle, and the uncertainty principle.

2.1.1 Energy Quanta

One experiment that demonstrates an inconsistency between experimental results and the classical theory of light is called the photoelectric effect. If monochromatic light is incident on a clean surface of a material, then under certain conditions, electrons (photoelectrons) are emitted from the surface. According to classical physics, if the intensity of the light is large enough, the work function of the material will be overcome and an electron will be emitted from the surface independent of the incident frequency. This result is not observed. The observed effect is that at a constant incident intensity, the maximum kinetic energy of the photoelectron varies linearly with frequency with a limiting frequency $\nu = \nu_0$, below which no photoelectron is produced. This result is shown in Figure 2.1. If the incident intensity varies at a constant frequency, the rate of photoelectron emission changes, but the maximum kinetic energy remains the same.

Planck postulated in 1900 that thermal radiation is emitted from a heated surface in discrete packets of energy called *quanta*. The energy of these quanta is given by $E = h\nu$, where ν is the frequency of the radiation and h is a constant now known as Planck's constant ($h = 6.625 \times 10^{-34}$ J-s). Then in 1905, Einstein interpreted the photoelectric results by suggesting that the energy in a light wave is also contained

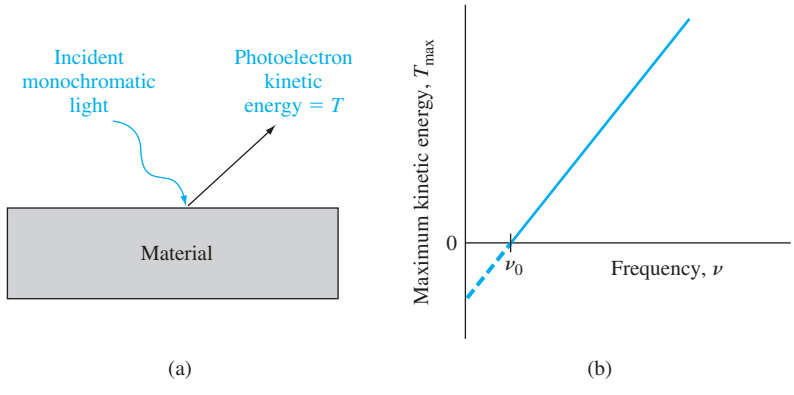


Figure 2.1 | (a) The photoelectric effect and (b) the maximum kinetic energy of the photoelectron as a function of incident frequency.

in discrete packets or bundles. The particle-like packet of energy is called a *photon,* whose energy is also given by $E = h\nu$. A photon with sufficient energy, then, can knock an electron from the surface of the material. The minimum energy required to remove an electron is called the *work function* of the material and any excess photon energy goes into the kinetic energy of the photoelectron. This result was confirmed experimentally as demonstrated in Figure 2.1. The photoelectric effect shows the discrete nature of the photon and demonstrates the particle-like behavior of the photon.

The maximum kinetic energy of the photoelectron can be written as

$$T = \frac{1}{2}mv^2 = h\nu - \Phi = h\nu - h\nu_0 \qquad (\nu \geq \nu_0) \tag{2.1}$$

$$\Phi = h\nu_0$$

where $h\nu$ is the incident photon energy and $\Phi = h\nu_0$ is the minimum energy, or work function, required to remove an electron from the surface.

EXAMPLE 2.1

Objective: Calculate the photon energy corresponding to a particular wavelength.

Consider an x-ray with a wavelength of $\lambda = 0.708 \times 10^{-8}$ cm.

■ Solution

The energy is

$$E = h\nu = \frac{hc}{\lambda} = \frac{(6.625 \times 10^{-34})(3 \times 10^{10})}{0.708 \times 10^{-8}} = 2.81 \times 10^{-15}\ \text{J}$$

This value of energy may be given in the more common unit of electron-volt (see Appendix D). We have

$$E = \frac{2.81 \times 10^{-15}}{1.6 \times 10^{-19}} = 1.75 \times 10^{4}\ \text{eV}$$

■ Comment

The reciprocal relation between photon energy and wavelength is demonstrated: A large energy corresponds to a short wavelength.

■ EXERCISE PROBLEM

Ex 2.1 Determine the energy (in eV) of a photon having a wavelength of (*a*) $\lambda = 100$ Å and (*b*) $\lambda = 4500$ Å. [Ans. (*a*) 124 eV; (*b*) 2.76 eV]

2.1.2 Wave–Particle Duality

We have seen in the last section that light waves, in the photoelectric effect, behave as if they are particles. The particle-like behavior of electromagnetic waves was also instrumental in the explanation of the Compton effect. In this experiment, an x-ray beam was incident on a solid. A portion of the x-ray beam was deflected and the frequency of the deflected wave had shifted compared with the incident wave. The observed change in frequency and the deflected angle corresponded exactly to the expected results of a "billiard ball" collision between an x-ray quanta, or photon, and an electron in which both energy and momentum are conserved.

In 1924, de Broglie postulated the existence of matter waves. He suggested that since waves exhibit particle-like behavior, particles should be expected to show wave-like properties. The hypothesis of de Broglie was the existence of a *wave–particle duality principle*. The momentum of a photon is given by

$$p = \frac{h}{\lambda} \tag{2.2}$$

where λ is the wavelength of the light wave. Then, de Broglie hypothesized that the wavelength of a particle can be expressed as

$$\lambda = \frac{h}{p} \tag{2.3}$$

where p is the momentum of the particle and λ is known as the *de Broglie wavelength* of the matter wave.

The wave nature of electrons has been tested in several ways. In one experiment by Davisson and Germer in 1927, electrons from a heated filament were accelerated at normal incidence onto a single crystal of nickel. A detector measured the scattered electrons as a function of angle. Figure 2.2 shows the experimental setup and Figure 2.3 shows the results. The existence of a peak in the density of scattered electrons can be explained as a constructive interference of waves scattered by the periodic atoms in the planes of the nickel crystal. The angular distribution is very similar to an interference pattern produced by light diffracted from a grating.

Exemplo do comportamento ondulatório dos elétrons:
1) MeV;
2) FIB.

In order to gain some appreciation of the frequencies and wavelengths involved in the wave-particle duality principle, Figure 2.4 shows the electromagnetic frequency spectrum. We see that a wavelength of 72.7 Å obtained in the next example is in the ultraviolet range. Typically, we will be considering wavelengths in the

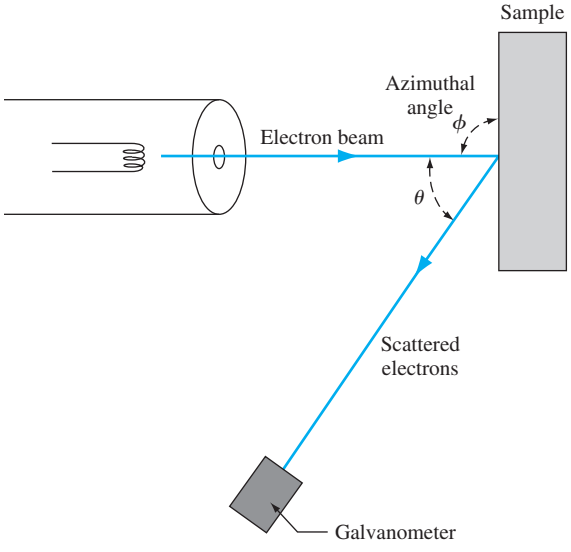


Figure 2.2 | Experimental arrangement of the Davisson–Germer experiment.

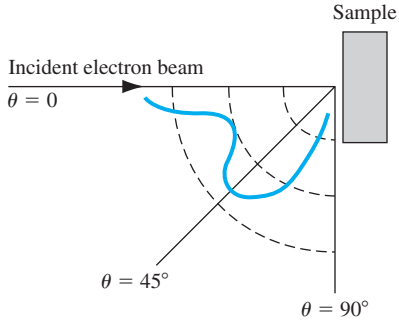


Figure 2.3 | Scattered electron flux as a function of scattering angle for the Davisson–Germer experiment.

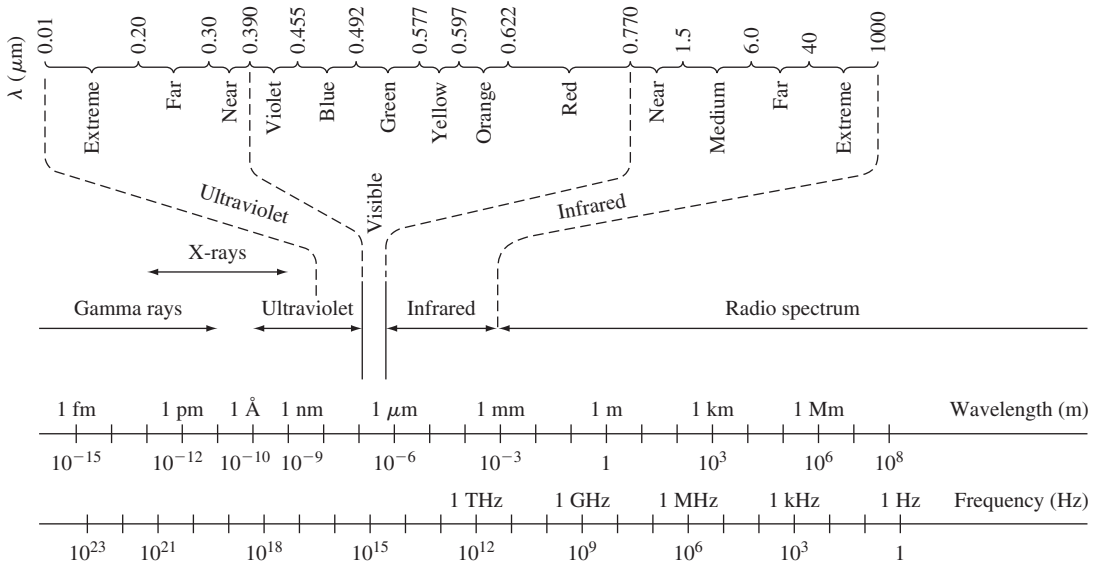


Figure 2.4 | The electromagnetic frequency spectrum.

ultraviolet and visible range. These wavelengths are very short compared to the usual radio spectrum range.[1]

EXAMPLE 2.2

Objective: Calculate the de Broglie wavelength of a particle.

Consider an electron traveling at a velocity of 10^7 cm/s $= 10^5$ m/s.

■ Solution

The momentum is given by

$$p = mv = (9.11 \times 10^{-31})(10^5) = 9.11 \times 10^{-26} \text{ kg-m/s}$$

Then, the de Broglie wavelength is

$$\lambda = \frac{h}{p} = \frac{6.625 \times 10^{-34}}{9.11 \times 10^{-26}} = 7.27 \times 10^{-9} \text{ m}$$

or

$$\lambda = 72.7 \text{ Å}$$

■ Comment

This calculation shows the order of magnitude of the de Broglie wavelength for a "typical" electron.

■ EXERCISE PROBLEM

Ex 2.2 (*a*) An electron has a kinetic energy of 12 meV. Determine the de Broglie wavelength (in Å), (*b*) A particle with mass 2.2×10^{-31} kg has a de Broglie wavelength of 112 Å. Determine the momentum and kinetic energy of the particle. [Ans. (*a*) $\lambda = 112$ Å; (*b*) $p = 5.915 \times 10^{-26}$ kg-m/s, $E = 4.97 \times 10^{-2}$ eV]

[1]An electron microscope is a microscope that produces a magnified image of a specimen. The electron microscope has a magnification approximately 1000 times that of an optical microscope, because the electrons have wavelengths on the order of 100,000 times shorter than the light waves.

In some cases, electromagnetic waves behave as if they are particles (photons) and sometimes particles behave as if they are waves. This wave–particle duality principle of quantum mechanics applies primarily to small particles such as electrons, but it has also been shown to apply to protons and neutrons. For very large particles, we can show that the relevant equations reduce to those of classical mechanics. The wave–particle duality principle is the basis on which we will use wave theory to describe the motion and behavior of electrons in a crystal.

2.1.3 The Uncertainty Principle

The Heisenberg uncertainty principle, given in 1927, also applies primarily to very small particles and states that we cannot describe with absolute accuracy the behavior of these subatomic particles. The uncertainty principle describes a fundamental relationship between conjugate variables, including position and momentum and also energy and time.

The first statement of the uncertainty principle is that it is impossible to simultaneously describe with absolute accuracy the position and momentum of a particle. If the uncertainty in the momentum is Δp and the uncertainty in the postion is Δx, then the uncertainty principle is stated as[2]

$$\Delta p \, \Delta x \geq \hbar \tag{2.4}$$

where $\hbar$ is defined as $\hbar = h/2\pi = 1.054 \times 10^{-34}$ J-s and is called a modified Planck's constant. This statement may be generalized to include angular position and angular momentum.

The second statement of the uncertainty principle is that it is impossible to simultaneously describe with absolute accuracy the energy of a particle and the instant of time the particle has this energy. Again, if the uncertainty in the energy is given by ΔE and the uncertainty in the time is given by Δt, then the uncertainty principle is stated as

$$\Delta E \, \Delta t \geq \hbar \tag{2.5}$$

One way to visualize the uncertainty principle is to consider the simultaneous measurement of position and momentum, and the simultaneous measurement of energy and time. The uncertainty principle implies that these simultaneous measurements are in error to a certain extent. However, the modified Planck's constant $\hbar$ is very small; the uncertainty principle is only significant for subatomic particles. We must keep in mind nevertheless that the uncertainty principle is a fundamental statement and does not deal only with measurements.

One consequence of the uncertainty principle is that we cannot, for example, determine the exact position of an electron. We will, instead, determine the *probability* of finding an electron at a particular position. In later chapters, we will develop a

[2]In some texts, the uncertainty principle is stated as $\Delta p \, \Delta x \geq \hbar/2$. We are interested here in the order of magnitude and will not be concerned with small differences.

probability density function that will allow us to determine the probability that an electron has a particular energy. So in describing electron behavior, we will be dealing with probability functions.

TEST YOUR UNDERSTANDING

TYU 2.1 The uncertainty in position of an electron is $\Delta x = 8$ Å. (*a*) Determine the minimum uncertainly in momentum. (*b*) If the nominal value of momentum is $p = 1.2 \times 10^{-23}$ kg-m/s, determine the corresponding uncertainty in kinetic energy. (The uncertainty in kinetic energy can be found from $\Delta E = (dE/dp)\Delta p = (p\,\Delta p/m)$ [Ans. (*a*) $\Delta p = 1.318 \times 10^{-25}$ kg-m/s; (*b*) $\Delta E = 10.85$ eV]

TYU 2.2 (*a*) A proton's energy is measured with an uncertainty of 0.8 eV. Determine the minimum uncertainty in time over which this energy is measured. (*b*) Repeat part (*a*) for an electron. [Ans. (*a*) $\Delta t = 8.23 \times 10^{-16}$ s; (*b*) same as part (*a*)]

2.2 | SCHRODINGER'S WAVE EQUATION

The various experimental results involving electromagnetic waves and particles, which could not be explained by classical laws of physics, showed that a revised formulation of mechanics was required. Schrodinger, in 1926, provided a formulation called *wave mechanics,* which incorporated the principles of quanta introduced by Planck, and the wave–particle duality principle introduced by de Broglie. On the basis of wave–particle duality principle, we will describe the motion of electrons in a crystal by wave theory. This wave theory is described by Schrodinger's wave equation.

2.2.1 The Wave Equation

The one-dimensional, nonrelativistic Schrodinger's wave equation is given by

$$\frac{-\hbar^2}{2m}\cdot\frac{\partial^2\Psi(x,t)}{\partial x^2}+V(x)\Psi(x,t)=j\hbar\frac{\partial\Psi(x,t)}{\partial t} \tag{2.6}$$

where $\Psi(x, t)$ is the wave function, $V(x)$ is the potential function assumed to be independent of time, m is the mass of the particle, and j is the imaginary constant $\sqrt{-1}$. There are theoretical arguments that justify the form of Schrodinger's wave equation, but the equation is a basic postulate of quantum mechanics. The wave function $\Psi(x, t)$ will be used to describe the behavior of the system and, mathematically, $\Psi(x, t)$ can be a complex quantity.

We may determine the time-dependent portion of the wave function and the position-dependent, or time-independent, portion of the wave function by using the technique of separation of variables. Assume that the wave function can be written in the form

$$\Psi(x,t)=\psi(x)\phi(t) \tag{2.7}$$

where $\psi(x)$ is a function of the position x only and $\phi(t)$ is a function of time t only. Substituting this form of the solution into Schrodinger's wave equation, we obtain

$$\frac{-\hbar^2}{2m}\phi(t)\frac{\partial^2\psi(x)}{\partial x^2} + V(x)\psi(x)\phi(t) = j\hbar\psi(x)\frac{\partial\phi(t)}{\partial t} \tag{2.8}$$

If we divide by the total wave function, Equation (2.8) becomes

$$\frac{-\hbar^2}{2m}\frac{1}{\psi(x)}\frac{\partial^2\psi(x)}{\partial x^2} + V(x) = j\hbar\cdot\frac{1}{\phi(t)}\cdot\frac{\partial\phi(t)}{\partial t} \tag{2.9}$$

Since the left side of Equation (2.9) is a function of position x only and the right side of the equation is a function of time t only, each side of this equation must be equal to a constant. We will denote this separation of variables constant by η.

The time-dependent portion of Equation (2.9) is then written as

$$\eta = j\hbar\cdot\frac{1}{\phi(t)}\cdot\frac{\partial\phi(t)}{\partial t} \tag{2.10}$$

where again the parameter η is called a separation constant. The solution of Equation (2.10) can be written in the form

$$\phi(t) = e^{-j(\eta/\hbar)t} \tag{2.11a}$$

The form of this solution is the classical exponential form of a sinusoidal wave. We have that $E = h\nu$ or $E = h\omega/2\pi$. Then $\omega = \eta/\hbar = E/\hbar$ so that the separation constant is equal to the total energy E of the particle.

We can then write

$$\phi(t) = e^{-j(E/\hbar)t} = e^{-j\omega t} \tag{2.11b}$$

We see that $\omega = E/\hbar$ and is the radian or angular frequency of the sinusoidal wave.

The time-independent portion of Schrodinger's wave equation can now be written from Equation (2.9) as

$$\frac{-\hbar^2}{2m}\cdot\frac{1}{\psi(x)}\cdot\frac{\partial^2\psi(x)}{\partial x^2} + V(x) = E \tag{2.12}$$

where the separation constant is the total energy E of the particle. Equation (2.12) may be written as

$$\boxed{\frac{\partial^2\psi(x)}{\partial x^2} + \frac{2m}{\hbar^2}(E - V(x))\psi(x) = 0} \tag{2.13}$$

where again m is the mass of the particle, $V(x)$ is the potential experienced by the particle, and E is the total energy of the particle. This time-independent Schrodinger's wave equation can also be justified on the basis of the classical wave equation as shown in Appendix E. The pseudo-derivation in the appendix is a simple approach but shows the plausibility of the time-independent Schrodinger's equation.

2.2.2 Physical Meaning of the Wave Function

We are ultimately trying to use the wave function $\Psi(x, t)$ to describe the behavior of an electron in a crystal. The function $\Psi(x, t)$ is a wave function, so it is reasonable to

ask what the relation is between the function and the electron. The total wave function is the product of the position-dependent, or time-independent, function and the time-dependent function. From Equation (2.7) we have

$$\Psi(x, t) = \psi(x)\phi(t) = \psi(x)e^{-j(E/\hbar)t} = \psi(x)e^{-j\omega t} \tag{2.14}$$

Since the total wave function $\Psi(x, t)$ is a complex function, it cannot by itself represent a real physical quantity.

Max Born postulated in 1926 that the function $|\Psi(x, t)|^2\,dx$ is the probability of finding the particle between x and $x + dx$ at a given time, or that $|\Psi(x, t)|^2$ is a probability density function. We have that

$$|\Psi(x, t)|^2 = \Psi(x, t) \cdot \Psi^*(x, t) \tag{2.15}$$

where $\Psi^*(x, t)$ is the complex conjugate function. Therefore

$$\Psi^*(x, t) = \psi^*(x) \cdot e^{+j(E/\hbar)t}$$

Then the product of the total wave function and its complex conjugate is given by

$$\Psi(x, t)\Psi^*(x, t) = \left[\psi(x)e^{-j(E/\hbar)t}\right]\left[\psi^*(x)e^{+j(E/\hbar)t}\right] = \psi(x)\psi^*(x) \tag{2.16}$$

Therefore, we have that

$$|\Psi(x, t)|^2 = \psi(x)\psi^*(x) = |\psi(\chi)|^2 \tag{2.17}$$

is the probability density function and is independent of time. One major difference between classical and quantum mechanics is that in classical mechanics, the position of a particle or body can be determined precisely, whereas in quantum mechanics, the position of a particle is found in terms of a probability. We will determine the probability density function for several examples, and since this property is independent of time, we will, in general, only be concerned with the time-independent wave function.

2.2.3 Boundary Conditions

Since the function $|\psi(x)|^2$ represents the probability density function, then for a single particle, we must have

$$\int_{-\infty}^{\infty} |\psi(x)|^2\,dx = 1 \tag{2.18}$$

The probability of finding the particle somewhere is certain. Equation (2.18) allows us to normalize the wave function and is one boundary condition that is used to determine some wave function coefficients.

The remaining boundary conditions imposed on the wave function and its derivative are postulates. However, we may state the boundary conditions and present arguments that justify why they must be imposed. The wave function and its first derivative must have the following properties if the total energy E and the potential $V(x)$ are finite everywhere.

Condition 1. $\psi(x)$ must be finite, single-valued, and continuous.

Condition 2. $\partial\psi(x)/\partial x$ must be finite, single-valued, and continuous.

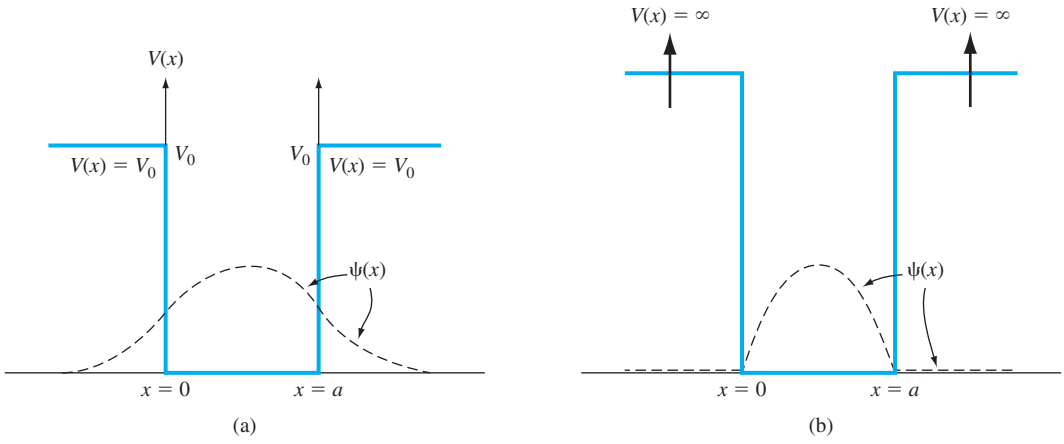


Figure 2.5 | Potential functions and corresponding wave function solutions for the case (a) when the potential function is finite everywhere and (b) when the potential function is infinite in some regions.

Since $|\psi(x)|^2$ is a probability density, then $\psi(x)$ must be finite and single-valued. If the probability density were to become infinite at some point in space, then the probability of finding the particle at this position would be certain and the uncertainty principle would be violated. If the total energy E and the potential $V(x)$ are finite everywhere, then from Equation (2.13), the second derivative must be finite, which implies that the first derivative must be continuous. The first derivative is related to the particle momentum, which must be finite and single-valued. Finally, a finite first derivative implies that the function itself must be continuous. In some of the specific examples that we will consider, the potential function will become infinite in particular regions of space. For these cases, the first derivative will not necessarily be continuous, but the remaining boundary conditions will still hold.

Figure 2.5 shows two possible examples of potential functions and the corresponding wave solutions. In Figure 2.5a, the potential function is finite everywhere. The wave function as well as its first derivative is continuous. In Figure 2.5b, the potential function is infinite for $x < 0$ and for $x > a$. The wave function is continuous at the boundaries, but the first derivative is discontinuous. We will actually determine the wave functions in the following sections and in end-of-chapter problems.

2.3 | APPLICATIONS OF SCHRODINGER'S WAVE EQUATION

We will now apply Schrodinger's wave equation in several examples using various potential functions. These examples will demonstrate the techniques used in the solution of Schrodinger's differential equation and the results of these examples will provide an indication of the electron behavior under these various potentials.

We will utilize the resulting concepts later in the discussion of semiconductor properties.

2.3.1 Electron in Free Space

As a first example of applying the Schrodinger's wave equation, consider the motion of an electron in free space. If there is no force acting on the particle, then the potential function $V(x)$ will be constant and we must have $E > V(x)$. Assume, for simplicity, that the potential function $V(x) = 0$ for all x. Then, the time-independent wave equation can be written from Equation (2.13) as

$$\frac{\partial^2\psi(x)}{\partial x^2} + \frac{2mE}{\hbar^2}\psi(x) = 0 \tag{2.19}$$

The solution to this differential equation can be written in the form

$$\psi(x) = A\exp\left[\frac{jx\sqrt{2mE}}{\hbar}\right] + B\exp\left[\frac{-jx\sqrt{2mE}}{\hbar}\right] \tag{2.20a}$$

or

$$\psi(x) = A\exp(jkx) + B\exp(-jkx) \tag{2.20b}$$

where

$$k = \sqrt{\frac{2mE}{\hbar^2}} \tag{2.21}$$

and is called a wave number.

Recall that the time-dependent portion of the solution is

$$\phi(t) = e^{-j(E/\hbar)t} = e^{-j\omega t} \tag{2.22}$$

Then the total solution for the wave function is given by

$$\Psi(x, t) = A\exp[j(kx - \omega t)] + B\exp[-j(kx + \omega t)] \tag{2.23}$$

This wave function solution is a traveling wave, which means that a particle moving in free space is represented by a traveling wave. The first term, with the coefficient A, is a wave traveling in the $+x$ direction, while the second term, with the coefficient B, is a wave traveling in the $-x$ direction. The value of these coefficients will be determined from boundary conditions. We will again see the traveling-wave solution for an electron in a crystal or semiconductor material.

Assume, for a moment, that we have a particle traveling in the $+x$ direction, which will be described by the $+x$ traveling wave. The coefficient $B = 0$. We can write the traveling-wave solution in the form

$$\Psi(x, t) = A\exp[j(kx - \omega t)] \tag{2.24}$$

where k is the wave number given by

$$k = \sqrt{\frac{2mE}{\hbar^2}} = \sqrt{\frac{p^2}{\hbar^2}} = \frac{p}{\hbar} \tag{2.25a}$$

or

$$p = \hbar k \tag{2.25b}$$

Also recall that the de Broglie wavelength was given by

$$\lambda = \frac{h}{p} = \frac{2\pi\hbar}{p} \tag{2.26}$$

Combining Equations (2.25a) and (2.26), the wavelength can also be written in terms of the wave number as

$$\lambda = \frac{2\pi}{k} \tag{2.27a}$$

or

$$k = \frac{2\pi}{\lambda} \tag{2.27b}$$

A free particle with a well-defined energy will also have a well-defined wavelength and momentum.

The probability density function is $\Psi(x, t)\Psi^*(x, t) = AA^*$, which is a constant independent of position. A free particle with a well-defined momentum can be found anywhere with equal probability. This result is in agreement with the Heisenberg uncertainty principle in which a precise momentum implies an undefined position.

A localized free particle is defined by a wave packet, formed by a superposition of wave functions with different momentum or k values. We will not consider the wave packet here.

2.3.2 The Infinite Potential Well

The problem of a particle in the infinite potential well is a classic example of a bound particle. The potential $V(x)$ as a function of position for this problem is shown in Figure 2.6. The particle is assumed to exist in region II, so the particle is contained within a finite region of space. The time-independent Schrodinger's wave equation is again given by Equation (2.13) as

$$\frac{\partial^2\psi(x)}{\partial x^2} + \frac{2m}{\hbar^2}(E - V(x))\psi(x) = 0 \tag{2.13}$$

where E is the total energy of the particle. If E is finite, the wave function must be zero, or $\psi(x) = 0$, in both regions I and III. A particle cannot penetrate these infinite potential barriers, so the probability of finding the particle in regions I and III is zero.

The time-independent Schrodinger's wave equation in region II, where $V = 0$, becomes

$$\frac{\partial^2\psi(x)}{\partial x^2} + \frac{2mE}{\hbar^2}\psi(x) = 0 \tag{2.28}$$

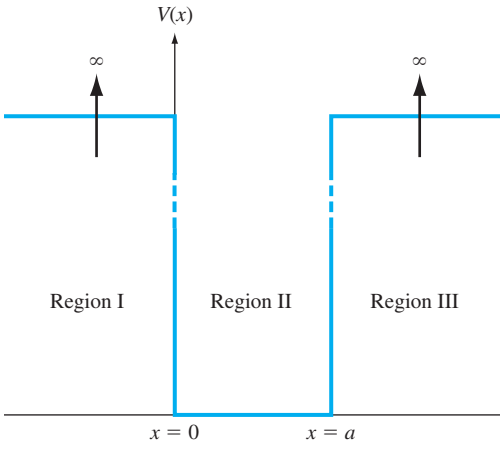


Figure 2.6 | Potential function of the infinite potential well.

A particular form of solution to this equation is given by

$$\psi(x) = A_1 \cos kx + A_2 \sin kx \tag{2.29}$$

where

$$k = \sqrt{\frac{2mE}{\hbar^2}} \tag{2.30}$$

One boundary condition is that the wave function $\psi(x)$ must be continuous so that

$$\psi(x = 0) = \psi(x = a) = 0 \tag{2.31}$$

Applying the boundary condition at $x = 0$, we must have that $A_1 = 0$. At $x = a$, we have

$$\psi(x = a) = 0 = A_2 \sin ka \tag{2.32}$$

This equation is valid if $ka = n\pi$, where the parameter n is a positive integer, or $n = 1, 2, 3, \ldots$. The parameter n is referred to as a quantum number. We can write

$$k = \frac{n\pi}{a} \tag{2.33}$$

Negative values of n simply introduce a negative sign in the wave function and yield redundant solutions for the probability density function. We cannot physically distinguish any difference between $+n$ and $-n$ solutions. Because of this redundancy, negative values of n are not considered.

The coefficient A_2 can be found from the normalization boundary condition that was given by Equation (2.18) as $\int_{-\infty}^{\infty} \psi(x)\psi^*(x)dx = 1$. If we assume that the wave

function solution $\psi(x)$ is a real function, then $\psi(x) = \psi^*(x)$. Substituting the wave function into Equation (2.18), we have

$$\int_0^a A_2^2 \sin^2 kx\, dx = 1 \tag{2.34}$$

Evaluating this integral gives[3]

$$A_2 = \sqrt{\frac{2}{a}} \tag{2.35}$$

Finally, the time-independent wave solution is given by

$$\psi(x) = \sqrt{\frac{2}{a}} \sin\left(\frac{n\pi x}{a}\right) \qquad \text{where } n = 1, 2, 3, \ldots \tag{2.36}$$

This solution represents the electron in the infinite potential well and is a standing wave solution. The free electron was represented by a traveling wave, and now the bound particle is represented by a standing wave.

The parameter k in the wave solution was defined by Equations (2.30) and (2.33). Equating these two expressions for k, we obtain

$$k^2 \rightarrow k_n^2 = \frac{2mE_n}{\hbar^2} = \frac{n^2\pi^2}{a^2} \tag{2.37}$$

The total energy can then be written as

$$E = E_n = \frac{\hbar^2 n^2 \pi^2}{2ma^2} \qquad \text{where } n = 1, 2, 3, \ldots \tag{2.38}$$

For the particle in the infinite potential well, the wave function is now given by

$$\psi(x) = \sqrt{\frac{2}{a}} \sin k_n x \tag{2.39}$$

where the constant k_n, from Equation (2.37), must have discrete values, implying that the total energy of the particle can only have discrete values. *This result means that the energy of the particle is* quantized. *That is, the energy of the particle can only have particular discrete values.* The quantization of the particle energy is contrary to results from classical physics, which would allow the particle to have continuous energy values. The discrete energies lead to quantum states that will be considered in more detail in this and later chapters. The quantization of the energy of a bound particle is an important result.

This quantization of electron energy will be observed again at the end of the chapter for an electron bound to an ion forming an atom.

[3]A more thorough analysis shows that $|A_2|^2 = 2/a$, so solutions for the coefficient A_2 include $+\sqrt{2/a}$. $-\sqrt{2/a}$, $+ j\sqrt{2/a}$, $-j\sqrt{2/a}$, or any complex number whose magnitude is $\sqrt{2/a}$. Since the wave function itself has no physical meaning, the choice of which coefficient to use is immaterial: They all produce the same probability density function.

EXAMPLE 2.3

Objective: Calculate the first three energy levels of an electron in an infinite potential well.

Consider an electron in an infinite potential well of width 5 Å.

■ Solution

From Equation (2.38) we have[4]

$$E_n = \frac{\hbar^2 n^2 \pi^2}{2ma^2} = \frac{n^2(1.054 \times 10^{-34})^2 \pi^2}{2(9.11 \times 10^{-31})(5 \times 10^{-10})^2} = n^2(2.41 \times 10^{-19}) \text{ J}$$

or

$$E_n = \frac{n^2(2.41 \times 10^{-19})}{1.6 \times 10^{-19}} = n^2(1.51) \text{ eV}$$

Then,

$$E_1 = 1.51 \text{ eV}, \quad E_2 = 6.04 \text{ eV}, \quad E_3 = 13.59 \text{ eV}$$

■ Comment

This calculation shows the order of magnitude of the energy levels of a bound electron.

■ EXERCISE PROBLEM

Ex 2.3 (*a*) The width of an infinite potential well is 12 Å. Determine the first three allowed energy levels (in eV) for an electron. (*b*) Repeat part (*a*) for a proton.

[Ans. (*a*) 0.2615eV, 1.045 eV, 2.351 eV; (*b*) 1.425×10^{-4} eV, 5.70×10^{-4} eV, 1.28×10^{-3} eV]

Figure 2.7a shows the first four allowed energies for the particle in the infinite potential well, and Figure 2.7b,c shows the corresponding wave functions and probability functions. We may note that as the energy increases, the probability of finding the particle at any given value of x becomes more uniform.

2.3.3 The Step Potential Function

Consider now a step potential function as shown in Figure 2.8. In the previous section, we considered a particle being confined between two potential barriers. In this example, we will assume that a flux of particles is incident on the potential barrier. We will assume that the particles are traveling in the $+x$ direction and that they originated at $x = -\infty$. A particularly interesting result is obtained for the case when the total energy of the particle is less than the barrier height, or $E < V_0$.

We again need to consider the time-independent wave equation in each of the two regions. This general equation was given in Equation (2.13) as $\partial^2\psi(x)/\partial x^2 + 2m/\hbar^2(E - V(x))\psi(x) = 0$. The wave equation in region I, in which $V = 0$, is

$$\frac{\partial^2 \psi_1(x)}{\partial x^2} + \frac{2mE}{\hbar^2}\psi_1(x) = 0 \tag{2.40}$$

[4]See Appendix D for a discussion of the electron-volt (eV) as a unit of energy.

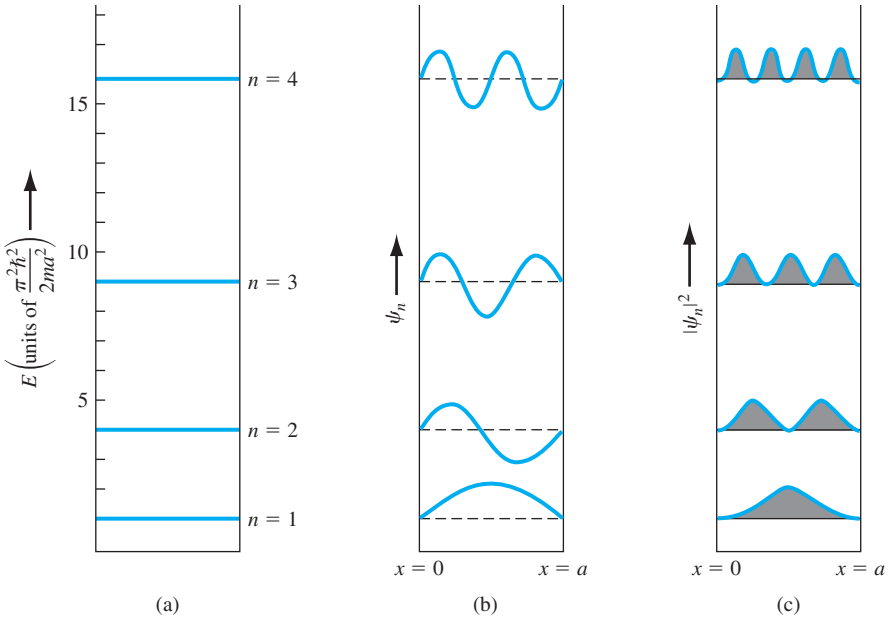


Figure 2.7 | Particle in an infinite potential well: (a) four lowest discrete energy levels, (b) corresponding wave functions, and (c) corresponding probability functions. *(From Pierret [10].)*

The general solution to this equation can be written in the form

$$\psi_1(x) = A_1 e^{jk_1 x} + B_1 e^{-jk_1 x} \quad (x \leq 0) \tag{2.41}$$

where the constant k_1 is

$$k_1 = \sqrt{\frac{2mE}{\hbar^2}} \tag{2.42}$$

The first term in Equation (2.41) is a traveling wave in the $+x$ direction that represents the incident wave, and the second term is a traveling wave in the $-x$ direction

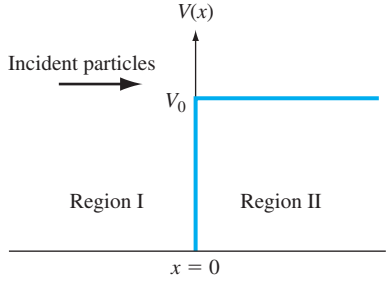


Figure 2.8 | The step potential function.

that represents a reflected wave. As in the case of a free particle, the incident and reflected particles are represented by traveling waves.

For the incident wave, $A_1 \cdot A_1^*$ is the probability density function of the incident particles. If we multiply this probability density function by the incident velocity, then $v_i \cdot A_1 \cdot A_1^*$ is the flux of incident particles in units of #/cm^2-s. Likewise, the quantity $v_r \cdot B_1 \cdot B_1^*$ is the flux of the reflected particles, where v_r is the velocity of the reflected wave. (The parameters v_i and v_r in these terms are actually the magnitudes of the velocity only.)

In region II, the potential is $V = V_0$. If we assume that $E < V_0$, then the differential equation describing the wave function in region II can be written as

$$\frac{\partial^2 \psi_2(x)}{\partial x^2} - \frac{2m}{\hbar^2}(V_0 - E)\psi_2(x) = 0 \tag{2.43}$$

The general solution may then be written in the form

$$\psi_2(x) = A_2 e^{-k_2 x} + B_2 e^{+k_2 x} \quad (x \geq 0) \tag{2.44}$$

where

$$k_2 = \sqrt{\frac{2m(V_0 - E)}{\hbar^2}} \tag{2.45}$$

One boundary condition is that the wave function $\psi_2(x)$ must remain finite, which means that the coefficient $B_2 = 0$. The wave function is now given by

$$\psi_2(x) = A_2 e^{-k_2 x} \quad (x \geq 0) \tag{2.46}$$

The wave function at $x = 0$ must be continuous so that

$$\psi_1(0) = \psi_2(0) \tag{2.47}$$

Then from Equations (2.41), (2.46), and (2.47), we obtain

$$A_1 + B_1 = A_2 \tag{2.48}$$

Since the potential function is everywhere finite, the first derivative of the wave function must also be continuous so that

$$\left.\frac{\partial \psi_1}{\partial x}\right|_{x=0} = \left.\frac{\partial \psi_2}{\partial x}\right|_{x=0} \tag{2.49}$$

Using Equations (2.41), (2.46), and (2.49), we obtain

$$jk_1 A_1 - jk_1 B_1 = -k_2 A_2 \tag{2.50}$$

We can solve Equations (2.48) and (2.50) to determine the coefficients B_1 and A_2 in terms of the incident wave coefficient A_1. The results are

$$B_1 = \frac{-(k_2^2 + 2jk_1k_2 - k_1^2)}{k_2^2 + k_1^2} \cdot A_1 \tag{2.51a}$$

and

$$A_2 = \frac{2k_1(k_1 - jk_2)}{k_2^2 + k_1^2} \cdot A_1 \tag{2.51b}$$

The reflected probability density function is given by

$$B_1 \cdot B_1^* = \frac{(k_2^2 - k_1^2 + 2jk_1k_2)(k_2^2 - k_1^2 - 2jk_1k_2)}{(k_2^2 + k_1^2)^2} \cdot A_1 \cdot A_1^* \tag{2.52}$$

We can define a reflection coefficient, R, as the ratio of the reflected flux to the incident flux, which is written as

$$R = \frac{v_r \cdot B_1 \cdot B_1^*}{v_1 \cdot A_1 \cdot A_1^*} \tag{2.53}$$

where v_i and v_r are the incident and reflected velocities, respectively, of the particles. In region I, $V = 0$ so that $E = T$, where T is the kinetic energy of the particle. The kinetic energy is given by

$$T = \frac{1}{2}mv^2 \tag{2.54}$$

so that the constant k_1, from Equation (2.42) may be written as

$$k_1 = \sqrt{\frac{2m}{\hbar^2}\left(\frac{1}{2}mv^2\right)} = \sqrt{m^2\frac{v^2}{\hbar^2}} = \frac{mv}{\hbar} \tag{2.55}$$

The incident velocity can then be written as

$$v_i = \frac{\hbar}{m} \cdot k_1 \tag{2.56}$$

Since the reflected particle also exists in region I, the reflected velocity (magnitude) is given by

$$v_r = \frac{\hbar}{m} \cdot k_1 \tag{2.57}$$

The incident and reflected velocities (magnitudes) are equal. The reflection coefficient is then

$$R = \frac{v_r \cdot B_1 \cdot B_1^*}{v_i \cdot A_1 \cdot A_1^*} = \frac{B_1 \cdot B_1^*}{A_1 \cdot A_1^*} \tag{2.58}$$

Substituting the expression from Equation (2.52) into Equation (2.58), we obtain

$$R = \frac{B_1 \cdot B_1^*}{A_1 \cdot A_1^*} = \frac{k_2^2 - k_1^2 + 4k_1^2k_2^2}{k_2^2 + k_1^2} = 1.0 \tag{2.59}$$

The result of $R = 1$ implies that all of the particles incident on the potential barrier for $E < V_0$ are eventually reflected. Particles are not absorbed or transmitted through the potential barrier. This result is entirely consistent with classical physics and one might ask why we should consider this problem in terms of quantum mechanics. The interesting result is in terms of what happens in region II.

The wave solution in region II was given by Equation (2.46) as $\psi_2(x) = A_2e^{-k_2x}$. The coefficient A_2 from Equation (2.48) is $A_2 = A_1 + B_1$, which we derived from the boundary conditions. For the case of $E < V_0$, the coefficient A_2 is not zero. If A_2 is not zero, then the probability density function $\psi_2(x) \cdot \psi_2^*(x)$ of the particle being found in region II is not equal to zero. *This result implies that there is a finite*

probability that the incident particle will penetrate the potential barrier and exist in region II. The probability of a particle penetrating the potential barrier is another difference between classical and quantum mechanics: The quantum mechanical penetration is classically not allowed. Although there is a finite probability that the particle may penetrate the barrier, since the reflection coefficient in region I is unity, the particle in region II must eventually turn around and move back into region I.

EXAMPLE 2.4

Objective: Calculate the penetration depth of a particle impinging on a potential barrier.

Consider an incident electron that is traveling at a velocity of 1×10^5 m/s in region I.

■ Solution

With $V(x) = 0$, the total energy is also equal to the kinetic energy so that

$$E = T = \frac{1}{2}mv^2 = 4.56 \times 10^{-21}\ \text{J} = 2.85 \times 10^{-2}\ \text{eV}$$

Now, assume that the potential barrier at $x = 0$ is twice as large as the total energy of the incident particle, or that $V_0 = 2E$. The wave function solution in region II is $\psi_2(x) = A_2 e^{-k_2}$, where the constant k_2 is given by $k_2 = \sqrt{2m(V_0 - E)/\hbar^2}$.

In this example, we want to determine the distance $x = d$ at which the wave function magnitude has decayed to e^{-1} of its value at $x = 0$. Then, for this case, we have $k_2 d = 1$ or

$$1 = d\sqrt{\frac{2m(2E - E)}{\hbar^2}} = d\sqrt{\frac{2mE}{\hbar^2}}$$

The distance is then given by

$$d = \sqrt{\frac{\hbar^2}{2mE}} = \frac{1.054 \times 10^{-34}}{\sqrt{2(9.11 \times 10^{-31})\,(4.56 \times 10^{-21})}} = 11.6 \times 10^{-10}\ \text{m}$$

or

$$d = 11.6\ \text{Å}$$

■ Comment

This penetration distance corresponds to approximately two lattice constants of silicon. The numbers used in this example are rather arbitrary. We used a distance at which the wave function decayed to e^{-1} of its initial value. We could have arbitrarily used e^{-2}, for example, but the results give an indication of the magnitude of penetration depth.

■ EXERCISE PROBLEM

Ex 2.4 The probability of finding a particle at a distance d in region II compared with that at $x = 0$ is given by $\exp(-2k_2 d)$. Consider an electron traveling in region I at a velocity of 10^5 m/s incident on a potential barrier whose height is three times the kinetic energy of the electron. Find the probability of finding the electron at a distance d compared with $x = 0$ where d is (*a*) 10 Å and (*b*) 100 Å into the potential barrier. [(*a*) 8.089%; (*b*) 2.43×10^{-6}%]

The case when the total energy of a particle, which is incident on the potential barrier, is greater than the barrier height, or $E > V_0$, is left as an exercise at the end of the chapter.

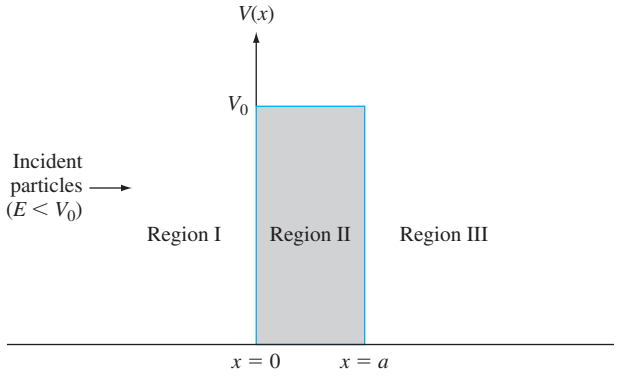


Figure 2.9 | The potential barrier function.

2.3.4 The Potential Barrier and Tunneling

We now want to consider the potential barrier function, which is shown in Figure 2.9. The more interesting problem, again, is in the case when the total energy of an incident particle is $E < V_0$. Again assume that we have a flux of incident particles originating on the negative x-axis traveling in the $+x$ direction. As before, we need to solve Schrodinger's time-independent wave equation in each of the three regions. The solutions of the wave equation in regions I, II, and III are given, respectively, as

$$\psi_1(x) = A_1 e^{jk_1 x} + B_1 e^{-jk_1 x} \tag{2.60a}$$

$$\psi_2(x) = A_2 e^{k_2 x} + B_2 e^{-k_2 x} \tag{2.60b}$$

$$\psi_3(x) = A_3 e^{jk_1 x} + B_3 e^{-jk_1 x} \tag{2.60c}$$

where

$$k_1 = \sqrt{\frac{2mE}{\hbar^2}} \tag{2.61a}$$

and

$$k_2 = \sqrt{\frac{2m}{\hbar^2}(V_0 - E)} \tag{2.61b}$$

The coefficient B_3 in Equation (2.60c) represents a negative traveling wave in region III. However, once a particle gets into region III, there are no potential changes to cause a reflection; therefore, the coefficient B_3 must be zero. We must keep both exponential terms in Equation (2.60b) since the potential barrier width is finite; that is, neither term will become unbounded. We have four boundary relations for the boundaries at $x = 0$ and $x = a$ corresponding to the wave function and its first derivative being continuous. We can solve for the four coefficients B_1, A_2, B_2, and A_3 in terms of A_1. The wave solutions in the three regions are shown in Figure 2.10.

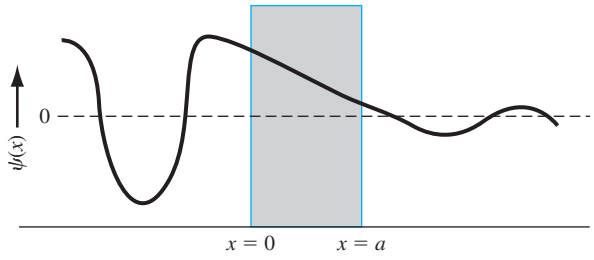


Figure 2.10 | The wave functions through the potential barrier.

One particular parameter of interest is the transmission coefficient, in this case defined as the ratio of the transmitted flux in region III to the incident flux in region I. Then the transmission coefficient T is

$$T = \frac{v_t \cdot A_3 \cdot A_3^*}{v_i \cdot A_1 \cdot A_i^*} = \frac{A_3 \cdot A_3^*}{A_1 \cdot A_1^*} \tag{2.62}$$

where v_t and v_i are the velocities of the transmitted and incident particles, respectively. Since the potential $V = 0$ in both regions I and III, the incident and transmitted velocities are equal. The transmission coefficient may be determined by solving the boundary condition equations. For the special case when $E \ll V_0$, we find that

$$T \approx 16\left(\frac{E}{V_0}\right)\left(1 - \frac{E}{V_0}\right)\exp\left(-2k_2 a\right) \tag{2.63}$$

Equation (2.63) implies that there is a finite probability that a particle impinging a potential barrier will penetrate the barrier and will appear in region III. This phenomenon is called tunneling *and it, too, contradicts classical mechanics.* We will see later how this quantum mechanical tunneling phenomenon can be applied to semiconductor device characteristics, such as in the tunnel diode.

EXAMPLE 2.5

Objective: Calculate the probability of an electron tunneling through a potential barrier.

Consider an electron with an energy of 2 eV impinging on a potential barrier with $V_0 = 20$ eV and a width of 3 Å.

■ Solution

Equation (2.63) is the tunneling probability. The factor k_2 is

$$k_2 = \sqrt{\frac{2m(V_0 - E)}{\hbar^2}} = \sqrt{\frac{2(9.11 \times 10^{-31})(20 - 2)(1.6 \times 10^{-19})}{(1.054 \times 10^{-34})^2}}$$

or

$$k_2 = 2.17 \times 10^{10}\ \text{m}^{-1}$$

Then

$$T = 16(0.1)(1 - 0.1)\exp\left[-2(2.17 \times 10^{10})(3 \times 10^{-10})\right]$$

and finally

$$T = 3.17 \times 10^{-6}$$

Comment

The tunneling probability may appear to be a small value, but the value is not zero. If a large number of particles impinge on a potential barrier, a significant number can penetrate the barrier.

EXERCISE PROBLEM

Ex 2.5 (a)Estimate the probability of an electron with energy $E = 0.12$ eV tunneling through a rectangular potential barrier with a height of $V_0 = 1.2$ eV and a width of $a = 5$ Å. (b) Repeat part (a) for a barrier width of $a = 25$ Å.
[Ans. (a) $T = 7.02 \times 10^{-3}$; (b) $T = 3.97 \times 10^{-12}$]

Additional applications of Schrodinger's wave equation with various one-dimensional potential functions are found in problems at the end of the chapter. Several of these potential functions represent quantum well structures that are found in modern semiconductor devices.

TEST YOUR UNDERSTANDING

TYU 2.3 (a) Estimate the probability of an electron with energy $E = 0.10$ eV tunneling through a rectangular potential barrier with a barrier height of $V_0 = 0.8$ eV and width $a = 12$ Å. (b) Repeat part (a) for a barrier height of $V_0 = 1.5$ eV.
[Ans. (a) $T = 5.97 \times 10^{-5}$; (b) $T = 4.79 \times 10^{-7}$]

TYU 2.4 A certain semiconductor device requires a tunneling probability of $T = 5 \times 10^{-6}$ for an electron tunneling through a rectangular barrier with a barrier height of $V_0 = 0.8$ eV. The electron energy is $E = 0.08$ eV.
Determine the maximum barrier width. (Ans. $a = 14.49$ Å)

2.4 | EXTENSIONS OF THE WAVE THEORY TO ATOMS[5]

So far in this chapter, we have considered several one-dimensional potential energy functions and solved Schrodinger's time-independent wave equation to obtain the probability function of finding a particle at various positions. Consider now the one-electron, or hydrogen, atom potential function. We will only briefly consider the mathematical details and wave function solutions, but the results are interesting and important.

2.4.1 The One-Electron Atom

The nucleus is a heavy, positively charged proton and the electron is a light, negatively charged particle that, in the classical Bohr theory, is revolving around the

[5]The detailed mathematical analysis is beyond the scope of this text, but the results, which are emphasized in this section, are important in the following discussions of semiconductor physics.

nucleus. The potential function is due to the coulomb attraction between the proton and electron and is given by

$$V(r) = \frac{-e^2}{4\pi\epsilon_0 r} \tag{2.64}$$

where e is the magnitude of the electronic charge and ϵ_0 is the permittivity of free space. This potential function, although spherically symmetric, leads to a three-dimensional problem in spherical coordinates.

We may generalize the time-independent Schrodinger's wave equation to three dimensions by writing

$$\nabla^2\psi(r, \theta, \phi) + \frac{2m_0}{\hbar^2}(E - V(r))\psi(r, \theta, \phi) = 0 \tag{2.65}$$

where ∇^2 is the Laplacian operator and must be written in spherical coordinates for this case. The parameter m_0 is the rest mass of the electron.[6] In spherical coordinates, Schrodinger's wave equation may be written as

$$\frac{1}{r^2}\cdot\frac{\partial}{\partial r}\left(r^2\frac{\partial\psi}{\partial r}\right) + \frac{1}{r^2\sin^2\theta}\cdot\frac{\partial^2\psi}{\partial\phi^2} + \frac{1}{r^2\sin\theta}\cdot\frac{\partial}{\partial\theta}\left(\sin\theta\cdot\frac{\partial\psi}{\partial\theta}\right) + \frac{2m_0}{\hbar^2}(E - V(r))\psi = 0 \tag{2.66}$$

The solution to Equation (2.66) can be determined by the separation-of-variables technique. We will assume that the solution to the time-independent wave equation can be written in the form

$$\psi(r, \theta, \phi) = R(r)\cdot\Theta(\theta)\cdot\Phi(\phi) \tag{2.67}$$

where R, Θ, and Φ, are functions only of r, θ, and ϕ, respectively. Substituting this form of solution into Equation (2.66), we will obtain

$$\frac{\sin^2\theta}{R}\cdot\frac{\partial}{\partial r}\left(r^2\frac{\partial R}{\partial r}\right) + \frac{1}{\Phi}\cdot\frac{\partial^2\phi}{\partial\phi^2} + \frac{\sin\theta}{\Theta}\cdot\frac{\partial}{\partial\theta}\left(\sin\theta\cdot\frac{\partial\Theta}{\partial\theta}\right) + r^2\sin^2\theta\cdot\frac{2m_0}{\hbar^2}(E - V) = 0 \tag{2.68}$$

We may note that the second term in Equation (2.68) is a function of ϕ only, whereas all the other terms are functions of either r or ϕ. We may then write that

$$\frac{1}{\Phi}\cdot\frac{\partial^2\Phi}{\partial\phi^2} = -m^2 \tag{2.69}$$

where m is a separation of variables constant.[7] The solution to Equation (2.69) is of the form

$$\phi = e^{jm\phi} \tag{2.70}$$

[6]The mass should be the rest mass of the two-particle system, but since the proton mass is much greater than the electron mass, the equivalent mass reduces to that of the electron.

[7]Where m means the separation-of-variables constant developed historically. That meaning will be retained here even though there may be some confusion with the electron mass. In general, the mass parameter will be used in conjunction with a subscript.

Since the wave function must be single-valued, we impose the condition that m is an integer, or

$$m = 0, \pm 1, \pm 2, \pm 3, \ldots \tag{2.71}$$

Incorporating the separation-of-variables constant, we can further separate the variables θ and r and generate two additional separation-of-variables constants l and n. The separation-of-variables constants n, l, and m are known as *quantum numbers*. The parameter n is referred to as the principal quantum number, l is the azimuthal or angular quantum number, and m is the magnetic quantum number. The quantum numbers are related by

$$\begin{aligned} n &= 1, 2, 3, \ldots \\ l &= n-1, n-2, n-3, \ldots 0 \\ |m| &= l, l-1, \ldots, 0 \end{aligned} \tag{2.72}$$

Each set of quantum numbers corresponds to a quantum state that the electron may occupy.

The electron energy may be written in the form

$$E_n = \frac{-m_0 e^4}{(4\pi\epsilon_0)^2 2\hbar^2 n^2} \tag{2.73}$$

where again n is the principal quantum number. The negative energy indicates that the electron is bound to the nucleus and we again see that the energy of the bound electron is quantized. If the energy were to become positive, then the electron would no longer be a bound particle and the total energy would no longer be quantized. Since the parameter n in Equation (2.73) is an integer, the total energy of the electron can take on only discrete values. The quantized energy is again a result of the particle being bound in a finite region of space.

EXAMPLE 2.6

Objective: Calculate the first three energy levels of an electron for the one-electron atom.

■ Solution

We have

$$E_n = \frac{-m_0 e^4}{(4\pi\epsilon_0)^2 2\hbar^2 n^2} = \frac{-(9.11 \times 10^{-31})(1.6 \times 10^{-19})^4}{[4\pi(8.85 \times 10^{-12})]^2 2(1.054 \times 10^{-34})^2 n^2}$$

$$= \frac{-21.726 \times 10^{-19}}{n^2}\ \text{J} \quad \text{or} \quad = \frac{-13.58}{n^2}\ \text{eV}$$

For $n = 1$; $E_1 = -13.58$ eV

$n = 2$; $E_2 = -3.39$ eV

$n = 3$; $E_3 = -1.51$ eV

■ Comment

As the energy levels increase, the energy becomes less negative, which means that the electron is becoming less tightly bound to the atom.

■ EXERCISE PROBLEM

Ex 2.6 In Example 2.6, assume the permittivity of free space, ϵ_0, is replaced by the permittivity of a material where $\epsilon = \epsilon_r \epsilon_0$. Repeat the calculations in Example 2.6 if $\epsilon_r = 11.7$ (silicon).

(Ans. $E_1 = -99.2$ meV, $E_2 = -24.8$ meV, $E_3 = -11.0$ meV)

The solution of the wave equation may be designated by ψ_{nlm}, where n, l, and m are again the various quantum numbers. For the lowest energy state, $n = 1$, $l = 0$, and $m = 0$, and the wave function is given by

$$\psi_{100} = \frac{1}{\sqrt{\pi}} \cdot \left(\frac{1}{a_0}\right)^{3/2} e^{-r/a_0} \tag{2.74}$$

This function is spherically symmetric, and the parameter a_0 is given by

$$a_0 = \frac{4\pi\epsilon_0 \hbar^2}{m_0 e^2} = 0.529 \text{ Å} \tag{2.75}$$

and is equal to the Bohr radius.

The radial probability density function, or the probability of finding the electron at a particular distance from the nucleus, is proportional to the product $\psi_{100} \cdot \psi_{100}^*$ and also to the differential volume of the shell around the nucleus. The probability density function for the lowest energy state is plotted in Figure 2.11a. The most probable distance from the nucleus is at $r = a_0$, which is the same as the Bohr theory. Considering this spherically symmetric probability function, we may now begin to conceive the concept of an electron cloud, or energy shell, surrounding the nucleus rather than a discrete particle orbiting around the nucleus.

The radial probability density function for the next higher, spherically symmetric wave function, corresponding to $n = 2$, $l = 0$, and $m = 0$, is shown in Figure 2.11b. This figure shows the idea of the next-higher energy shell of the electron. The second energy shell is at a greater radius from the nucleus than the first energy shell. As indicated in the figure, though, there is still a small probability that the electron will exist at the smaller radius. For the case of $n = 2$ and $l = 1$, there are three possible states corresponding to the three allowed values of the quantum number m. These wave functions are no longer spherically symmetric.

Although we have not gone into a great deal of mathematical detail for the one-electron atom, three results are important for the further analysis of semiconductor materials.

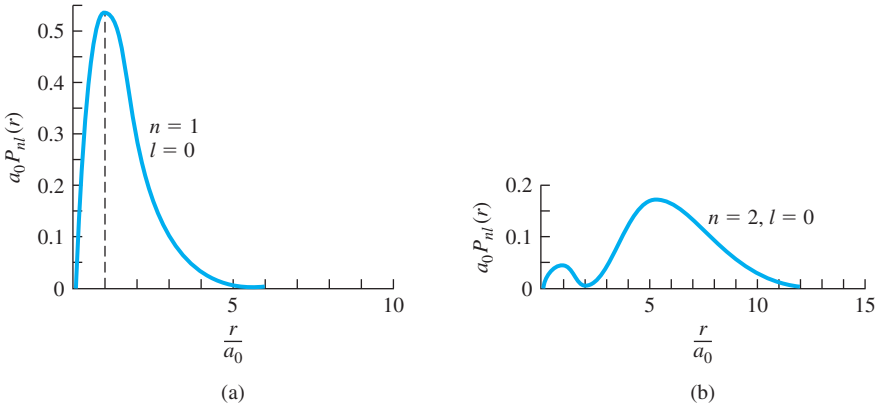


Figure 2.11 | The radial probability density function for the one-electron atom in the (a) lowest energy state and (b) next-higher energy state.
(From Eisberg and Resnick [5].)

The first is the solution of Schrodinger's wave equation, which again yields electron probability functions, as it did for the simpler potential functions. In developing the physics of semiconductor materials in later chapters, we will also be considering electron probability functions. The second result is the quantization of allowed energy levels for the bound electron. The third is the concept of quantum numbers and quantum states, which evolved from the separation-of-variables technique. We will consider this concept again in the next section and in later chapters when we deal with the semiconductor material physics.

2.4.2 The Periodic Table

The initial portion of the periodic table of elements may be determined by using the results of the one-electron atom plus two additional concepts. The first concept needed is that of *electron spin*. The electron has an intrinsic angular momentum, or spin, which is quantized and may take on one of two possible values. The spin is designated by a quantum number s, which has a value of $s = +\frac{1}{2}$ or $s = -\frac{1}{2}$. We now have four basic quantum numbers: n, l, m, and s.

The second concept needed is the *Pauli exclusion principle*. The Pauli exclusion principle states that, in any given system (an atom, molecule, or crystal), no two electrons may occupy the same quantum state. In an atom, the exclusion principle means that no two electrons may have the same set of quantum numbers. We will see that the exclusion principle is also an important factor in determining the distribution of electrons among available energy states in a crystal.

Table 2.1 shows the first few elements of the periodic table. For the first element, hydrogen, we have one electron in the lowest energy state corresponding to $n = 1$. From Equation (2.72) both quantum numbers l and m must be zero. However, the electron can take on either spin factor $+\frac{1}{2}$ or $-\frac{1}{2}$. For helium, two electrons may exist in the lowest energy state. For this case, $l = m = 0$, so now both electron spin states are occupied and the lowest energy shell is full. The chemical activity of an element is determined primarily by the valence, or outermost, electrons. Since the valence energy shell of helium is full, helium does not react with other elements and is an inert element.

Table 2.1 | Initial portion of the periodic table

Element	Notation	n	l	m	s
Hydrogen	$1s^1$	1	0	0	$+\frac{1}{2}$ or $-\frac{1}{2}$
Helium	$1s^2$	1	0	0	$+\frac{1}{2}$ and $-\frac{1}{2}$
Lithium	$1s^22s^1$	2	0	0	$+\frac{1}{2}$ or $-\frac{1}{2}$
Beryllium	$1s^22s^2$	2	0	0	$+\frac{1}{2}$ and $-\frac{1}{2}$
Boron	$1s^22s^22p^1$	2	1		
Carbon	$1s^22s^22p^2$	2	1		
Nitrogen	$1s^22s^22p^3$	2	1		$m = 0, -1, +1$
Oxygen	$1s^22s^22p^4$	2	1		$s = +\frac{1}{2}, -\frac{1}{2}$
Fluorine	$1s^22s^22p^5$	2	1		
Neon	$1s^22s^22p^6$	2	1		

The third element, lithium, has three electrons. The third electron must go into the second energy shell corresponding to $n = 2$. When $n = 2$, the quantum number l may be 0 or 1, and when $l = 1$, the quantum number m may be -1, 0, or $+1$. In each case, the electron spin factor may be $+\frac{1}{2}$ or $-\frac{1}{2}$. For $n = 2$, then, there are eight possible quantum states. Neon has 10 electrons. Two electrons are in the $n = 1$ energy shell and eight electrons are in the $n = 2$ energy shell. The second energy shell is now full, which means that neon is also an inert element.

From the solution of Schrodinger's wave equation for the one-electron atom, plus the concepts of electron spin and the Pauli exclusion principle, we can begin to build up the periodic table of elements. As the atomic numbers of the elements increase, electrons will begin to interact with each other, so that the buildup of the periodic table will deviate somewhat from the simple method.

2.5 | SUMMARY

- A few basic concepts of quantum mechanics, which can be used to describe the behavior of electrons under various potential functions, were considered. The understanding of electron behavior is crucial in understanding semiconductor physics.
- The wave–particle duality principle is an important element in quantum mechanics. Particles can have wave-like behavior and waves can have particle-like behavior.
- Schrodinger's wave equation forms the basis for describing and predicting the behavior of electrons.
- Max Born postulated that $|\psi(x)|^2$ is a probability density function.
- A result of applying Schrodinger's wave equation to a bound particle is that the energy of the bound particle is *quantized.*
- A result of applying Schrodinger's wave equation to an electron incident on a potential barrier is that there is a finite probability of *tunneling*.
- The concept of quantum numbers was developed from the results of applying Schrodinger's wave equation to the one-electron atom.
- The basic structure of the periodic table is predicted by applying Schrodinger's wave equation to the one-electron atom and using the Pauli exclusion principle.

GLOSSARY OF IMPORTANT TERMS

de Broglie wavelength The wavelength of a particle given as the ratio of Planck's constant to momentum.

Heisenberg uncertainty principle The principle that states that we cannot describe with absolute accuracy the relationship between sets of conjugate variables that describe the behavior of particles, such as momentum and position.

Pauli exclusion principle The principle that states that no two electrons can occupy the same quantum state.

photon The particle-like packet of electromagnetic energy.

quanta The particle-like packet of thermal radiation.

quantized energies The allowed discrete energy levels that bound particles may occupy.

quantum numbers A set of numbers that describes the quantum state of a particle, such as an electron in an atom.

quantum state A particular state of an electron that may be described, for example, by a set of quantum numbers.

tunneling The quantum mechanical phenomenon by which a particle may penetrate through a thin potential barrier.

wave–particle duality The characteristic by which electromagnetic waves sometimes exhibit particle-like behavior and particles sometimes exhibit wave-like behavior.

CHECKPOINT

After studying this chapter, the reader should have the ability to:

- Discuss the principle of energy quanta, the wave–particle duality principle, and the uncertainty principle.
- Apply Schrodinger's wave equation and boundary conditions to problems with various potential functions.
- Determine quantized energy levels of bound particles.
- Determine the approximate tunneling probability of aaaaaaaaaaaa particle incident on a potential barrier.
- State Pauli exclusion principle.
- Discuss the results of the one-electron atom analysis, including quantum numbers and their interrelationship as well as the initial formation of the periodic table.

REVIEW QUESTIONS

1. State the wave–particle duality principle and state the relationship between momentum and wavelength.
2. What is the physical meaning of Schrodinger's wave function?
3. What is meant by a probability density function?
4. List the boundary conditions for solutions to Schrodinger's wave equation.
5. What is meant by quantized energy levels? Can an electron contained in a potential well have an arbitrary energy?
6. Describe the concept of tunneling.
7. List the quantum numbers of the one-electron atom and discuss how they were developed.
8. State the interrelationship between the quantum numbers of the one-electron atom and how this result leads to, for example, the development of inert elements.

PROBLEMS

2.1 The classical wave equation for a two-wire transmission line is given by $\partial^2 V(x, t)/\partial x^2 = LC \cdot \partial^2 V(x, t)/\partial t^2$. One possible solution is given by $V(x, t) = (\sin Kx) \cdot (\sin \omega t)$ where $K = n\pi/a$ and $\omega = K/\sqrt{LC}$. Sketch, on the same graph, the function $V(x, t)$ as a function of x for $0 \leq x \leq a$ and $n = 1$ when (i) $\omega t = 0$, (ii) $\omega t = \pi/2$, (iii) $\omega t = \pi$, (iv) $\omega t = 3\pi/2$, and (v) $\omega t = 2\pi$.

2.2 The function $V(x, t) = \cos(2\pi x/\lambda - \omega t)$ is also a solution to the classical wave equation. Sketch on the same graph the function $V(x, t)$ as a function of x for $0 \leq x \leq 3\lambda$ when: (i) $\omega t = 0$, (ii) $\omega t = 0.25\pi$, (iii) $\omega t = 0.5\pi$, (iv) $\omega t = 0.75\pi$, and (v) $\omega t = \pi$.

2.3 Repeat Problem 2.2 for the function $V(x, t) = \cos(2\pi x/\lambda + \omega t)$.

2.4 Determine the phase velocities of the traveling waves described in Problems 2.2 and 2.3.

Section 2.1 Principles of Quantum Mechanics

2.5 The work function of a material refers to the minimum energy required to remove an electron from the material. Assume that the work function of gold is 4.90 eV and that of cesium is 1.90 eV. Calculate the maximum wavelength of light for the photoelectric emission of electrons for gold and cesium.

2.6 (*a*) The wavelength of green light is $\lambda = 550$ nm. If an electron has the same wavelength, determine the electron velocity and momentum. (*b*) Repeat part (*a*) for red light with a wavelength of $\lambda = 440$ nm. (*c*) For parts (*a*) and (*b*), is the momentum of the photon equal to the momentum of the electron?

2.7 Determine the de Broglie wavelength for (*a*) an electron with kinetic energy of (*i*) 1.2 eV, (*ii*) 12 eV, (*iii*) 120 eV; and for (*b*) a hydrogen atom with a kinetic energy of 1.2 eV.

2.8 According to classical physics, the average energy of an electron in an electron gas at thermal equilibrium is $3kT/2$. Determine, for $T = 300$ K, the average electron energy (in eV), average electron momentum, and the de Broglie wavelength.

2.9 An electron and a photon have the same energy. At what value of energy (in eV) will the wavelength of the photon be 10 times that of the electron?

2.10 (*a*) The de Broglie wavelength of an electron is 85 Å. Determine the electron energy (eV), momentum, and velocity. (*b*) An electron is moving with a velocity of 8×10^5 cm/s. Determine the electron energy (eV), momentum, and de Broglie wavelength (in Å).

2.11 It is desired to produce x-ray radiation with a wavelength of 1 Å. (*a*) Through what potential voltage difference must the electron be accelerated in vacuum so that it can, upon colliding with a target, generate such a photon? (Assume that all of the electron's energy is transferred to the photon.) (*b*) What is the de Broglie wavelength of the electron in part (*a*) just before it hits the target?

2.12 When the uncertainty principle is considered, it is not possible to locate a photon in space more precisely than about one wavelength. Consider a photon with wavelength $\lambda = 1$ μm. What is the uncertainty in the photon's momentum?

2.13 (*a*) The uncertainty in position is 12 Å for a particle of mass 9×10^{-31} kg. The nominal energy of the particle is 16 eV. Determine the minimum uncertainty in (*i*) momentum and (*ii*) kinetic energy of the particle. (*b*) Repeat part (*a*) for a particle of mass 5×10^{-28} kg.

2.14 An automobile has a mass of 1500 kg. What is the uncertainty in the velocity (in miles per hour) when its center of mass is located with an uncertainty no greater than 1 cm?

2.15 (*a*) The electron's energy is measured with an uncertainty no greater than 0.8 eV. Determine the minimum uncertainty in the time over which the measurement is made. (*b*) The uncertainty in the position of an electron is no greater than 1.5 Å. Determine the minimum uncertainty in its momentum.

Section 2.2 Schrodinger's Wave Equation

2.16 Assume that $\Psi_1(x, t)$ and $\Psi_2(x, t)$ are solutions of the one-dimensional time-dependent Schrodinger's wave equation. (*a*) Show that $\Psi_1 + \Psi_2$ is a solution. (*b*) Is $\Psi_1 \cdot \Psi_2$ a solution of the Schrodinger's equation in general? Why or why not?

2.17 Consider the wave function $\Psi(x, t) = A\left(\cos\left(\frac{\pi x}{2}\right)\right) e^{-j\omega t}$ for $-1 \le x \le +3$. Determine A so that $\int_{-1}^{+3} |\Psi(x, t)|^2 dx = 1$.

2.18 Consider the wave function $\Psi(x, t) = A(\cos n\pi x)e^{-j\omega t}$ for $-1/2 \le x \le +1/2$, where n is an integer. Determine A so that $\int_{-1/2}^{+1/2} |\Psi(x, t)|^2 dx = 1$.

2.19 The solution to Schrodinger's wave equation for a particular situation is given by $\psi(x) = \sqrt{2/a_0} \cdot e^{-x/a_0}$. Determine the probability of finding the particle between the following limits: (*a*) $0 \le x \le a_0/4$, (*b*) $a_0/4 \le x \le a_0/2$, and (*c*) $0 \le x \le a_0$.

2.20 An electron is described by a wave function given by $\psi(x) = \sqrt{\frac{2}{a}}\ \cos\left(\frac{\pi x}{a}\right)$ for $\frac{-a}{2} < x < \frac{+a}{2}$. The wave function is zero elsewhere. Calculate the probability of finding the electron between (*a*) $0 < x < \frac{a}{4}$, (*b*) $\frac{a}{4} < x < \frac{a}{2}$, and (*c*) $\frac{-a}{2} < x < \frac{+a}{2}$.

2.21 Repeat Problem 2.20 if the wave function is given by $\psi(x) = \sqrt{\frac{2}{a}} \sin\left(\frac{2\pi x}{a}\right)$.

Section 2.3 Applications of Schrodinger's Wave Equation

2.22 (*a*) An electron in free space is described by a plane wave given by $\Psi(x, t) = Ae^{j(kx-\omega t)}$. If $k = 8 \times 10^8\ m^{-1}$ and $\omega\ 8 \times 10^{12}$ rad/s, determine the (*i*) phase velocity and wavelength of the plane wave, and the (*ii*) momentum and kinetic energy (in eV) of the electron. (*b*) Repeat part (*a*) for $k = -1.5 \times 10^9\ \text{m}^{-1}$ and $\omega = 1.5 \times 10^{13}$ rad/s.

2.23 An electron is traveling in the negative x direction with a kinetic energy of 0.025 eV. (*a*) Write the equation of a plane wave that describes this particle. (*b*) What is the wave number, wavelength, and angular frequency of the wave that describes this electron.

2.24 Determine the wave number, wavelength, angular frequency, and period of a wave function that describes an electron traveling in free space at a velocity of (*a*) $v = 5 \times 10^6$ cm/s and (*b*) $v = 10^8$ cm/s.

2.25 An elecron is bound in a one-dimesional infinite potential well with a width of 75 Å. Determine the electron energy levels (in eV) for $n = 1, 2, 3$.

2.26 An electron is bound in a one-dimensional infinite potential well with a width of 10 Å. (*a*) Calculate the first three energy levels that the electron may occupy. (*b*) If the electron drops from the third to the second energy level, what is the wavelength of a photon that might be emitted?

2.27 A particle with a mass of 15 mg is bound in a one-dimensional infinite potential well that is 1.2 cm wide. (*a*) If the energy of the particle is 15 mJ, determine the value of n for that state. (*b*) What is the energy of the $(n+1)$ state? (*c*) Would quantum effects be observable for this particle?

2.28 Calculate the lowest energy level for a neutron in a nucleus, by treating it as if it were in an infinite potential well of width equal to 10^{-14} m. Compsare this with the lowest energy level for an electron in the same infinite potential well.

2.29 Consider the particle in the infinite potential well as shown in Figure P2.29. Derive and sketch the wave functions corresponding to the four lowest energy levels. (Do not normalize the wave functions.)

***2.30** Consider a three-dimensional infinite potential well. The potential function is given by $V(x) = 0$ for $0 < x < a$, $0 < y < a$, $0 < z < a$, and $V(x) = \infty$ elsewhere. Start with

*Asterisks next to problems indicate problems that are more difficult.

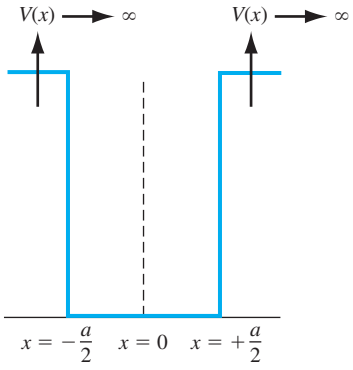


Figure P2.29 | Potential function for Problem 2.29.

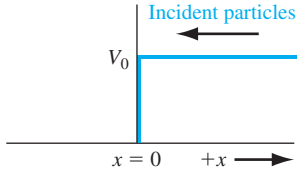


Figure P2.33 | Potential function for Problem 2.33.

Schrodinger's wave equation, use the separation of variables technique, and show that the energy is quantized and is given by

$$E_{n_x n_y n_z} = \frac{\hbar^2 \pi^2}{2ma^2}\left(n_x^2 + n_y^2 + n_z^2\right)$$

where $n_x = 1, 2, 3, \ldots$, $n_y = 1, 2, 3, \ldots$, $n_z = 1, 2, 3, \ldots$.

2.31 Consider a free electron bound within a two-dimensional infinite potential well defined by $V = 0$ for $0 < x < 40$ Å, $0 < y < 20$ Å, and $V = \infty$ elsewhere. (*a*) Determine the expression for the allowed electron energies. (*b*) Describe any similarities and any differences with the results of the one-dimensional infinite potential well.

2.32 Consider a proton in a one-dimensional infinite potential well shown in Figure 2.6. (*a*) Derive the expression for the allowed energy states of the proton. (*b*) Calculate the energy difference (in units of eV) between the lowest possible energy and the next higher energy state for (*i*) $a = 4$ Å, and (*ii*) $a = 0.5$ cm.

2.33 For the step potential function shown in Figure P2.33, assume that $E > V_0$ and that particles are incident from the $+x$ direction traveling in the $-x$ direction. (*a*) Write the wave solutions for each region. (*b*) Derive expressions for the transmission and reflection coefficients.

2.34 Consider an electron with a kinetic energy of 2.8 eV incident on a step potential function of height 3.5 eV. Determine the relative probability of finding the electron at a distance (*a*) 5 Å beyond the barrier, (*b*) 15 Å beyond the barrier, and (*c*) 40 Å beyond the barrier compared with the probability of finding the incident particle at the barrier edge.

2.35 (*a*) Calculate the transmission coefficient of an electron with a kinetic energy of 0.1 eV impinging on a potential barrier of height 1.0 eV and a width of 4 Å (*b*) Repeat part (*a*) for a barrier width of 12 Å. (*c*) Using the results of part (*a*), determine the density of electrons per second that impinge the barrier if the tunneling current density is 1.2 mA/cm^2.

2.36 (*a*) Estimate the tunneling probability of a particle with an effective mass of $0.067m_0$ (an electron in gallium arsenide), where m_0 is the mass of an electron, tunneling through a rectangular potential barrier of height $V_0 = 0.8$ eV and width 15 Å. The particle kinetic energy is 0.20 eV. (*b*) Repeat part (*a*) if the effective mass of the particle is $1.08m_0$ (an electron in silicon).

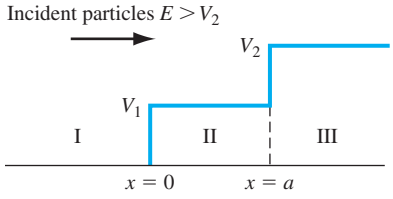


Figure P2.39 | Potential function for Problem 2.39.

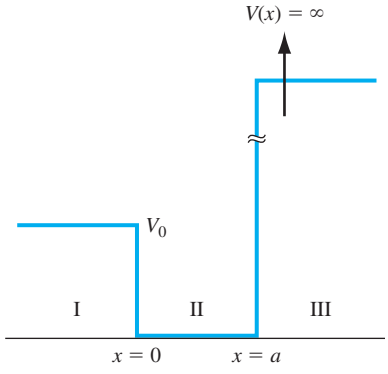


Figure P2.40 | Potential function for Problem 2.40.

2.37 (*a*) A proton with a kinetic energy of 1 MeV is incident on a potential barrier of height 12 MeV and width 10^{-14} m. What is the tunneling probability. (*b*) The width of the potential barrier in part (*a*) is to be decreased so that the tunneling probability is increased by a factor of 10. What is the width of the potential barrier?

***2.38** An electron with energy E is incident on a rectangular potential barrier as shown in Figure 2.9. The potential barrier is of width a and height $V_0 \gg E$. (*a*) Write the form of the wave function in each of the three regions. (*b*) For this geometry, determine what coefficient in the wave function solutions is zero. (*c*) Derive the expression for the transmission coefficient for the electron (tunneling probability). (*d*) Sketch the wave function for the electron in each region.

***2.39** A potential function is shown in Figure P2.39 with incident particles coming from $-\infty$ with a total energy $E > V_2$. The constants k are defined as

$$k_1 = \sqrt{\frac{2mE}{\hbar^2}} \qquad k_2 = \sqrt{\frac{2m}{\hbar^2}(E - V_1)} \qquad k_3 = \sqrt{\frac{2m}{\hbar^2}(E - V_2)}$$

Assume a special case for which $k_2 a = 2n\pi$, $n = 1, 2, 3, \ldots$. Derive the expression, in terms of the constants, k_1, k_2, and k_3, for the transmission coefficient. The transmission coefficient is defined as the ratio of the flux of particles in region III to the incident flux in region I.

***2.40** Consider the one-dimensional potential function shown in Figure P2.40. Assume the total energy of an electron is $E < V_0$. (*a*) Write the wave solutions that apply in each region. (*b*) Write the set of equations that result from applying the boundary conditions. (*c*) Show explicitly why, or why not, the energy levels of the electron are quantized.

Section 2.4 Extensions of the Wave Theory to Atoms

2.41 Calculate the energy of the electron in the hydrogen atom (in units of eV) for the first four allowed energy levels.

2.42 Show that the most probable value of the radius r for the 1s electron in a hydrogen atom is equal to the Bohr radius a_0.

2.43 Show that the wave function for ψ_{100} given by Equation (2.74) is a solution to the differential equation given by Equation (2.65).

2.44 What property do H, Li, Na, and K have in common?

READING LIST

***1.** Datta, S. *Quantum Phenomena*. Vol. 8 of *Modular Series on Solid State Devices.* Reading, MA: Addison–Wesley, 1989.

***2.** deCogan, D. *Solid State Devices: A Quantum Physics Approach*. New York: Springer-Verlag, 1987.

3. Dimitrijev, S. *Principles of Semiconductor Devices.* New York: Oxford University, 2006.

4. Eisberg, R. M. *Fundamentals of Modern Physics.* New York: Wiley, 1961.

5. Eisberg, R., and R. Resnick. *Quantum Physics of Atoms, Molecules, Solids, Nuclei, and Particles*. New York: Wiley, 1974.

6. Kano, K. *Semiconductor Devices*. Upper Saddle River, NJ: Prentice Hall, 1998.

7. Kittel, C. *Introduction to Solid State Physics,* 7th ed. Berlin: Springer-Verlag, 1993.

8. McKelvey, J. P. *Solid State Physics for Engineering and Materials Science*. Malabar, FL: Krieger Publishing, 1993.

9. Pauling, L., and E. B. Wilson. *Introduction to Quantum Mechanics*. New York: McGraw-Hill, 1935.

10. Pierret, R. F. *Semiconductor Device Fundamentals*. Reading, MA: Addison-Wesley Publishing Co., 1996.

11. Pohl, H. A. *Quantum Mechanics for Science and Engineering*. Englewood Cliffs, NJ: Prentice Hall, 1967.

12. Schiff, L. I. *Quantum Mechanics.* New York: McGraw-Hill, 1955.

13. Shur, M. *Introduction to Electronic Devices.* New York: John Wiley and Sons, 1996.

*Indicates references that are at an advanced level compared to this text.

CHAPTER 3

Introduction to the Quantum Theory of Solids

In the last chapter, we applied quantum mechanics and Schrodinger's wave equation to determine the behavior of electrons in the presence of various potential functions. We found one important characteristic of an electron bound to an atom or bound within a finite space to be that the electron can take on only discrete values of energy; that is, the energies are quantized. We also discussed the Pauli exclusion principle, which stated that only one electron is allowed to occupy any given quantum state. In this chapter, we will generalize these concepts to the electron in a crystal lattice.

One of our goals is to determine the electrical properties of a semiconductor material, which we will then use to develop the current–voltage characteristics of semiconductor devices. Toward this end, we have two tasks in this chapter: to determine the properties of electrons in a crystal lattice and to determine the statistical characteristics of the very large number of electrons in a crystal. ■

3.0 | PREVIEW

In this chapter, we will:

- Develop the concept of allowed and forbidden electron energy bands in a single-crystal material, and describe conduction and valence energy bands in a semiconductor material.
- Discuss the concept of negatively charged electrons and positively charged holes as two distinct charge carriers in a semiconductor material.
- Develop electron energy versus momentum curves in a single-crystal material, which yields the concept of direct and indirect bandgap semiconductor materials.
- Discuss the concept of effective mass of an electron and a hole.
- Derive the density of quantum states in the allowed energy bands.

- Develop the Fermi-Dirac probability function, which describes the statistical distribution of electrons among the allowed energy levels, and define the Fermi energy level.

3.1 | ALLOWED AND FORBIDDEN ENERGY BANDS

In the last chapter, we considered the one-electron, or hydrogen, atom. That analysis showed that the energy of the bound electron is quantized: Only discrete values of electron energy are allowed. The radial probability density for the electron was also determined. This function gives the probability of finding the electron at a particular distance from the nucleus and shows that the electron is not localized at a given radius. We can extrapolate these single-atom results to a crystal and qualitatively derive the concepts of allowed and forbidden energy bands. We will then apply quantum mechanics and Schrodinger's wave equation to the problem of an electron in a single crystal. We find that the electronic energy states occur in bands of allowed states that are separated by forbidden energy bands.

3.1.1 Formation of Energy Bands

Figure 3.1a shows the radial probability density function for the lowest electron energy state of the single, noninteracting hydrogen atom, and Figure 3.1b shows the same probability curves for two atoms that are in close proximity to each other. The wave functions of the electrons of the two atoms overlap, which means that the two electrons will interact. This interaction or perturbation results in the discrete quantized energy level splitting into two discrete energy levels, schematically shown in Figure 3.1c. The splitting of the discrete state into two states is consistent with the Pauli exclusion principle.

A simple analogy of the splitting of energy levels by interacting particles is the following. Two identical race cars and drivers are far apart on a race track. There is no interaction between the cars, so they both must provide the same power to achieve a given speed. However, if one car pulls up close behind the other car, there is an interaction called *draft.* The second car will be pulled to an extent by the lead car. The lead car will therefore require more power to achieve the same speed since it is pulling the second car, and the second car will require less power since it is being

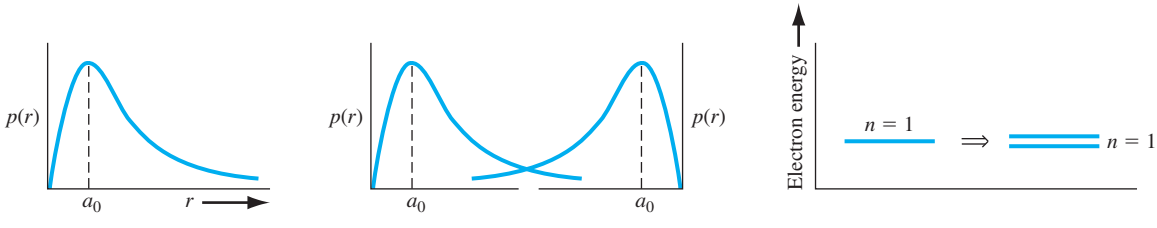


Figure 3.1 | (a) Probability density function of an isolated hydrogen atom. (b) Overlapping probability density functions of two adjacent hydrogen atoms. (c) The splitting of the $n = 1$ state.

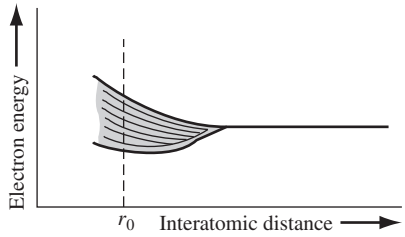


Figure 3.2 | The splitting of an energy state into a band of allowed energies.

pulled by the lead car. So there is a "splitting" of power (energy) of the two interacting race cars. (Keep in mind not to take analogies too literally.)

Now, if we somehow start with a regular periodic arrangement of hydrogen-type atoms that are initially very far apart, and begin pushing the atoms together, the initial quantized energy level will split into a band of discrete energy levels. This effect is shown schematically in Figure 3.2, where the parameter r_0 represents the equilibrium interatomic distance in the crystal. At the equilibrium interatomic distance, there is a band of allowed energies, but within the allowed band, the energies are at discrete levels. The Pauli exclusion principle states that the joining of atoms to form a system (crystal) does not alter the total number of quantum states regardless of size. However, since no two electrons can have the same quantum number, the discrete energy must split into a band of energies in order that each electron can occupy a distinct quantum state.

We have seen previously that, at any energy level, the number of allowed quantum states is relatively small. In order to accommodate all of the electrons in a crystal, we must have many energy levels within the allowed band. As an example, suppose that we have a system with 10^{19} one-electron atoms and also suppose that, at the equilibrium interatomic distance, the width of the allowed energy band is 1 eV. For simplicity, we assume that each electron in the system occupies a different energy level and, if the discrete energy states are equidistant, then the energy levels are separated by 10^{-19} eV. This energy difference is extremely small, so that for all practical purposes, we have a quasi-continuous energy distribution through the allowed energy band. The fact that 10^{-19} eV is a very small difference between two energy states can be seen from the following example.

EXAMPLE 3.1

Objective: Calculate the change in kinetic energy of an electron when the velocity changes by a small amount.

Consider an electron traveling at a velocity of 10^7 cm/s. Assume that the velocity increases by a value of 1 cm/s. The increase in kinetic energy is given by

$$\Delta E = \frac{1}{2}mv_2^2 - \frac{1}{2}mv_1^2 = \frac{1}{2}m(v_2^2 - v_1^2)$$

Let $v_2 = v_1 + \Delta v$. Then

$$v_2^2 = (v_1 + \Delta v)^2 = v_1^2 + 2v_1\Delta v + (\Delta v)^2$$

But $\Delta v \ll v_1$, so we have that

$$\Delta E \approx \frac{1}{2} m(2v_1 \Delta v) = mv_1 \Delta v$$

■ Solution

Substituting the number into this equation, we obtain

$$\Delta E = (9.11 \times 10^{-31})(10^5)(0.01) = 9.11 \times 10^{-28} \text{ J}$$

which may be converted to units of electron volts as

$$\Delta E = \frac{9.11 \times 10^{-28}}{1.6 \times 10^{-19}} = 5.7 \times 10^{-9} \text{ eV}$$

■ Comment

A change in velocity of 1 cm/s compared with 10^7 cm/s results in a change in energy of 5.7×10^{-9} eV, which is orders of magnitude larger than the change in energy of 10^{-19} eV between energy states in the allowed energy band. This example serves to demonstrate that a difference in adjacent energy states of 10^{-19} eV is indeed very small, so that the discrete energies within an allowed band may be treated as a quasi-continuous distribution.

■ EXERCISE PROBLEM

Ex 3.1 The initial velocity of an electron is 10^7 cm/s. If the kinetic energy of the electron increases by $\Delta E = 10^{-12}$ eV, determine the increase in velocity.
(Ans. $\Delta v = 1.76 \times 10^{-4}$ cm/s)

Consider again a regular periodic arrangement of atoms, in which each atom now contains more than one electron. Suppose the atom in this imaginary crystal contains electrons up through the $n = 3$ energy level. If the atoms are initially very far apart, the electrons in adjacent atoms will not interact and will occupy the discrete energy levels. If these atoms are brought closer together, the outermost electrons in the $n = 3$ energy shell will begin to interact initially, so that this discrete energy level will split into a band of allowed energies. If the atoms continue to move closer together, the electrons in the $n = 2$ shell may begin to interact and will also split into a band of allowed energies. Finally, if the atoms become sufficiently close together, the innermost electrons in the $n = 1$ level may interact, so that this energy level may also split into a band of allowed energies. The splitting of these discrete energy levels is qualitatively shown in Figure 3.3. If the equilibrium interatomic distance is r_0, then we have bands of allowed energies that the electrons may occupy separated by bands of forbidden energies. This energy-band splitting and the formation of allowed and forbidden bands is the energy-band theory of single-crystal materials.

The actual band splitting in a crystal is much more complicated than indicated in Figure 3.3. A schematic representation of an isolated silicon atom is shown in Figure 3.4a. Ten of the 14 silicon atom electrons occupy deep-lying energy levels close to the nucleus. The four remaining valence electrons are relatively weakly bound and are the electrons involved in chemical reactions. Figure 3.4b shows the band splitting of silicon. We need only consider the $n = 3$ level for the valence

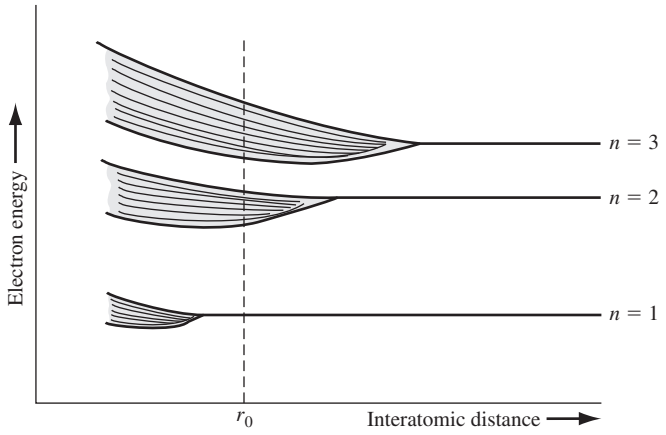


Figure 3.3 | Schematic showing the splitting of three energy states into allowed bands of energies.

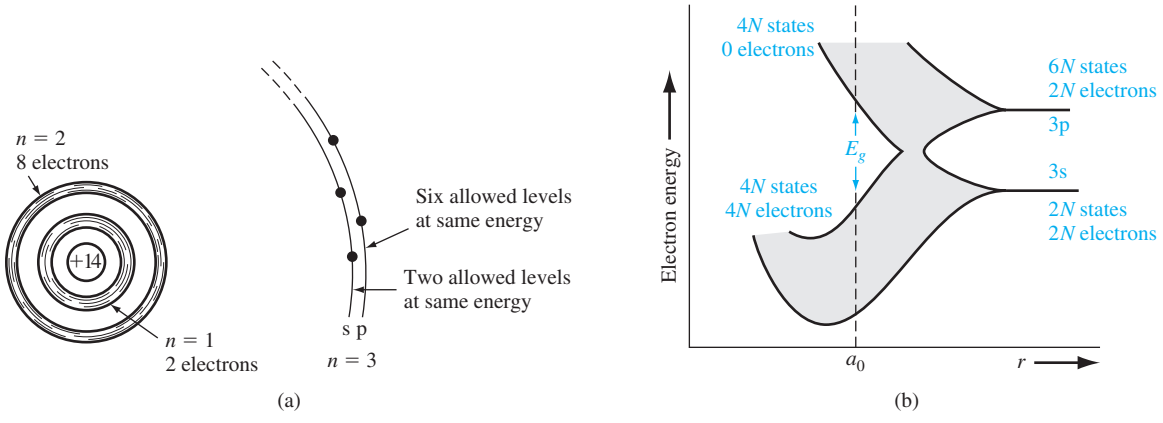


Figure 3.4 | (a) Schematic of an isolated silicon atom. (b) The splitting of the 3s and 3p states of silicon into the allowed and forbidden energy bands.
(From Shockley [6].)

electrons, since the first two energy shells are completely full and are tightly bound to the nucleus. The 3s state co rresponds to $n = 3$ and $l = 0$ and contains two quantum states per atom. This state will contain two electrons at $T = 0$ K. The 3p state corresponds to $n = 3$ and $l = 1$ and contains six quantum states per atom. This state will contain the remaining two electrons in the individual silicon atom.

As the interatomic distance decreases, the 3s and 3p states interact and overlap. At the equilibrium interatomic distance, the bands have again split, but now four quantum states per atom are in the lower band and four quantum states per atom are in the upper band. At absolute zero degrees, electrons are in the lowest energy state, 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 bandgap energy E_g between the top of the valence band and the bottom of the conduction band is the width of the forbidden energy band.

We have discussed qualitatively how and why bands of allowed and forbidden energies are formed in a crystal. The formation of these energy bands is directly related to the electrical characteristics of the crystal, as we will see later in our discussion.

*3.1.2 The Kronig–Penney Model[1]

In the previous section, we discussed qualitatively the splitting of allowed electron energies as atoms are brought together to form a crystal. The concept of allowed and forbidden energy bands can be developed more rigorously by considering quantum mechanics and Schrodinger's wave equation. It may be easy for the reader to "get lost" in the following derivation, but the result forms the basis for the energy-band theory of semiconductors.

The potential function of a single, noninteracting, one-electron atom is shown in Figure 3.5a. Also indicated on the figure are the discrete energy levels allowed for the electron. Figure 3.5b shows the same type of potential function for the case when several atoms in close proximity are arranged in a one-dimensional array. The potential functions of adjacent atoms overlap, and the net potential function for this case is shown in Figure 3.5c. It is this potential function we would need to use in Schrodinger's wave equation to model a one-dimensional single-crystal material.

The solution to Schrodinger's wave equation, for this one-dimensional single-crystal lattice, is made more tractable by considering a simpler potential function. Figure 3.6 is the one-dimensional Kronig–Penney model of the periodic potential function, which is used to represent a one-dimensional single-crystal lattice. We need to solve Schrodinger's wave equation in each region. As with previous quantum mechanical problems, the more interesting solution occurs for the case when $E < V_0$, which corresponds to a particle being bound within the crystal. The electrons are contained in the potential wells, but we have the possibility of tunneling between wells. The Kronig–Penney model is an idealized periodic potential representing a one-dimensional single crystal, but the results will illustrate many of the important features of the quantum behavior of electrons in a periodic lattice.

To obtain the solution to Schrodinger's wave equation, we make use of a mathematical theorem by Bloch. The theorem states that all one-electron wave functions, for problems involving periodically varying potential energy functions, must be of the form

$$\psi(x) = u(x)e^{jkx} \tag{3.1}$$

*Indicates sections that will aid in the total summation of understanding of semiconductor devices, but may be skipped the first time through the text without loss of continuity.

[1]Other techniques, such as the nearly free electron model, can be used to predict the energy-band theory of semiconductor materials. See, for example, Kittel [3] or Wolfe et al. [14].

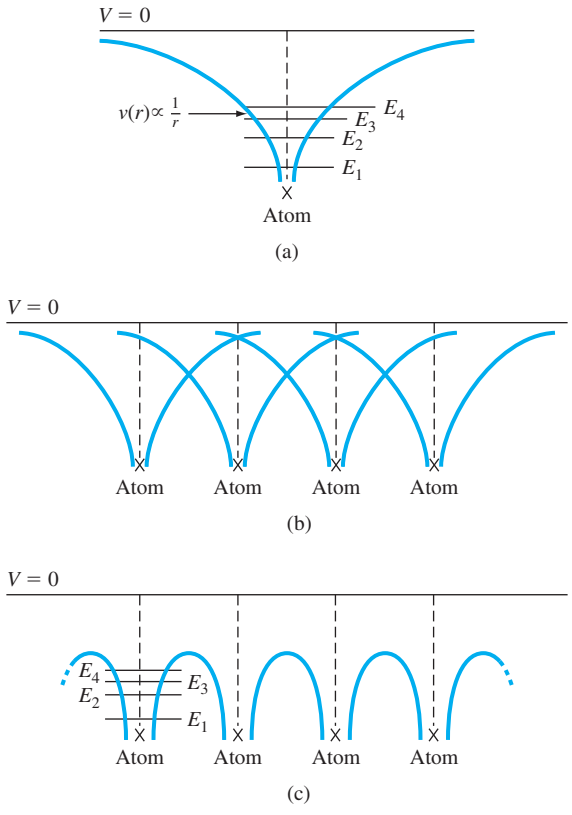


Figure 3.5 | (a) Potential function of a single isolated atom. (b) Overlapping potential functions of adjacent atoms. (c) Net potential function of a one-dimensional single crystal.

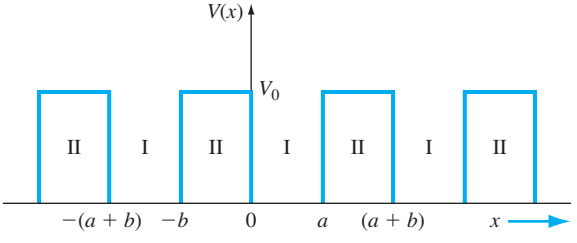


Figure 3.6 | The one-dimensional periodic potential function of the Kronig–Penney model.

The parameter k is called a constant of motion and will be considered in more detail as we develop the theory. The function $u(x)$ is a periodic function with period $(a + b)$.

We stated in Chapter 2 that the total solution to the wave equation is the product of the time-independent solution and the time-dependent solution, or

$$\Psi(x, t) = \psi(x)\phi(t) = u(x)e^{jkx} \cdot e^{-j(E/\hbar)t} \tag{3.2}$$

which may be written as

$$\Psi(x, t) = u(x)e^{j(kx-(E/\hbar)t)} \tag{3.3}$$

This traveling-wave solution represents the motion of an electron in a single-crystal material. The amplitude of the traveling wave is a periodic function and the parameter k is also referred to as a wave number.

We can now begin to determine a relation between the parameter k, the total energy E, and the potential V_0. If we consider region I in Figure 3.6 $(0 < x < a)$ in which $V(x) = 0$, take the second derivative of Equation (3.1), and substitute this result into the time-independent Schrodinger's wave equation given by Equation (2.13), we obtain the relation

$$\frac{d^2u_1(x)}{dx^2} + 2jk\frac{du_1(x)}{dx} - (k^2 - \alpha^2)u_1(x) = 0 \tag{3.4}$$

The function $u_1(x)$ is the amplitude of the wave function in region I and the parameter α is defined as

$$\alpha^2 = \frac{2mE}{\hbar^2} \tag{3.5}$$

Consider now a specific region II, $-b < x < 0$, in which $V(x) = V_0$, and apply Schrodinger's wave equation. We obtain the relation

$$\frac{d^2u_2(x)}{dx^2} + 2jk\frac{du_2(x)}{dx} - \left(k^2 - \alpha^2 + \frac{2mV_0}{\hbar^2}\right)u_2(x) = 0 \tag{3.6}$$

where $u_2(x)$ is the amplitude of the wave function in region II. We may define

$$\frac{2m}{\hbar^2}(E - V_0) = \alpha^2 - \frac{2mV_0}{\hbar^2} = \beta^2 \tag{3.7}$$

so that Equation (3.6) may be written as

$$\frac{d^2u_2(x)}{dx^2} + 2jk\frac{du_2(x)}{dx} - (k^2 - \beta^2)u_2(x) = 0 \tag{3.8}$$

Note that from Equation (3.7), if $E > V_0$, the parameter β is real, whereas if $E < V_0$, then β is imaginary.

The solution to Equation (3.4), for region I, is of the form

$$u_1(x) = Ae^{j(\alpha-k)x} + Be^{-j(\alpha+k)x} \quad \text{for } (0 < x < a) \tag{3.9}$$

and the solution to Equation (3.8), for region II, is of the form

$$u_2(x) = Ce^{j(\beta-k)x} + De^{-j(\beta+k)x} \quad \text{for } (-b < x < 0) \tag{3.10}$$

Since the potential function $V(x)$ is everywhere finite, both the wave function $\psi(x)$ and its first derivative $\partial\psi(x)/\partial x$ must be continuous. This continuity condition implies that the wave amplitude function $u(x)$ and its first derivative $\partial u(x)/\partial x$ must also be continuous.

If we consider the boundary at $x = 0$ and apply the continuity condition to the wave amplitude, we have

$$u_1(0) = u_2(0) \tag{3.11}$$

Substituting Equations (3.9) and (3.10) into Equation (3.11), we obtain

$$A + B - C - D = 0 \tag{3.12}$$

Now applying the condition that

$$\left.\frac{du_1}{dx}\right|_{x=0} = \left.\frac{du_2}{dx}\right|_{x=0} \tag{3.13}$$

we obtain

$$(\alpha - k)A - (\alpha + k)B - (\beta - k)C + (\beta + k)D = 0 \tag{3.14}$$

We have considered region I as $0 < x < a$ and region II as $-b < x < 0$. The periodicity and the continuity condition mean that the function u_1, as $x \to a$, is equal to the function u_2, as $x \to -b$. This condition may be written as

$$u_1(a) = u_2(-b) \tag{3.15}$$

Applying the solutions for $u_1(x)$ and $u_2(x)$ to the boundary condition in Equation (3.15) yields

$$Ae^{j(\alpha-k)a} + Be^{-j(\alpha+k)a} - Ce^{-j(\beta-k)b} - De^{j(\beta+k)b} = 0 \tag{3.16}$$

The last boundary condition is

$$\left.\frac{du_1}{dx}\right|_{x=a} = \left.\frac{du_2}{dx}\right|_{x=-b} \tag{3.17}$$

which gives

$$\begin{aligned}(\alpha - k)Ae^{j(\alpha-k)a} - (\alpha + k)Be^{-j(\alpha+k)a} - (\beta - k)Ce^{-j(\beta-k)b} \\ + (\beta + k)De^{j(\beta+k)b} = 0\end{aligned} \tag{3.18}$$

We now have four homogeneous equations, Equations (3.12), (3.14), (3.16), and (3.18), with four unknowns as a result of applying the four boundary conditions. In a set of simultaneous, linear, homogeneous equations, there is a nontrivial solution if, and only if, the determinant of the coefficients is zero. In our case, the coefficients in question are the coefficients of the parameters A, B, C, and D.

The evaluation of this determinant is extremely laborious and will not be considered in detail. The result is

$$\frac{-(\alpha^2 + \beta^2)}{2\alpha\beta}(\sin \alpha a)(\sin \beta b) + (\cos \alpha a)(\cos \beta b) = \cos k(a + b) \tag{3.19}$$

Equation (3.19) relates the parameter k to the total energy E (through the parameter α) and the potential function V_0 (through the parameter β).

As we mentioned, the more interesting solutions occur for $E < V_0$, which applies to the electron bound within the crystal. From Equation (3.7), the parameter β is then an imaginary quantity. We may define

$$\beta = j\gamma \tag{3.20}$$

where γ is a real quantity. Equation (3.19) can be written in terms of γ as

$$\frac{\gamma^2 - \alpha^2}{2\alpha\gamma}(\sin \alpha a)(\sinh \gamma b) + (\cos \alpha a)(\cosh \gamma b) = \cos k(a + b) \tag{3.21}$$

Equation (3.21) does not lend itself to an analytical solution, but must be solved using numerical or graphical techniques to obtain the relation between k, E, and V_0. The solution of Schrodinger's wave equation for a single bound particle resulted in discrete allowed energies. The solution of Equation (3.21) will result in a band of allowed energies.

To obtain an equation that is more susceptible to a graphical solution and thus will illustrate the nature of the results, let the potential barrier width $b \to 0$ and the barrier height $V_0 \to \infty$, but such that the product bV_0 remains finite. Equation (3.21) then reduces to

$$\left(\frac{mV_0ba}{\hbar^2}\right)\frac{\sin \alpha a}{\alpha a} + \cos \alpha a = \cos ka \tag{3.22}$$

We may define a parameter P' as

$$P' = \frac{mV_0ba}{\hbar^2} \tag{3.23}$$

Then, finally, we have the relation

$$P'\frac{\sin \alpha a}{\alpha a} + \cos \alpha a = \cos ka \tag{3.24}$$

Equation (3.24) again gives the relation between the parameter k, total energy E (through the parameter α), and the potential barrier bV_0. We may note that Equation (3.24) is not a solution of Schrodinger's wave equation but gives the conditions for which Schrodinger's wave equation will have a solution. If we assume that the crystal is infinitely large, then k in Equation (3.24) can assume a continuum of values and must be real.

3.1.3 The *k*-Space Diagram

To begin to understand the nature of the solution, initially consider the special case for which $V_0 = 0$. In this case $P' = 0$, which corresponds to a free particle since there are no potential barriers. From Equation (3.24), we have that

$$\cos \alpha a = \cos ka \tag{3.25}$$

or

$$\alpha = k \tag{3.26}$$

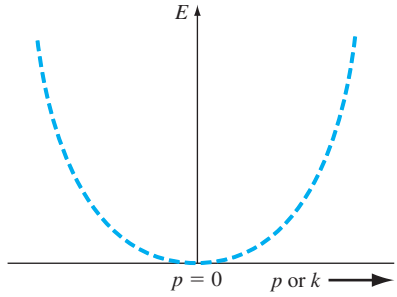


Figure 3.7 | The parabolic E versus k curve for the free electron.

Since the potential is equal to zero, the total energy E is equal to the kinetic energy, so that, from Equation (3.5), Equation (3.26) may be written as

$$\alpha = \sqrt{\frac{2mE}{\hbar^2}} = \sqrt{\frac{2m\left(\frac{1}{2}mv^2\right)}{\hbar^2}} = \frac{p}{\hbar} = k \tag{3.27}$$

where p is the particle momentum. The constant of the motion parameter k is related to the particle momentum for the free electron. The parameter k is also referred to as a wave number.

We can also relate the energy and momentum as

$$E = \frac{p^2}{2m} = \frac{k^2\hbar^2}{2m} \tag{3.28}$$

Figure 3.7 shows the parabolic relation of Equation (3.28) between the energy E and momentum p for the free particle. Since the momentum and wave number are linearly related, Figure 3.7 is also the E versus k curve for the free particle.

We now want to consider the relation between E and k from Equation (3.24) for the particle in the single-crystal lattice. As the parameter P' increases, the particle becomes more tightly bound to the potential well or atom. We may define the left side of Equation (3.24) to be a function $f(\alpha a)$, so that

$$f(\alpha a) = P'\frac{\sin \alpha a}{\alpha a} + \cos \alpha a \tag{3.29}$$

Figure 3.8a is a plot of the first term of Equation (3.29) versus αa. Figure 3.8b shows a plot of the cos αa term and Figure 3.8c is the sum of the two terms, or $f(\alpha a)$.

Now from Equation (3.24), we also have that

$$f(\alpha a) = \cos ka \tag{3.30}$$

For Equation (3.30) to be valid, the allowed values of the $f(\alpha a)$ function must be bounded between $+1$ and -1. Figure 3.8c shows the allowed values of $f(\alpha a)$ and the allowed values of αa in the shaded areas. Also shown on the figure are the values of ka from the right side of Equation (3.30), which correspond to the allowed values of $f(\alpha a)$.

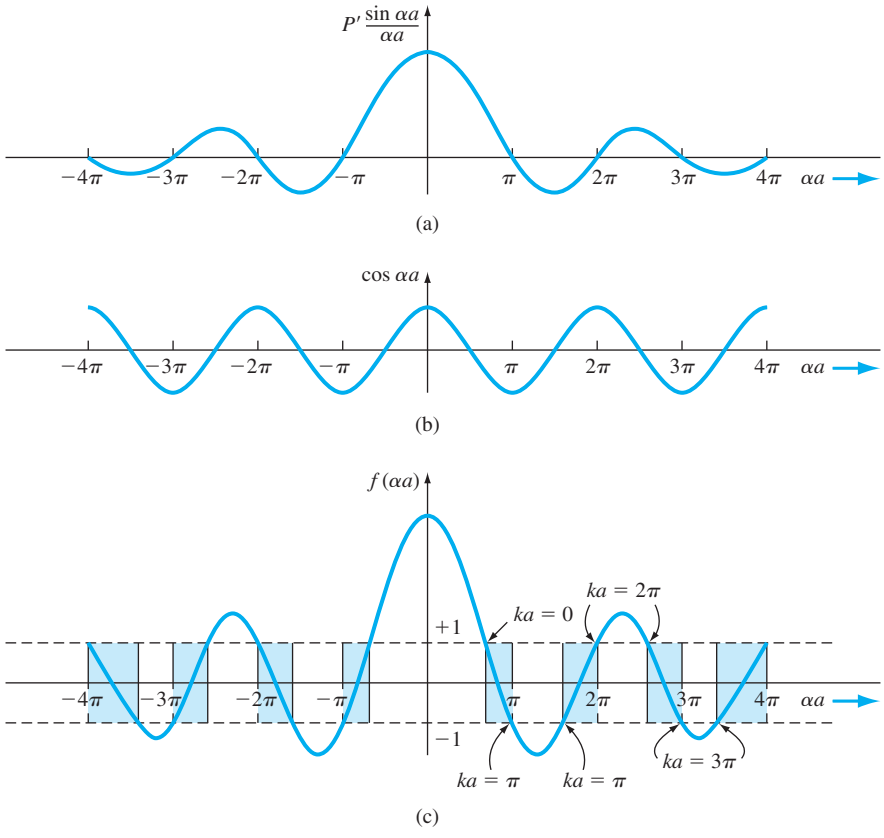


Figure 3.8 | A plot of (a) the first term in Equation (3.29), (b) the second term in Equation (3.29), and (c) the entire $f(\alpha a)$ function. The shaded areas show the allowed values of (αa) corresponding to real values of k.

The parameter α is related to the total energy E of the particle through Equation (3.5), which is $\alpha^2 = 2mE/\hbar^2$. A plot of the energy E of the particle as a function of the wave number k can be generated from Figure 3.8c. Figure 3.9 shows this plot and thus shows the concept of allowed energy bands for the particle propagating in the crystal lattice. Since the energy E has discontinuities, we also have the concept of forbidden energies for the particles in the crystal.

EXAMPLE 3.2

Objective: Determine the width (in eV) of a forbidden energy band.

Determine the width of the forbidden bandgap that exists at $ka = \pi$ (see Figure 3.9). Assume that the coefficient $P' = 8$ and the potential width is $a = 4.5$ Å.

■ **Solution**

Combining Equations (3.29) and (3.30), we have

$$\cos ka = P'\frac{\sin \alpha a}{\alpha a} + \cos \alpha a$$

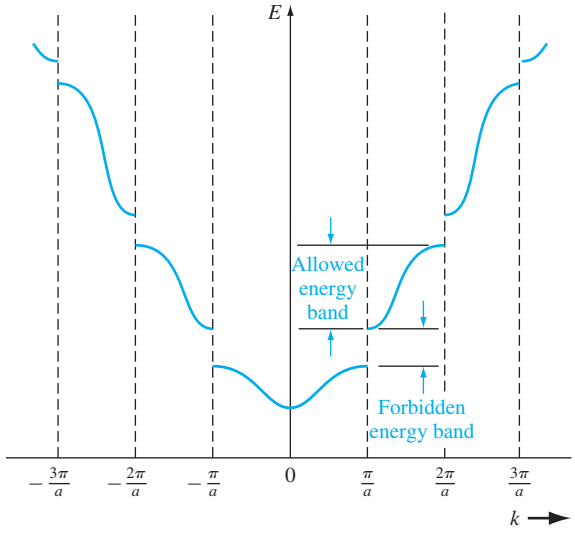


Figure 3.9 | The E versus k diagram generated from Figure 3.8. The allowed energy bands and forbidden energy bandgaps are indicated.

At $ka = \pi$ and using $P' = 8$, we have

$$-1 = 8\frac{\sin \alpha a}{\alpha a} + \cos \alpha a$$

We need to find the smallest values of αa that satisfy this equation and then relate α to the energy E to find the bandgap energy. From Figure 3.8, we see that, at one value of $ka = \pi$, we have $\alpha a = \pi \equiv \alpha_1 a$. Then

$$\alpha_1 a = \sqrt{\frac{2mE_1}{\hbar^2}} \cdot a = \pi$$

or

$$E_1 = \frac{\pi^2\hbar^2}{2ma^2} = \frac{\pi^2(1.054 \times 10^{-34})^2}{2(9.11 \times 10^{-31})(4.5 \times 10^{-10})^2} = 2.972 \times 10^{-19}\ \text{J}$$

From Figure 3.8, we see that, at the other value of $ka = \pi$, αa is in the range $\pi < \alpha a < 2\pi$. By trial and error, we find $\alpha a = 5.141 \equiv \alpha_2 a$. Then

$$\alpha_2 a = \sqrt{\frac{2mE_2}{\hbar^2}} \cdot a = 5.141$$

or

$$E_2 = \frac{(5.141)^2\hbar^2}{2ma^2} = \frac{(5.141)^2(1.054 \times 10^{-34})^2}{2(9.11 \times 10^{-31})(4.5 \times 10^{-10})^2} = 7.958 \times 10^{-19}\ \text{J}$$

The bandgap energy is then

$$E_g = E_2 - E_1 = 7.958 \times 10^{-19} - 2.972 \times 10^{-19} = 4.986 \times 10^{-19}\ \text{J}$$

or

$$E_g = \frac{4.986 \times 10^{-19}}{1.6 \times 10^{-19}} = 3.12 \text{ eV}$$

Comment

The results of this example give an order of magnitude of forbidden energy band widths.

EXERCISE PROBLEM

Ex 3.2 Using the parameters given in Example 3.2, determine the width of the allowed energy band in the range $\pi < ka < 2\pi$. (Ans. $\Delta E = 2.49$ eV)

Consider again the right side of Equation (3.24), which is the function cos *ka*. The cosine function is periodic so that

$$\cos ka = \cos(ka + 2n\pi) = \cos(ka - 2n\pi) \tag{3.31}$$

where n is a positive integer. We may consider Figure 3.9 and displace portions of the curve by 2π. Mathematically, Equation (3.24) is still satisfied. Figure 3.10 shows how various segments of the curve can be displaced by the 2π factor. Figure 3.11 shows the case in which the entire E versus k plot is contained within $-\pi/a < k < \pi/a$. This plot is referred to as a reduced k-space diagram, or a reduced-zone representation.

We noted in Equation (3.27) that for a free electron, the particle momentum and the wave number k are related by $p = \hbar k$. Given the similarity between the free

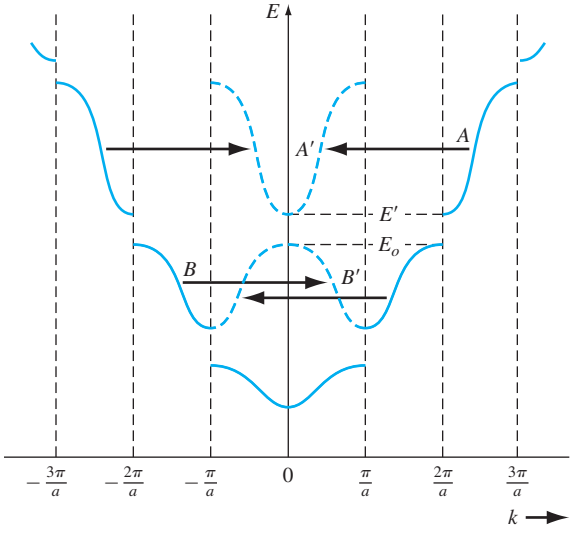


Figure 3.10 | The E versus k diagram showing 2π displacements of several sections of allowed energy bands.

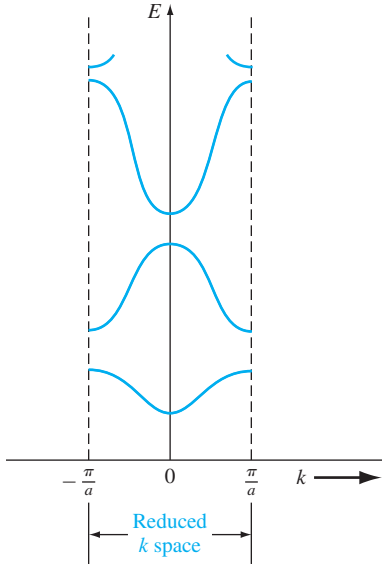


Figure 3.11 | The E versus k diagram in the reduced-zone representation.

electron solution and the results of the single crystal shown in Figure 3.9, the parameter $\hbar k$ in a single crystal is referred to as the *crystal momentum.* This parameter is not the actual momentum of the electron in the crystal but is a constant of the motion that includes the crystal interaction.

We have been considering the Kronig–Penney model, which is a one-dimensional periodic potential function used to model a single-crystal lattice. The principal result of this analysis, so far, is that electrons in the crystal occupy certain allowed energy bands and are excluded from the forbidden energy bands. For real three-dimensional single-crystal materials, a similar energy-band theory exists. We will obtain additional electron properties from the Kronig–Penney model in the next sections.

TEST YOUR UNDERSTANDING

TYU 3.1 Using the parameters given in Example 3.2, determine the width (in eV) of the second forbidden energy band existing at $ka = 2\pi$ (see Figure 3.8(c)).
(Ans. $E_g = 4.23$ eV)

TYU 3.2 Using the parameters given in Example 3.2, determine the width (in eV) of the allowed energy band in the range $0 < ka < \pi$ (see Figure 3.8(c)).
(Ans. $E = 0.6594$ eV)

3.2 | ELECTRICAL CONDUCTION IN SOLIDS

Again, we are eventually interested in determining the current–voltage characteristics of semiconductor devices. We will need to consider electrical conduction in solids as it relates to the band theory we have just developed. Let us begin by considering the motion of electrons in the various allowed energy bands.

3.2.1 The Energy Band and the Bond Model

In Chapter 1, we discussed the covalent bonding of silicon. Figure 3.12 shows a two-dimensional representation of the covalent bonding in a single-crystal silicon lattice. This figure represents silicon at $T = 0$ K in which each silicon atom is surrounded by eight valence electrons that are in their lowest energy state and are directly involved in the covalent bonding. Figure 3.4b represented the splitting of the discrete silicon energy states into bands of allowed energies as the silicon crystal is formed. At $T = 0$ K, the $4N$ states in the lower band, the valence band, are filled with the valence electrons. All of the valence electrons schematically shown in Figure 3.12 are in the valence band. The upper energy band, the conduction band, is completely empty at $T = 0$ K.

As the temperature increases above 0 K, a few valence band electrons may gain enough thermal energy to break the covalent bond and jump into the conduction band. Figure 3.13a shows a two-dimensional representation of this bond-breaking

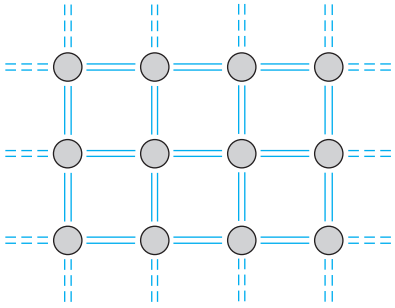


Figure 3.12 | Two-dimensional representation of the covalent bonding in a semiconductor at $T = 0$ K.

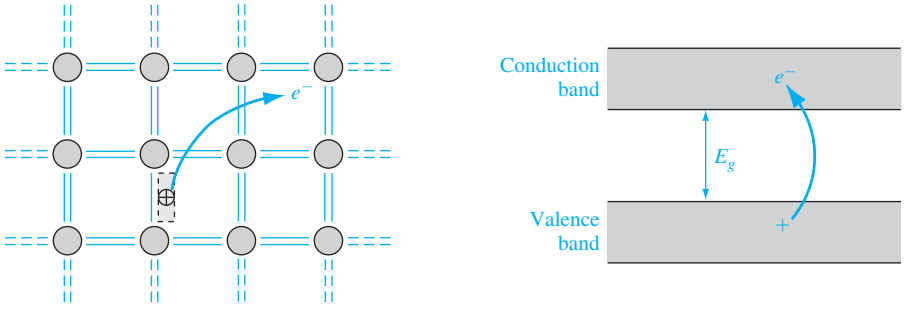


Figure 3.13 | (a) Two-dimensional representation of the breaking of a covalent bond. (b) Corresponding line representation of the energy band and the generation of a negative and positive charge with the breaking of a covalent bond.

effect and Figure 3.13b, a simple line representation of the energy-band model, shows the same effect.

The semiconductor is neutrally charged. This means that, as the negatively charged electron breaks away from its covalent bonding position, a positively charged "empty state" is created in the original covalent bonding position in the valence band. As the temperature further increases, more covalent bonds are broken, more electrons jump to the conduction band, and more positive "empty states" are created in the valence band.

We can also relate this bond breaking to the E versus k energy bands. Figure 3.14a shows the E versus k diagram of the conduction and valence bands at $T = 0$ K. The energy states in the valence band are completely full and the states in the conduction band are empty. Figure 3.14b shows these same bands for $T > 0$ K, in which some electrons have gained enough energy to jump to the conduction band and have left empty states in the valence band. We are assuming at this point that no external forces are applied so the electron and "empty state" distributions are symmetrical with k.

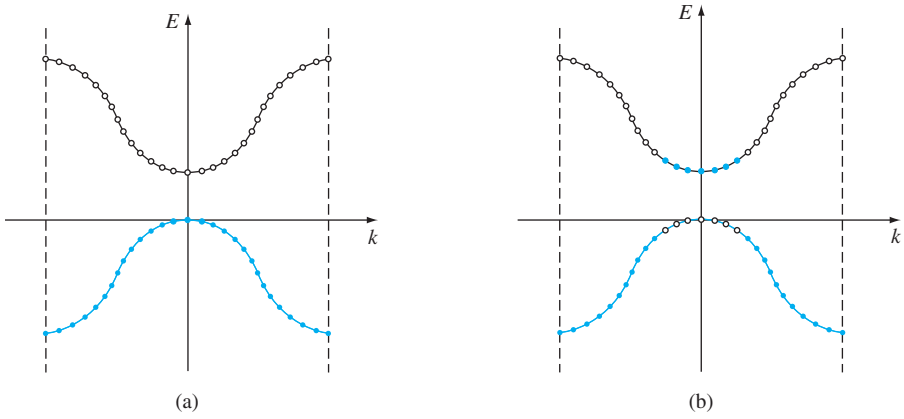


Figure 3.14 | The E versus k diagram of the conduction and valence bands of a semiconductor at (a) $T = 0$ K and (b) $T > 0$ K.

3.2.2 Drift Current

Current is due to the net flow of charge. If we had a collection of positively charged ions with a volume density N (cm^{-3}) and an average drift velocity v_d (cm/s), then the drift current density would be

$$J = qNv_d \qquad \text{A/cm}^2 \tag{3.32}$$

If, instead of considering the average drift velocity, we considered the individual ion velocities, then we could write the drift current density as

$$J = q\sum_{i=1}^{N} v_i \tag{3.33}$$

where v_i is the velocity of the ith ion. The summation in Equation (3.33) is taken over a unit volume so that the current density J is still in units of A/cm^2.

Since electrons are charged particles, a net drift of electrons in the conduction band will give rise to a current. The electron distribution in the conduction band, as shown in Figure 3.14b, is an even function of k when no external force is applied. Recall that k for a free electron is related to momentum so that, since there are as many electrons with a $+|k|$ value as there are with a $-|k|$ value, the net drift current density due to these electrons is zero. This result is certainly expected since there is no externally applied force.

If a force is applied to a particle and the particle moves, it must gain energy. This effect is expressed as

$$dE = F\,dx = F\,v\,dt \tag{3.34}$$

where F is the applied force, dx is the differential distance the particle moves, v is the velocity, and dE is the increase in energy. If an external force is applied to the electrons in the conduction band, there are empty energy states into which the electrons can move; therefore, because of the external force, electrons can gain energy and a

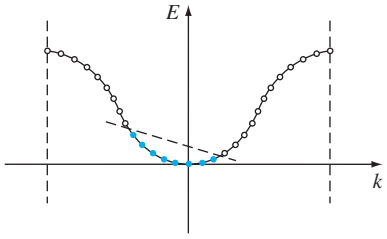


Figure 3.15 | The asymmetric distribution of electrons in the E versus k diagram when an external force is applied.

net momentum. The electron distribution in the conduction band may look like that shown in Figure 3.15, which implies that the electrons have gained a net momentum.

We may write the drift current density due to the motion of electrons as

$$J = -e \sum_{i=1}^{n} v_i \tag{3.35}$$

where e is the magnitude of the electronic charge and n is the number of electrons per unit volume in the conduction band. Again, the summation is taken over a unit volume so that the current density is still in units of A/cm^2. We may note from Equation (3.35) that the current is directly related to the electron velocity; that is, the current is related to how well the electron can move in the crystal.

3.2.3 Electron Effective Mass

The movement of an electron in a lattice will, in general, be different from that of an electron in free space. In addition to an externally applied force, there are internal forces in the crystal due to positively charged ions or protons and negatively charged electrons, which will influence the motion of electrons in the lattice. We can write

$$F_{\text{total}} = F_{\text{ext}} + F_{\text{int}} = ma \tag{3.36}$$

where F_{total}, F_{ext}, and F_{int} are the total force, the externally applied force, and the internal forces, respectively, acting on a particle in a crystal. The parameter a is the acceleration and m is the rest mass of the particle.

Since it is difficult to take into account all of the internal forces, we will write the equation

$$F_{\text{ext}} = m^* a \tag{3.37}$$

where the acceleration a is now directly related to the external force. The parameter m^*, called the *effective mass,* takes into account the particle mass and also takes into account the effect of the internal forces.

To use an analogy for the effective mass concept, consider the difference in motion between a glass marble in a container filled with water and in a container filled with oil. In general, the marble will drop through the water at a faster rate than through the oil. The external force in this example is the gravitational force and the internal

forces are related to the viscosity of the liquids. Because of the difference in motion of the marble in these two cases, the mass of the marble would appear to be different in water than in oil. (As with any analogy, we must be careful not to be too literal.)

We can also relate the effective mass of an electron in a crystal to the E versus k curves, such as is shown in Figure 3.11. In a semiconductor material, we will be dealing with allowed energy bands that are almost empty of electrons and other energy bands that are almost full of electrons.

To begin, consider the case of a free electron whose E versus k curve is shown in Figure 3.7. Recalling Equation (3.28), the energy and momentum are related by $E = p^2/2m = \hbar^2k^2/2m$, where m is the mass of the electron. The momentum and wave number k are related by $p = \hbar k$. If we take the derivative of Equation (3.28) with respect to k, we obtain

$$\frac{dE}{dk} = \frac{\hbar^2 k}{m} = \frac{\hbar p}{m} \tag{3.38}$$

Relating momentum to velocity, Equation (3.38) can be written as

$$\frac{1}{\hbar}\frac{dE}{dk} = \frac{p}{m} = v \tag{3.39}$$

where v is the velocity of the particle. The first derivative of E with respect to k is related to the velocity of the particle.

If we now take the second derivative of E with respect to k, we have

$$\frac{d^2E}{dk^2} = \frac{\hbar^2}{m} \tag{3.40}$$

We may rewrite Equation (3.40) as

$$\boxed{\frac{1}{\hbar^2}\frac{d^2E}{dk^2} = \frac{1}{m}} \tag{3.41}$$

The second derivative of E with respect to k is inversely proportional to the mass of the particle. For the case of a free electron, the mass is a constant (nonrelativistic effect), so the second derivative function is a constant. We may also note from Figure 3.7 that d^2E/dk^2 is a positive quantity, which implies that the mass of the electron is also a positive quantity.

If we apply an electric field to the free electron and use Newton's classical equation of motion, we can write

$$F = ma = -e\mathrm{E} \tag{3.42}$$

where a is the acceleration, E is the applied electric field, and e is the magnitude of the electronic charge. Solving for the acceleration, we have

$$a = \frac{-e\mathrm{E}}{m} \tag{3.43}$$

The motion of the free electron is in the opposite direction to the applied electric field because of the negative charge.

We may now apply the results to the electron in the bottom of an allowed energy band. Consider the allowed energy band in Figure 3.16a. The energy near the bottom

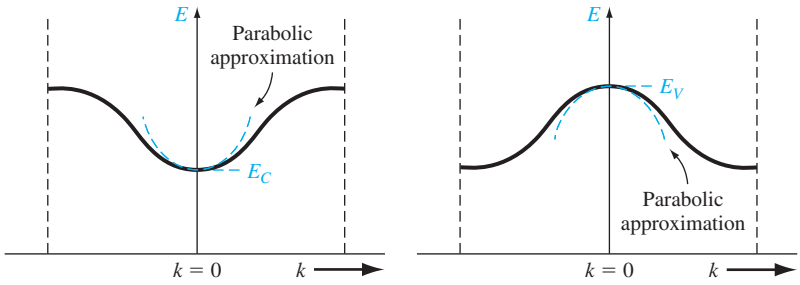


Figure 3.16 | (a) The conduction band in reduced k space, and the parabolic approximation. (b) The valence band in reduced k space, and the parabolic approximation.

of this energy band may be approximated by a parabola, just as that of a free particle. We may write

$$E - E_c = C_1(k)^2 \tag{3.44}$$

The energy E_c is the energy at the bottom of the band. Since $E > E_c$, the parameter C_1 is a positive quantity.

Taking the second derivative of E with respect to k from Equation (3.44), we obtain

$$\frac{d^2E}{dk^2} = 2C_1 \tag{3.45}$$

We may put Equation (3.45) in the form

$$\frac{1}{\hbar^2}\frac{d^2E}{dk^2} = \frac{2C_1}{\hbar^2} \tag{3.46}$$

Comparing Equation (3.46) with Equation (3.41), we may equate $\hbar^2/2C_1$ to the mass of the particle. However, the curvature of the curve in Figure 3.16a will not, in general, be the same as the curvature of the free-particle curve. We may write

$$\frac{1}{\hbar^2}\frac{d^2E}{dk^2} = \frac{2C_1}{\hbar^2} = \frac{1}{m^*} \tag{3.47}$$

where m^* is called the effective mass. Since $C_1 > 0$, we have that $m^* > 0$ also.

The effective mass is a parameter that relates the quantum mechanical results to the classical force equations. In most instances, the electron in the bottom of the conduction band can be thought of as a classical particle whose motion can be modeled by Newtonian mechanics, provided that the internal forces and quantum mechanical properties are taken into account through the effective mass. If we apply an electric field to the electron in the bottom of the allowed energy band, we may write the acceleration as

$$a = \frac{-e\mathrm{E}}{m_n^*} \tag{3.48}$$

where m_n^* is the effective mass of the electron. The effective mass m_n^* of the electron near the bottom of the conduction band is a constant.

3.2.4 Concept of the Hole

In considering the two-dimensional representation of the covalent bonding shown in Figure 3.13a, a positively charged "empty state" was created when a valence electron was elevated into the conduction band. For $T > 0$ K, all valence electrons may gain thermal energy; if a valence electron gains a small amount of thermal energy, it may hop into the empty state. The movement of a valence electron into the empty state is equivalent to the movement of the positively charged empty state itself. Figure 3.17 shows the movement of valence electrons in the crystal, alternately filling one empty state and creating a new empty state—a motion equivalent to a positive charge moving in the valence band. The crystal now has a second equally important charge carrier that can give rise to a current. This charge carrier is called a *hole* and, as we will see, can also be thought of as a classical particle whose motion can be modeled using Newtonian mechanics.

The drift current density due to electrons in the valence band, such as shown in Figure 3.14b, can be written as

$$J = -e \sum_{i\,(\text{filled})} v_i \tag{3.49}$$

where the summation extends over all filled states. This summation is inconvenient since it extends over a nearly full valence band and takes into account a very large number of states. We may rewrite Equation (3.49) in the form

$$J = -e \sum_{i\,(\text{total})} v_i + e \sum_{i\,(\text{empty})} v_i \tag{3.50}$$

If we consider a band that is totally full, all available states are occupied by electrons. The individual electrons can be thought of as moving with a velocity as given by Equation (3.39):

$$v(E) = \left(\frac{1}{\hbar}\right)\left(\frac{dE}{dk}\right) \tag{3.39}$$

The band is symmetric in k and each state is occupied so that, for every electron with a velocity $|v|$, there is a corresponding electron with a velocity $-|v|$. Since the band is full, the distribution of electrons with respect to k cannot be changed with an

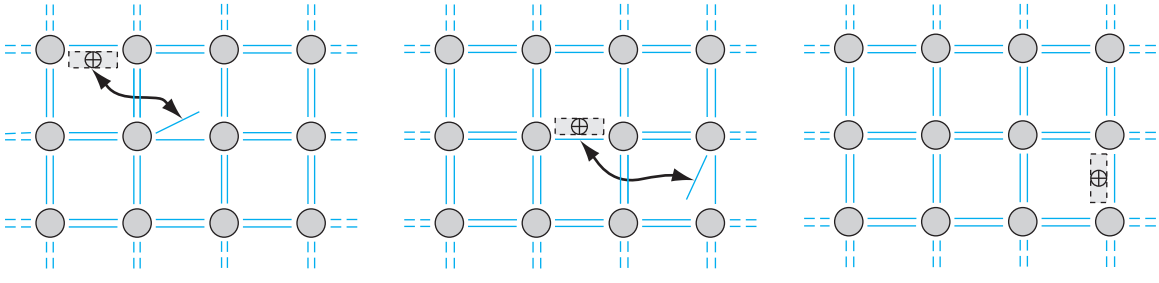


Figure 3.17 | Visualization of the movement of a hole in a semiconductor.

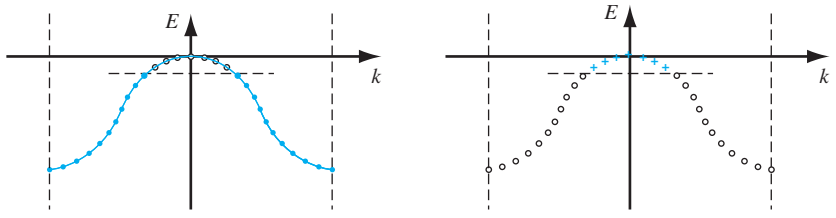


Figure 3.18 | (a) Valence band with conventional electron-filled states and empty states. (b) Concept of positive charges occupying the original empty states.

externally applied force. The net drift current density generated from a completely full band, then, is zero, or

$$-e\sum_{i\,(\text{total})} v_i \equiv 0 \tag{3.51}$$

We can now write the drift current density from Equation (3.50) for an almost full band as

$$J = +e\sum_{i\,(\text{empty})} v_i \tag{3.52}$$

where the v_i in the summation is the

$$v(E) = \left(\frac{1}{\hbar}\right)\left(\frac{dE}{dk}\right)$$

associated with the empty state. Equation (3.52) is entirely equivalent to placing a positively charged particle in the empty states and assuming all other states in the band are empty, or neutrally charged. This concept is shown in Figure 3.18. Figure 3.18a shows the valence band with the conventional electron-filled states and empty states, whereas Figure 3.18b shows the new concept of positive charges occupying the original empty states. This concept is consistent with the discussion of the positively charged "empty state" in the valence band, as shown in Figure 3.17.

The v_i in the summation of Equation (3.52) is related to how well this positively charged particle moves in the semiconductor. Now consider an electron near the top of the allowed energy band shown in Figure 3.16b. The energy near the top of the allowed energy band may again be approximated by a parabola so that we may write

$$(E - E_v) = -C_2(k)^2 \tag{3.53}$$

The energy E_v is the energy at the top of the energy band. Since $E < E_v$ for electrons in this band, the parameter C_2 must be a positive quantity.

Taking the second derivative of E with respect to k from Equation (3.53), we obtain

$$\frac{d^2E}{dk^2} = -2C_2 \tag{3.54}$$

We may rearrange this equation so that

$$\frac{1}{\hbar^2}\frac{d^2E}{dk^2} = \frac{-2C_2}{\hbar^2} \tag{3.55}$$

Comparing Equation (3.55) with Equation (3.41), we may write

$$\frac{1}{\hbar^2}\frac{d^2E}{dk^2} = \frac{-2C_2}{\hbar^2} = \frac{1}{m^*} \tag{3.56}$$

where m^* is again an effective mass. We have argued that C_2 is a positive quantity, which now implies that m^* is a negative quantity. An electron moving near the top of an allowed energy band behaves as if it has a negative mass.

We must keep in mind that the effective mass parameter is used to relate quantum mechanics and classical mechanics. The attempt to relate these two theories leads to this strange result of a negative effective mass. However, we must recall that solutions to Schrodinger's wave equation also led to results that contradicted classical mechanics. The negative effective mass is another such example.

In discussing the concept of effective mass in the previous section, we used an analogy of marbles moving through two liquids. Now consider placing an ice cube in the center of a container filled with water: the ice cube will move upward toward the surface in a direction opposite to the gravitational force. The ice cube appears to have a negative effective mass since its acceleration is opposite to the external force. The effective mass parameter takes into account all internal forces acting on the particle.

If we again consider an electron near the top of an allowed energy band and use Newton's force equation for an applied electric field, we will have

$$F = m^*a = -e\mathrm{E} \tag{3.57}$$

However, m^* is now a negative quantity, so we may write

$$a = \frac{-e\mathrm{E}}{-|m^*|} = \frac{+e\mathrm{E}}{|m^*|} \tag{3.58}$$

An electron moving near the top of an allowed energy band moves in the same direction as the applied electric field.

The net motion of electrons in a nearly full band can be described by considering just the empty states, provided that a positive electronic charge is associated with each state and that the negative of m^* from Equation (3.56) is associated with each state. We now can model this band as having particles with a positive electronic charge and a positive effective mass. The density of these particles in the valence band is the same as the density of empty electronic energy states. This new particle is the *hole*. The hole, then, has a positive effective mass denoted by m_p^* and a positive electronic charge, so it will move in the same direction as an applied field.

3.2.5 Metals, Insulators, and Semiconductors

Each crystal has its own energy-band structure. We noted that the splitting of the energy states in silicon, for example, to form the valence and conduction bands, was complex. Complex band splitting occurs in other crystals, leading to large variations in band structures between various solids and to a wide range of electrical characteristics observed in these various materials. We can qualitatively begin to understand some basic differences in electrical characteristics caused by variations in band structure by considering some simplified energy bands.

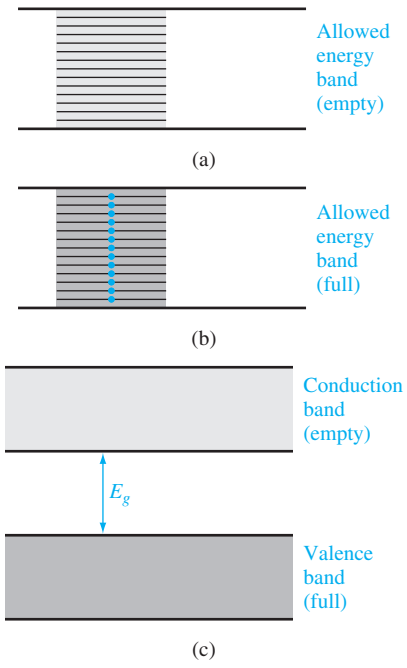


Figure 3.19 | Allowed energy bands showing (a) an empty band, (b) a completely full band, and (c) the bandgap energy between the two allowed bands.

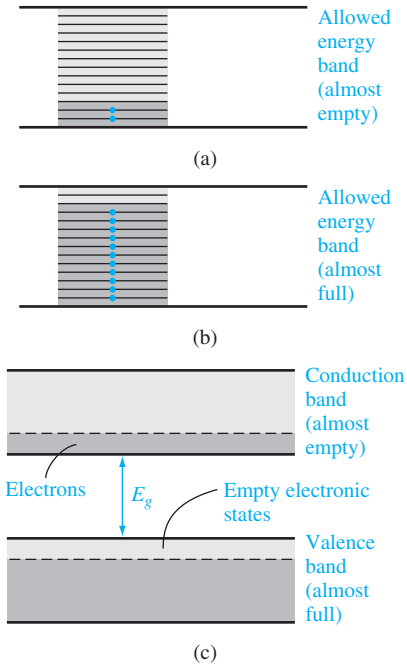


Figure 3.20 | Allowed energy bands showing (a) an almost empty band, (b) an almost full band, and (c) the bandgap energy between the two allowed bands.

There are several possible energy-band conditions to consider. Figure 3.19a shows an allowed energy band that is completely empty of electrons. If an electric field is applied, there are no particles to move, so there will be no current. Figure 3.19b shows another allowed energy band whose energy states are completely full of electrons. We argued in the previous section that a completely full energy band will also not give rise to a current. A material that has energy bands either completely empty or completely full is an insulator. The resistivity of an insulator is very large or, conversely, the conductivity of an insulator is very small. There are essentially no charged particles that can contribute to a drift current. Figure 3.19c shows a simplified energy-band diagram of an insulator. The bandgap energy E_g of an insulator is usually on the order of 3.5 to 6 eV or larger, so that at room temperature, there are essentially no electrons in the conduction band and the valence band remains completely full. There are very few thermally generated electrons and holes in an insulator.

Figure 3.20a shows an energy band with relatively few electrons near the bottom of the band. Now, if an electric field is applied, the electrons can gain energy, move to higher energy states, and move through the crystal. The net flow of charge is a current. Figure 3.20b shows an allowed energy band that is almost full of electrons, which means that we can consider the holes in this band. If an electric field is applied, the holes can move and give rise to a current. Figure 3.20c shows the simplified energy-band diagram for this case. The bandgap energy may be on the order of 1 eV.

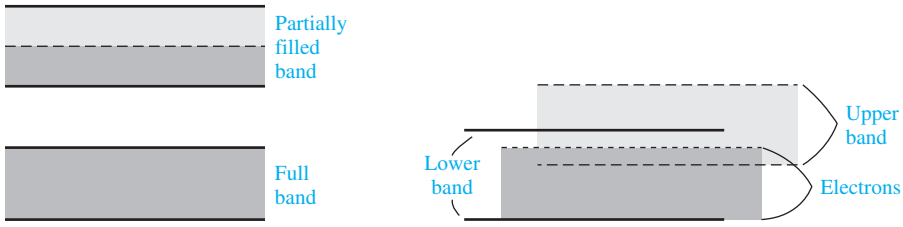


Figure 3.21 | Two possible energy bands of a metal showing (a) a partially filled band and (b) overlapping allowed energy bands.

This energy-band diagram represents a semiconductor for $T > 0$ K. The resistivity of a semiconductor, as we will see in the next chapter, can be controlled and varied over many orders of magnitude.

The characteristics of a metal include a very low resistivity. The energy-band diagram for a metal may be in one of two forms. Figure 3.21a shows the case of a partially full band in which there are many electrons available for conduction, so that the material can exhibit a large electrical conductivity. Figure 3.21b shows another possible energy-band diagram of a metal. The band splitting into allowed and forbidden energy bands is a complex phenomenon, and Figure 3.21b shows a case in which the conduction and valence bands overlap at the equilibrium interatomic distance. As in the case shown in Figure 3.21a, there are large numbers of electrons as well as large numbers of empty energy states into which the electrons can move, so this material can also exhibit a very high electrical conductivity.

TEST YOUR UNDERSTANDING

TYU 3.3 A simplified E versus k curve for an electron in the conduction band is given. The value of a is 10 Å. Determine the relative effective mass m^*/m_0.
(Ans. $m^*/m_0 = 1.175$)

TYU 3.4 A simplified E versus k curve for a hole in the valence band is given. Assume a value of $a = 12$ Å. Determine the relative effective mass $|m^*/m_0|$.
(Ans. $|m^*/m_0| = 0.2985$)

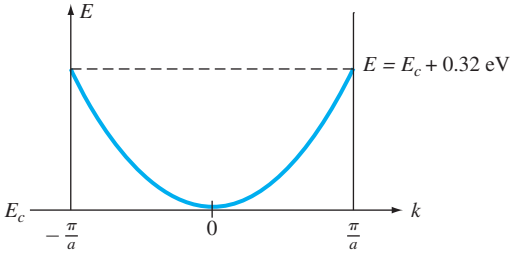


Figure 3.22 | Figure for Exercise TYU 3.3.

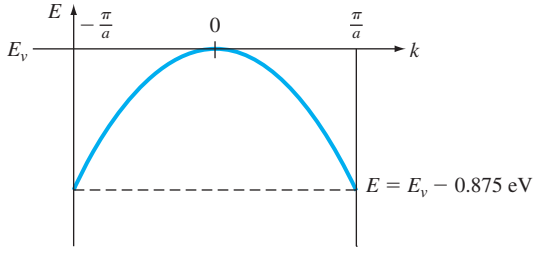


Figure 3.23 | Figure for Exercise TYU 3.4.

3.3 | EXTENSION TO THREE DIMENSIONS

The basic concepts of allowed and forbidden energy bands and effective mass have been developed in the previous sections. In this section, we will extend these concepts to three dimensions and to real crystals. We will qualitatively consider particular characteristics of the three-dimensional crystal in terms of the E versus k plots, bandgap energy, and effective mass. We must emphasize that we will only briefly touch on the basic three-dimensional concepts; therefore, many details will not be considered.

One problem encountered in extending the potential function to a three-dimensional crystal is that the distance between atoms varies as the direction through the crystal changes. Figure 3.24 shows a face-centered cubic structure with the [100] and [110] directions indicated. Electrons traveling in different directions encounter different potential patterns and therefore different k-space boundaries. The E versus k diagrams are, in general, a function of the k-space direction in a crystal.

3.3.1 The k-Space Diagrams of Si and GaAs

Figure 3.25 shows an E versus k diagram of gallium arsenide (GaAs) and of silicon (Si). These simplified diagrams show the basic properties considered in this text but do not show many of the details more appropriate for advanced-level courses.

Note that in place of the usual positive and negative k axes, we now show two different crystal directions. The E versus k diagram for the one-dimensional model was symmetric in k so that no new information is obtained by displaying the negative axis. It is normal practice to plot the [100] direction along the normal $+k$ axis and to plot the [111] portion of the diagram so the $+k$ points to the left. In the case of diamond or zincblende lattices, the maxima in the valence band energy and minima in the conduction band energy occur at $k = 0$ or along one of these two directions.

Figure 3.25a shows the E versus k diagram for GaAs. The valence band maximum and the conduction band minimum both occur at $k = 0$. The electrons in the

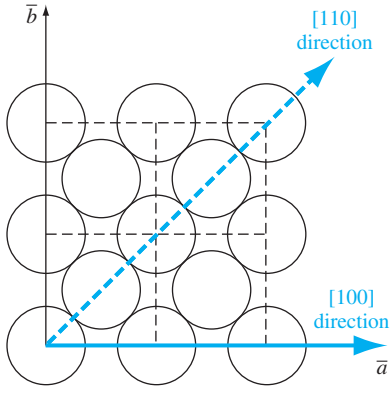


Figure 3.24 | The (100) plane of a face-centered cubic crystal showing the [100] and [110] directions.

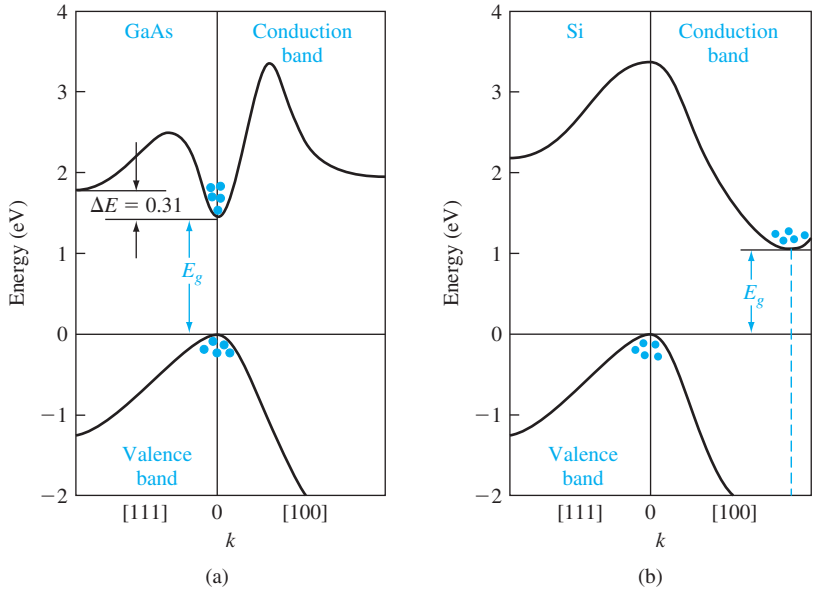


Figure 3.25 | Energy-band structures of (a) GaAs and (b) Si. *(From Sze [12].)*

conduction band tend to settle at the minimum conduction band energy that is at $k = 0$. Similarly, holes in the valence band tend to congregate at the uppermost valence band energy. In GaAs, the minimum conduction band energy and maximum valence band energy occur at the same k value. A semiconductor with this property is said to be a *direct* bandgap semiconductor; transitions between the two allowed bands can take place with no change in crystal momentum. This direct nature has significant effect on the optical properties of the material. GaAs and other direct bandgap materials are ideally suited for use in semiconductor lasers and other optical devices.

The E versus k diagram for silicon is shown in Figure 3.25b. The maximum in the valence band energy occurs at $k = 0$ as before. The minimum in the conduction band energy occurs not at $k = 0$, but along the [100] direction. The difference between the minimum conduction band energy and the maximum valence band energy is still defined as the bandgap energy E_g. A semiconductor whose maximum valence band energy and minimum conduction band energy do not occur at the same k value is called an *indirect* bandgap semiconductor. When electrons make a transition between the conduction and valence bands, we must invoke the law of conservation of momentum. A transition in an indirect bandgap material must necessarily include an interaction with the crystal so that crystal momentum is conserved.

Germanium is also an indirect bandgap material, whose valence band maximum occurs at $k = 0$ and whose conduction band minimum occurs along the [111] direction. GaAs is a direct bandgap semiconductor, but other compound semiconductors, such as GaP and AlAs, have indirect bandgaps.

3.3.2 Additional Effective Mass Concepts

The curvature of the E versus k diagrams near the minimum of the conduction band energy is related to the effective mass of the electron. We may note from Figure 3.25 that the curvature of the conduction band at its minimum value for GaAs is larger than that of silicon, so the effective mass of an electron in the conduction band of GaAs will be smaller than that in silicon.

For the one-dimensional E versus k diagram, the effective mass was defined by Equation (3.41) as $1/m^* = 1/\hbar^2 \cdot d^2E/dk^2$. A complication occurs in the effective mass concept in a real crystal. A three-dimensional crystal can be described by three k vectors. The curvature of the E versus k diagram at the conduction band minimum may not be the same in the three k directions. In later sections and chapters, the effective mass parameters used in calculations will be a kind of statistical average that is adequate for most device calculations.[2]

3.4 | DENSITY OF STATES FUNCTION

As we have stated, we eventually wish to describe the current–voltage characteristics of semiconductor devices. Since current is due to the flow of charge, an important step in the process is to determine the number of electrons and holes in the semiconductor that will be available for conduction. The number of carriers that can contribute to the conduction process is a function of the number of available energy or quantum states since, by the Pauli exclusion principle, only one electron can occupy a given quantum state. When we discussed the splitting of energy levels into bands of allowed and forbidden energies, we indicated that the band of allowed energies was actually made up of discrete energy levels. We must determine the density of these allowed energy states as a function of energy in order to calculate the electron and hole concentrations.

3.4.1 Mathematical Derivation

To determine the density of allowed quantum states as a function of energy, we need to consider an appropriate mathematical model. Electrons are allowed to move relatively freely in the conduction band of a semiconductor but are confined to the crystal. As a first step, we will consider a free electron confined to a three-dimensional infinite potential well, where the potential well represents the crystal. The potential of the infinite potential well is defined as

$$\begin{aligned} V(x, y, z) &= 0 \quad \text{for } 0 < x < a \\ &\qquad\quad 0 < y < a \\ &\qquad\quad 0 < z < a \\ V(x, y, z) &= \infty \quad \text{elsewhere} \end{aligned} \tag{3.59}$$

where the crystal is assumed to be a cube with length a. Schrodinger's wave equation in three dimensions can be solved by using the separation of variables technique.

[2]See Appendix F for further discussion of effective mass concepts.

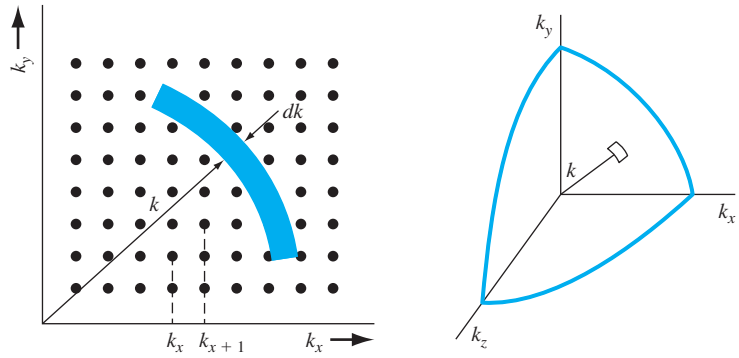


Figure 3.26 | (a) A two-dimensional array of allowed quantum states in k space. (b) The positive one-eighth of the spherical k space.

Extrapolating the results from the one-dimensional infinite potential well, we can show (see Problem 3.23) that

$$\frac{2mE}{\hbar^2} = k^2 = k_x^2 + k_y^2 + k_z^2 = \left(n_x^2 + n_y^2 + n_z^2\right)\left(\frac{\pi^2}{a^2}\right) \tag{3.60}$$

where n_x, n_y, and n_z are positive integers. (Negative values of n_x, n_y, and n_z yield the same wave function, except for the sign, as the positive integer values, resulting in the same probability function and energy; therefore the negative integers do not represent a different quantum state.)

We can schematically plot the allowed quantum states in k space. Figure 3.26a shows a two-dimensional plot as a function of k_x and k_y. Each point represents an allowed quantum state corresponding to various integral values of n_x and n_y. Positive and negative values of k_x, k_y, or k_z have the same energy and represent the same energy state. Since negative values of k_x, k_y, or k_z do not represent additional quantum states, the density of quantum states will be determined by considering only the positive one-eighth of the spherical k space as shown in Figure 3.26b.

The distance between two quantum states in the k_x direction, for example, is given by

$$k_{x+1} - k_x = (n_x + 1)\left(\frac{\pi}{a}\right) - n_x\left(\frac{\pi}{a}\right) = \frac{\pi}{a} \tag{3.61}$$

Generalizing this result to three dimensions, the volume V_k of a single quantum state is

$$V_k = \left(\frac{\pi}{a}\right)^3 \tag{3.62}$$

We can now determine the density of quantum states in k space. A differential volume in k space is shown in Figure 3.26b and is given by $4\pi k^2\,dk$, so the differential density of quantum states in k space can be written as

$$g_T(k)\,dk = 2\left(\frac{1}{8}\right)\frac{4\pi k^2\,dk}{\left(\frac{\pi}{a}\right)^3} \tag{3.63}$$

The first factor, 2, takes into account the two spin states allowed for each quantum state; the next factor, $\frac{1}{8}$, takes into account that we are considering only the quantum states for positive values of k_x, k_y, and k_z. The factor $4\pi k^2\,dk$ is again the differential volume and the factor $(\pi/a)^3$ is the volume of one quantum state. Equation (3.63) may be simplified to

$$g_T(k)\,dk = \frac{\pi k^2\,dk}{\pi^3}\cdot a^3 \tag{3.64}$$

Equation (3.64) gives the density of quantum states as a function of momentum, through the parameter k. We can now determine the density of quantum states as a function of energy E. For a free electron, the parameters E and k are related by

$$k^2 = \frac{2mE}{\hbar^2} \tag{3.65a}$$

or

$$k = \frac{1}{\hbar}\sqrt{2mE} \tag{3.65b}$$

The differential dk is

$$dk = \frac{1}{\hbar}\sqrt{\frac{m}{2E}}\,dE \tag{3.66}$$

Then, substituting the expressions for k^2 and dk into Equation (3.64), the number of energy states between E and $E + dE$ is given by

$$g_T(E)\,dE = \frac{\pi a^3}{\pi^3}\left(\frac{2mE}{\hbar^2}\right)\cdot\frac{1}{\hbar}\sqrt{\frac{m}{2E}}\,dE \tag{3.67}$$

Since $\hbar = h/2\pi$, Equation (3.67) becomes

$$g_T(E)\,dE = \frac{4\pi a^3}{h^3}\cdot(2m)^{3/2}\cdot\sqrt{E}\,dE \tag{3.68}$$

Equation (3.68) gives the total number of quantum states between the energy E and $E + dE$ in the crystal space volume of a^3. If we divide by the volume a^3, then we will obtain the density of quantum states per unit volume of the crystal. Equation (3.68) then becomes

$$g(E) = \frac{4\pi(2m)^{3/2}}{h^3}\sqrt{E} \tag{3.69}$$

The density of quantum states is a function of energy E. As the energy of this free electron becomes small, the number of available quantum states decreases. This density function is really a double density, in that the units are given in terms of states per unit energy per unit volume.

EXAMPLE 3.3

Objective: Calculate the density of states per unit volume over a particular energy range.

Consider the density of states for a free electron given by Equation (3.69). Calculate the density of states per unit volume with energies between 0 and 1 eV.

■ Solution

The volume density of quantum states, from Equation (3.69), is

$$N = \int_0^{1\text{ eV}} g(E)\,dE = \frac{4\pi(2m)^{3/2}}{h^3} \cdot \int_0^{1\text{ eV}} \sqrt{E}\,dE$$

or

$$N = \frac{4\pi(2m)^{3/2}}{h^3} \cdot \frac{2}{3} \cdot E^{3/2}$$

The density of states is now

$$N = \frac{4\pi[2(9.11 \times 10^{-31})]^{3/2}}{(6.625 \times 10^{-34})^3} \cdot \frac{2}{3} \cdot (1.6 \times 10^{-19})^{3/2} = 4.5 \times 10^{27}\ \text{m}^{-3}$$

or

$$N = 4.5 \times 10^{21}\ \text{states/cm}^3$$

■ Comment

The density of quantum states is typically a large number. An effective density of states in a semiconductor, as we will see in the following sections and in the next chapter, is also a large number but is usually less than the density of atoms in the semiconductor crystal.

■ EXERCISE PROBLEM

Ex 3.3 For a free electron, calculate the density of quantum states (#/cm^3) over the energy range of (a) $0 \le E \le 2.0$ eV and (b) $1 \le E \le 2$ eV.

[Ans. (a) $N = 1.28 \times 10^{22}\ \text{cm}^{-3}$; (b) $N = 8.29 \times 10^{21}\ \text{cm}^{-3}$]

3.4.2 Extension to Semiconductors

In the previous section, we derived a general expression for the density of allowed electron quantum states using the model of a free electron with mass m bounded in a three-dimensional infinite potential well. We can extend this same general model to a semiconductor to determine the density of quantum states in the conduction band and the density of quantum states in the valence band. Electrons and holes are confined within the semiconductor crystal, so we will again use the basic model of the infinite potential well.

The parabolic relationship between energy and momentum of a free electron is given in Equation (3.28) as $E = p^2/2m = \hbar^2k^2/2m$. Figure 3.16a shows the conduction energy band in the reduced k space. The E versus k curve near $k = 0$ at the bottom of the conduction band can be approximated as a parabola, so we may write

$$E = E_c + \frac{\hbar^2k^2}{2m_n^*} \tag{3.70}$$

where E_c is the bottom edge of the conduction band and m_n^* is the electron density of states effective mass.[3] Equation (3.70) may be rewritten to give

$$E - E_c = \frac{\hbar^2k^2}{2m_n^*} \tag{3.71}$$

[3]Again, see Appendix F for further discussion of effective mass concepts.

The general form of the E versus k relation for an electron in the bottom of a conduction band is the same as the free electron, except the mass is replaced by the effective mass. We can then think of the electron in the bottom of the conduction band as being a "free" electron with its own particular mass. The right side of Equation (3.71) is of the same form as the right side of Equation (3.28), which was used in the derivation of the density of states function. Because of this similarity, which yields the "free" conduction electron model, we may generalize the free electron results of Equation (3.69) and write the density of allowed electronic energy states in the conduction band as

$$g_c(E) = \frac{4\pi(2m_n^*)^{3/2}}{h^3}\sqrt{E - E_c} \tag{3.72}$$

Equation (3.72) is valid for $E \geq E_c$. As the energy of the electron in the conduction band decreases, the number of available quantum states also decreases.

The density of quantum states in the valence band can be obtained by using the same infinite potential well model, since the hole is also confined in the semiconductor crystal and can be treated as a "free" particle. The density of states effective mass of the hole is m_p^*. Figure 3.16b shows the valence energy band in the reduced k space. We may also approximate the E versus k curve near $k = 0$ by a parabola for a "free" hole, so that

$$E = E_v - \frac{\hbar^2 k^2}{2m_p^*} \tag{3.73}$$

Equation (3.73) may be rewritten to give

$$E_v - E = \frac{\hbar^2 k^2}{2m_p^*} \tag{3.74}$$

Again, the right side of Equation (3.74) is of the same form used in the general derivation of the density of states function. We may then generalize the density of states function from Equation (3.69) to apply to the valence band, so that

$$g_v(E) = \frac{4\pi(2m_p^*)^{3/2}}{h^3}\sqrt{E_v - E} \tag{3.75}$$

Equation (3.75) is valid for $E \leq E_v$.

We have argued that quantum states do not exist within the forbidden energy band, so $g(E) = 0$ for $E_v < E < E_c$. Figure 3.27 shows the plot of the density of quantum states as a function of energy. If the electron and hole effective masses were equal, then the functions $g_c(E)$ and $g_v(E)$ would be symmetrical about the energy midway between E_c and E_v, or the midgap energy, E_{midgap}.

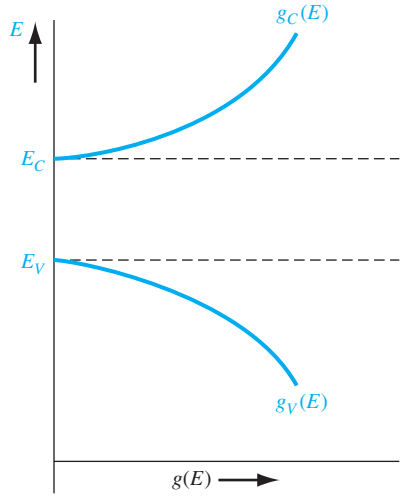


Figure 3.27 | The density of energy states in the conduction band and the density of energy states in the valence band as a function of energy.

EXAMPLE 3.4

Objective: Determine the number (#/cm^3) of quantum states in silicon between E_c and $E_c + kT$ at $T = 300$ K.

■ Solution

Using Equation (3.72), we can write

$$N = \int_{E_c}^{E_c + kT} \frac{4\pi(2m_n^*)^{3/2}}{h^3}\sqrt{E - E_c}\cdot dE$$

$$= \frac{4\pi(2m_n^*)^{3/2}}{h^3}\cdot\frac{2}{3}\cdot(E - E_c)^{3/2}\Big|_{E_c}^{E_c + kT}$$

$$= \frac{4\pi[2(1.08)(9.11\times10^{-31})]^{3/2}}{(6.625\times10^{-34})^3}\cdot\frac{2}{3}\cdot[(0.0259)(1.6\times10^{-19})]^{3/2}$$

$$= 2.12\times10^{25}\ \text{m}^{-3}$$

or

$$N = 2.12\times10^{19}\ \text{cm}^{-3}$$

■ Comment

The result of this example shows the order of magnitude of the density of quantum states in a semiconductor.

■ EXERCISE PROBLEM

Ex3.4 Determine the number (#/cm^3) of quantum states in silicon between $(E_v - kT)$ and E_v at $T = 300$ K. $(N = 2.72\times10^{18}\ \text{cm}^{-3})$

3.5 | STATISTICAL MECHANICS

In dealing with large numbers of particles, we are interested only in the statistical behavior of the group as a whole rather than in the behavior of each individual particle. For example, gas within a container will exert an average pressure on the walls of the vessel. The pressure is actually due to the collisions of the individual gas molecules with the walls, but we do not follow each individual molecule as it collides with the wall. Likewise in a crystal, the electrical characteristics will be determined by the statistical behavior of a large number of electrons.

3.5.1 Statistical Laws

In determining the statistical behavior of particles, we must consider the laws that the particles obey. There are three distribution laws determining the distribution of particles among available energy states.

One distribution law is the Maxwell–Boltzmann probability function. In this case, the particles are considered to be distinguishable by being numbered, for example, from 1 to N, with no limit to the number of particles allowed in each energy state. The behavior of gas molecules in a container at fairly low pressure is an example of this distribution.

A second distribution law is the Bose–Einstein function. The particles in this case are indistinguishable and, again, there is no limit to the number of particles permitted in each quantum state. The behavior of photons, or black body radiation, is an example of this law.

The third distribution law is the Fermi–Dirac probability function. In this case, the particles are again indistinguishable, but now only one particle is permitted in each quantum state. Electrons in a crystal obey this law. In each case, the particles are assumed to be noninteracting.

3.5.2 The Fermi–Dirac Probability Function

Figure 3.28 shows the ith energy level with g_i quantum states. A maximum of one particle is allowed in each quantum state by the Pauli exclusion principle. There are g_i ways of choosing where to place the first particle, $(g_i - 1)$ ways of choosing where to place the second particle, $(g_i - 2)$ ways of choosing where to place the third particle, and so on. Then the total number of ways of arranging N_i particles in the ith energy level (where $N_i \leq g_i$) is

$$(g_i)(g_i - 1) \cdots (g_i - (N_i - 1)) = \frac{g_i!}{(g_i - N_i)!} \tag{3.76}$$

This expression includes all permutations of the N_i particles among themselves.

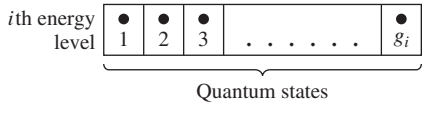


Figure 3.28 | The ith energy level with g_i quantum states.

However, since the particles are indistinguishable, the $N_i!$ number of permutations that the particles have among themselves in any given arrangement do not count as separate arrangements. The interchange of any two electrons, for example, does not produce a new arrangement. Therefore, the actual number of independent ways of realizing a distribution of N_i particles in the ith level is

$$W_i = \frac{g_i!}{N_i!(g_i - N_i)!} \tag{3.77}$$

EXAMPLE 3.5

Objective: Determine the possible number of ways of realizing a particular distribution for (a) $g_i = N_i = 10$ and (b) $g_i = 10$, $N_i = 9$.

■ Solution

(a) $g_i = N_i = 10$: We may note that $(g_i - N_i)! = 0! = 1$. Then, from Equation (3.77), we find

$$\frac{g_i!}{N_i!(g_i - N_i)!} = \frac{10!}{10!} = 1$$

(b) $g_i = 10$, $N_i = 9$: We may note that $(g_i - N_i)! = 1! = 1$. Then, we find

$$\frac{g_i!}{N_i!(g_i - N_i)!} = \frac{10!}{(9!)(1)} = \frac{(10)(9!)}{(9!)} = 10$$

■ Comment

In part (a), we have 10 particles to be arranged in 10 quantum states. There is only one possible arrangement. Each quantum state contains one particle. In part (b), we have 9 particles to be arranged in 10 quantum states. There is one empty quantum state, and there are 10 possible positions in which that empty state may occur. Thus, there are 10 possible arrangements for this case.

■ EXERCISE PROBLEM

Ex3.5 Determine the possible number of ways of realizing a particular distribution if $g_i = 10$ and $N_i = 8$. (45 ·snA)

Equation (3.77) gives the number of independent ways of realizing a distribution of N_i particles in the ith level. The total number of ways of arranging $(N_1, N_2, N_3, \ldots, N_n)$ indistinguishable particles among n energy levels is the product of all distributions, or

$$W = \prod_{i=1}^{n} \frac{g_i!}{N_i!(g_i - N_i)!} \tag{3.78}$$

The parameter W is the total number of ways in which N electrons can be arranged in this system, where $N = \sum_{i=1}^{n} N_i$ is the total number of electrons in the system. We want to find the most probable distribution, which means that we want to find the maximum W. The maximum W is found by varying N_i among the E_i levels, which varies the distribution, but at the same time, we will keep the total number of particles and total energy constant.

We may write the most probable distribution function as

$$\frac{N(E)}{g(E)} = f_F(E) = \frac{1}{1 + \exp\left(\dfrac{E - E_F}{kT}\right)} \tag{3.79}$$

The number density $N(E)$ is the number of particles per unit volume per unit energy and the function $g(E)$ is the number of quantum states per unit volume per unit energy. The function $f_F(E)$ is called the *Fermi–Dirac distribution* or probability function and gives the probability that a quantum state at the energy E will be occupied by an electron. The energy E_F is called the *Fermi energy*. Another interpretation of the distribution function is that $f_F(E)$ is the ratio of filled to total quantum states at any energy E.

3.5.3 The Distribution Function and the Fermi Energy

To begin to understand the meaning of the distribution function and the Fermi energy, we can plot the distribution function versus energy. Initially, let $T = 0$ K and consider the case when $E < E_F$. The exponential term in Equation (3.79) becomes $\exp[(E - E_F)/kT] \rightarrow \exp(-\infty) = 0$. The resulting distribution function is $f_F(E < E_F) = 1$. Again let $T = 0$ K and consider the case when $E > E_F$. The exponential term in the distribution function becomes $\exp[(E - E_F)/kT] \rightarrow \exp(+\infty) \rightarrow +\infty$. The resulting Fermi–Dirac distribution function now becomes $f_F(E > E_F) = 0$.

The Fermi–Dirac distribution function for $T = 0$ K is plotted in Figure 3.29. This result shows that, for $T = 0$ K, the electrons are in their lowest possible energy states. The probability of a quantum state being occupied is unity for $E < E_F$ and the probability of a state being occupied is zero for $E > E_F$. All electrons have energies below the Fermi energy at $T = 0$ K.

Figure 3.30 shows discrete energy levels of a particular system as well as the number of available quantum states at each energy. If we assume, for this case, that the system contains 13 electrons, then Figure 3.30 shows how these electrons are distributed among the various quantum states at $T = 0$ K. The electrons will be in the lowest possible energy state, so the probability of a quantum state being occupied in energy levels E_1 through E_4 is unity, and the probability of a quantum state being occupied in energy level E_5 is zero. The Fermi energy, for this case, must be above E_4 but less than E_5. The Fermi energy determines the statistical distribution of electrons and does not have to correspond to an allowed energy level.

Now consider a case in which the density of quantum states $g(E)$ is a continuous function of energy, as shown in Figure 3.31. If we have N_0 electrons in this system,

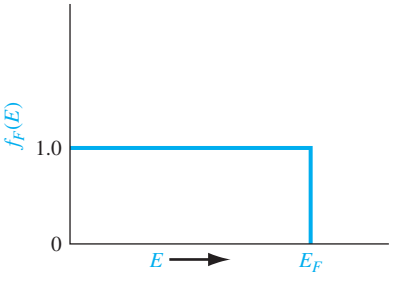


Figure 3.29 | The Fermi probability function versus energy for $T = 0$ K.

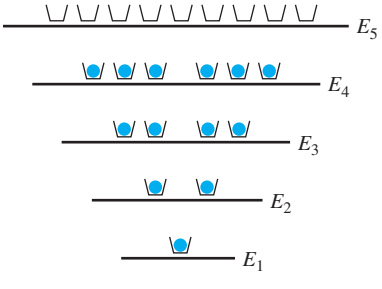


Figure 3.30 | Discrete energy states and quantum states for a particular system at $T = 0$ K.

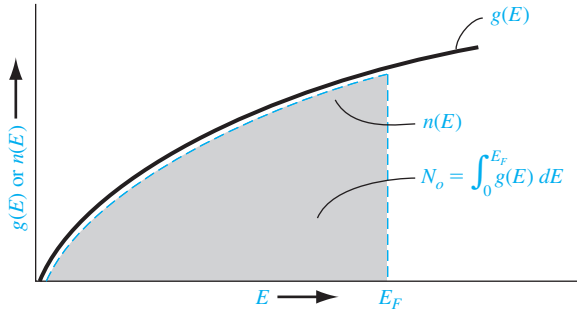


Figure 3.31 | Density of quantum states and electrons in a continuous energy system at $T = 0$ K.

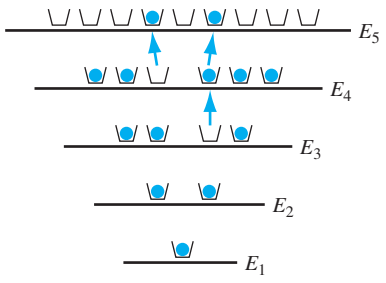


Figure 3.32 | Discrete energy states and quantum states for the same system shown in Figure 3.30 for $T > 0$ K.

then the distribution of these electrons among the quantum states at $T = 0$ K is shown by the dashed line. The electrons are in the lowest possible energy state so that all states below E_F are filled and all states above E_F are empty. If $g(E)$ and N_0 are known for this particular system, then the Fermi energy E_F can be determined.

Consider the situation when the temperature increases above $T = 0$ K. Electrons gain a certain amount of thermal energy so that some electrons can jump to higher energy levels, which means that the distribution of electrons among the available energy states will change. Figure 3.32 shows the same discrete energy levels and quantum states as in Figure 3.30. The distribution of electrons among the quantum states has changed from the $T = 0$ K case. Two electrons from the E_4 level have gained enough energy to jump to E_5, and one electron from E_3 has jumped to E_4. As the temperature changes, the distribution of electrons versus energy changes.

The change in the electron distribution among energy levels for $T > 0$ K can be seen by plotting the Fermi–Dirac distribution function. If we let $E = E_F$ and $T > 0$ K, then Equation (3.79) becomes

$$f_F(E = E_F) = \frac{1}{1 + \exp(0)} = \frac{1}{1 + 1} = \frac{1}{2}$$

The probability of a state being occupied at $E = E_F$ is $\frac{1}{2}$. Figure 3.33 shows the Fermi–Dirac distribution function plotted for several temperatures, assuming that the Fermi energy is independent of temperature.

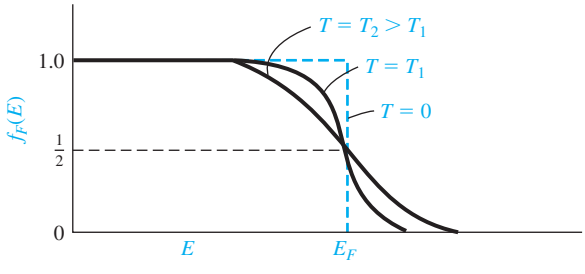


Figure 3.33 | The Fermi probability function versus energy for different temperatures.

We can see that for temperatures above absolute zero, there is a nonzero probability that some energy states above E_F will be occupied by electrons and some energy states below E_F will be empty. This result again means that some electrons have jumped to higher energy levels with increasing thermal energy.

EXAMPLE 3.6

Objective: Calculate the probability that an energy state above E_F is occupied by an electron.

Let $T = 300$ K. Determine the probability that an energy level $3kT$ above the Fermi energy is occupied by an electron.

■ Solution

From Equation (3.79), we can write

$$f_F(E) = \frac{1}{1 + \exp\left(\dfrac{E - E_F}{kT}\right)} = \frac{1}{1 + \exp\left(\dfrac{3kT}{kT}\right)}$$

which becomes

$$f_F(E) = \frac{1}{1 + 20.09} = 0.0474 = 4.74\%$$

■ Comment

At energies above E_F, the probability of a state being occupied by an electron can become significantly less than unity, or the ratio of electrons to available quantum states can be quite small.

■ EXERCISE PROBLEM

Ex3.6 Assume the Fermi energy level is 0.30 eV below the conduction band energy E_c. Assume $T = 300$ K. (a) Determine the probability of a state being occupied by an electron at $E = E_c + kT/4$. (b) Repeat part (a) for an energy state at $E = E_c + kT$.

[Ans. (a) 7.26×10^{-6}; (b) 3.43×10^{-6}]

We can see from Figure 3.33 that the probability of an energy above E_F being occupied increases as the temperature increases and the probability of a state below E_F being empty increases as the temperature increases.

EXAMPLE 3.7

Objective: Determine the temperature at which there is 1 percent probability that an energy state is empty.

Assume that the Fermi energy level for a particular material is 6.25 eV and that the electrons in this material follow the Fermi–Dirac distribution function. Calculate the temperature at which there is a 1 percent probability that a state 0.30 eV below the Fermi energy level will not contain an electron.

■ Solution

The probability that a state is empty is

$$1 - f_F(E) = 1 - \frac{1}{1 + \exp\left(\dfrac{E - E_F}{kT}\right)}$$

Then

$$0.01 = 1 - \frac{1}{1 + \exp\left(\dfrac{5.95 - 6.25}{kT}\right)}$$

Solving for kT, we find $kT = 0.06529$ eV, so that the temperature is $T = 756$ K.

Comment

The Fermi probability function is a strong function of temperature.

EXERCISE PROBLEM

Ex 3.7 Assume that E_F is 0.3 eV below E_c. Determine the temperature at which the probability of an electron occupying an energy state at $E = (E_c + 0.025)$ eV is 8×10^{-6}.

(Ans. $T = 321$ K)

We may note that the probability of a state a distance dE above E_F being occupied is the same as the probability of a state a distance dE below E_F being empty. The function $f_F(E)$ is symmetrical with the function $1 - f_F(E)$ about the Fermi energy, E_F. This symmetry effect is shown in Figure 3.34 and will be used in the next chapter.

Consider the case when $E - E_F \gg kT$, where the exponential term in the denominator of Equation (3.79) is much greater than unity. We may neglect the 1 in the denominator, so the Fermi–Dirac distribution function becomes

$$f_F(E) \approx \exp\left[\frac{-(E - E_F)}{kT}\right] \tag{3.80}$$

Equation (3.80) is known as the Maxwell–Boltzmann approximation, or simply the Boltzmann approximation, to the Fermi–Dirac distribution function. Figure 3.35 shows the Fermi–Dirac probability function and the Boltzmann approximation. This figure gives an indication of the range of energies over which the approximation is valid.

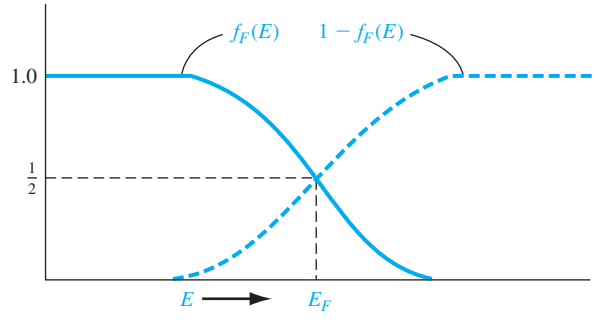


Figure 3.34 | The probability of a state being occupied, $f_F(E)$, and the probability of a state being empty, $1 - f_F(E)$.

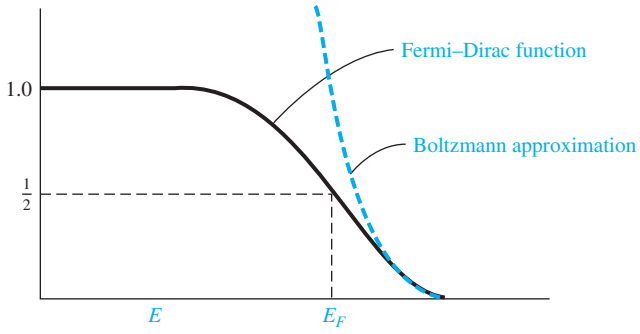


Figure 3.35 | The Fermi–Dirac probability function and the Maxwell–Boltzmann approximation.

EXAMPLE 3.8

Objective: Determine the energy at which the Boltzmann approximation may be considered valid.

Calculate the energy, in terms of kT and E_F, at which the difference between the Boltzmann approximation and the Fermi–Dirac function is 5 percent of the Fermi function.

■ Solution

We can write

$$\frac{\exp\left[\dfrac{-(E-E_F)}{kT}\right] - \dfrac{1}{1+\exp\left(\dfrac{E-E_F}{kT}\right)}}{\dfrac{1}{1+\exp\left(\dfrac{E-E_F}{kT}\right)}} = 0.05$$

If we multiply both numerator and denominator by the $1 + \exp(\ )$ function, we have

$$\exp\left[\frac{-(E-E_F)}{kT}\right] \cdot \left\{1+\exp\left[\frac{E-E_F}{kT}\right]\right\} - 1 = 0.05$$

which becomes

$$\exp\left[\frac{-(E-E_F)}{kT}\right] = 0.05$$

or

$$(E-E_F) = kT\,ln\left(\frac{1}{0.05}\right) \approx 3kT$$

■ Comment

As seen in this example and in Figure 3.35, the $E - E_F \gg kT$ notation is somewhat misleading. The Maxwell–Boltzmann and Fermi–Dirac functions are within 5 percent of each other when $E - E_F \approx 3kT$.

■ EXERCISE PROBLEM

Ex 3.8 Repeat Example 3.8 for the case when the difference between the Boltzmann approximation and the Fermi–Dirac function is 2 percent of the Fermi function.

(Ans. $E - E_F = 3.9kT$)

The actual Boltzmann approximation is valid when $\exp[(E - E_,)/kT] \gg 1$. However, it is still common practice to use the $E - E_F \gg kT$ notation when applying the Boltzmann approximation. We will use this Boltzmann approximation in our discussion of semiconductors in the next chapter.

TEST YOUR UNDERSTANDING

TYU 3.5 Assume that the Fermi energy level is 0.35 eV above the valence band energy. Let $T = 300$ K. (a) Determine the probability of a state being empty of an electron at $E = E_v - kT/2$. (b) Repeat part (a) for an energy state at $E = E_v - 3kT/2$.
[Ans. (a) 8.20×10^{-7}; (b) 3.02×10^{-7}]

TYU 3.6 Repeat Exercise Problem Ex 3.6 for $T = 400$ K.
[Ans. (a) 1.31×10^{-4}; (b) 6.21×10^{-5}]

TYU 3.7 Repeat Exercise Problem TYU 3.5 for $T = 400$ K.
[Ans. (a) 2.41×10^{-5}; (b) 8.85×10^{-6}]

3.6 | SUMMARY

- Discrete allowed electron energies split into a band of allowed energies as atoms are brought together to form a crystal.
- The concept of allowed and forbidden energy bands was developed more rigorously by considering quantum mechanics and Schrodinger's wave equation using the Kronig–Penney model representing the potential function of a single-crystal material. This result forms the basis of the energy-band theory of semiconductors.
- The concept of effective mass was developed. Effective mass relates the motion of a particle in a crystal to an externally applied force and takes into account the effect of the crystal lattice on the motion of the particle.
- Two charged particles exist in a semiconductor. An electron is a negatively charged particle with a positive effective mass existing at the bottom of an allowed energy band. A hole is a positively charged particle with a positive effective mass existing at the top of an allowed energy band.
- The E versus k diagrams of silicon and gallium arsenide were given and the concept of direct and indirect bandgap semiconductors was discussed.
- Energies within an allowed energy band are actually at discrete levels and each contains a finite number of quantum states. The density per unit energy of quantum states was determined by using the three-dimensional infinite potential well as a model.
- In dealing with large numbers of electrons and holes, we must consider the statistical behavior of these particles. The Fermi–Dirac probability function was developed, which gives the probability of a quantum state at an energy E of being occupied by an electron. The Fermi energy was defined.

GLOSSARY OF IMPORTANT TERMS

allowed energy band A band or range of energy levels that an electron in a crystal is allowed to occupy based on quantum mechanics.

density of states function The density of available quantum states as a function of energy, given in units of number per unit energy per unit volume.

electron effective mass The parameter that relates the acceleration of an electron in the conduction band of a crystal to an external force; a parameter that takes into account the effect of internal forces in the crystal.

Fermi–Dirac probability function The function describing the statistical distribution of electrons among available energy states and the probability that an allowed energy state is occupied by an electron.

fermi energy In the simplest definition, the energy below which all states are filled with electrons and above which all states are empty at $T = 0$ K.

forbidden energy band A band or range of energy levels that an electron in a crystal is not allowed to occupy based on quantum mechanics.

hole The positively charged "particle" associated with an empty state in the top of the valence band.

hole effective mass The parameter that relates the acceleration of a hole in the valence band of a crystal to an applied external force (a positive quantity); a parameter that takes into account the effect of internal forces in a crystal.

***k*-space diagram** The plot of electron energy in a crystal versus k, where k is the momentum-related constant of the motion that incorporates the crystal interaction.

Kronig–Penney model The mathematical model of a periodic potential function representing a one-dimensional single-crystal lattice by a series of periodic step functions.

Maxwell–Boltzmann approximation The condition in which the energy is several kT above the Fermi energy or several kT below the Fermi energy so that the Fermi–Dirac probability function can be approximated by a simple exponential function.

CHECKPOINT

After studying this chapter, the reader should have the ability to

- Discuss the concept of allowed and forbidden energy bands in a single crystal both qualitatively and more rigorously from the results of using the Kronig–Penney model.
- Discuss the splitting of energy bands in silicon.
- State the definition of effective mass from the E versus k diagram and discuss its meaning in terms of the movement of a particle in a crystal.
- Discuss the concept of a hole.
- Discuss the characteristics of a direct and an indirect bandgap semiconductor.
- Qualitatively, in terms of energy bands, discuss the difference between a metal, an insulator, and semiconductor.
- What is meant by the density of states function?
- Understand the meaning of the Fermi–Dirac distribution function and the Fermi energy.

REVIEW QUESTIONS

1. What is the Kronig–Penney model? What does it represent?

2. State two results of using the Kronig–Penney model with Schrodinger's wave equation.

3. What is effective mass? How is effective mass defined in terms of the E versus k diagram?

4. What is a direct bandgap semiconductor? What is an indirect bandgap semiconductor?

5. What is the meaning of the density of states function?

6. What was the mathematical model used in deriving the density of states function?
7. In general, what is the relation between density of states and energy?
8. What is the meaning of the Fermi–Dirac probability function?
9. What is the Fermi energy?

PROBLEMS

Section 3.1 Allowed and Forbidden Energy Bands

3.1 Consider Figure 3.4b, which shows the energy-band splitting of silicon. If the equilibrium lattice spacing were to change by a small amount, discuss how you would expect the electrical properties of silicon to change. Determine at what point the material would behave like an insulator or like a metal.

3.2 Show that Equations (3.4) and (3.6) are derived from Schrodinger's wave equation, using the form of solution given by Equation (3.3).

3.3 Show that Equations (3.9) and (3.10) are solutions of the differential equations given by Equations (3.4) and (3.8), respectively.

3.4 Show that Equations (3.12), (3.14), (3.16), and (3.18) result from the boundary conditions in the Kronig–Penney model.

3.5 (*a*) Plot the function $f(\alpha a) = 12(\sin \alpha a)/\alpha a + \cos \alpha a$ for $0 \leq \alpha a \leq 4\pi$. Also, given the function $f(\alpha a) = \cos ka$, indicate the allowed values of αa that will satisfy this equation. (*b*) Determine the values of αa at (*i*) $ka = \pi$ and (*ii*) $ka = 2\pi$.

3.6 Repeat Problem 3.5 for the function $f(\alpha a) = 5(\sin \alpha a)/\alpha a + \cos \alpha a = \cos ka$.

3.7 Using Equation (3.24), show that $dE/dk = 0$ at $k = n\pi/a$, where $n = 0, 1, 2, \ldots$.

3.8 Using the parameters of Problem 3.5 for a free electron and letting $a = 4.2$ Å, determine the width (in eV) of the forbidden energy bands that exist at (*a*) $ka = \pi$ and (*b*) $ka = 2\pi$. (Refer to Figure 3.8c).

3.9 Using the parameters in Problem 3.5 for a free electron and letting $a = 4.2$ Å, determine the width (in eV) of the allowed energy bands that exist for (*a*) $0 < ka < \pi$ and (*b*) $\pi < ka < 2\pi$.

3.10 Repeat Problem 3.8 using the parameters in Problem 3.6.

3.11 Repeat Problem 3.9 using the parameters in Problem 3.6.

3.12 The bandgap energy in a semiconductor is usually a slight function of temperature. In some cases, the bandgap energy versus temperature can be modeled by

$$E_g = E_g(0) - \frac{\alpha T^2}{(\beta + T)}$$

where $E_g(0)$ is the value of the bandgap energy at $T = 0$ K. For silicon, the parameter values are $E_g(0) = 1.170$ eV, $\alpha = 4.73 \times 10^{-4}$ eV/K, and $\beta = 636$ K. Plot E_g versus T over the range $0 \leq T \leq 600$ K. In particular, note the value at $T = 300$ K.

Section 3.2 Electrical Conduction in Solids

3.13 Two possible conduction bands are shown in the E versus k diagram given in Figure P3.13. State which band will result in the heavier electron effective mass; state why.

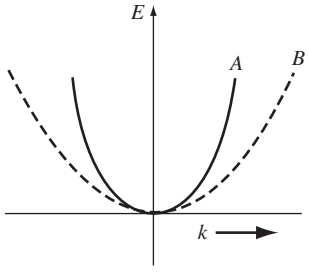


Figure P3.13 | Conduction bands for Problem 3.13.

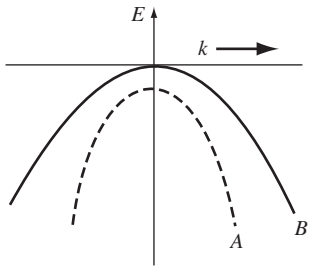


Figure P3.14 | Valence bands for Problem 3.14.

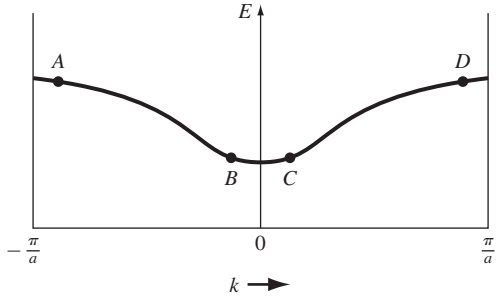


Figure P3.15 | Figure for Problem 3.15.

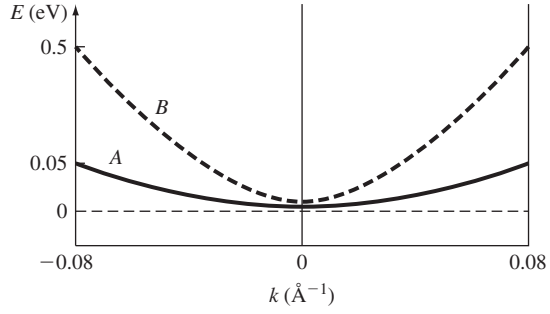


Figure P3.16 | Figure for Problem 3.16.

3.14 Two possible valence bands are shown in the E versus k diagram given in Figure P3.14. State which band will result in the heavier hole effective mass; state why.

3.15 The E versus k diagram for a particular allowed energy band is shown in Figure P3.15. Determine (*a*) the sign of the effective mass and (*b*) the direction of velocity for a particle at each of the four positions shown.

3.16 Figure P3.16 shows the parabolic E versus k relationship in the conduction band for an electron in two particular semiconductor materials. Determine the effective mass (in units of the free electron mass) of the two electrons.

3.17 Figure P3.17 shows the parabolic E versus k relationship in the valence band for a hole in two particular semiconductor materials. Determine the effective mass (in units of the free electron mass) of the two holes.

3.18 (*a*) The forbidden bandgap energy in GaAs is 1.42 eV. (*i*) Determine the minimum frequency of an incident photon that can interact with a valence electron and elevate the electron to the conduction band. (*ii*) What is the corresponding wavelength? (*b*) Repeat part (*a*) for silicon with a bandgap energy of 1.12 eV.

3.19 The E versus k diagrams for a free electron (curve A) and for an electron in a semiconductor (curve B) are shown in Figure P3.19. Sketch (*a*) dE/dk versus k and (*b*) d^2E/dk^2 versus k for each curve. (*c*) What conclusion can you make concerning a comparison in effective masses for the two cases?

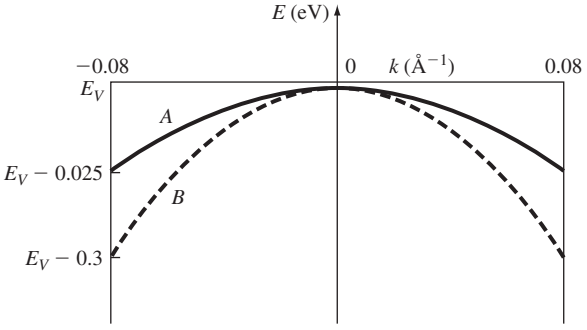


Figure P3.17 | Figure for Problem 3.17.

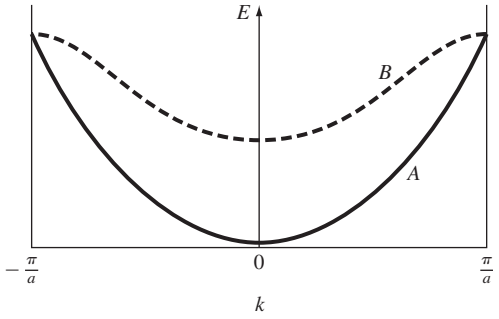


Figure P3.19 | Figure for Problem 3.19.

Section 3.3 Extension to Three Dimensions

3.20 The energy-band diagram for silicon is shown in Figure 3.25b. The minimum energy in the conduction band is in the [100] direction. The energy in this one-dimensional direction near the minimum value can be approximated by

$$E = E_0 - E_1 \cos \alpha(k - k_0)$$

where k_0 is the value of k at the minimum energy. Determine the effective mass of the particle at $k = k_0$ in terms of the equation parameters.

3.21 The constant kinetic energy curves for electrons in the conduction band of germanium consists of four ellipsoids similar to those in silicon (see Appendix F). The longitudinal and transverse effective masses are $m_l = 1.64m_0$ and $m_t = 0.082m_0$, respectively. Determine the (*a*) density of states effective mass and (*b*) conductivity effective mass.

3.22 Heavy and light holes exist in GaAs with effective masses $m_{hh} = 0.45m_0$ and $m_{lh} = 0.082m_0$, respectively. Determine the (*a*) density of states effective mass and (*b*) conductivity effective mass.

Section 3.4 Density of States Function

3.23 Starting with the three-dimensional infinite potential well function given by Equation (3.59) and using the separation of variables technique, derive Equation (3.60).

3.24 Show that Equation (3.69) can be derived from Equation (3.64).

3.25 Derive the density of states function for a one-dimensional electron gas in GaAs ($m_n^* = 0.067m_0$). Note that the kinetic energy may be written as $E = (\pm p)^2/2m_n^*$, which means that there are two momentum states for each energy level.

3.26 (*a*) Determine the total number (#/cm^3) of energy states in silicon between E_c and $E_c + 2kT$ at (*i*) $T = 300$ K and (*ii*) $T = 400$ K. (*b*) Repeat part (*a*) for GaAs.

3.27 (*a*) Determine the total number (#/cm^3) of energy states in silicon between E_v and $E_v - 3\,kT$ at (*i*) $T = 300$ K and (*ii*) $T = 400$ K. (*b*) Repeat part (*a*) for GaAs.

3.28 (*a*) Plot the density of states in the conduction band of silicon over the range $E_c < E < E_c + 0.4$ eV. (*b*) Repeat part (*a*) for the density of states in the valence band over the range $E_v - 0.4$ eV $< E < E_v$.

3.29 (*a*) For silicon, find the ratio of the density of states in the conduction band at $E = E_c + kT$ to the density of states in the valence band at $E = E_v - kT$. (*b*) Repeat part (*a*) for GaAs.

Section 3.5 Statistical Mechanics

3.30 Plot the Fermi–Dirac probability function, given by Equation (3.79), over the range $-0.2 \le (E - E_F) \le 0.2$ eV for (*a*) $T = 200$ K, (*b*) $T = 300$ K, and (*c*) $T = 400$ K.

3.31 (*a*) Repeat Example 3.5 for the case when $g_i = 10$ and $N_i = 7$. (*b*) Repeat part (*a*) for (*i*) $g_i = 12$, $N_i = 10$ and (*ii*) $g_i = 12$, $N_i = 8$.

3.32 Determine the probability that an energy level is occupied by an electron if the state is above the Fermi level by (*a*) kT, (*b*) $5kT$, and (*c*) $10kT$.

3.33 Determine the probability that an energy level is empty of an electron if the state is below the Fermi level by (*a*) kT, (*b*) $5kT$, and (*c*) $10kT$.

3.34 (*a*) The Fermi energy in silicon is 0.30 eV below the conduction band energy E_c at $T = 300$ K. Plot the probability of a state being occupied by an electron in the conduction band over the range $E_c \le E \le E_c + 2kT$. (*b*) The Fermi energy in silicon is 0.25 eV above the valence band energy E_v. Plot the probability of a state being empty by an electron in the valence band over the range $E_v - 2kT$ eV $\le E \le E_v$.

3.35 The probability that a state at $E_c + kT$ is occupied by an electron is equal to the probability that a state at $E_v - kT$ is empty. Determine the position of the Fermi energy level as a function of E_c and E_v.

3.36 Six free electrons exist in a one-dimensional infinite potential well of width $a = 12$ Å. Determine the Fermi energy level at $T = 0$ K.

3.37 (*a*) Five free electrons exist in a three-dimensional infinite potential well with all three widths equal to $a = 12$ Å. Determine the Fermi energy level at $T = 0$ K. (*b*) Repeat part (*a*) for 13 electrons.

3.38 Show that the probability of an energy state being occupied ΔE above the Fermi energy is the same as the probability of a state being empty ΔE below the Fermi level.

3.39 (*a*) Determine for what energy above E_F (in terms of kT) the Fermi–Dirac probability function is within 1 percent of the Boltzmann approximation. (*b*) Give the value of the probability function at this energy.

3.40 The Fermi energy level for a particular material at $T = 300$ K is 5.50 eV. The electrons in this material follow the Fermi–Dirac distribution function. (*a*) Find the probability of an electron occupying an energy at 5.80 eV. (*b*) Repeat part (*a*) if the temperature is increased to $T = 700$ K. (Assume that E_F is a constant.) (*c*) Determine the temperature at which there is a 2 percent probability that a state 0.25 eV below the Fermi level will be empty of an electron.

3.41 The Fermi energy for copper at $T = 300$ K is 7.0 eV. The electrons in copper follow the Fermi–Dirac distribution function. (*a*) Find the probability of an energy level at 7.15 eV being occupied by an electron. (*b*) Repeat part (*a*) for $T = 1000$ K. (Assume that E_F is a constant.) (*c*) Repeat part (*a*) for $E = 6.85$ eV and $T = 300$ K. (*d*) Determine the probability of the energy state at $E = E_F$ being occupied at $T = 300$ K and at $T = 1000$ K.

3.42 Consider the energy levels shown in Figure P3.42. Let $T = 300$ K. (*a*) If $E_1 - E_F = 0.30$ eV, determine the probability that an energy state at $E = E_1$ is occupied by

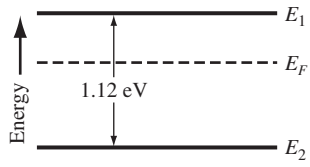


Figure P3.42 | Energy levels for Problem 3.42.

an electron and the probability that an energy state at $E = E_2$ is empty. (*b*) Repeat part (*a*) if $E_F - E_2 = 0.40$ eV.

3.43 Repeat problem 3.42 for the case when $E_1 - E_2 = 1.42$ eV.

3.44 Determine the derivative with respect to energy of the Fermi–Dirac distribution function. Plot the derivative with respect to energy for (*a*) $T = 0$ K, (*b*) $T = 300$ K, and (*c*) $T = 500$ K.

3.45 Assume that the Fermi energy level is exactly in the center of the bandgap energy of a semiconductor at $T = 300$ K. (*a*) Calculate the probability that an energy state in the bottom of the conduction band is occupied by an electron for Si, Ge, and GaAs. (*b*) Calculate the probability that an energy state in the top of the valence band is empty for Si, Ge, and GaAs.

3.46 (*a*) Calculate the temperature at which there is a 10^{-8} probability that an energy state 0.60 eV above the Fermi energy level is occupied by an electron. (*b*) Repeat part (*a*) for a probability of 10^{-6}.

3.47 Calculate the energy range (in eV) between $f_F = 0.95$ and $f_F = 0.05$ for $E_F = 5.0$ eV at (*a*) $T = 200$ K and (*b*) $T = 400$ K.

READING LIST

1. Dimitrijev, S. *Principles of Semiconductor Devices*. New York: Oxford University, 2006.
2. Kano, K. *Semiconductor Devices.* Upper Saddle River, NJ: Prentice Hall, 1998.
3. Kittel, C. *Introduction to Solid State Physics,* 7th ed. Berlin: Springer-Verlag, 1993.
4. McKelvey, J. P. *Solid State Physics for Engineering and Materials Science.* Malabar, FL: Krieger, 1993.
5. Pierret, R. F. *Semiconductor Device Fundamentals.* Reading, MA: Addison-Wesley, 1996.

*6. Shockley, W. *Electrons and Holes in Semiconductors.* New York: D. Van Nostrand, 1950.

7. Shur, M. *Introduction to Electronic Devices.* New York: John Wiley and Sons, 1996.

*8. Shur, M. *Physics of Semiconductor Devices.* Englewood Cliffs, NJ: Prentice Hall, 1990.

9. Singh, J. *Semiconductor Devices: An Introduction.* New York: McGraw-Hill, 1994.
10. Singh, J. *Semiconductor Devices: Basic Principles.* New York: John Wiley and Sons, 2001.

11. Streetman, B. G., and S. K. Banerjee. *Solid State Electronic Devices,* 6th ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2006.

12. Sze, S. M. *Semiconductor Devices: Physics and Technology,* 2nd ed. New York: John Wiley and Sons, 2001.

*__13.__ Wang, S. *Fundamentals of Semiconductor Theory and Device Physics.* Englewood Cliffs, NJ: Prentice Hall, 1988.

14. Wolfe, C. M., N. Holonyak, Jr., and G. E. Stillman. *Physical Properties of Semiconductors.* Englewood Cliffs, NJ: Prentice Hall, 1989.

*Indicates references that are at an advanced level compared to this text.

CHAPTER 4

The Semiconductor in Equilibrium

So far, we have been considering a general crystal and applying to it the concepts of quantum mechanics in order to determine a few of the characteristics of electrons in a single-crystal lattice. In this chapter, we apply these concepts specifically to a semiconductor material. In particular, we use the density of quantum states in the conduction band and the density of quantum states in the valence band along with the Fermi–Dirac probability function to determine the concentration of electrons and holes in the conduction and valence bands, respectively. We also apply the concept of the Fermi energy to the semiconductor material.

This chapter deals with the semiconductor in equilibrium. Equilibrium, or thermal equilibrium, implies that no external forces such as voltages, electric fields, magnetic fields, or temperature gradients are acting on the semiconductor. All properties of the semiconductor will be independent of time in this case. ■

4.0 | PREVIEW

In this chapter, we will:

- Derive the thermal-equilibrium concentrations of electrons and holes in a semiconductor as a function of the Fermi energy level.
- Discuss the process by which the properties of a semiconductor material can be favorably altered by adding specific impurity atoms to the semiconductor.
- Determine the thermal-equilibrium concentrations of electrons and holes in a semiconductor as a function of the concentration of dopant atoms added to the semiconductor.
- Determine the position of the Fermi energy level as a function of the concentrations of dopant atoms added to the semiconductor.

4.1 | CHARGE CARRIERS IN SEMICONDUCTORS

Current is the rate at which charge flows. In a semiconductor, two types of charge carrier, the electron and the hole, can contribute to a current. Since the current in a semiconductor is determined largely by the number of electrons in the conduction band and the number of holes in the valence band, an important characteristic of the semiconductor is the density of these charge carriers. The density of electrons and holes is related to the density of states function and the Fermi distribution function, both of which we have considered. A qualitative discussion of these relationships will be followed by a more rigorous mathematical derivation of the thermal-equilibrium concentration of electrons and holes.

4.1.1 Equilibrium Distribution of Electrons and Holes

The distribution (with respect to energy) of electrons in the conduction band is given by the density of allowed quantum states times the probability that a state is occupied by an electron. This statement is written in equation form as

$$n(E) = g_c(E)f_F(E) \tag{4.1}$$

where $f_F(E)$ is the Fermi–Dirac probability function and $g_c(E)$ is the density of quantum states in the conduction band. The total electron concentration per unit volume in the conduction band is then found by integrating Equation (4.1) over the entire conduction-band energy.

Similarly, the distribution (with respect to energy) of holes in the valence band is the density of allowed quantum states in the valence band multiplied by the probability that a state is *not* occupied by an electron. We may express this as

$$p(E) = g_v(E)[1 - f_F(E)] \tag{4.2}$$

The total hole concentration per unit volume is found by integrating this function over the entire valence-band energy.

To find the thermal-equilibrium electron and hole concentrations, we need to determine the position of the Fermi energy E_F with respect to the bottom of the conduction-band energy E_c and the top of the valence-band energy E_v. To address this question, we will initially consider an intrinsic semiconductor. An ideal intrinsic semiconductor is a pure semiconductor with no impurity atoms and no lattice defects in the crystal (e.g., pure silicon). We have argued in the previous chapter that, for an intrinsic semiconductor at $T = 0$ K, all energy states in the valence band are filled with electrons and all energy states in the conduction band are empty of electrons. The Fermi energy must, therefore, be somewhere between E_c and E_v. (The Fermi energy does not need to correspond to an allowed energy.)

As the temperature begins to increase above 0 K, the valence electrons will gain thermal energy. A few electrons in the valence band may gain sufficient energy to jump to the conduction band. As an electron jumps from the valence band to the conduction band, an empty state, or hole, is created in the valence band. In an intrinsic

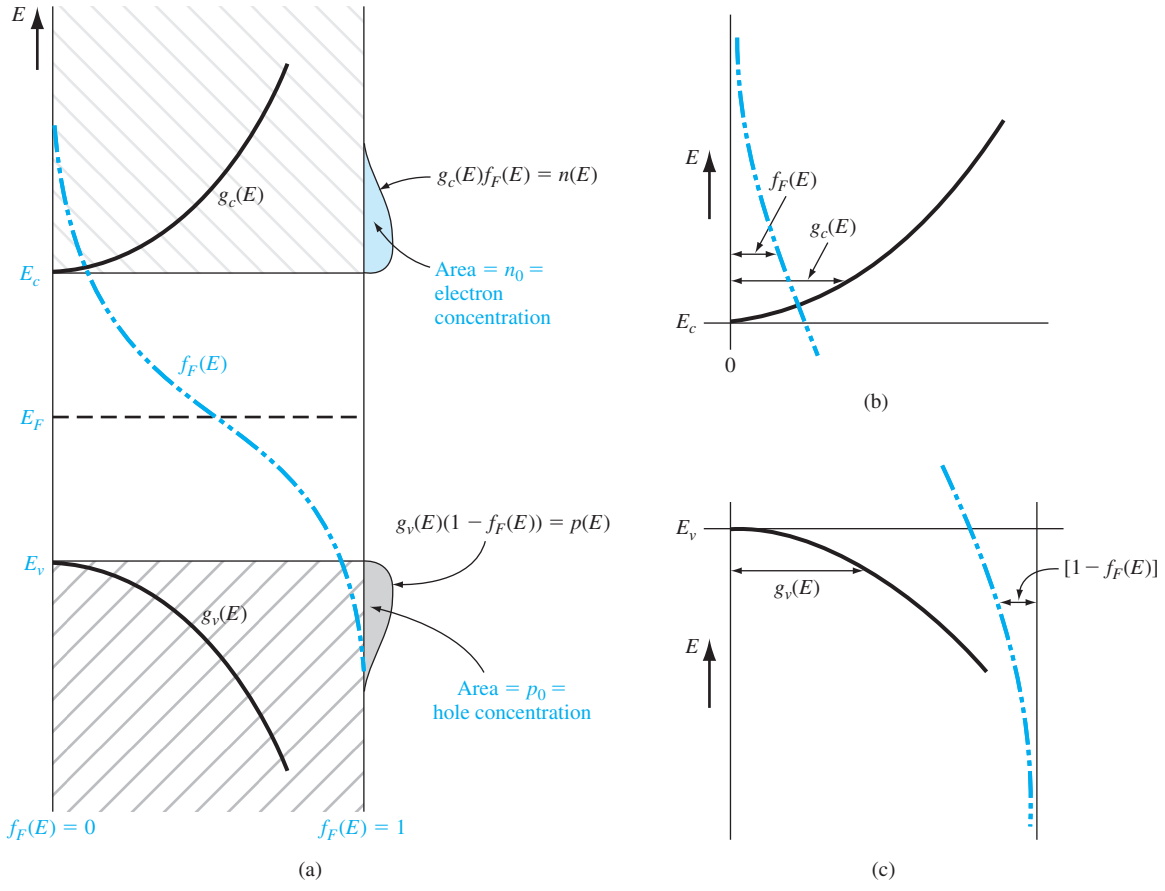


Figure 4.1 | (a) Density of states functions, Fermi–Dirac probability function, and areas representing electron and hole concentrations for the case when E_F is near the midgap energy; (b) expanded view near the conduction-band energy; and (c) expanded view near the valence-band energy.

semiconductor, then, electrons and holes are created in pairs by the thermal energy so that the number of electrons in the conduction band is equal to the number of holes in the valence band.

Figure 4.1a shows a plot of the density of states function in the conduction-band $g_c(E)$, the density of states function in the valence-band $g_v(E)$, and the Fermi–Dirac probability function for $T > 0$ K when E_F is approximately halfway between E_c and E_v. If we assume, for the moment, that the electron and hole effective masses are equal, then $g_c(E)$ and $g_v(E)$ are symmetrical functions about the midgap energy (the energy midway between E_c and E_v). We noted previously that the function $f_F(E)$ for $E > E_F$ is symmetrical to the function $1 - f_F(E)$ for $E < E_F$ about the energy $E = E_F$. This also means that the function $f_F(E)$ for $E = E_F + dE$ is equal to the function $1 - f_F(E)$ for $E = E_F - dE$.

Figure 4.1b is an expanded view of the plot in Figure 4.1a showing $f_F(E)$ and $g_c(E)$ above the conduction-band energy E_c. The product of $g_c(E)$ and $f_F(E)$ is the distribution of electrons $n(E)$ in the conduction band given by Equation (4.1). This product is plotted in Figure 4.1a. Figure 4.1c, which is an expanded view of the plot in Figure 4.1a shows $[1 - f_F(E)]$ and $g_v(E)$ below the valence-band energy E_v. The product of $g_v(E)$ and $[1 - f_F(E)]$ is the distribution of holes $p(E)$ in the valence band given by Equation (4.2). This product is also plotted in Figure 4.1a. The areas under these curves are then the total density of electrons in the conduction band and the total density of holes in the valence band. From this we see that if $g_c(E)$ and $g_v(E)$ are symmetrical, the Fermi energy must be at the midgap energy in order to obtain equal electron and hole concentrations. If the effective masses of the electron and hole are not exactly equal, then the effective density of states functions $g_c(E)$ and $g_v(E)$ will not be exactly symmetrical about the midgap energy. The Fermi level for the intrinsic semiconductor will then shift slightly from the midgap energy in order to obtain equal electron and hole concentrations.

4.1.2 The n_0 and p_0 Equations

We have argued that the Fermi energy for an intrinsic semiconductor is near midgap. In deriving the equations for the thermal-equilibrium concentration of electrons n_0 and the thermal-equilibrium concentration of holes p_0, we will not be quite so restrictive. We will see later that, in particular situations, the Fermi energy can deviate from this midgap energy. We will assume initially, however, that the Fermi level remains within the bandgap energy.

Thermal-Equilibrium Electron Concentration The equation for the thermal-equilibrium concentration of electrons may be found by integrating Equation (4.1) over the conduction band energy, or

$$n_0 = \int g_c(E) f_F(E)\, dE \tag{4.3}$$

The lower limit of integration is E_c and the upper limit of integration should be the top of the allowed conduction band energy. However, since the Fermi probability function rapidly approaches zero with increasing energy as indicated in Figure 4.1a, we can take the upper limit of integration to be infinity.

We are assuming that the Fermi energy is within the forbidden-energy bandgap. For electrons in the conduction band, we have $E > E_c$. If $(E_c - E_F) \gg kT$, then $(E - E_F) \gg kT$, so that the Fermi probability function reduces to the Boltzmann approximation,[1] which is

$$f_F(E) = \frac{1}{1 + \exp\dfrac{(E - E_F)}{kT}} \approx \exp\left[\frac{-(E - E_F)}{kT}\right] \tag{4.4}$$

[1]The Maxwell–Boltzmann and Fermi–Dirac distribution functions are within 5 percent of each other when $E - E_F \approx 3kT$ (see Figure 3.35). The $\gg$ notation is then somewhat misleading to indicate when the Boltzmann approximation is valid, although it is commonly used.

Applying the Boltzmann approximation to Equation (4.3), the thermal-equilibrium density of electrons in the conduction band is found from

$$n_0 = \int_{E_c}^{\infty} \frac{4\pi(2m_n^*)^{3/2}}{h^3}\sqrt{E - E_c}\,\exp\left[\frac{-(E - E_F)}{kT}\right] dE \tag{4.5}$$

The integral of Equation (4.5) may be solved more easily by making a change of variable. If we let

$$\eta = \frac{E - E_c}{kT} \tag{4.6}$$

then Equation (4.5) becomes

$$n_0 = \frac{4\pi(2m_n^* kT)^{3/2}}{h^3}\exp\left[\frac{-(E_c - E_F)}{kT}\right]\int_0^{\infty} \eta^{1/2}\exp(-\eta)\,d\eta \tag{4.7}$$

The integral is the gamma function, with a value of

$$\int_0^{\infty} \eta^{1/2}\exp(-\eta)\,d\eta = \frac{1}{2}\sqrt{\pi} \tag{4.8}$$

Then Equation (4.7) becomes

$$n_0 = 2\left(\frac{2\pi m_n^* kT}{h^2}\right)^{3/2}\exp\left[\frac{-(E_c - E_F)}{kT}\right] \tag{4.9}$$

We may define a parameter N_c as

$$N_c = 2\left(\frac{2\pi m_n^* kT}{h^2}\right)^{3/2} \tag{4.10}$$

The parameter m_n^* is the density of states effective mass of the electron. The thermal-equilibrium electron concentration in the conduction band can be written as

$$n_0 = N_c\exp\left[\frac{-(E_c - E_F)}{kT}\right] \tag{4.11}$$

The parameter N_c is called the *effective density of states function in the conduction band.* If we were to assume that $m_n^* = m_0$, then the value of the effective density of states function at $T = 300$ K is $N_c = 2.5 \times 10^{19}$ cm^{-3}, which is the order of magnitude of N_c for most semiconductors. If the effective mass of the electron is larger or smaller than m_0, then the value of the effective density of states function changes accordingly, but is still of the same order of magnitude.

EXAMPLE 4.1

Objective: Calculate the probability that a quantum state in the conduction band at $E = E_c + kT/2$ is occupied by an electron, and calculate the thermal-equilibrium electron concentration in silicon at $T = 300$ K.

Assume the Fermi energy is 0.25 eV below the conduction band. The value of N_c for silicon at $T = 300$ K is $N_c = 2.8 \times 10^{19}$ cm^{-3} (see Appendix B).

■ Solution

The probability that a quantum state at $E = E_c + kT/2$ is occupied by an electron is given by

$$f_F(E) = \frac{1}{1 + \exp\left(\dfrac{E - E_F}{kT}\right)} \cong \exp\left[\frac{-(E - E_F)}{kT}\right] = \exp\left[\frac{-(E_c + (kT/2) - E_F)}{kT}\right]$$

or

$$f_F(E) = \exp\left[\frac{-(0.25 + (0.0259/2))}{0.0259}\right] = 3.90 \times 10^{-5}$$

The electron concentration is given by

$$n_0 = N_c \exp\left[\frac{-(E_c - E_F)}{kT}\right] = (2.8 \times 10^{19}) \exp\left[\frac{-0.25}{0.0259}\right]$$

or

$$n_0 = 1.80 \times 10^{15}\ \text{cm}^{-3}$$

Comment

The probability of a state being occupied can be quite small, but the fact that there are a large number of states means that the electron concentration is a reasonable value.

EXERCISE PROBLEM

Ex 4.1 Determine the probability that a quantum state at energy $E = E_c + kT$ is occupied by an electron, and calculate the electron concentration in GaAs at $T = 300$ K if the Fermi energy level is 0.25 eV below E_c.

[Ans. $f_F(E) = 2.36 \times 10^{-5}$, $n_0 = 3.02 \times 10^{13}\ \text{cm}^{-3}$]

Thermal-Equilibrium Hole Concentration The thermal-equilibrium concentration of holes in the valence band is found by integrating Equation (4.2) over the valence-band energy, or

$$p_0 = \int g_v(E)[1 - f_F(E)]\,dE \tag{4.12}$$

We may note that

$$1 - f_F(E) = \frac{1}{1 + \exp\left(\dfrac{E_F - E}{kT}\right)} \tag{4.13a}$$

For energy states in the valence band, $E < E_v$. If $(E_F - E_v) \gg kT$ (the Fermi function is still assumed to be within the bandgap), then we have a slightly different form of the Boltzmann approximation. Equation (4.13a) may be written as

$$1 - f_F(E) = \frac{1}{1 + \exp\left(\dfrac{E_F - E}{kT}\right)} \approx \exp\left[\frac{-(E_F - E)}{kT}\right] \tag{4.13b}$$

Applying the Boltzmann approximation of Equation (4.13b) to Equation (4.12), we find the thermal-equilibrium concentration of holes in the valence band is

$$p_0 = \int_{-\infty}^{E_v} \frac{4\pi\,(2m_p^*)^{3/2}}{h^3} \sqrt{E_v - E}\, \exp\left[\frac{-(E_F - E)}{kT}\right] dE \tag{4.14}$$

where the lower limit of integration is taken as minus infinity instead of the bottom of the valence band. The exponential term decays fast enough so that this approximation is valid.

Equation (4.14) may be solved more easily by again making a change of variable. If we let

$$\eta' = \frac{E_v - E}{kT} \tag{4.15}$$

then Equation (4.14) becomes

$$p_0 = \frac{-4\pi(2m_p^* kT)^{3/2}}{h^3} \exp\left[\frac{-(E_F - E_v)}{kT}\right] \int_{+\infty}^{0} (\eta')^{1/2} \exp(-\eta')\, d\eta' \tag{4.16}$$

where the negative sign comes from the differential $dE = -kTd\eta'$. Note that the lower limit of η' becomes $+\infty$ when $E = -\infty$. If we change the order of integration, we introduce another minus sign. From Equation (4.8), Equation (4.16) becomes

$$p_0 = 2\left(\frac{2\pi m_p^* kT}{h^2}\right)^{3/2} \exp\left[\frac{-(E_F - E_v)}{kT}\right] \tag{4.17}$$

We may define a parameter N_v as

$$N_v = 2\left(\frac{2\pi m_p^* kT}{h^2}\right)^{3/2} \tag{4.18}$$

which is called the *effective density of states function in the valence band*. The parameter m_p^* is the density of states effective mass of the hole. The thermal-equilibrium concentration of holes in the valence band may now be written as

$$p_0 = N_v \exp\left[\frac{-(E_F - E_v)}{kT}\right] \tag{4.19}$$

The magnitude of N_v is also on the order of 10^{19} cm^{-3} at $T = 300$ K for most semiconductors.

EXAMPLE 4.2

Objective: Calculate the thermal-equilibrium hole concentration in silicon at $T = 400$ K.

Assume that the Fermi energy is 0.27 eV above the valence-band energy. The value of N_v for silicon at $T = 300$ K is $N_v = 1.04 \times 10^{19}$ cm^{-3}. (See Appendix B)

■ **Solution**

The parameter values at $T = 400$ K are found as:

$$N_v = (1.04 \times 10^{19})\left(\frac{400}{300}\right)^{3/2} = 1.60 \times 10^{19} \text{ cm}^{-3}$$

and

$$kT = (0.0259)\left(\frac{400}{300}\right) = 0.03453 \text{ eV}$$

The hole concentration is then

$$p_0 = N_v \exp\left[\frac{-(E_F - E_v)}{kT}\right] = (1.60 \times 10^{19}) \exp\left(\frac{-0.27}{0.03453}\right)$$

or

$$p_0 = 6.43 \times 10^{15} \text{ cm}^{-3}$$

Comment

The parameter values at any temperature can easily be found by using the 300 K values and the temperature dependence.

EXERCISE PROBLEM

Ex 4.2 (*a*) Repeat Example 4.2 at $T = 250$ K. (*b*) What is the ratio of p_0 at $T = 250$ K to that at $T = 400$ K? [Ans. (*a*) 4.54×10^{-3} cm^{-3}; (*b*) 2.92×10^{13} cm^{-3}]

The effective density of states functions, N_c and N_v, are constant for a given semiconductor material at a fixed temperature. Table 4.1 gives the values of the density of states function and of the density of states effective masses for silicon, gallium arsenide, and germanium. Note that the value of N_c for gallium arsenide is smaller than the typical 10^{19} cm^{-3} value. This difference is due to the small electron effective mass in gallium arsenide.

The thermal-equilibrium concentrations of electrons in the conduction band and of holes in the valence band are directly related to the effective density of states constants and to the Fermi energy level.

TEST YOUR UNDERSTANDING

TYU 4.1 Calculate the thermal equilibrium electron and hole concentration in silicon at $T = 300$ K for the case when the Fermi energy level is 0.22 eV below the conduction-band energy E_c. The value of E_g is given in Appendix B.4. (Ans. $n_0 = 5.73 \times 10^{15}$ cm^{-3}, $p_0 = 8.43 \times 10^{3}$ cm^{-3})

TYU 4.2 Determine the thermal equilibrium electron and hole concentration in GaAs at $T = 300$ K for the case when the Fermi energy level is 0.30 eV above the valence-band energy E_v. The value of E_g is given in Appendix B.4. (Ans. $n_0 = 0.0779$ cm^{-3}, $p_0 = 6.53 \times 10^{13}$ cm^{-3})

4.1.3 The Intrinsic Carrier Concentration

For an intrinsic semiconductor, the concentration of electrons in the conduction band is equal to the concentration of holes in the valence band. We may denote n_i and p_i as the electron and hole concentrations, respectively, in the intrinsic semiconductor. These parameters are usually referred to as the intrinsic electron concentration and intrinsic hole concentration. However, $n_i = p_i$, so normally we simply use the parameter n_i as the intrinsic carrier concentration, which refers to either the intrinsic electron or hole concentration.

Table 4.1 | Effective density of states function and density of states effective mass values

	N_c (cm^{-3})	N_v (cm^{-3})	m_n^*/m_0	m_p^*/m_0
Silicon	2.8×10^{19}	1.04×10^{19}	1.08	0.56
Gallium arsenide	4.7×10^{17}	7.0×10^{18}	0.067	0.48
Germanium	1.04×10^{19}	6.0×10^{18}	0.55	0.37

The Fermi energy level for the intrinsic semiconductor is called the intrinsic Fermi energy, or $E_F = E_{Fi}$. If we apply Equations (4.11) and (4.19) to the intrinsic semiconductor, then we can write

$$n_0 = n_i = N_c \exp\left[\frac{-(E_c - E_{Fi})}{kT}\right] \tag{4.20}$$

and

$$p_0 = p_i = n_i = N_v \exp\left[\frac{-(E_{Fi} - E_v)}{kT}\right] \tag{4.21}$$

If we take the product of Equations (4.20) and (4.21), we obtain

$$n_i^2 = N_c N_v \exp\left[\frac{-(E_c - E_{Fi})}{kT}\right] \cdot \exp\left[\frac{-(E_{Fi} - E_v)}{kT}\right] \tag{4.22}$$

or

$$\boxed{n_i^2 = N_c N_v \exp\left[\frac{-(E_c - E_v)}{kT}\right] = N_c N_v \exp\left[\frac{-E_g}{kT}\right]} \tag{4.23}$$

where E_g is the bandgap energy. For a given semiconductor material at a constant temperature, the value of n_i is a constant, and independent of the Fermi energy.

The intrinsic carrier concentration for silicon at $T = 300$ K may be calculated by using the effective density of states function values from Table 4.1. The value of n_i calculated from Equation (4.23) for $E_g = 1.12$ eV is $n_i = 6.95 \times 10^9$ cm^{-3}. The commonly accepted value[2] of n_i for silicon at $T = 300$ K is approximately 1.5×10^{10} cm^{-3}. This discrepancy may arise from several sources. First, the values of the effective masses are determined at a low temperature where the cyclotron resonance experiments are performed. Since the effective mass is an experimentally determined parameter, and since the effective mass is a measure of how well a particle moves in a crystal, this parameter may be a slight function of temperature. Next, the density of states function for a semiconductor was obtained by generalizing the model of an electron in a three-dimensional infinite potential well. This theoretical function may also not agree exactly with experiment. However, the difference between the theoretical value and the experimental value of n_i is approximately a factor of 2, which, in many cases, is not significant. Table 4.2 lists the commonly accepted values of n_i for silicon, gallium arsenide, and germanium at $T = 300$ K.

The intrinsic carrier concentration is a very strong function of temperature.

Table 4.2 | Commonly accepted values of n_i at $T = 300$ K

Silicon	$n_i = 1.5 \times 10^{10}$ cm^{-3}
Gallium arsenide	$n_i = 1.8 \times 10^{6}$ cm^{-3}
Germanium	$n_i = 2.4 \times 10^{13}$ cm^{-3}

[2]Various references may list slightly different values of the intrinsic silicon concentration at room temperature. In general, they are all between 1×10^{10} and 1.5×10^{10} cm^{-3}. This difference is, in most cases, not significant.

EXAMPLE 4.3

Objective: Calculate the intrinsic carrier concentration in silicon at $T = 250$ K and at $T = 400$ K.

The values of N_c and N_v for silicon at $T = 300$ K are 2.8×10^{19} cm^{-3} and 1.04×10^{19} cm^{-3}, respectively. Both N_c and N_v vary as $T^{3/2}$. Assume the bandgap energy of silicon is 1.12 eV and does not vary over this temperature range.

■ Solution

Using Equation (4.23), we find, at $T = 250$ K

$$n_i^2 = (2.8 \times 10^{19})(1.04 \times 10^{19})\left(\frac{250}{300}\right)^3 \exp\left[\frac{-1.12}{(0.0259)(250/300)}\right]$$

$$= 4.90 \times 10^{15}$$

or

$$n_i = 7.0 \times 10^7 \text{ cm}^{-3}$$

At $T = 400$ K, we find

$$n_i^2 = (2.8 \times 10^{19})(1.04 \times 10^{19})\left(\frac{400}{300}\right)^3 \exp\left[\frac{-1.12}{(0.0259)(400/300)}\right]$$

$$= 5.67 \times 10^{24}$$

or

$$n_i = 2.38 \times 10^{12} \text{ cm}^{-3}$$

■ Comment

We may note from this example that the intrinsic carrier concentration increased by over 4 orders of magnitude as the temperature increased by 150°C.

■ EXERCISE PROBLEM

Ex 4.3 (*a*) Calculate the intrinsic carrier concentration in GaAs at $T = 400$ K and at $T = 250$ K. Assume that $E_g = 1.42$ eV is constant over this temperature range. (*b*) What is the ratio of n_i at $T = 400$ K to that at $T = 250$ K?

[Ans. (*a*) $n_i(400) = 3.29 \times 10^9$ cm^{-3}; $n_i(250) = 7.13 \times 10^3$ cm^{-3}; (*b*) 4.61×10^5]

Figure 4.2 is a plot of n_i from Equation (4.23) for silicon, gallium arsenide, and germanium as a function of temperature. As seen in the figure, the value of n_i for these semiconductors may easily vary over several orders of magnitude as the temperature changes over a reasonable range.

TEST YOUR UNDERSTANDING

TYU 4.3 Calculate the intrinsic concentration in silicon at (*a*) $T = 200$ K and (*b*) $T = 450$ K. (*c*) Determine the ratio of n_i at $T = 450$ K to that at $T = 200$ K.

[Ans. (*a*) $n_i = 7.63 \times 10^4$ cm^{-3}; (*b*) $n_i = 1.72 \times 10^{13}$ cm^{-3}; (*c*) 2.26×10^8]

TYU 4.4 Repeat TYU 4.3 for GaAs.

[Ans. (*a*) $n_i = 1.37$ cm^{-3}; (*b*) $n_i = 3.85 \times 10^{10}$ cm^{-3}; (*c*) 2.81×10^{10}]

TYU 4.5 Repeat TYU 4.3 for Ge.

[Ans. (*a*) $n_i = 2.15 \times 10^{10}$ cm^{-3}; (*b*) $n_i = 2.97 \times 10^{13}$ cm^{-3}; (*c*) 1.38×10^3]

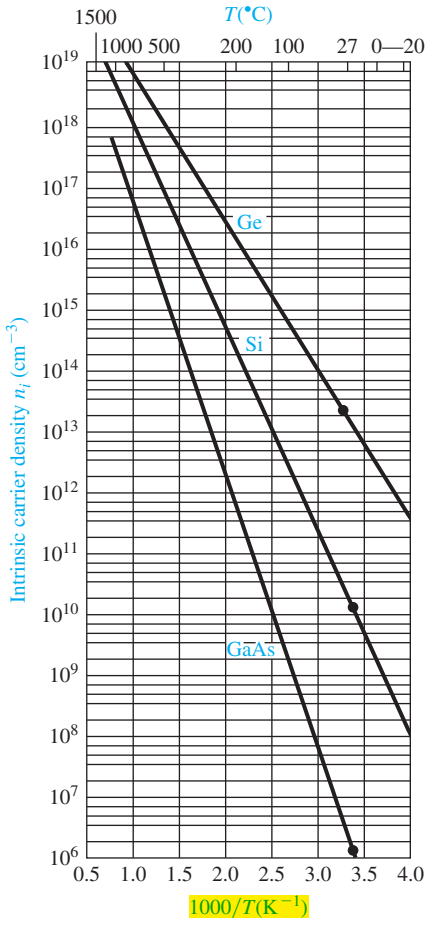


Figure 4.2 | The intrinsic carrier concentration of Ge, Si, and GaAs as a function of temperature.
(From Sze [14].)

4.1.4 The Intrinsic Fermi-Level Position

We have qualitatively argued that the Fermi energy level is located near the center of the forbidden bandgap for the intrinsic semiconductor. We can specifically calculate the intrinsic Fermi-level position. Since the electron and hole concentrations are equal, setting Equations (4.20) and (4.21) equal to each other, we have

$$N_c \exp\left[\frac{-(E_c - E_{Fi})}{kT}\right] = N_v \exp\left[\frac{-(E_{Fi} - E_v)}{kT}\right] \tag{4.24}$$

If we take the natural log of both sides of this equation and solve for E_{Fi}, we obtain

$$E_{Fi} = \frac{1}{2}(E_c + E_v) + \frac{1}{2}kT\ln\left(\frac{N_v}{N_c}\right) \tag{4.25}$$

From the definitions for N_c and N_v given by Equations (4.10) and (4.18), respectively, Equation (4.25) may be written as

$$E_{Fi} = \frac{1}{2}(E_c + E_v) + \frac{3}{4}kT\ln\left(\frac{m_p^*}{m_n^*}\right) \tag{4.26a}$$

The first term, $\frac{1}{2}(E_c + E_v)$, is the energy exactly midway between E_c and E_v, or the midgap energy. We can define

$$\frac{1}{2}(E_c + E_v) = E_{\text{midgap}}$$

so that

$$E_{Fi} - E_{\text{midgap}} = \frac{3}{4}kT\ln\left(\frac{m_p^*}{m_n^*}\right) \tag{4.26b}$$

If the electron and hole effective masses are equal so that $m_p^* = m_n^*$, then the intrinsic Fermi level is exactly in the center of the bandgap. If $m_p^* > m_n^*$, the intrinsic Fermi level is slightly above the center, and if $m_p^* < m_n^*$, it is slightly below the center of the bandgap. The density of states function is directly related to the carrier effective mass; thus, a larger effective mass means a larger density of states function. The intrinsic Fermi level must shift away from the band with the larger density of states in order to maintain equal numbers of electrons and holes.

EXAMPLE 4.4

Objective: Calculate the position of the intrinsic Fermi level with respect to the center of the bandgap in silicon at $T = 300$ K.

The density of states effective carrier masses in silicon are $m_n^* = 1.08m_0$ and $m_p^* = 0.56m_0$.

■ Solution

The intrinsic Fermi level with respect to the center of the bandgap is

$$E_{Fi} - E_{\text{midgap}} = \frac{3}{4}kT\ln\left(\frac{m_p^*}{m_n^*}\right) = \frac{3}{4}(0.0259)\ln\left(\frac{0.56}{1.08}\right)$$

or

$$E_{Fi} - E_{\text{midgap}} = -0.0128 \text{ eV} = -12.8 \text{ meV}$$

■ Comment

The intrinsic Fermi level in silicon is 12.8 meV below the midgap energy. If we compare 12.8 meV to 560 meV, which is one-half of the bandgap energy of silicon, we can, in many applications, simply approximate the intrinsic Fermi level to be in the center of the bandgap.

■ EXERCISE PROBLEM

Ex 4.4 Determine the position of the intrinsic Fermi level at $T = 300$ K with respect to the center of the bandgap for (*a*) GaAs and (*b*) Ge. [Ans. (*a*) +38.25 meV; (*b*) −7.70 meV]

TEST YOUR UNDERSTANDING

TYU 4.6 Determine the position of the intrinsic Fermi level with respect to the center of the bandgap in silicon at (*a*) $T = 200$ K and (*b*) $T = 400$ K. Assume the effective masses are constant over this temperature range. [Ans. (*a*) −8.505 meV; (*b*) −17.01meV]

4.2 | DOPANT ATOMS AND ENERGY LEVELS

The intrinsic semiconductor may be an interesting material, but the real power of semiconductors is realized by adding small, controlled amounts of specific dopant, or impurity, atoms. This doping process, described briefly in Chapter 1, can greatly alter the electrical characteristics of the semiconductor. The doped semiconductor, called an *extrinsic* material, is the primary reason we can fabricate the various semiconductor devices that we will consider in later chapters.

4.2.1 Qualitative Description

In Chapter 3, we discussed the covalent bonding of silicon and considered the simple two-dimensional representation of the single-crystal silicon lattice as shown in Figure 4.3. Now consider adding a group V element, such as phosphorus, as a substitutional impurity. The group V element has five valence electrons. Four of these will contribute to the covalent bonding with the silicon atoms, leaving the fifth more loosely bound to the phosphorus atom. This effect is schematically shown in Figure 4.4. We refer to the fifth valence electron as a donor electron.

The phosphorus atom without the donor electron is positively charged. At very low temperatures, the donor electron is bound to the phosphorus atom. However, by intuition, it should seem clear that the energy required to elevate the donor electron into the conduction band is considerably less than that for the electrons involved in the covalent bonding. Figure 4.5 shows the energy-band diagram that we would expect. The energy level, E_d, is the energy state of the donor electron.

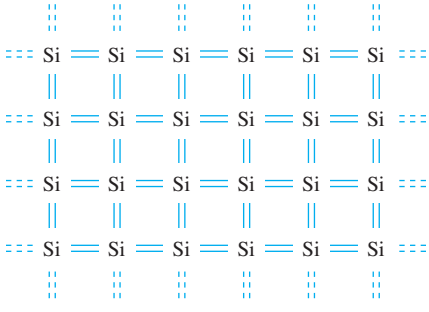


Figure 4.3 | Two-dimensional representation of the intrinsic silicon lattice.

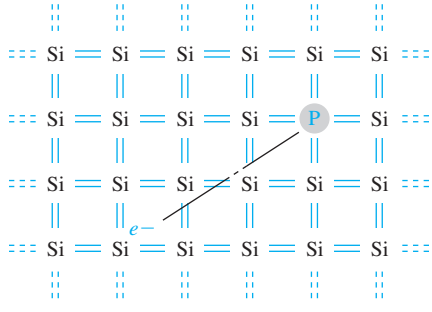


Figure 4.4 | Two-dimensional representation of the silicon lattice doped with a phosphorus atom.

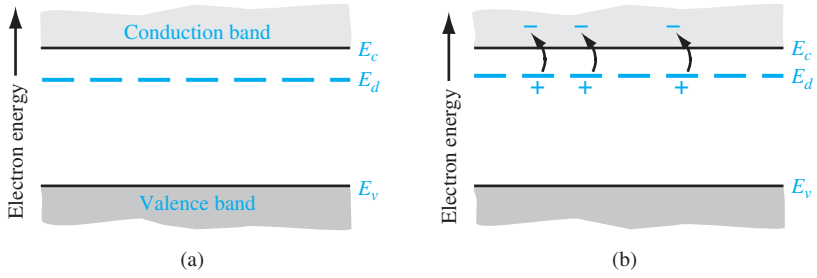


Figure 4.5 | The energy-band diagram showing (a) the discrete donor energy state and (b) the effect of a donor state being ionized.

If a small amount of energy, such as thermal energy, is added to the donor electron, it can be elevated into the conduction band, leaving behind a positively charged phosphorus ion. The electron in the conduction band can now move through the crystal generating a current, while the positively charged ion is fixed in the crystal. This type of impurity atom donates an electron to the conduction band and so is called a *donor impurity atom*. The donor impurity atoms add electrons to the conduction band without creating holes in the valence band. The resulting material is referred to as an *n-type* semiconductor (*n* for the negatively charged electron).

Now consider adding a group III element, such as boron, as a substitutional impurity to silicon. The group III element has three valence electrons, which are all taken up in the covalent bonding. As shown in Figure 4.6a, one covalent bonding position appears to be empty. If an electron were to occupy this "empty" position, its energy would have to be greater than that of the valence electrons, since the net charge state of the boron atom would now be negative. However, the electron occupying this "empty" position does not have sufficient energy to be in the conduction band, so its energy is far smaller than the conduction-band energy. Figure 4.6b shows how valence electrons may gain a small amount of thermal energy and move about in

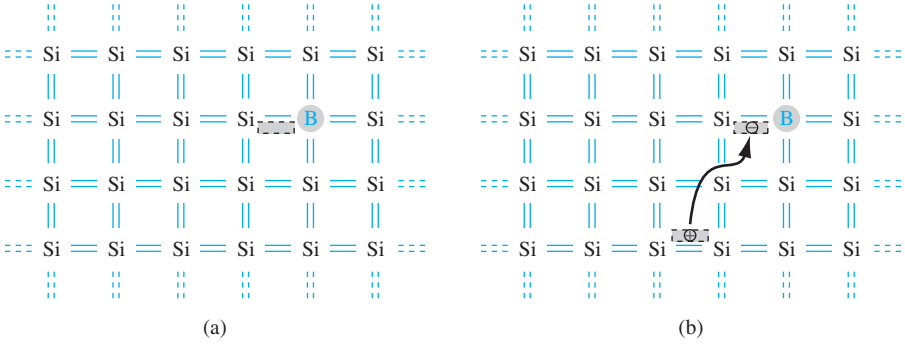


Figure 4.6 | Two-dimensional representation of a silicon lattice (a) doped with a boron atom and (b) showing the ionization of the boron atom resulting in a hole.

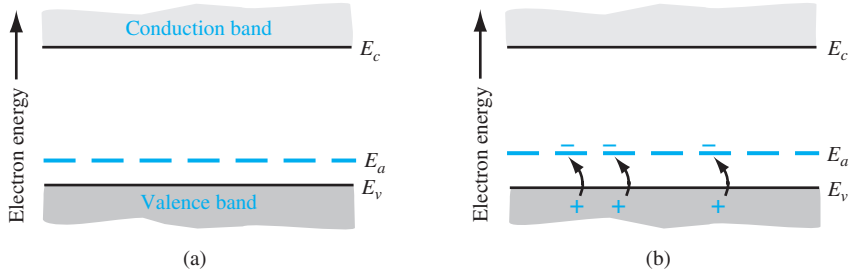


Figure 4.7 | The energy-band diagram showing (a) the discrete acceptor energy state and (b) the effect of an acceptor state being ionized.

the crystal. The "empty" position associated with the boron atom becomes occupied, and other valence electron positions become vacated. These other vacated electron positions can be thought of as holes in the semiconductor material.

Figure 4.7 shows the expected energy state of the "empty" position and also the formation of a hole in the valence band. The hole can move through the crystal generating a current, while the negatively charged boron atom is fixed in the crystal. The group III atom accepts an electron from the valence band and so is referred to as an *acceptor impurity atom*. The acceptor atom can generate holes in the valence band without generating electrons in the conduction band. This type of semiconductor material is referred to as a *p type* material (*p* for the positively charged hole).

The pure single-crystal semiconductor material is called an intrinsic material. Adding controlled amounts of dopant atoms, either donors or acceptors, creates a material called an *extrinsic semiconductor.* An extrinsic semiconductor will have either a preponderance of electrons (n type) or a preponderance of holes (p type).

4.2.2 Ionization Energy

We can calculate the approximate distance of the donor electron from the donor impurity ion, and also the approximate energy required to elevate the donor electron into the conduction band. This energy is referred to as the ionization energy. We will use the Bohr model of the atom for these calculations. The justification for using this model is that the most probable distance of an electron from the nucleus in a hydrogen atom, determined from quantum mechanics, is the same as the Bohr radius. The energy levels in the hydrogen atom determined from quantum mechanics are also the same as obtained from the Bohr theory.

In the case of the donor impurity atom, we may visualize the donor electron orbiting the donor ion, which is embedded in the semiconductor material. We will need to use the permittivity of the semiconductor material in the calculations rather than the permittivity of free space as is used in the case of the hydrogen atom. We will also use the effective mass of the electron in the calculations.

The analysis begins by setting the coulomb force of attraction between the electron and ion equal to the centripetal force of the orbiting electron. This condition will give a steady orbit. We have

$$\frac{e^2}{4\pi \epsilon r_n^2} = \frac{m^* v^2}{r_n} \tag{4.27}$$

where v is the magnitude of the velocity and r_n is the radius of the orbit. If we assume the angular momentum is also quantized, then we can write

$$m^* r_n v = n\hbar \tag{4.28}$$

where n is a positive integer. Solving for v from Equation (4.28), substituting into Equation (4.27), and solving for the radius, we obtain

$$r_n = \frac{n^2 \hbar^2 4\pi \epsilon}{m^* e^2} \tag{4.29}$$

The assumption of the angular momentum being quantized leads to the radius also being quantized.

The Bohr radius is defined as

$$a_0 = \frac{4\pi\epsilon_0 \hbar^2}{m_0 e^2} = 0.53 \text{ Å} \tag{4.30}$$

We can normalize the radius of the donor orbital to that of the Bohr radius, which gives

$$\frac{r_n}{a_0} = n^2 \epsilon_r \left(\frac{m_0}{m^*}\right) \tag{4.31}$$

where ϵ_r is the relative dielectric constant of the semiconductor material, m_0 is the rest mass of an electron, and m^* is the conductivity effective mass of the electron in the semiconductor.[3]

If we consider the lowest energy state in which $n = 1$, and if we consider silicon in which $\epsilon_r = 11.7$ and the conductivity effective mass is $m^*/m_0 = 0.26$, then we have that

$$\frac{r_1}{a_0} = 45 \tag{4.32}$$

or $r_1 = 23.9$ Å. This radius corresponds to approximately four lattice constants of silicon. Recall that one unit cell in silicon effectively contains eight atoms, so the radius of the orbiting donor electron encompasses many silicon atoms. The donor electron is not tightly bound to the donor atom.

The total energy of the orbiting electron is given by

$$E = T + V \tag{4.33}$$

where T is the kinetic energy and V is the potential energy of the electron. The kinetic energy is

$$T = \frac{1}{2} m^* v^2 \tag{4.34}$$

[3]The conductivity effective mass is used when electrons and holes are in motion. See Appendix F for a discussion of effective mass concepts.

Using the velocity v from Equation (4.28) and the radius r_n from Equation (4.29), the kinetic energy becomes

$$T = \frac{m^* e^4}{2(n\hbar)^2(4\pi\epsilon)^2} \tag{4.35}$$

The potential energy is

$$V = \frac{-e^2}{4\pi\epsilon r_n} = \frac{-m^* e^4}{(n\hbar)^2(4\pi\epsilon)^2} \tag{4.36}$$

The total energy is the sum of the kinetic and potential energies, so that

$$E = T + V = \frac{-m^* e^4}{2(n\hbar)^2(4\pi\epsilon)^2} \tag{4.37}$$

For the hydrogen atom, $m^* = m_0$ and $\epsilon = \epsilon_0$. The ionization energy of the hydrogen atom in the lowest energy state is then $E = -13.6$ eV. If we consider silicon, the ionization energy is $E = -25.8$ meV, much less than the bandgap energy of silicon. This energy is the approximate ionization energy of the donor atom, or the energy required to elevate the donor electron into the conduction band.

For ordinary donor impurities such as phosphorus or arsenic in silicon or germanium, this hydrogenic model works quite well and gives some indication of the magnitudes of the ionization energies involved. Table 4.3 lists the actual experimentally measured ionization energies for a few impurities in silicon and germanium. Germanium and silicon have different relative dielectric constants and effective masses; thus, we expect the ionization energies to differ.

4.2.3 Group III–V Semiconductors

In the previous sections, we have discussed the donor and acceptor impurities in a group IV semiconductor, such as silicon. The situation in the group III–V compound semiconductors, such as gallium arsenide, is more complicated. Group II elements, such as beryllium, zinc, and cadmium, can enter the lattice as substitutional impurities, replacing the group III gallium element to become acceptor impurities. Similarly, group VI elements, such as selenium and tellurium, can enter the lattice substitutionally, replacing the group V arsenic element to become donor impurities. The corresponding ionization energies for these impurities are smaller than those for the impurities in silicon. The ionization energies for the donors in gallium arsenide

Table 4.3 | Impurity ionization energies in silicon and germanium

Impurity	Ionization energy (eV) Si	Ge
Donors		
Phosphorus	0.045	0.012
Arsenic	0.05	0.0127
Acceptors		
Boron	0.045	0.0104
Aluminum	0.06	0.0102

Table 4.4 | Impurity ionization energies in gallium arsenide

Impurity	Ionization energy (eV)
Donors	
Selenium	0.0059
Tellurium	0.0058
Silicon	0.0058
Germanium	0.0061
Acceptors	
Beryllium	0.028
Zinc	0.0307
Cadmium	0.0347
Silicon	0.0345
Germanium	0.0404

are also smaller than those for the acceptors, because of the smaller effective mass of the electron compared to that of the hole.

Group IV elements, such as silicon and germanium, can also be impurity atoms in gallium arsenide. If a silicon atom replaces a gallium atom, the silicon impurity will act as a donor, but if the silicon atom replaces an arsenic atom, then the silicon impurity will act as an acceptor. The same is true for germanium as an impurity atom. Such impurities are called *amphoteric*. Experimentally in gallium arsenide, it is found that germanium is predominantly an acceptor and silicon is predominantly a donor. Table 4.4 lists the ionization energies for the various impurity atoms in gallium arsenide.

TEST YOUR UNDERSTANDING

TYU 4.7 (*a*) Calculate the ionization energy and the radius (normalized to the Bohr radius) of a donor electron in its lowest energy state in GaAs. (*b*) Repeat part (*a*) for Ge.

[Ans. (*a*) −5.30 meV, r_1/a_0 = 195.5; (*b*) −6.37 meV, r_1/a_0 133.3]

4.3 | THE EXTRINSIC SEMICONDUCTOR

We defined an intrinsic semiconductor as a material with no impurity atoms present in the crystal. An *extrinsic semiconductor* is defined as a semiconductor in which controlled amounts of specific dopant or impurity atoms have been added so that the thermal-equilibrium electron and hole concentrations are different from the intrinsic carrier concentration. One type of carrier will predominate in an extrinsic semiconductor.

4.3.1 Equilibrium Distribution of Electrons and Holes

Adding donor or acceptor impurity atoms to a semiconductor will change the distribution of electrons and holes in the material. Since the Fermi energy is related to the distribution function, the Fermi energy will change as dopant atoms are added.

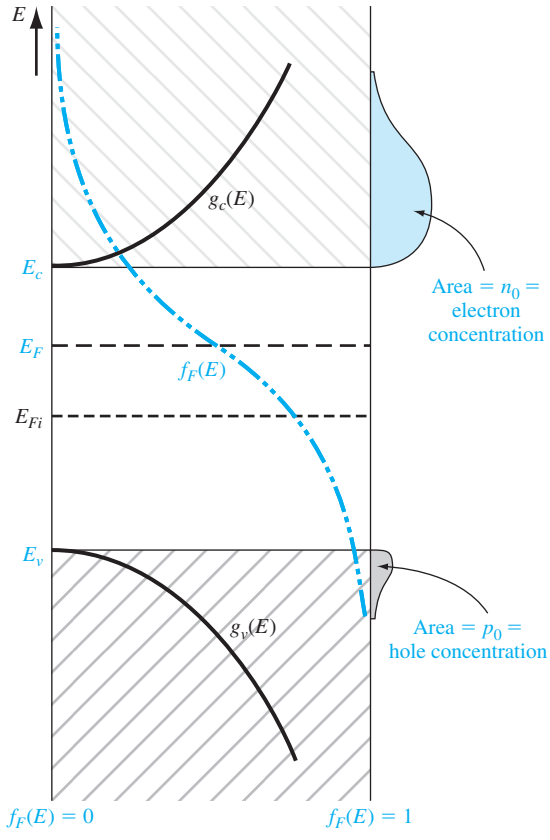


Figure 4.8 | Density of states functions, Fermi–Dirac probability function, and areas representing electron and hole concentrations for the case when E_F is above the intrinsic Fermi energy.

If the Fermi energy changes from near the midgap value, the density of electrons in the conduction band and the density of holes in the valence band will change. These effects are shown in Figures 4.8 and 4.9. Figure 4.8 shows the case for $E_F > E_{Fi}$ and Figure 4.9 shows the case for $E_F < E_{Fi}$. When $E_F > E_{Fi}$, the electron concentration is larger than the hole concentration, and when $E_F < E_{Fi}$, the hole concentration is larger than the electron concentration. When the density of electrons is greater than the density of holes, the semiconductor is n type; donor impurity atoms have been added. When the density of holes is greater than the density of electrons, the semiconductor is p type; acceptor impurity atoms have been added. The Fermi energy level in a semiconductor changes as the electron and hole concentrations change and, again, the Fermi energy changes as donor or acceptor impurities are added. The change in the Fermi level as a function of impurity concentrations is considered in Section 4.6.

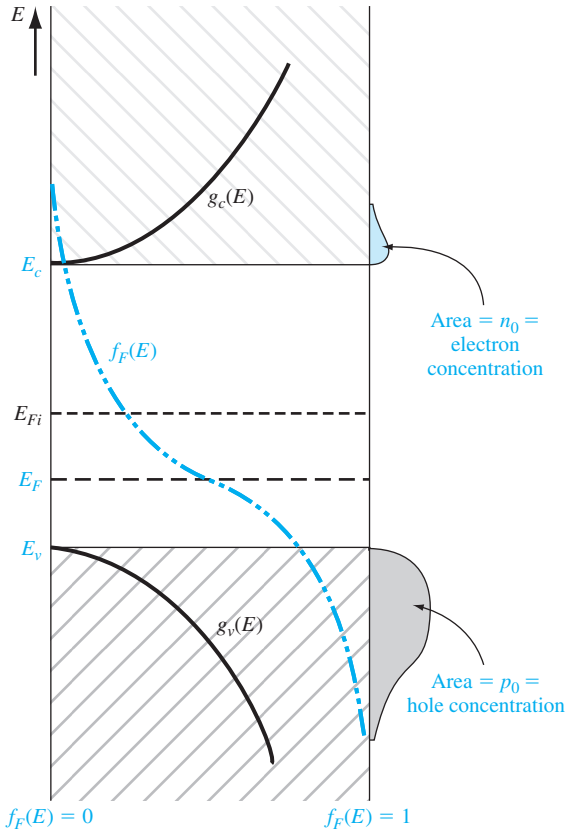


Figure 4.9 | Density of states functions, Fermi–Dirac probability function, and areas representing electron and hole concentrations for the case when E_F is below the intrinsic Fermi energy.

The expressions previously derived for the thermal-equilibrium concentration of electrons and holes, given by Equations (4.11) and (4.19), are general equations for n_0 and p_0 in terms of the Fermi energy. These equations are again given as

$$n_0 = N_c \exp\left[\frac{-(E_c - E_F)}{kT}\right]$$

and

$$p_0 = N_v \exp\left[\frac{-(E_F - E_v)}{kT}\right]$$

As we just discussed, the Fermi energy may vary through the bandgap energy, which will then change the values of n_0 and p_0.

EXAMPLE 4.5

Objective: Calculate the thermal equilibrium concentrations of electrons and holes for a given Fermi energy.

Consider silicon at $T = 300$ K so that $N_c = 2.8 \times 10^{19}$ cm^{-3} and $N_v = 1.04 \times 10^{19}$ cm^{-3}. Assume that the Fermi energy is 0.25 eV below the conduction band. If we assume that the bandgap energy of silicon is 1.12 eV, then the Fermi energy will be 0.87 eV above the valence band.

Solution

Using Equation (4.11), we have

$$n_0 = (2.8 \times 10^{19}) \exp\left(\frac{-0.25}{0.0259}\right) = 1.8 \times 10^{15} \text{ cm}^{-3}$$

From Equation (4.19), we can write

$$p_0 = (1.04 \times 10^{19}) \exp\left(\frac{-0.87}{0.0259}\right) = 2.7 \times 10^{4} \text{ cm}^{-3}$$

Comment

The change in the Fermi level is actually a function of the donor or acceptor impurity concentrations that are added to the semiconductor. However, this example shows that electron and hole concentrations change by orders of magnitude from the intrinsic carrier concentration as the Fermi energy changes by a few tenths of an electron-volt.

EXERCISE PROBLEM

Ex 4.5 Determine the thermal-equilibrium concentrations of electrons and holes in silicon at $T = 300$ K if the Fermi energy level is 0.215 eV above the valence-band energy E_v. (Ans. $p_0 = 2.58 \times 10^{15}$ cm^{-3}, $n_0 = 1.87 \times 10^{4}$ cm^{-3})

In the previous example, since $n_0 > p_0$, the semiconductor is n type. In an n-type semiconductor, electrons are referred to as the majority carrier and holes as the minority carrier. By comparing the relative values of n_0 and p_0 in the example, it is easy to see how this designation came about. Similarly, in a p-type semiconductor where $p_0 > n_0$, holes are the majority carrier and electrons are the minority carrier.

We may derive another form of the equations for the thermal-equilibrium concentrations of electrons and holes. If we add and subtract an intrinsic Fermi energy in the exponent of Equation (4.11), we can write

$$n_0 = N_c \exp\left[\frac{-(E_c - E_{Fi}) + (E_F - E_{Fi})}{kT}\right] \tag{4.38a}$$

or

$$n_0 = N_c \exp\left[\frac{-(E_c - E_{Fi})}{kT}\right] \exp\left[\frac{(E_F - E_{Fi})}{kT}\right] \tag{4.38b}$$

The intrinsic carrier concentration is given by Equation (4.20) as

$$n_i = N_c \exp\left[\frac{-(E_c - E_{Fi})}{kT}\right]$$

so that the thermal-equilibrium electron concentration can be written as

$$n_0 = n_i \exp\left[\frac{E_F - E_{Fi}}{kT}\right] \tag{4.39}$$

Similarly, if we add and subtract an intrinsic Fermi energy in the exponent of Equation (4.19), we will obtain

$$p_0 = n_i \exp\left[\frac{-(E_F - E_{Fi})}{kT}\right] \tag{4.40}$$

As we will see, the Fermi level changes when donors and acceptors are added, but Equations (4.39) and (4.40) show that, as the Fermi level changes from the intrinsic Fermi level, n_0 and p_0 change from the n_i value. If $E_F > E_{Fi}$, then we will have $n_0 > n_i$ and $p_0 < n_i$. One characteristic of an n-type semiconductor is that $E_F > E_{Fi}$ so that $n_0 > p_0$. Similarly, in a p-type semiconductor, $E_F < E_{Fi}$ so that $p_0 > n_i$ and $n_0 < n_i$; thus, $p_0 > n_0$.

We can see the functional dependence of n_0 and p_0 with E_F in Figures 4.8 and 4.9. As E_F moves above or below E_{Fi}, the overlapping probability function with the density of states functions in the conduction band and valence band changes. As E_F moves above E_{Fi}, the probability function in the conduction band increases, while the probability, $1 - f_F(E)$, of an empty state (hole) in the valence band decreases. As E_F moves below E_{Fi}, the opposite occurs.

4.3.2 The $n_0 p_0$ Product

We may take the product of the general expressions for n_0 and p_0 as given in Equations (4.11) and (4.19), respectively. The result is

$$n_0 p_0 = N_c N_v \exp\left[\frac{-(E_c - E_F)}{kT}\right] \exp\left[\frac{-(E_F - E_v)}{kT}\right] \tag{4.41}$$

which may be written as

$$n_0 p_0 = N_c N_v \exp\left[\frac{-E_g}{kT}\right] \tag{4.42}$$

As Equation (4.42) was derived for a general value of Fermi energy, the values of n_0 and p_0 are not necessarily equal. However, Equation (4.42) is exactly the same as Equation (4.23), which we derived for the case of an intrinsic semiconductor. We then have that, for the semiconductor in thermal equilibrium,

$$n_0 p_0 = n_i^2 \tag{4.43}$$

Equation (4.43) states that the product of n_0 and p_0 is always a constant for a given semiconductor material at a given temperature. Although this equation seems

very simple, it is one of the fundamental principles of semiconductors in thermal equilibrium. The significance of this relation will become more apparent in the chapters that follow. It is important to keep in mind that Equation (4.43) was derived using the Boltzmann approximation. If the Boltzmann approximation is not valid, then likewise, Equation (4.43) is not valid.

An extrinsic semiconductor in thermal equilibrium does not, strictly speaking, contain an intrinsic carrier concentration, although some thermally generated carriers are present. The intrinsic electron and hole carrier concentrations are modified by the donor or acceptor impurities. However, we may think of the intrinsic concentration n_i in Equation (4.43) simply as a parameter of the semiconductor material.

*4.3.3 The Fermi–Dirac Integral

In the derivation of the Equations (4.11) and (4.19) for the thermal equilibrium electron and hole concentrations, we assumed that the Boltzmann approximation was valid. If the Boltzmann approximation does not hold, the thermal equilibrium electron concentration is written from Equation (4.3) as

$$n_0 = \frac{4\pi}{h^3}(2m_n^*)^{3/2}\int_{E_c}^{\infty}\frac{(E-E_c)^{1/2}\,dE}{1+\exp\left(\dfrac{E-E_F}{kT}\right)} \tag{4.44}$$

If we again make a change of variable and let

$$\eta = \frac{E-E_c}{kT} \tag{4.45a}$$

and also define

$$\eta_F = \frac{E_F-E_c}{kT} \tag{4.45b}$$

then we can rewrite Equation (4.44) as

$$n_0 = 4\pi\left(\frac{2m_n^*kT}{h^2}\right)^{3/2}\int_0^{\infty}\frac{\eta^{1/2}\,d\eta}{1+\exp(\eta-\eta_F)} \tag{4.46}$$

The integral is defined as

$$F_{1/2}(\eta_F) = \int_0^{\infty}\frac{\eta^{1/2}\,d\eta}{1+\exp(\eta-\eta_F)} \tag{4.47}$$

This function, called the Fermi–Dirac integral, is a tabulated function of the variable η_F. Figure 4.10 is a plot of the Fermi–Dirac integral. Note that if $\eta_F > 0$, then $E_F > E_c$; thus, the Fermi energy is actually in the conduction band.

EXAMPLE 4.6

Objective: Calculate the electron concentration using the Fermi–Dirac integral.

Let $\eta_F = 2$ so that the Fermi energy is above the conduction band by approximately 52 meV at $T = 300$ K.

■ Solution

Equation (4.46) can be written as

$$n_0 = \frac{2}{\sqrt{\pi}}N_cF_{1/2}(\eta_F)$$

For silicon at $T = 300$ K, $N_c = 2.8 \times 10^{19}$ cm^{-3} and, from Figure 4.10, the Fermi–Dirac integral has a value or $F_{1/2}(2) = 2.7$. Then

$$n_0 = \frac{2}{\sqrt{\pi}}(2.8 \times 10^{19})(2.7) = 8.53 \times 10^{19} \text{ cm}^{-3}$$

Comment

Note that if we had used Equation (4.11), the thermal equilibrium value of n_0 would be $n_0 = 2.08 \times 10^{20}$ cm^{-3}, which is incorrect since the Boltzmann approximation is not valid for this case.

EXERCISE PROBLEM

Ex 4.6 If $n_0 = 1.5 \times 10^{20}$ cm^{-3} in silicon at $T = 300$ K, determine the position of the Fermi level relative to the conduction-band energy E_c. (Ans. $E_F - E_c \cong 0.08288$ eV)

We may use the same general method to calculate the thermal equilibrium concentration of holes. We obtain

$$p_0 = 4\pi\left(\frac{2m_p^* kT}{h^2}\right)^{3/2} \int_0^\infty \frac{(\eta')^{1/2}\, d\eta'}{1 + \exp(\eta' - \eta_F')} \tag{4.48}$$

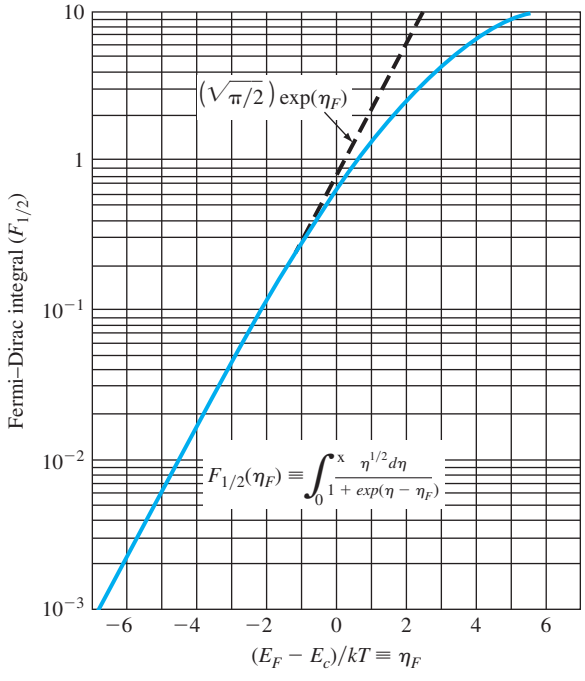


Figure 4.10 | The Fermi–Dirac integral $F_{1/2}$ as a function of the Fermi energy.
(*From Sze [14].*)

where

$$\eta' = \frac{E_v - E}{kT} \tag{4.49a}$$

and

$$\eta'_F = \frac{E_v - E_F}{kT} \tag{4.49b}$$

The integral in Equation (4.48) is the same Fermi–Dirac integral defined by Equation (4.47), although the variables have slightly different definitions. We may note that if $\eta'_F > 0$, then the Fermi level is in the valence band.

TEST YOUR UNDERSTANDING

TYU 4.8 (*a*) Calculate the thermal-equilibrium electron concentration in silicon at $T = 300$ K for the case when $E_F = E_c$. (*b*) Calculate the thermal-equilibrium hole concentration in silicon at $T = 300$ K for the case when $E_F = E_v$.

[Ans. (*a*) $n_0 = 2.05 \times 10^{19}$ cm^{-3}; (*b*) $p_0 = 7.63 \times 10^{18}$ cm^{-3}]

4.3.4 Degenerate and Nondegenerate Semiconductors

In our discussion of adding dopant atoms to a semiconductor, we have implicitly assumed that the concentration of dopant atoms added is small when compared to the density of host or semiconductor atoms. The small number of impurity atoms are spread far enough apart so that there is no interaction between donor electrons, for example, in an n-type material. We have assumed that the impurities introduce discrete, noninteracting donor energy states in the n-type semiconductor and discrete, noninteracting acceptor states in the p-type semiconductor. These types of semiconductors are referred to as nondegenerate semiconductors.

If the impurity concentration increases, the distance between the impurity atoms decreases and a point will be reached when donor electrons, for example, will begin to interact with each other. When this occurs, the single discrete donor energy will split into a band of energies. As the donor concentration further increases, the band of donor states widens and may overlap the bottom of the conduction band. This overlap occurs when the donor concentration becomes comparable with the effective density of states. When the concentration of electrons in the conduction band exceeds the density of states N_c, the Fermi energy lies within the conduction band. This type of semiconductor is called a degenerate n-type semiconductor.

In a similar way, as the acceptor doping concentration increases in a p-type semiconductor, the discrete acceptor energy states will split into a band of energies and may overlap the top of the valence band. The Fermi energy will lie in the valence band when the concentration of holes exceeds the density of states N_v. This type of semiconductor is called a degenerate p-type semiconductor.

Schematic models of the energy-band diagrams for a degenerate n-type and degenerate p-type semiconductor are shown in Figure 4.11. The energy states below E_F

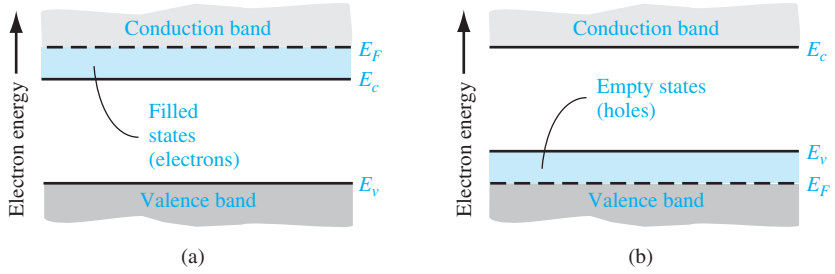


Figure 4.11 | Simplified energy-band diagrams for degenerately doped (a) n-type and (b) p-type semiconductors.

are mostly filled with electrons and the energy states above E_F are mostly empty. In the degenerate n-type semiconductor, the states between E_F and E_c are mostly filled with electrons; thus, the electron concentration in the conduction band is very large. Similarly, in the degenerate p-type semiconductor, the energy states between E_v and E_F are mostly empty; thus, the hole concentration in the valence band is very large.

4.4 | STATISTICS OF DONORS AND ACCEPTORS

In the previous chapter, we discussed the Fermi–Dirac distribution function, which gives the probability that a particular energy state will be occupied by an electron. We need to reconsider this function and apply the probability statistics to the donor and acceptor energy states.

4.4.1 Probability Function

One postulate used in the derivation of the Fermi–Dirac probability function was the Pauli exclusion principle, which states that only one particle is permitted in each quantum state. The Pauli exclusion principle also applies to the donor and acceptor states.

Suppose we have N_i electrons and g_i quantum states, where the subscript i indicates the ith energy level. There are g_i ways of choosing where to put the first particle. Each donor level has two possible spin orientations for the donor electron; thus, each donor level has two quantum states. The insertion of an electron into one quantum state, however, precludes putting an electron into the second quantum state. By adding one electron, the vacancy requirement of the atom is satisfied, and the addition of a second electron in the donor level is not possible. The distribution function of donor electrons in the donor energy states is then slightly different than the Fermi–Dirac function.

The probability function of electrons occupying the donor state is

$$n_d = \frac{N_d}{1 + \frac{1}{2}\exp\left(\frac{E_d - E_F}{kT}\right)} \tag{4.50}$$

where n_d is the density of electrons occupying the donor level and E_d is the energy of the donor level. The factor $\frac{1}{2}$ in this equation is a direct result of the spin factor just mentioned. The $\frac{1}{2}$ factor is sometimes written as $1/g$, where g is called a degeneracy factor.

Equation (4.50) can also be written in the form

$$n_d = N_d - N_d^+ \tag{4.51}$$

where N_d^+ is the concentration of ionized donors. In many applications, we will be interested more in the concentration of ionized donors than in the concentration of electrons remaining in the donor states.

If we do the same type of analysis for acceptor atoms, we obtain the expression

$$p_a = \frac{N_a}{1 + \frac{1}{g}\exp\left(\frac{E_F - E_a}{kT}\right)} = N_a - N_a^- \tag{4.52}$$

where N_a is the concentration of acceptor atoms, E_a is the acceptor energy level, p_a is the concentration of holes in the acceptor states, and N_a^- is the concentration of ionized acceptors. A hole in an acceptor state corresponds to an acceptor atom that is neutrally charged and still has an "empty" bonding position as we have discussed in Section 4.2.1. The parameter g is, again, a degeneracy factor. The ground state degeneracy factor g is normally taken as 4 for the acceptor level in silicon and gallium arsenide because of the detailed band structure.

4.4.2 Complete Ionization and Freeze-Out

The probability function for electrons in the donor energy state was just given by Equation (4.50). If we assume that $(E_d - E_F) \gg kT$, then

$$n_d \approx \frac{N_d}{\frac{1}{2}\exp\left(\frac{E_d - E_F}{kT}\right)} = 2N_d \exp\left[\frac{-(E_d - E_F)}{kT}\right] \tag{4.53}$$

If $(E_d - E_F) \gg kT$, then the Boltzmann approximation is also valid for the electrons in the conduction band so that, from Equation (4.11),

$$n_0 = N_c \exp\left[\frac{-(E_c - E_F)}{kT}\right]$$

We can determine the relative number of electrons in the donor state compared with the total number of electrons; therefore, we can consider the ratio of electrons in the donor state to the total number of electrons in the conduction band plus donor state. Using the expressions of Equations (4.53) and (4.11), we write

$$\frac{n_d}{n_d + n_0} = \frac{2N_d \exp\left[\frac{-(E_d - E_F)}{kT}\right]}{2N_d \exp\left[\frac{-(E_d - E_F)}{kT}\right] + N_c \exp\left[\frac{-(E_c - E_F)}{kT}\right]} \tag{4.54}$$

The Fermi energy cancels out of this expression. Dividing by the numerator term, we obtain

$$\frac{n_d}{n_d + n_0} = \frac{1}{1 + \dfrac{N_c}{2N_d}\exp\left[\dfrac{-(E_c - E_d)}{kT}\right]} \tag{4.55}$$

The factor $(E_c - E_d)$ is just the ionization energy of the donor electrons.

EXAMPLE 4.7

Objective: Determine the fraction of total electrons still in the donor states at $T = 300$ K.

Consider phosphorus doping in silicon, for $T = 300$ K, at a concentration of $N_d = 10^{16}$ cm^{-3}.

■ Solution

Using Equation (4.55), we find

$$\frac{n_d}{n_0 + n_d} = \frac{1}{1 + \dfrac{2.8 \times 10^{19}}{2(10^{16})}\exp\left(\dfrac{-0.045}{0.0259}\right)} = 0.0041 = 0.41\%$$

■ Comment

This example shows that there are very few electrons in the donor state compared with the conduction band. Essentially all of the electrons from the donor states are in the conduction band and, since only about 0.4 percent of the donor states contain electrons, the donor states are said to be completely ionized.

■ EXERCISE PROBLEM

Ex 4.7 Repeat Example 4.7 for (*a*) $T = 250$ K and (*b*) $T = 200$ K. (*c*) What can be said about the fraction as the temperature decreases?

[Ans. (*a*) 7.50×10^{-3}; (*b*) 1.75×10^{-2}; (*c*) Fraction increases as temperature decreases.]

At room temperature, then, the donor states are essentially completely ionized and, for a typical doping of 10^{16} cm^{-3}, almost all donor impurity atoms have donated an electron to the conduction band.

At room temperature, there is also essentially *complete ionization* of the acceptor atoms. This means that each acceptor atom has accepted an electron from the valence band so that p_a is zero. At typical acceptor doping concentrations, a hole is created in the valence band for each acceptor atom. This ionization effect and the creation of electrons and holes in the conduction band and valence band, respectively, are shown in Figure 4.12.

The opposite of complete ionization occurs at $T = 0$ K. At absolute zero degrees, all electrons are in their lowest possible energy state; that is, for an n-type semiconductor, each donor state must contain an electron, therefore $n_d = N_d$ or $N_d^+ = 0$. We must have, then, from Equation (4.50) that $\exp[(E_d - E_F)/kT] = 0$. Since $T = 0$ K, this will occur for $\exp(-\infty) = 0$, which means that $E_F > E_d$. The Fermi energy level must be above the donor energy level at absolute zero. In the case of a p-type semiconductor at absolute zero temperature, the impurity atoms will not contain any electrons, so that the Fermi energy level must be below the acceptor energy state. The distribution of electrons among the various energy states, and hence the Fermi energy, is a function of temperature.

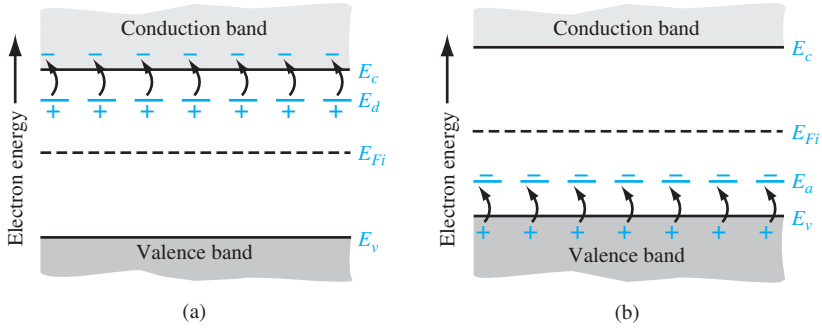


Figure 4.12 | Energy-band diagrams showing complete ionization of (a) donor states and (b) acceptor states.

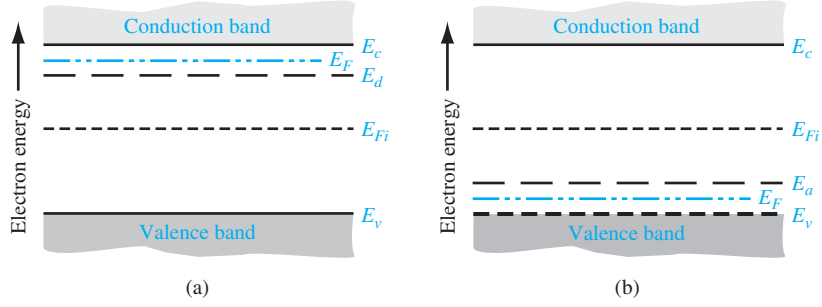


Figure 4.13 | Energy-band diagram at $T = 0$ K for (a) n-type and (b) p-type semiconductors.

A detailed analysis, not given in this text, shows that at $T = 0$ K, the Fermi energy is halfway between E_c and E_d for the n-type material and halfway between E_a and E_v for the p-type material. Figure 4.13 shows these effects. No electrons from the donor state are thermally elevated into the conduction band; this effect is called *freeze-out*. Similarly, when no electrons from the valance band are elevated into the acceptor states, the effect is also called freeze-out.

Between $T = 0$ K, freeze-out, and $T = 300$ K, complete ionization, we have partial ionization of donor or acceptor atoms.

EXAMPLE 4.8

Objective: Determine the temperature at which 90 percent of acceptor atoms are ionized.

Consider p-type silicon doped with boron at a concentration of $N_a = 10^{16}$ cm^{-3}.

■ **Solution**

Find the ratio of holes in the acceptor state to the total number of holes in the valence band plus acceptor state. Taking into account the Boltzmann approximation and assuming the degeneracy factor is $g = 4$, we write

$$\frac{p_a}{p_0 + p_a} = \frac{1}{1 + \dfrac{N_v}{4N_a} \cdot \exp\left[\dfrac{-(E_a - E_v)}{kT}\right]}$$

For 90 percent ionization,

$$\frac{p_a}{p_0 + p_a} = 0.10 = \frac{1}{1 + \dfrac{(1.04 \times 10^{19})\left(\dfrac{T}{300}\right)^{3/2}}{4(10^{16})} \cdot \exp\left[\dfrac{-0.045}{0.0259\left(\dfrac{T}{300}\right)}\right]}$$

Using trial and error, we find that $T = 193$ K.

■ Comment

This example shows that at approximately 100°C below room temperature, we still have 90 percent of the acceptor atoms ionized; in other words, 90 percent of the acceptor atoms have "donated" a hole to the valence band.

■ EXERCISE PROBLEM

Ex 4.8 Determine the fraction of total holes still in the acceptor states in silicon for $N_a = 10^{16}$ cm^{-3} at (*a*) $T = 250$ K and (*b*) $T = 200$ K.

[Ans. (*a*) 3.91 × 10^{-3}; (*b*) 8.73 × 10^{-2}]

TEST YOUR UNDERSTANDING

TYU 4.9 Determine the fraction of total holes still in the acceptor states in silicon at $T = 300$ K for a boron impurity concentration of $N_a = 10^{17}$ cm^{-3}. (Ans. 0.176)

TYU 4.10 Consider silicon with a phosphorus impurity concentration of $N_d = 10^{15}$ cm^{-3}. Determine the percent of ionized phosphorus atoms at (*a*) $T = 100$ K, (*b*) $T = 200$ K, (*c*) $T = 300$ K, and (*d*) $T = 400$ K.

[Ans. (*a*) 93.62%; (*b*) 99.82%; (*c*) 99.96%; (*d*) 99.98%]

4.5 | CHARGE NEUTRALITY

In thermal equilibrium, the semiconductor crystal is electrically neutral. The electrons are distributed among the various energy states, creating negative and positive charges, but the net charge density is zero. This charge-neutrality condition is used to determine the thermal-equilibrium electron and hole concentrations as a function of the impurity doping concentration. We will define a compensated semiconductor and then determine the electron and hole concentrations as a function of the donor and acceptor concentrations.

4.5.1 Compensated Semiconductors

A *compensated semiconductor* is one that contains both donor and acceptor impurity atoms in the same region. A compensated semiconductor can be formed, for example, by diffusing acceptor impurities into an n-type material or by diffusing donor impurities into a p-type material. An n-type compensated semiconductor occurs when $N_d > N_a$, and a p-type compensated semiconductor occurs when $N_a > N_d$. If $N_a = N_d$, we have a completely compensated semiconductor that has, as we will show, the characteristics of an intrinsic material. Compensated semiconductors are created quite naturally during device fabrication as we will see later.

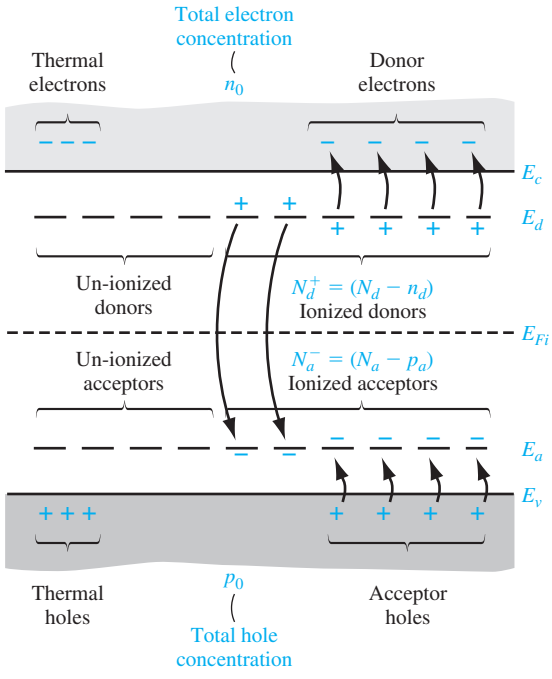


Figure 4.14 | Energy-band diagram of a compensated semiconductor showing ionized and un-ionized donors and acceptors.

4.5.2 Equilibrium Electron and Hole Concentrations

Figure 4.14 shows the energy-band diagram of a semiconductor when both donor and acceptor impurity atoms are added to the same region to form a compensated semiconductor. The figure shows how the electrons and holes can be distributed among the various states.

The charge neutrality condition is expressed by equating the density of negative charges to the density of positive charges. We then have

$$n_0 + N_a^- = p_0 + N_d^+ \tag{4.56}$$

or

$$n_0 + (N_a - p_a) = p_0 + (N_d - n_d) \tag{4.57}$$

where n_0 and p_0 are the thermal-equilibrium concentrations of electrons and holes in the conduction band and valence band, respectively. The parameter n_d is the concentration of electrons in the donor energy states, so $N_d^+ = N_d - n_d$ is the concentration of positively charged donor states. Similarly, p_a is the concentration of holes in the acceptor states, so $N_a^- = N_a - p_a$ is the concentration of negatively charged acceptor states. We have expressions for n_0, p_0, n_d, and p_a in terms of the Fermi energy and temperature.

Thermal-Equilibrium Electron Concentration If we assume complete ionization, n_d and p_a are both zero, and Equation (4.57) becomes

$$n_0 + N_a = p_0 + n_d \quad N_d \tag{4.58}$$

If we express p_0 as n_i^2/n_0, then Equation (4.58) can be written as

$$n_0 + N_a = \frac{n_i^2}{n_0} + N_d \tag{4.59a}$$

which in turn can be written as

$$n_0^2 - (N_d - N_a)n_0 - n_i^2 = 0 \tag{4.59b}$$

The electron concentration n_0 can be determined using the quadratic formula, or

$$n_0 = \frac{(N_d - N_a)}{2} + \sqrt{\left(\frac{N_d - N_a}{2}\right)^2 + n_i^2} \tag{4.60}$$

The positive sign in the quadratic formula must be used, since, in the limit of an intrinsic semiconductor when $N_a = N_d = 0$, the electron concentration must be a positive quantity, or $n_0 = n_i$.

Equation (4.60) is used to calculate the electron concentration in an n-type semiconductor, or when $N_d > N_a$. Although Equation (4.60) was derived for a compensated semiconductor, the equation is also valid for $N_a = 0$.

EXAMPLE 4.9

Objective: Determine the thermal-equilibrium electron and hole concentrations in silicon at $T = 300$ K for given doping concentrations. (*a*) Let $N_d = 10^{16}$ cm^{-3} and $N_a = 0$. (*b*) Let $N_d = 5 \times 10^{15}$ cm^{-3} and $N_a = 2 \times 10^{15}$ cm^{-3}.

Recall that $n_i = 1.5 \times 10^{10}$ cm^{-3} in silicon at $T = 300$ K.

■ **Solution**

(*a*) From Equation (4.60), the majority carrier electron concentration is

$$n_0 = \frac{10^{16}}{2} + \sqrt{\left(\frac{10^{16}}{2}\right)^2 + (1.5 \times 10^{10})^2} \cong 10^{16} \text{ cm}^{-3}$$

The minority carrier hole concentration is found to be

$$p_0 = \frac{n_i^2}{n_0} = \frac{(1.5 \times 10^{10})^2}{10^{16}} = 2.25 \times 10^4 \text{ cm}^{-3}$$

(*b*) Again, from Equation (4.60), the majority carrier electron concentration is

$$n_0 = \frac{5 \times 10^{15} - 2 \times 10^{15}}{2} + \sqrt{\left(\frac{5 \times 10^{15} - 2 \times 10^{15}}{2}\right)^2 + (1.5 \times 10^{10})^2} \cong 3 \times 10^{15} \text{ cm}^{-3}$$

The minority carrier hole concentration is

$$p_0 = \frac{n_i^2}{n_0} = \frac{(1.5 \times 10^{10})^2}{3 \times 10^{15}} = 7.5 \times 10^4 \text{ cm}^{-3}$$

Comment

In both parts of this example, $(N_d - N_a) \gg n_i$, so the thermal-equilibrium majority carrier electron concentration is essentially equal to the difference between the donor and acceptor concentrations. Also, in both cases, the majority carrier electron concentration is orders of magnitude larger than the minority carrier hole concentration.

EXERCISE PROBLEM

Ex 4.9 Find the thermal-equilibrium electron and hole concentrations in silicon with doping concentrations of $N_a = 7 \times 10^{15}$ cm^{-3} and $N_a = 3 \times 10^{15}$ cm^{-3} for (*a*) $T = 250$ K and (*b*) $T = 400$ K.

[Ans. (*a*) $n_0 = 4 \times 10^{15}$ cm^{-3}, $p_0 = 1.225$ cm^{-3}; (*b*) $n_0 = 4 \times 10^{15}$ cm^{-3}, $p_0 = 1.416 \times 10^9$ cm^{-3}]

We have argued in our discussion and we may note from the results of Example 4.9 that the concentration of electrons in the conduction band increases above the intrinsic carrier concentration as we add donor impurity atoms. At the same time, the minority carrier hole concentration decreases below the intrinsic carrier concentration as we add donor atoms. We must keep in mind that as we add donor impurity atoms and the corresponding donor electrons, there is a redistribution of electrons among available energy states. Figure 4.15 shows a schematic of this physical redistribution. A few of the donor electrons will fall into the empty states in the valence band and, in doing so, will annihilate some of the intrinsic holes. The minority carrier hole concentration will therefore decrease as we have seen in Example 4.9. At the

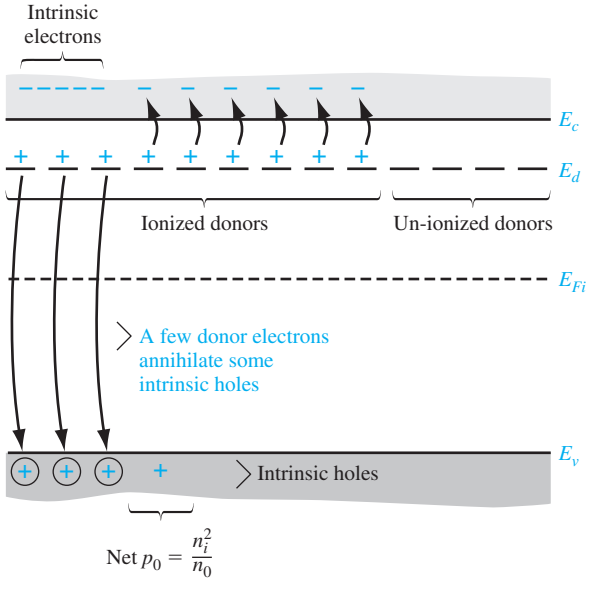


Figure 4.15 | Energy-band diagram showing the redistribution of electrons when donors are added.

same time, because of this redistribution, the net electron concentration in the conduction band is *not* simply equal to the donor concentration plus the intrinsic electron concentration.

EXAMPLE 4.10

Objective: Calculate the thermal-equilibrium electron and hole concentrations in germanium for a given doping concentration.

Consider a germanium sample at $T = 300$ K in which $N_d = 2 \times 10^{14}$ cm^{-3} and $N_a = 0$. Assume that $n_i = 2.4 \times 10^{13}$ cm^{-3}.

■ Solution

Again, from Equation (4.60), the majority carrier electron concentration is

$$n_0 = \frac{2 \times 10^{14}}{2} + \sqrt{\left(\frac{2 \times 10^{14}}{2}\right)^2 + (2.4 \times 10^{13})^2} \cong 2.028 \times 10^{14} \text{ cm}^{-3}$$

The minority carrier hole concentration is

$$p_0 = \frac{n_i^2}{n_0} = \frac{(2.4 \times 10^{13})^2}{2.028 \times 10^{14}} = 2.84 \times 10^{12} \text{ cm}^{-3}$$

■ Comment

If the donor impurity concentration is not too different in magnitude from the intrinsic carrier concentration, then the thermal-equilibrium majority carrier electron concentration is influenced by the intrinsic concentration.

■ EXERCISE PROBLEM

Ex 4.10 Repeat Example 4.10 for (*a*) $T = 250$ K and (*b*) $T = 350$ K. (*c*) What can be said about a very low-doped material as the temperature increases?

[Ans. (*a*) $n_0 \cong 2 \times 10^{14}$ cm^{-3}, $p_0 = 9.47 \times 10^9$ cm^{-3}; (*b*) $n_0 = 3.059 \times 10^{14}$ cm^{-3}, $p_0 = 1.059 \times 10^{14}$ cm^{-3}; (*c*) Material approaches an intrinsic semiconductor]

We have seen that the intrinsic carrier concentration n_i is a very strong function of temperature. As the temperature increases, additional electron–hole pairs are thermally generated so that the n_i^2 term in Equation (4.60) may begin to dominate. The semiconductor will eventually lose its extrinsic characteristics. Figure 4.16 shows the electron concentration versus temperature in silicon doped with 5×10^{14} donors per cm^3 . As the temperature increases, we can see where the intrinsic concentration begins to dominate. Also shown is the partial ionization, or the onset of freeze-out, at the low temperature.

Thermal-Equilibrium Hole Concentration If we reconsider Equation (4.58) and express n_0 as n_i^2/p_0, then we have

$$\frac{n_i^2}{p_0} + N_a = p_0 + N_d \tag{4.61a}$$

which we can write as

$$p_0^2 - (N_a - N_d)p_0 - n_i^2 = 0 \tag{4.61b}$$

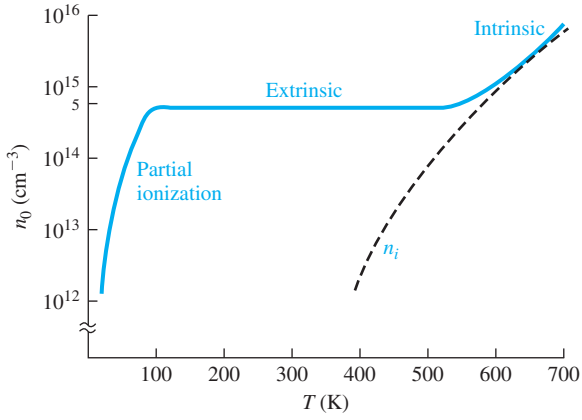


Figure 4.16 | Electron concentration versus temperature showing the three regions: partial ionization, extrinsic, and intrinsic.

Using the quadratic formula, the hole concentration is given by

$$p_0 = \frac{N_a - N_d}{2} + \sqrt{\left(\frac{N_a - N_d}{2}\right)^2 + n_i^2} \tag{4.62}$$

where the positive sign, again, must be used. Equation (4.62) is used to calculate the thermal-equilibrium majority carrier hole concentration in a p-type semiconductor, or when $N_a > N_d$. This equation also applies for $N_d = 0$.

EXAMPLE 4.11

Objective: Calculate the thermal-equilibrium electron and hole concentrations in a compensated p-type semiconductor.

Consider a silicon semiconductor at $T = 300$ K in which $N_a = 10^{16}$ cm^{-3} and $N_d = 3 \times 10^{15}$ cm^{-3}. Assume $n_i = 1.5 \times 10^{10}$ cm^{-3}.

■ Solution

Since $N_a > N_d$, the compensated semiconductor is p-type and the thermal-equilibrium majority carrier hole concentration is given by Equation (4.62) as

$$p_0 = \frac{10^{16} - 3 \times 10^{15}}{2} + \sqrt{\left(\frac{10^{16} - 3 \times 10^{15}}{2}\right)^2 + (1.5 \times 10^{10})^2}$$

so that

$$p_0 \approx 7 \times 10^{15} \text{ cm}^{-3}$$

The minority carrier electron concentration is

$$n_0 = \frac{n_i^2}{p_0} = \frac{(1.5 \times 10^{10})^2}{7 \times 10^{15}} = 3.21 \times 10^4 \text{ cm}^{-3}$$

■ Comment

If we assume complete ionization and if $(N_a - N_d) \gg n_i$, then the majority carrier hole concentration is, to a very good approximation, just the difference between the acceptor and donor concentrations.

EXERCISE PROBLEM

Ex 4.11 Consider silicon at $T = 300$ K. Calculate the thermal-equilibrium electron and hole concentrations for impurity concentrations of (*a*) $N_a = 4 \times 10^{16}$ cm^{-3}, $N_d = 8 \times 10^{15}$ cm^{-3} and (*b*) $N_a = N_d = 3 \times 10^{15}$ cm^{-3}.

[Ans. (*a*) $n_0 = 3.2 \times 10^{16}$ cm^{-3}, $p_0 = 7.03 \times 10^{3}$ cm^{-3}; (*b*) $n_0 = p_0 = 1.5 \times 10^{10}$ cm^{-3}]

We may note that, for a compensated p-type semiconductor, the minority carrier electron concentration is determined from

$$n_0 = \frac{n_i^2}{p_0} = \frac{n_i^2}{(N_a - N_d)}$$

Equations (4.60) and (4.62) are used to calculate the majority carrier electron concentration in an n-type semiconductor and majority carrier hole concentration in a p-type semiconductor, respectively. The minority carrier hole concentration in an n-type semiconductor could, theoretically, be calculated from Equation (4.62). However, we would be subtracting two numbers on the order of 10^{16} cm^{-3}, for example, to obtain a number on the order of 10^4 cm^{-3}, which from a practical point of view is not possible. The minority carrier concentrations are calculated from $n_0 p_0 = n_i^2$ once the majority carrier concentration has been determined.

TEST YOUR UNDERSTANDING

TYU 4.11 Consider a compensated GaAs semiconductor at $T = 300$ K doped at $N_d = 5 \times 10^{15}$ cm^{-3} and $N_a = 2 \times 10^{16}$ cm^{-3}. Calculate the thermal equilibrium electron and hole concentrations. (Ans. $p_0 = 1.5 \times 10^{16}$ cm^{-3}, $n_0 = 2.16 \times 10^{-4}$ cm^{-3})

TYU 4.12 Silicon is doped at $N_d = 10^{15}$ cm^{-3} and $N_a = 0$. (*a*) Plot the concentration of electrons versus temperature over the range $300 \leq T \leq 600$ K. (*b*) Calculate the temperature at which the electron concentration is equal to 1.1×10^{15} cm^{-3}. (Ans. $T \approx 552$ K)

TYU 4.13 A silicon device with n-type material is to be operated at $T = 550$ K. At this temperature, the intrinsic carrier concentration must contribute no more than 5 percent of the total electron concentration. Determine the minimum donor concentration required to meet this specification. (Ans. $N_D = 1.40 \times 10^{15}$ cm^{-3})

4.6 | POSITION OF FERMI ENERGY LEVEL

We have discussed qualitatively in Section 4.3.1 how the electron and hole concentrations change as the Fermi energy level moves through the bandgap energy. Then, in Section 4.5, we calculated the electron and hole concentrations as a function of donor and acceptor impurity concentrations. We can now determine the position of the Fermi energy level as a function of the doping concentrations and as a function of temperature. The relevance of the Fermi energy level will be further discussed after the mathematical derivations.

4.6.1 Mathematical Derivation

The position of the Fermi energy level within the bandgap can be determined by using the equations already developed for the thermal-equilibrium electron and hole concentrations. If we assume the Boltzmann approximation to be valid, then from Equation (4.11) we have $n_0 = N_c \exp[-(E_c - E_F)/kT]$. We can solve for $E_c - E_F$ from this equation and obtain

$$E_c - E_F = kT \ln\left(\frac{N_c}{n_0}\right) \tag{4.63}$$

where n_0 is given by Equation (4.60). If we consider an n-type semiconductor in which $N_d \gg n_i$, then $n_0 \approx N_d$, so that

$$E_c - E_F = kT \ln\left(\frac{N_c}{N_d}\right) \tag{4.64}$$

The distance between the bottom of the conduction band and the Fermi energy is a logarithmic function of the donor concentration. As the donor concentration increases, the Fermi level moves closer to the conduction band. Conversely, if the Fermi level moves closer to the conduction band, then the electron concentration in the conduction band is increasing. We may note that if we have a compensated semiconductor, then the N_d term in Equation (4.64) is simply replaced by $N_d - N_a$, or- the net effective donor concentration.

DESIGN EXAMPLE 4.12

Objective: Determine the required donor impurity concentration to obtain a specified Fermi energy.

Silicon 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.

■ Solution

From Equation (4.64), we have

$$E_c - E_F = kT \ln\left(\frac{N_c}{N_d - N_a}\right)$$

which can be rewritten as

$$N_d - N_a = N_c \exp\left[\frac{-(E_c - E_F)}{kT}\right]$$

Then

$$N_d - N_a = 2.8 \times 10^{19} \exp\left[\frac{-0.20}{0.0259}\right] = 1.24 \times 10^{16} \text{ cm}^{-3}$$

or

$$N_d = 1.24 \times 10^{16} + N_a = 2.24 \times 10^{16} \text{ cm}^{-3}$$

■ Comment

A compensated semiconductor can be fabricated to provide a specific Fermi energy level.

EXERCISE PROBLEM

Ex 4.12 Consider silicon at $T = 300$ K with doping concentrations of $N_d = 8 \times 10^{15}$ cm^{-3} and $N_a = 5 \times 10^{15}$ cm^{-3}. Determine the position of the Fermi energy level with respect to E_c. (Ans. $E_c - E_F = 0.2368$ eV)

We may develop a slightly different expression for the position of the Fermi level. We had from Equation (4.39) that $n_0 = n_i \exp[(E_F - E_{Fi})/kT]$. We can solve for $E_F - E_{Fi}$ as

$$E_F - E_{Fi} = kT \ln\left(\frac{n_0}{n_i}\right) \tag{4.65}$$

Equation (4.65) can be used specifically for an n-type semiconductor, where n_0 is given by Equation (4.60), to find the difference between the Fermi level and the intrinsic Fermi level as a function of the donor concentration. We may note that, if the net effective donor concentration is zero, that is, $N_d - N_a = 0$, then $n_0 = n_i$ and $E_F = E_{Fi}$. A completely compensated semiconductor has the characteristics of an intrinsic material in terms of carrier concentration and Fermi-level position.

We can derive the same types of equations for a p-type semiconductor. From Equation (4.19), we have $p_0 = N_v \exp[-(E_F - E_v)/kT]$, so that

$$E_F - E_v = kT \ln\left(\frac{N_v}{p_0}\right) \tag{4.66}$$

If we assume that $N_a \gg n_i$, then Equation (4.66) can be written as

$$E_F - E_v = kT \ln\left(\frac{N_v}{N_a}\right) \tag{4.67}$$

The distance between the Fermi level and the top of the valence-band energy for a p-type semiconductor is a logarithmic function of the acceptor concentration: as the acceptor concentration increases, the Fermi level moves closer to the valence band. Equation (4.67) still assumes that the Boltzmann approximation is valid. Again, if we have a compensated p-type semiconductor, then the N_a term in Equation (4.67) is replaced by $N_a - N_d$, or the net effective acceptor concentration.

We can also derive an expression for the relationship between the Fermi level and the intrinsic Fermi level in terms of the hole concentration. We have from Equation (4.40) that $p_0 = n_i \exp[-(E_F - E_{Fi})/kT]$, which yields

$$E_{Fi} - E_F = kT \ln\left(\frac{p_0}{n_i}\right) \tag{4.68}$$

Equation (4.68) can be used to find the difference between the intrinsic Fermi level and the Fermi energy in terms of the acceptor concentration. The hole concentration p_0 in Equation (4.68) is given by Equation (4.62).

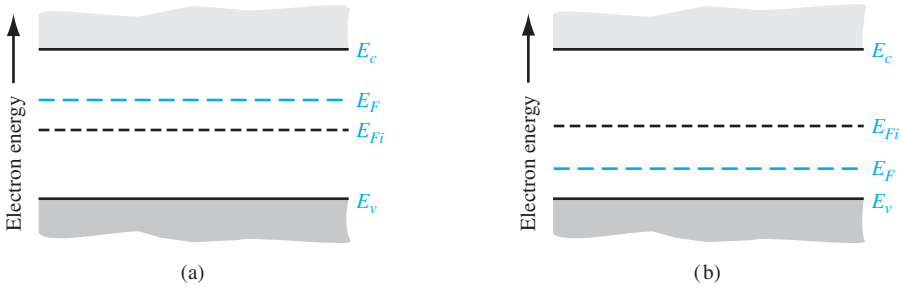


Figure 4.17 | Position of Fermi level for an (a) n-type ($N_d > N_a$) and (b) p-type ($N_d > N_a$) semiconductor.

We may again note from Equation (4.65) that, for an n-type semiconductor, $n_0 > n_i$ and $E_F > E_{Fi}$. The Fermi level for an n-type semiconductor is above E_{Fi}. For a p-type semiconductor, $p_0 > n_i$, and from Equation (4.68) we see that $E_{Fi} > E_F$. The Fermi level for a p-type semiconductor is below E_{Fi}. These results are shown in Figure 4.17.

4.6.2 Variation of E_F with Doping Concentration and Temperature

We may plot the position of the Fermi energy level as a function of the doping concentration. Figure 4.18 shows the Fermi energy level as a function of donor concentration (n type) and as a function of acceptor concentration (p type) for silicon at $T = 300$ K. As the doping levels increase, the Fermi energy level moves closer to the conduction band for the n-type material and closer to the valence band for the p-type material. Keep in mind that the equations for the Fermi energy level that we have derived assume that the Boltzmann approximation is valid.

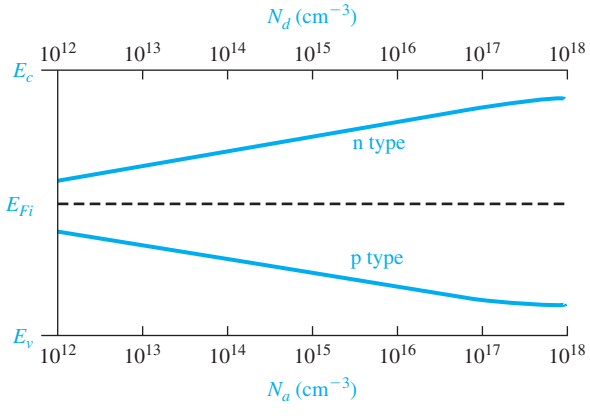


Figure 4.18 | Position of Fermi level as a function of donor concentration (n type) and acceptor concentration (p type).

EXAMPLE 4.13

Objective: Determine the Fermi energy level and the maximum doping concentration at which the Boltzmann approximation is still valid.

Consider p-type silicon, at $T = 300$ K, doped with boron. We may assume that the limit of the Boltzmann approximation occurs when $E_F - E_a = 3kT$. (See Section 4.1.2.)

■ Solution

From Table 4.3, we find the ionization energy is $E_a - E_v = 0.045$ eV for boron in silicon. If we assume that $E_{Fi} \approx E_{\text{midgap}}$, then from Equation (4.68), the position of the Fermi level at the maximum doping is given by

$$E_{Fi} - E_F = \frac{E_g}{2} - (E_a - E_v) - (E_F - E_a) = kT \ln\left(\frac{N_a}{n_i}\right)$$

or

$$0.56 - 0.045 - 3(0.0259) = 0.437 = (0.0259) \ln\left(\frac{N_a}{n_i}\right)$$

We can then solve for the doping as

$$N_a = n_i \exp\left(\frac{0.437}{0.0259}\right) = 3.2 \times 10^{17} \text{ cm}^{-3}$$

■ Comment

If the acceptor (or donor) concentration in silicon is greater than approximately 3×10^{17} cm^{-3}, then the Boltzmann approximation of the distribution function becomes less valid and the equations for the Fermi-level position are no longer quite as accurate.

■ EXERCISE PROBLEM

Ex 4.13 Consider n-type silicon at $T = 300$ K doped with arsenic. Determine the maximum doping at which the Boltzmann approximation is still valid. Assume the limit is such that $E_d - E_F = 3kT$. ($N_d = 2.02 \times 10^{17}$ cm^{-3})

The intrinsic carrier concentration n_i, in Equations (4.65) and (4.68), is a strong function of temperature, so that E_F is a function of temperature also. Figure 4.19 shows the variation of the Fermi energy level in silicon with temperature for several donor and acceptor concentrations. As the temperature increases, n_i increases, and E_F moves closer to the intrinsic Fermi level. At high temperature, the semiconductor material begins to lose its extrinsic characteristics and begins to behave more like an intrinsic semiconductor. At the very low temperature, freeze-out occurs; the Boltzmann approximation is no longer valid and the equations we derived for the Fermi-level position no longer apply. At the low temperature where freeze-out occurs, the Fermi level goes above E_d for the n-type material and below E_a for the p-type material. At absolute zero degrees, all energy states below E_F are full and all energy states above E_F are empty.

4.6.3 Relevance of the Fermi Energy

We have been calculating the position of the Fermi energy level as a function of doping concentrations and temperature. This analysis may seem somewhat arbitrary and fictitious. However, these relations do become significant later in our discussion of pn junctions and the other semiconductor devices we consider. An important

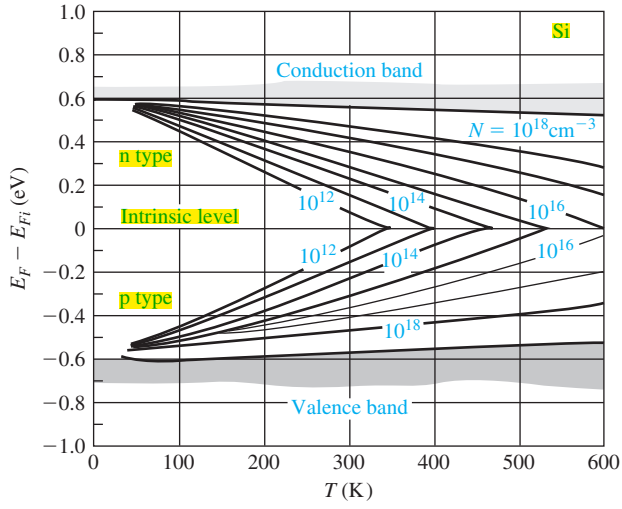


Figure 4.19 | Position of Fermi level as a function of temperature for various doping concentrations.
(From Sze [14].)

point is that, in thermal equilibrium, the Fermi energy level is a constant throughout a system. We will not prove this statement, but we can intuitively see its validity by considering the following example.

Suppose we have a particular material, A, whose electrons are distributed in the energy states of an allowed band as shown in Figure 4.20a. Most of the energy states below E_{FA} contain electrons and most of the energy states above E_{FA} are empty of electrons. Consider another material, B, whose electrons are distributed in the energy states of an allowed band as shown in Figure 4.20b. The energy states below E_{FB} are mostly full and the energy states above E_{FB} are mostly empty. If these two materials are brought into intimate contact, the electrons in the entire system will tend to seek the lowest possible energy. Electrons from material A will flow into the lower energy states of material B, as indicated in Figure 4.20c, until thermal equilibrium is reached. Thermal equilibrium occurs when the distribution of electrons, as a function of energy, is the same in the two materials. This equilibrium state occurs when the Fermi energy is the same in the two materials as shown in Figure 4.20d. The Fermi energy, important in the physics of the semiconductor, also provides a good pictorial representation of the characteristics of the semiconductor materials and devices.

TEST YOUR UNDERSTANDING

TYU 4.14 Determine the position of the Fermi level with respect to the valence-band energy in p-type GaAs at $T = 300$ K. The doping concentrations are $N_a = 5 \times 10^{16}$ cm^{-3} and $N_d = 4 \times 10^{15}$ cm^{-3}. (Ans. $E_F - E_a = 0.130$ eV)

TYU 4.15 Calculate the position of the Fermi energy level in n-type silicon at $T = 300$ K with respect to the intrinsic Fermi energy level. The doping concentrations are $N_d = 2 \times 10^{17}$ cm^{-3} and $N_a = 3 \times 10^{16}$ cm^{-3}. (Ans. $E_F - E_{Fi} = 0.421$ eV)

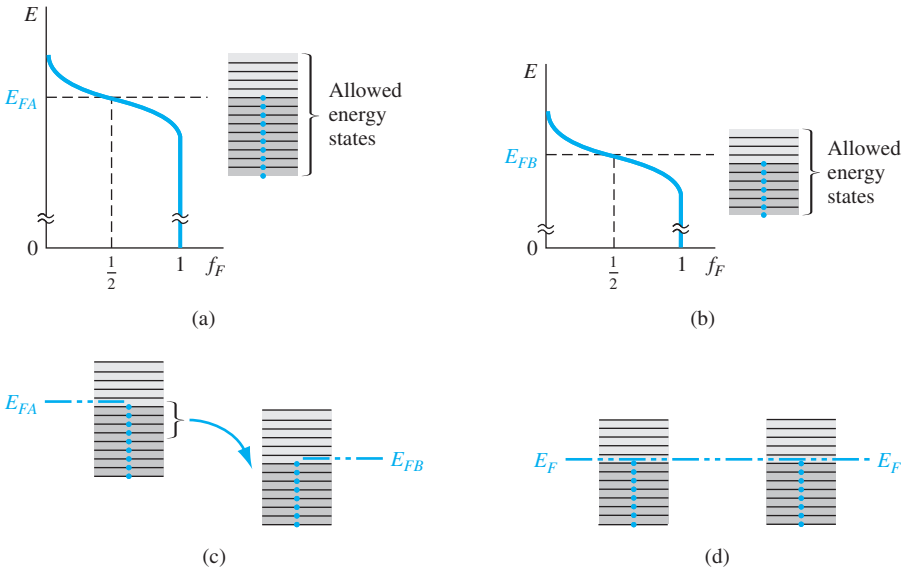


Figure 4.20 | The Fermi energy of (a) material A in thermal equilibrium, (b) material B in thermal equilibrium, (c) materials A and B at the instant they are placed in contact, and (d) materials A and B in contact at thermal equilibrium.

4.7 | SUMMARY

- The concentration of electrons in the conduction band is the integral over the conduction-band energy of the product of the density of states function in the conduction band and the Fermi–Dirac probability function.
- The concentration of holes in the valence band is the integral over the valence-band energy of the product of the density of states function in the valence band and the probability of a state being empty, which is $[1 - f_F(E)]$.
- Using the Maxwell–Boltzmann approximation, the thermal-equilibrium concentration of electrons in the conduction band is given by

$$n_0 = N_c \exp\left[\frac{-(E_c - E_F)}{kT}\right]$$

  where N_c is the effective density of states in the conduction band.
- Using the Maxwell–Boltzmann approximation, the thermal-equilibrium concentration of holes in the valence band is given by

$$p_0 = N_v \exp\left[\frac{-(E_F - E_v)}{kT}\right]$$

  where N_v is the effective density of states in the valence band.
- The intrinsic carrier concentration is found from

$$n_i^2 = N_c N_v \exp\left[\frac{-E_g}{kT}\right]$$

- The concept of doping the semiconductor with donor (group V elements) impurities and acceptor (group III elements) impurities to form n-type and p-type extrinsic semiconductors was discussed.
- The fundamental relationship of $n_0 p_0 = n_i^2$ was derived.
- Using the concepts of complete ionization and charge neutrality, equations for the electron and hole concentrations as a function of impurity doping concentrations were derived.
- The position of the Fermi energy level as a function of impurity doping concentrations was derived.
- The relevance of the Fermi energy was discussed. The Fermi energy is a constant throughout a semiconductor that is in thermal equilibrium.

GLOSSARY OF IMPORTANT TERMS

acceptor atoms Impurity atoms added to a semiconductor to create a p-type material.

charge carrier The electron and/or hole that moves inside the semiconductor and gives rise to electrical currents.

compensated semiconductor A semiconductor that contains both donors and acceptors in the same semiconductor region.

complete ionization The condition when all donor atoms are positively charged by giving up their donor electrons and all acceptor atoms are negatively charged by accepting electrons.

degenerate semiconductor A semiconductor whose electron concentration or hole concentration is greater than the effective density of states, so that the Fermi level is in the conduction band (n type) or in the valence band (p type).

donor atoms Impurity atoms added to a semiconductor to create an n-type material.

effective density of states The parameter N_c, which results from integrating the density of quantum states $g_c(E)$ times the Fermi function $f_F(E)$ over the conduction-band energy, and the parameter N_v, which results from integrating the density of quantum states $g_v(E)$ times $[1 - f_F(E)]$ over the valence-band energy.

extrinsic semiconductor A semiconductor in which controlled amounts of donors and/or acceptors have been added so that the electron and hole concentrations change from the intrinsic carrier concentration and a preponderance of either electrons (n type) or holes (p type) is created.

freeze-out The condition that occurs in a semiconductor when the temperature is lowered and the donors and acceptors become neutrally charged. The electron and hole concentrations become very small.

CHECKPOINT

After studying this chapter, the reader should have the ability to:

- Derive the equations for the thermal equilibrium concentrations of electrons and holes in terms of the Fermi energy.
- Derive the equation for the intrinsic carrier concentration.
- Disscuss what is meant by the effective density of states for electrons and holes.
- Describe the effect of adding donor and acceptor impurity atoms to a semiconductor.

- Understand the concept of complete ionization.
- Derive the fundamental relationship $n_0 p_0 = n_i^2$.
- Describe the meaning of degenerate and nondegenerate semiconductor materials.
- Discuss the concept of charge neutrality.
- Derive the equations for n_0 and p_0 in terms of impurity doping concentrations.
- Derive the equations for the Fermi energy in terms of the impurity doping concentrations.
- Discuss the variation of the Fermi energy with doping concentration and temperature.

REVIEW QUESTIONS

1. How does the electron concentration in the conduction band change with energy E for $E > E_c$?
2. In deriving the equation for n_0 in terms of the Fermi function, the upper limit of the integral should be the energy at the top of the conduction band. Justify using infinity instead.
3. Assuming the Boltzmann approximation applies, write the equations for n_0 and p_0 in terms of the Fermi energy.
4. What is the source of electrons and holes in an intrinsic semiconductor?
5. Under what condition would the intrinsic Fermi level be at the midgap energy?
6. What is a donor impurity? What is an acceptor impurity?
7. What is meant by complete ionization? What is meant by freeze-out?
8. What is the product of n_0 and p_0 equal to?
9. Write the equation for charge neutrality for the condition of complete ionization.
10. Sketch a graph of n_0 versus temperature for an n-type material.
11. Sketch graphs of the Fermi energy versus donor impurity concentration and versus temperature.
12. What is the relevance of the Fermi energy?

PROBLEMS

Section 4.1 Charge Carriers in Semiconductors

4.1 Calculate the intrinsic carrier concentration, n_i, at $T = 200$, 400, and 600 K for (*a*) silicon, (*b*) germanium, and (*c*) gallium arsenide.

4.2 Plot the intrinsic carrier concentration, n_i, for a temperature range of $200 \leq T \leq 600$ K for (*a*) silicon, (*b*) germanium, and (*c*) gallium arsenide. (Use a log scale for n_i.)

4.3 (*a*) The maximum intrinsic carrier concentration in a silicon device must be limited to 5×10^{11} cm^{-3}. Assume $E_g = 1.12$ eV. Determine the maximum temperature allowed for the device. (*b*) Repeat part (*a*) if the maximum intrinsic carrier concentration is limited to 5×10^{12} cm^{-3}.

4.4 In a particular semiconductor material, the effective density of states functions are given by $N_c = N_{c0} \cdot (T/300)^{3/2}$ and $N_v = N_{v0} \cdot (T/300)^{3/2}$ where N_{c0} and N_{v0} are constants independent of temperature. Experimentally determined intrinsic carrier concentrations are found to be $n_i = 1.40 \times 10^2$ cm^{-3} at $T = 200$ K and $n_i = 7.70 \times 10^{10}$ cm^{-2} at $T = 400$ K. Determine the product $N_{c0} \cdot N_{v0}$ and the bandgap energy E_g. (Assume E_g is constant over this temperature range.)

4.5 Two semiconductor materials have exactly the same properties except material A has a bandgap energy of 0.90 eV and material B has a bandgap energy of 1.10 eV. Determine the ratio of n_i of material B to that of material A for (*a*) $T = 200$ K, (*b*) $T = 300$ K, and (*c*) $T = 400$ K.

4.6 (*a*) The magnitude of the product $g_C(E)f_F(E)$ in the conduction band is a function of energy as shown in Figure 4.1. Assume the Boltzmann approximation is valid. Determine the energy with respect to E_c at which the maximum occurs. (*b*) Repeat part (*a*) for the magnitude of the product $g_v(E)[1 - f_F(E)]$ in the valence band.

4.7 Assume the Boltzmann approximation in a semiconductor is valid. Determine the ratio of $n(E) = g_C(E)f_F(E)$ at $E = E_c + 4kT$ to that at $E = E_c + kT/2$.

4.8 Assume that $E_c - E_F = 0.20$ eV in silicon. Plot $n(E) = g_C(E)f_F(E)$ over the range $E_c \leq E \leq E_c + 0.10$ eV for (*a*) $T = 200$ K and (*b*) $T = 400$ K.

4.9 (a) Consider silicon at $T = 300$ K. Plot the thermal equilibrium electron concentration n_0 (on a log scale) over the energy range $0.2 \leq E_c - E_F \leq 0.4$ eV. (*b*) Repeat part (*a*) for the hole concentration over the range $0.2 \leq E_F - E_v \leq 0.4$ eV.

4.10 Given the effective masses of electrons and holes in silicon, germanium, and gallium arsenide, calculate the position of the intrinsic Fermi energy level with respect to the center of the bandgap for each semiconductor at $T = 300$ K.

4.11 Calculate E_{Fi} with respect to the center of the bandgap in silicon for $T = 200$, 400, and 600 K.

4.12 (*a*) The carrier effective masses in a semiconductor are $m_n^* = 1.21\, m_0$ and $m_p^* = 0.70\, m_0$. Determine the position of the intrinsic Fermi level with respect to the center of the bandgap at $T = 300$ K. (*b*) Repeat part (*a*) if $m_n^* = 0.080\, m_0$ and $m_p^* = 0.75\, m_0$.

4.13 If the density of states function in the conduction band of a particular semiconductor is a constant equal to K, derive the expression for the thermal-equilibrium concentration of electrons in the conduction band, assuming Fermi–Dirac statistics and assuming the Boltzmann approximation is valid.

4.14 Repeat Problem 4.13 if the density of states function is given by $g_c(E) = C_1(E - E_c)$ for $E \geq E_c$ where C_1 is a constant.

Section 4.2 Dopant Atoms and Energy Levels

4.15 Calculate the ionization energy and radius of the donor electron in germanium using the Bohr theory.

4.16 Repeat Problem 4.15 for gallium arsenide.

Section 4.3 The Extrinsic Semiconductor

4.17 Silicon at $T = 300$ K is doped with arsenic atoms such that the concentration of electrons is $n_0 = 7 \times 10^{15}$ cm^{-3}. (*a*) Find $E_c - E_F$. (*b*) Determine $E_F - E_v$. (*c*) Calculate p_0. (*d*) Which carrier is the minority carrier? (*e*) Find $E_F - E_{Fi}$.

4.18 The value of p_0 in silicon at $T = 300$ K is 2×10^{16} cm^{-3}. (a) Determine $E_F - E_v$. (*b*) Calculate the value of $E_c - E_F$. (*c*) What is the value of n_0? (*d*) Determine $E_{Fi} - E_F$.

4.19 The electron concentration in silicon at $T = 300$ K is $n_0 = 2 \times 10^5$ cm^{-3}. (*a*) Determine the position of the Fermi level with respect to the valence band energy level. (*b*) Determine p_0. (*c*) Is this n- or p-type material?

4.20 (*a*) If $E_c - E_F = 0.28$ eV in gallium arsenide at $T = 375$ K, calculate the values of n_0 and p_0. (*b*) Assuming the value of n_0 in part (*a*) remains constant, determine $E_c - E_F$ and p_0 at $T = 300$ K.

4.21 Repeat Problem 4.20 for silicon.

4.22 The Fermi energy level in silicon at $T = 300$ K is as close to the top of the valence band as to the midgap energy. (*a*) Is the material n type or p type? (*b*) Calculate the values of n_0 and p_0.

4.23 (*a*) The Fermi energy level in silicon at $T = 300$ K is 0.22 eV above the intrinsic Fermi level. Determine n_0 and p_0. (*b*) Repeat part (*a*) for GaAs.

4.24 Silicon at $T = 300$ K is doped with boron atoms such that the concentration of holes is $p_0 = 5 \times 10^{15}$ cm^{-3}. (*a*) Find $E_F - E_v$. (*b*) Determine $E_c - E_F$. (*c*) Determine n_0. (*d*) Which carrier is the majority carrier? (*e*) Determine $E_{Fi} - E_F$.

4.25 Repeat Problem 4.24 for $T = 400$ K, assuming the hole concentration remains constant.

4.26 (*a*) Determine the values of n_0 and p_0 in GaAs at $T = 300$ K if $E_F - E_v = 0.25$ eV. (*b*) Assuming the value of p_0 in part (*a*) remains constant, determine the values of $E_F - E_v$ and n_0 at $T = 400$ K.

4.27 Repeat Problem 4.26 for silicon.

4.28 (*a*) Assume that $E_F = E_c + kT/2$ at $T = 300$ K in silicon. Determine n_0. (*b*) Repeat part (*a*) for GaAs.

4.29 Consider silicon at $T = 300$ K in which the hole concentration is $p_0 = 5 \times 10^{19}$ cm^{-3}. Determine $E_v - E_F$.

4.30 (*a*) In silicon at $T = 300$ K, we find that $E_F - E_C = 4\,kT$. Determine the electron concentration. (*b*) Repeat part (*a*) for GaAs.

Section 4.4 Statistics of Donors and Acceptors

***4.31** The electron and hole concentrations as a function of energy in the conduction band and valence band peak at a particular energy as shown in Figure 4.8. Consider silicon and assume $E_c - E_F = 0.20$ eV. Determine the energy, relative to the band edges, at which the concentrations peak.

***4.32** For the Boltzmann approximation to be valid for a semiconductor, the Fermi level must be at least 3 kT below the donor level in an n-type material and at least 3 kT above the acceptor level in a p-type material. If $T = 300$ K, determine the maximum electron concentration in an n-type semiconductor and the maximum hole concentration in a p-type semiconductor for the Boltzmann approximation to be valid in (*a*) silicon and (*b*) gallium arsenide.

4.33 Plot the ratio of un-ionized donor atoms to the total electron concentration versus temperature for silicon over the range $50 \leq T \leq 200$ K.

Section 4.5 Charge Neutrality

4.34 Determine the equilibrium electron and hole concentrations in silicon for the following conditions:

(*a*) $T = 300$ K, $N_d = 10^{15}$ cm^{-3}, $N_a = 4 \times 10^{15}$ cm^{-3}

(*b*) $T = 300$ K, $N_d = 3 \times 10^{16}$ cm^{-3}, $N_a = 0$

*Asterisks next to problems indicate problems that are more difficult.

(*c*) $T = 300$ K, $N_d = N_a = 2 \times 10^{15}$ cm^{-3}
(*d*) $T = 375$ K, $N_d = 0$, $N_a = 4 \times 10^{15}$ cm^{-3}
(*e*) $T = 450$ K, $N_d = 10^{14}$ cm^{-3}, $N_a = 0$

4.35 Repeat Problem 4.34 for GaAs.

4.36 (*a*) Consider a germanium semiconductor at $T = 300$ K. Calculate the thermal equilibrium electron and hole concentrations for (i) $N_d = 2 \times 10^{15}$ cm^{-3}, $N_a = 0$, and (ii) $N_a = 10^{16}$ cm^{-3}, $N_d = 7 \times 10^{15}$ cm^{-3}. (*b*) Repeat part (*a*) for GaAs. (*c*) For the case of GaAs in part (*b*), the minority carrier concentrations are on the order of 10^{-3} cm^{-3}. What does this result mean physically?

***4.37** The Fermi level in n-type silicon at $T = 300$ K is 245 meV below the conduction band and 200 meV below the donor level. Determine the probability of finding an electron (*a*) in the donor level and (*b*) in a state in the conduction-band kT above the conduction-band edge.

4.38 Assume that silicon, germanium, and gallium arsenide each have dopant concentrastions of $N_d = 1 \times 10^{13}$ cm^{-3} and $N_a = 2.5 \times 10^{13}$ cm^{-3} at $T = 300$ K. For each of the three materials: (*a*) Is this material n type or p type? (*b*) Calculate n_0 and p_0.

4.39 A silicon semiconductor material at $T = 300$ K is doped with arsenic atoms to a concentration of 2×10^{15} cm^{-3} and with boron atoms to a concentration of 1.2×10^{15} cm^{-3}. (*a*) Is the material n type or p type? (*b*) Determine n_0 and p_0. (*c*) Additional boron atoms are to be added such that the hole concentration is 4×10^{15} cm^{-3}. What concentration of boron atoms must be added and what is the new value of n_0?

4.40 The thermal equilibrium hole concentration in silicon at $T = 300$ K is $p_0 = 2 \times 10^5$ cm^{-3}. Determine the thermal-equilibrium electron concentration. Is the material n type or p type?

4.41 In a germanium sample at $T = 250$ K, it is found that $p_0 = 4n_0$ and that $N_d = 0$. Determine p_0, n_0, and N_a.

4.42 Consider a sample of silicon doped at $N_d = 0$ and $N_a = 10^{14}$ cm^{-3}. Plot the majority carrier concentration versus temperature over the range $200 \leq T \leq 500$ K.

4.43 The temperature of a sample of silicon is $T = 300$ K and the acceptor doping concentration is $N_a = 0$. Plot the minority carrier concentration (on a log–log plot) versus N_d over the range $10^{15} \leq N_d \leq 10^{18}$ cm^{-3}.

4.44 Repeat problem 4.43 for GaAs.

4.45 A particular semiconductor material is doped at $N_d = 2 \times 10^{14}$ cm^{-3} and $N_a = 1.2 \times 10^{14}$ cm^{-3}. The thermal equilibrium electron concentration is found to be $n_0 = 1.1 \times 10^{14}$ cm^{-3}. Assuming complete ionization, determine the intrinsic carrier concentration and the thermal equilibrium hole concentration.

4.46 (*a*) Silicon at $T = 300$ K is uniformly doped with boron atoms to a concentration of 3×10^{16} cm^{-3} and with arsenic atoms to a concentration of 1.5×10^{16} cm^{-3}. Is the material n type or p type? Calculate the thermal equilibrium concentrations of majority and minority carriers. (*b*) Additional impurity atoms are added such that holes are the majority carrier and the thermal equilibrium concentration is $p_0 = 5 \times 10^{16}$ cm^{-3}. What type and concentration of impurity atoms must be added? What is the new value of n_0?

4.47 In silicon at $T = 300$ K, it is found that $N_a = 7 \times 10^{15}$ cm^{-3} and $p_0 = 2 \times 10^4$ cm^{-3}. (*a*) Is the material n type or p type? (*b*) What are the majority and minority carrier concentrations? (*c*) What must be the concentration of donor impurities?

Section 4.6 Position of Fermi Energy Level

4.48 Consider germanium with an acceptor concentration of $N_a = 10^{15}$ cm^{-3} and a donor concentration of $N_d = 0$. Consider temperatures of $T = 200$, 400, and 600 K. Calculate the position of the Fermi energy with respect to the intrinsic Fermi level at these temperatures.

4.49 Consider silicon at $T = 300$ K with donor concentrations of $N_d = 10^{14}$, 10^{15}, 10^{16}, and 10^{17}, cm^{-3}. Assume $N_a = 0$. (*a*) Calculate the position of the Fermi energy level with respect to the conduction band for these donor concentrations. (*b*) Determine the position of the Fermi energy level with respect to the intrinsic Fermi energy level for the donor concentrations given in part (*a*).

4.50 A silicon device is doped with donor impurity atoms at a concentration of 10^{15} cm^{-3}. For the device to operate properly, the intrinsic carriers must contribute no more than 5 percent to the total electron concentration. (*a*) What is the maximum temperature that the device may operate? (*b*) What is the change in $E_c - E_F$ from the $T = 300$ K value to the maximum temperature value determined in part (*a*). (*c*) Is the Fermi level closer or further from the intrinsic value at the higher temperature?

4.51 Silicon is doped with acceptor impurity atoms at a concentration of $N_a = 3 \times 10^{15}$ cm^{-3}. Assume $N_d = 0$. Plot the position of the Fermi energy level with respect to the intrinsic Fermi energy level over the temperature range of $200 \leq T \leq 600$ K.

4.52 Consider GaAs at $T = 300$ K with $N_d = 0$. (*a*) Plot the position of the Fermi energy level with respect to the intrinsic Fermi energy level as a function of the acceptor impurity concentration over the range of $10^{14} \leq N_a \leq 10^{17}$ cm^{-3}. (*b*) Plot the position of the Fermi energy level with respect to the valence-band energy over the same acceptor impurity concentration as given in part (*a*).

4.53 For a particular semiconductor, $E_g = 1.50$ eV, $m_p^* = 10\, m_n^*$, $T = 300$ K, and $n_i = 1 \times 10^5$ cm^{-3}. (*a*) Determine the position of the intrinsic Fermi energy level with respect to the center of the bandgap. (*b*) Impurity atoms are added so that the Fermi energy level is 0.45 eV below the center of the bandgap. (*i*) Are acceptor or donor atoms added? (*ii*) What is the concentration of impurity atoms added?

4.54 Silicon at $T = 300$ K contains acceptor atoms at a concentration of $N_a = 5 \times 10^{15}$ cm^{-3}. Donor atoms are added forming an n-type compensated semiconductor such that the Fermi level is 0.215 eV below the conduction-band edge. What concentration of donor atoms are added?

4.55 (*a*) Silicon at $T = 300$ K is doped with donor impurity atoms at a concentration of $N_d = 6 \times 10^{15}$ cm^{-3}. (*i*) Determine $E_c - E_F$. (*ii*) Calculate the concentration of additional donor impurity atoms that must be added to move the Fermi energy level a distance kT closer to the conduction band edge. (*b*) Repeat part (*a*) for GaAs if the original donor impurity concentration is $N_d = 1 \times 10^{15}$ cm^{-3}.

4.56 (*a*) Determine the position of the Fermi energy level with respect to the intrinsic Fermi level in silicon at $T = 300$ K that is doped with boron atoms at a concentration of $N_a = 2 \times 10^{16}$ cm^{-3}. (*b*) Repeat part (*a*) if the silicon is doped with phosphorus atoms at a concentration of $N_d = 2 \times 10^{16}$ cm^{-3}. (*c*) Calculate n_0 and p_0 in parts (*a*) and (*b*).

4.57 GaAs at $T = 300$ K is doped with donor impurity atoms at a concentration of 7×10^{15} cm^{-3}. Additional impurity atoms are to be added such that the Fermi level is 0.55 eV above the intrinsic Fermi level. Determine the type (donor or acceptor) and concentration of impurity atoms to be added.

4.58 Determine the Fermi energy level with respect to the intrinsic Fermi level for each condition given in Problem 4.34.

4.59 Find the Fermi energy level with respect to the valence-band energy for the conditions given in Problem 4.35.

4.60 Calculate the position of the Fermi energy level with respect to the intrinsic Fermi for the conditions given in Problem 4.47.

Summary and Review

4.61 A new semiconductor material is to be "designed." The semiconductor is to be p type and doped with 5×10^{15} cm^{-3} acceptor atoms. Assume complete ionization and assume $N_d = 0$. The effective density of states functions are $N_c = 1.2 \times 10^{19}$ cm^{-3} and $N_v = 1.8 \times 10^{19}$ cm^{-3} at $T = 300$ K and vary as T^2. A special semiconductor device fabricated with this material requires that the hole concentration be no greater than 5.08×10^{15} cm^{-3} at $T = 350$ K. What is the minimum bandgap energy required in this new material?

4.62 Silicon atoms, at a concentration of 7×10^{15} cm^{-3}, are added to gallium arsenide. Assume that the silicon atoms act as fully ionized dopant atoms and that 5 percent of the concentration added replace gallium atoms and 95 percent replace arsenic atoms. Let $T = 300$ K. (*a*) Determine the donor and acceptor concentrations. (*b*) Is the material n type or p type? (*c*) Calculate the electron and hole concentrations. (*d*) Determine the position of the Fermi level with respect to E_{Fi}.

4.63 Defects in a semiconductor material introduce allowed energy states within the forbidden bandgap. Assume that a particular defect in silicon introduces two discrete levels: a donor level 0.25 eV above the top of the valence band and an acceptor level 0.65 eV above the top of the valence band. The charge state of each defect is a function of the position of the Fermi level. (*a*) Sketch the charge density of each defect as the Fermi level moves from E_v to E_c. Which defect level dominates in (*i*) heavily doped n type material and (*ii*) in heavily doped p-type material? (*b*) Determine the electron and hole concentrations and the location of the Fermi level in (*i*) an n type sample doped at $N_d = 10^{17}$ cm^{-3} and (*ii*) in a p-type sample doped at $N_a = 10^{17}$ cm^{-3} . (*c*) Determine the Fermi-level position if no dopant atoms are added. Is the material n type, p type, or intrinsic?

READING LIST

***1.** Hess, K. *Advanced Theory of Semiconductor Devices.* Englewood Cliffs, NJ: Prentice Hall, 1988.

2. Hu, C. C. *Modern Semiconductor Devices for Integrated Circuits.* Upper Saddle River, NJ: Pearson Prentice Hall, 2010.

3. Kano, K. *Semiconductor Devices.* Upper Saddle River, NJ: Prentice Hall, 1998.

***4.** Li, S. S. *Semiconductor Physical Electronics.* New York: Plenum Press, 1993.

5. McKelvey, J. P. *Solid State Physics for Engineering and Materials Science.* Malabar, FL.: Krieger Publishing, 1993.

6. Navon, D. H. *Semiconductor Microdevices and Materials.* New York: Holt, Rinehart & Winston, 1986.

7. Pierret, R. F. *Semiconductor Device Fundamentals.* Reading, MA: Addison-Wesley, 1996.

8. Shur, M. *Introduction to Electronic Devices.* New York: John Wiley and Sons, 1996.

***9.** Shur, M. *Physics of Semiconductor Devices.* Englewood Cliffs, NJ: Prentice Hall, 1990.

10. Singh, J. *Semiconductor Devices: An Introduction.* New York: McGraw-Hill, 1994.

11. Singh, J. *Semiconductor Devices: Basic Principles.* New York: John Wiley and Sons, 2001.

***12.** Smith, R. A. *Semiconductors.* 2nd ed. New York; Cambridge University Press, 1978.

13. Streetman, B. G., and S. Banerjee. *Solid State Electronic Devices,* 6th ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2006.

14. Sze, S. M., and K. K. Ng. *Physics of Semiconductor Devices.* 3rd ed. Hoboken, NJ: John Wiley and Sons, 2007.

***15.** Wang, S. *Fundamentals of Semiconductor Theory and Device Physics.* Englewood Cliffs, NJ: Prentice Hall, 1989.

***16.** Wolfe, C. M., N. Holonyak, Jr., and G. E. Stillman. *Physical Properties of Semiconductors.* Englewood Cliffs, NJ: Prentice Hall, 1989.

17. Yang, E. S. *Microelectronic Devices.* New York: McGraw-Hill, 1988.

*Indicates references that are at an advanced level compared to this text.

CHAPTER 5

Carrier Transport Phenomena

In the previous chapter, we considered the semiconductor in equilibrium and determined electron and hole concentrations in the conduction and valence bands, respectively. A knowledge of the densities of these charged particles is important toward an understanding of the electrical properties of a semiconductor material. The net flow of the electrons and holes in a semiconductor will generate currents. The process by which these charged particles move is called *transport*. In this chapter we consider two basic transport mechanisms in a semiconductor crystal: drift—the movement of charge due to electric fields, and diffusion—the flow of charge due to density gradients.

The carrier transport phenomena are the foundation for finally determining the current–voltage characteristics of semiconductor devices. We will implicitly assume in this chapter that, although there will be a net flow of electrons and holes due to the transport processes, thermal equilibrium will not be substantially disturbed. Nonequilibrium processes are considered in the next chapter. ■

5.0 | PREVIEW

In this chapter, we will:

- Describe the mechanism of carrier drift and induced drift current due to an applied electric field.
- Define and describe the characteristics of carrier mobility.
- Describe the mechanism of carrier diffusion and induced diffusion current due to a gradient in the carrier concentration.
- Define the carrier diffusion coefficient.
- Describe the effects of a nonuniform impurity doping concentration in a semiconductor material.
- Discuss and analyze the Hall effect in a semiconductor material.

5.1 | CARRIER DRIFT

An electric field applied to a semiconductor will produce a force on electrons and holes so that they will experience a net acceleration and net movement, provided there are available energy states in the conduction and valence bands. This net movement of charge due to an electric field is called *drift.* The net drift of charge gives rise to a *drift current.*

5.1.1 Drift Current Density

If we have a positive volume charge density ρ moving at an average drift velocity v_d, the drift current density is given by

$$J_{drf} = \rho v_d \tag{5.1a}$$

In terms of units, we have

$$J_{drf} = \left(\frac{Coul}{cm^3}\right) \cdot \left(\frac{cm}{s}\right) = \frac{Coul}{cm^2 - s} = \frac{A}{cm^2} \tag{5.1b}$$

If the volume charge density is due to positively charged holes, then

$$J_{p|drf} = (ep)v_{dp} \tag{5.2}$$

where $J_{p|drf}$ is the drift current density due to holes and v_{dp} is the average drift velocity of the holes.

The equation of motion of a positively charged hole in the presence of an electric field is

$$F = m_{ep}^* a = e\mathrm{E} \tag{5.3}$$

where e is the magnitude of the electronic charge, a is the acceleration, E is the electric field, and m_{cp}^* is the conductivity effective mass of the hole.[1] If the electric field is constant, then we expect the velocity to increase linearly with time. However, charged particles in a semiconductor are involved in collisions with ionized impurity atoms and with thermally vibrating lattice atoms. These collisions, or scattering events, alter the velocity characteristics of the particle.

As the hole accelerates in a crystal due to the electric field, the velocity increases. When the charged particle collides with an atom in the crystal, for example, the particle loses most, or all, of its energy. The particle will again begin to accelerate and gain energy until it is again involved in a scattering process. This continues over and over again. Throughout this process, the particle will gain an average drift velocity which, for low electric fields, is directly proportional to the electric field. We may then write

$$v_{dp} = \mu_p \mathrm{E} \tag{5.4}$$

where μ_p is the proportionality factor and is called the *hole mobility.* The mobility is an important parameter of the semiconductor since it describes how well a particle

[1]The conductivity effective mass is used when carriers are in motion. See Appendix F for further discussion of effective mass concepts.

Table 5.1 | Typical mobility values at $T = 300$ K and low doping concentrations

	μ_n (cm²/V-s)	μ_p (cm²/V-s)
Silicon	1350	480
Gallium arsenide	8500	400
Germanium	3900	1900

will move due to an electric field. The unit of mobility is usually expressed in terms of cm²/V-s.

By combining Equations (5.2) and (5.4), we may write the drift current density due to holes as

$$J_{p|drf} = (ep)v_{dp} = e\mu_p p\mathrm{E} \tag{5.5}$$

The drift current due to holes is in the same direction as the applied electric field.

The same discussion of drift applies to electrons. We may write

$$J_{n|drf} = \rho v_{dn} = (-en)v_{dn} \tag{5.6}$$

where $J_{n|drf}$ is the drift current density due to electrons and v_{dn} is the average drift velocity of electrons. The net charge density of electrons is negative.

The average drift velocity of an electron is also proportional to the electric field for small fields. However, since the electron is negatively charged, the net motion of the electron is opposite to the electric field direction. We can then write

$$v_{dn} = -\mu_n \mathrm{E} \tag{5.7}$$

where μ_n is the *electron mobility* and is a positive quantity. Equation (5.6) may now be written as

$$J_{n|drf} = (-en)(-\mu_n \mathrm{E}) = e\mu_n n\mathrm{E} \tag{5.8}$$

The conventional drift current due to electrons is also in the same direction as the applied electric field even though the electron movement is in the opposite direction.

Electron and hole mobilities are functions of temperature and doping concentrations, as we will see in the next section. Table 5.1 shows some typical mobility values at $T = 300$ K for low doping concentrations.

Since both electrons and holes contribute to the drift current, the total *drift current density* is the sum of the individual electron and hole drift current densities, so we may write

$$\boxed{J_{drf} = e(\mu_n n + \mu_p p)\mathrm{E}} \tag{5.9}$$

EXAMPLE 5.1

Objective: Calculate the drift current density in a semiconductor for a given electric field.

Consider a gallium arsenide sample at $T = 300$ K with doping concentrations of $N_a = 0$ and $N_d = 10^{16}$ cm^{-3}. Assume complete ionization and assume electron and hole mobilities given in Table 5.1. Calculate the drift current density if the applied electric field is E = 10 V/cm.

■ **Solution**

Since $N_d > N_a$, the semiconductor is n type and the majority carrier electron concentration, from Chapter 4 is given by

$$n = \frac{N_d - N_a}{2} + \sqrt{\left(\frac{N_d - N_a}{2}\right)^2 + n_i^2} \approx 10^{16}\ \text{cm}^{-3}$$

The minority carrier hole concentration is

$$p = \frac{n_i^2}{n} = \frac{(1.8 \times 10^6)^2}{10^{16}} = 3.24 \times 10^{-4}\ \text{cm}^{-3}$$

For this extrinsic n-type semiconductor, the drift current density is

$$J_{drf} = e(\mu_n n + \mu_p p)\text{E} \approx e\mu_n N_d \text{E}$$

Then

$$J_{drf} = (1.6 \times 10^{-19})(8500)(10^{16})(10) = 136\ \text{A/cm}^2$$

■ **Comment**

Significant drift current densities can be obtained in a semiconductor applying relatively small electric fields. We may note from this example that the drift current will usually be due primarily to the majority carrier in an extrinsic semiconductor.

■ **EXERCISE PROBLEM**

Ex 5.1 A drift current density of $J_{drf} = 75\ \text{A/cm}^2$ is required in a device using p-type silicon when an electric field of $\text{E} = 120\ \text{V/cm}$ is applied. Determine the required impurity doping concentration to achieve this specification. Assume that electron and hole mobilities given in Table 5.1 apply. ($N_a = 8.14 \times 10^{15}\ \text{cm}^{-3}$)

5.1.2 Mobility Effects

In the previous section, we defined mobility, which relates the average drift velocity of a carrier to the electric field. Electron and hole mobilities are important semiconductor parameters in the characterization of carrier drift, as seen in Equation (5.9).

Equation (5.3) related the acceleration of a hole to a force such as an electric field. We may write this equation as

$$F = m_{cp}^* \frac{dv}{dt} = e\text{E} \tag{5.10}$$

where v is the velocity of the particle due to the electric field and does not include the random thermal velocity. If we assume that the conductivity effective mass and electric field are constants, then we may integrate Equation (5.10) and obtain

$$v = \frac{e\text{E}t}{m_{cp}^*} \tag{5.11}$$

where we have assumed the initial drift velocity to be zero.

Figure 5.1a shows a schematic model of the random thermal velocity and motion of a hole in a semiconductor with zero electric field. There is a mean time

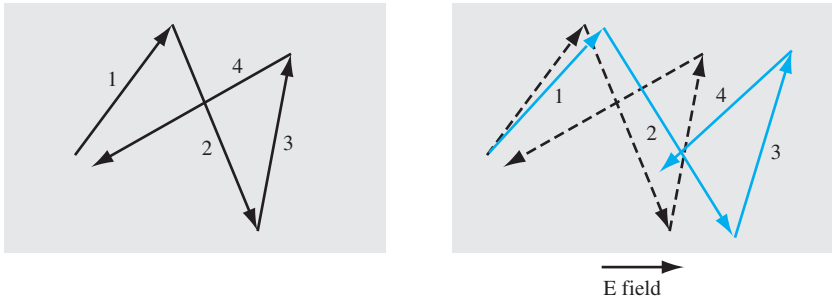


Figure 5.1 | Typical random behavior of a hole in a semiconductor (a) without an electric field and (b) with an electric field.

between collisions which may be denoted by τ_{cp}. If a small electric field (E-field) is applied as indicated in Figure 5.1b, there will be a net drift of the hole in the direction of the E-field, and the net drift velocity will be a small perturbation on the random thermal velocity, so the time between collisions will not be altered appreciably. If we use the mean time between collisions τ_{cp} in place of the time t in Equation (5.11), then the mean peak velocity just prior to a collision or scattering event is

$$v_{d|\text{peak}} = \left(\frac{e\tau_{cp}}{m_{cp}^*}\right)\mathrm{E} \tag{5.12a}$$

The average drift velocity is one half the peak value so that we can write

$$\langle v_d \rangle = \frac{1}{2}\left(\frac{e\tau_{cp}}{m_{cp}^*}\right)\mathrm{E} \tag{5.12b}$$

However, the collision process is not as simple as this model, but is statistical in nature. In a more accurate model including the effect of a statistical distribution, the factor $\frac{1}{2}$ in Equation (5.12b) does not appear. The hole mobility is then given by

$$\mu_p = \frac{v_{dp}}{\mathrm{E}} = \frac{e\tau_{cp}}{m_{cp}^*} \tag{5.13}$$

The same analysis applies to electrons; thus, we can write the electron mobility as

$$\mu_n = \frac{e\tau_{cn}}{m_{cn}^*} \tag{5.14}$$

where τ_{cn} is the mean time between collisions for an electron.

There are two collision or scattering mechanisms that dominate in a semiconductor and affect the carrier mobility: phonon or lattice scattering, and ionized impurity scattering.

The atoms in a semiconductor crystal have a certain amount of thermal energy at temperatures above absolute zero that causes the atoms to randomly vibrate about their lattice position within the crystal. The lattice vibrations cause a disruption in the perfect periodic potential function. A perfect periodic potential in a solid allows

electrons to move unimpeded, or with no scattering, through the crystal. But the thermal vibrations cause a disruption of the potential function, resulting in an interaction between the electrons or holes and the vibrating lattice atoms. This *lattice scattering* is also referred to as *phonon scattering*.

Since lattice scattering is related to the thermal motion of atoms, the rate at which the scattering occurs is a function of temperature. If we denote μ_L as the mobility that would be observed if only lattice scattering existed, then the scattering theory states that to first order

$$\mu_L \propto T^{-3/2} \tag{5.15}$$

Mobility that is due to lattice scattering increases as the temperature decreases. Intuitively, we expect the lattice vibrations to decrease as the temperature decreases, which implies that the probability of a scattering event also decreases, thus increasing mobility.

Figure 5.2 shows the temperature dependence of electron and hole mobilities in silicon. In lightly doped semiconductors, lattice scattering dominates and the carrier mobility decreases with temperature as we have discussed. The temperature dependence of mobility is proportional to T^{-n}. The inserts in the figure show that the parameter n is not equal to $\frac{3}{2}$ as the first-order scattering theory predicted. However, mobility does increase as the temperature decreases.

The second interaction mechanism affecting carrier mobility is called *ionized impurity scattering*. We have seen that impurity atoms are added to the semiconductor to control or alter its characteristics. These impurities are ionized at room temperature so that a coulomb interaction exists between the electrons or holes and the ionized impurities. This coulomb interaction produces scattering or collisions and also alters the velocity characteristics of the charge carrier. If we denote μ_I as the mobility that would be observed if only ionized impurity scattering existed, then to first order we have

$$\mu_I \propto \frac{T^{+3/2}}{N_I} \tag{5.16}$$

where $N_I = N_d^+ + N_a^-$ is the total ionized impurity concentration in the semiconductor. If temperature increases, the random thermal velocity of a carrier increases, reducing the time the carrier spends in the vicinity of the ionized impurity center. The less time spent in the vicinity of a coulomb force, the smaller the scattering effect and the larger the expected value of μ_I. If the number of ionized impurity centers increases, then the probability of a carrier encountering an ionized impurity center increases, implying a smaller value of μ_I.

Figure 5.3 is a plot of electron and hole mobilities in germanium, silicon, and gallium arsenide at $T = 300$ K as a function of impurity concentration. More accurately, these curves are of mobility versus ionized impurity concentration N_I. As the impurity concentration increases, the number of impurity scattering centers increases, thus reducing mobility.

If τ_L is the mean time between collisions due to lattice scattering, then dt/τ_L is the probability of a lattice scattering event occurring in a differential time dt. Likewise, if τ_I is the mean time between collisions due to ionized impurity scattering, then dt/τ_I is the probability of an ionized impurity scattering event occurring in the differential time dt.

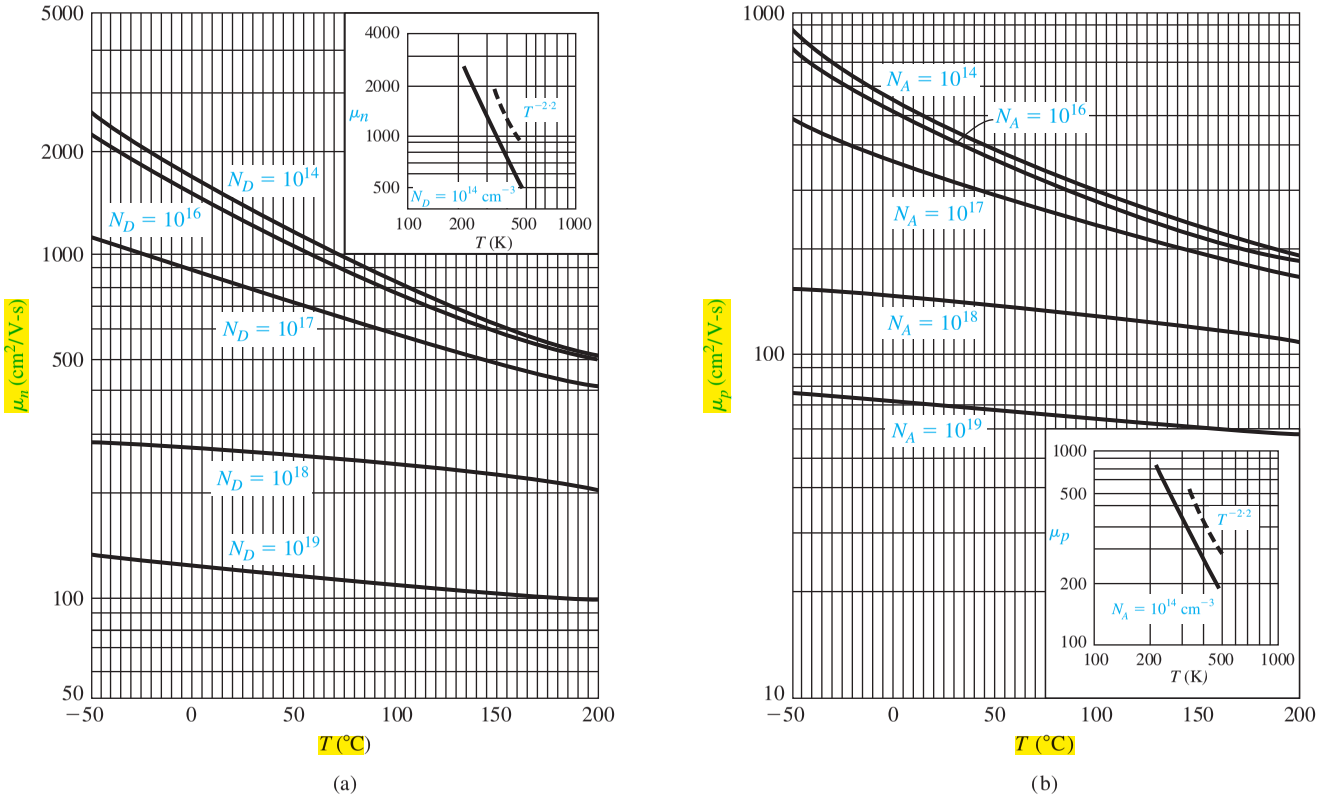


Figure 5.2 | (a) Electron and (b) hole mobilities in silicon versus temperature for various doping concentrations. Inserts show temperature dependence for "almost" intrinsic silicon.
(From Pierret [8].)

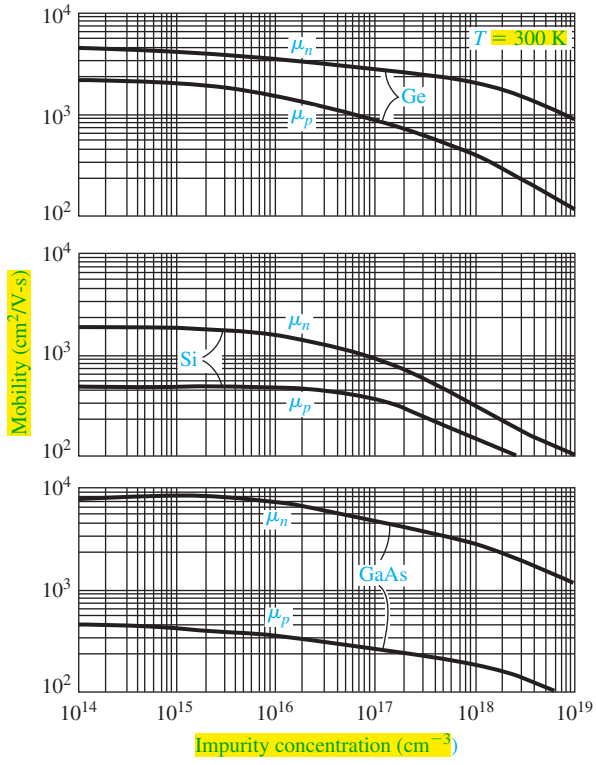


Figure 5.3 | Electron and hole mobilities versus impurity concentrations for germanium, silicon, and gallium arsenide at $T = 300$ K.
(From Sze [14].)

If these two scattering processes are independent, then the total probability of a scattering event occurring in the differential time dt is the sum of the individual events, or

$$\frac{dt}{\tau} = \frac{dt}{\tau_I} + \frac{dt}{\tau_L} \tag{5.17}$$

where τ is the mean time between any scattering event.

Comparing Equation (5.17) with the definitions of mobility given by Equation (5.13) or (5.14), we can write

$$\frac{1}{\mu} = \frac{1}{\mu_I} + \frac{1}{\mu_L} \tag{5.18}$$

where μ_I is the mobility due to the ionized impurity scattering process and μ_L is the mobility due to the lattice scattering process. The parameter μ is the net mobility. With two or more independent scattering mechanisms, the inverse mobilities add, which means that the net mobility decreases.

EXAMPLE 5.2

Objective: Determine the electron mobility in silicon at various doping concentrations and various temperatures.

Using Figure 5.2, find the electron mobility in silicon for:

(a) $T = 25°C$ for (*i*) $N_d = 10^{16}$ cm^{-3} and (*ii*) $N_d = 10^{17}$ cm^{-3}.

(b) $N_d = 10^{16}$ cm^{-3} for (*i*) $T = 0°C$ and (*ii*) $T = 100°C$.

■ **Solution:**

From Figure 5.2, we find the following:

(a) $T = 25°C$; (*i*) $N_d = 10^{16}$ cm$^{-3} \Rightarrow \mu_n \cong 1200$ cm^2/V-s.

(*ii*) $N_d = 10^{17}$ cm$^{-3} \Rightarrow \mu_n \cong 800$ cm^2/V-s.

(b) $N_d = 10^{16}$ cm^{-3}; (*i*) $T = 0°C \Rightarrow \mu_n \cong 1400$ cm^2/V-s.

(*ii*) $T = 100°C \Rightarrow \mu_n \cong 780$ cm^2/V-s.

■ **Comment**

The results of this example show that the mobility values are strong functions of the doping concentration and temperature. These variations must be taken into account in the design of semiconductor devices.

■ **EXERCISE PROBLEM**

Ex 5.2 Using Figure 5.2, find the hole mobility in silicon for:

(*a*) $T = 25°C$ for (*i*) $N_a = 10^{16}$ cm^{-3} and (*ii*) $N_a = 10^{18}$ cm^{-3}, and

(*b*) $N_a = 10^{14}$ cm^{-3} for (*i*) $T = 0°C$ and (*ii*) $T = 100°C$.

[(*a*) (*i*) $\mu_p \cong 410$ cm^2/V-s, (*ii*) $\mu_p \cong 130$ cm^2/V-s; (*b*) (*i*) $\mu_p \cong 550$ cm^2/V-s, (*ii*) $\mu_p \cong 300$ cm^2/V-s]

5.1.3 Conductivity

The drift current density, given by Equation (5.9), may be written as

$$J_{drf} = e(\mu_n n + \mu_p p)\mathrm{E} = \sigma \mathrm{E} \tag{5.19}$$

where σ is the *conductivity* of the semiconductor material. The conductivity is given in units of $(\Omega\text{-cm})^{-1}$ and is a function of the electron and hole concentrations and mobilities. We have just seen that the mobilities are functions of impurity concentrations; conductivity, then is a somewhat complicated function of impurity concentration.

The reciprocal of conductivity is *resistivity,* which is denoted by ρ and is given in units of ohm-cm. We can write the formula for resistivity as[2]

$$\rho = \frac{1}{\sigma} = \frac{1}{e(\mu_n n + \mu_p p)} \tag{5.20}$$

Figure 5.4 is a plot of resistivity as a function of impurity concentration in silicon, germanium, gallium arsenide, and gallium phosphide at $T = 300$ K. Obviously, the curves are not linear functions of N_d or N_a because of mobility effects.

[2]The symbol ρ is also used for volume charge density. The context in which ρ is used should make it clear whether it stands for charge density or resistivity.

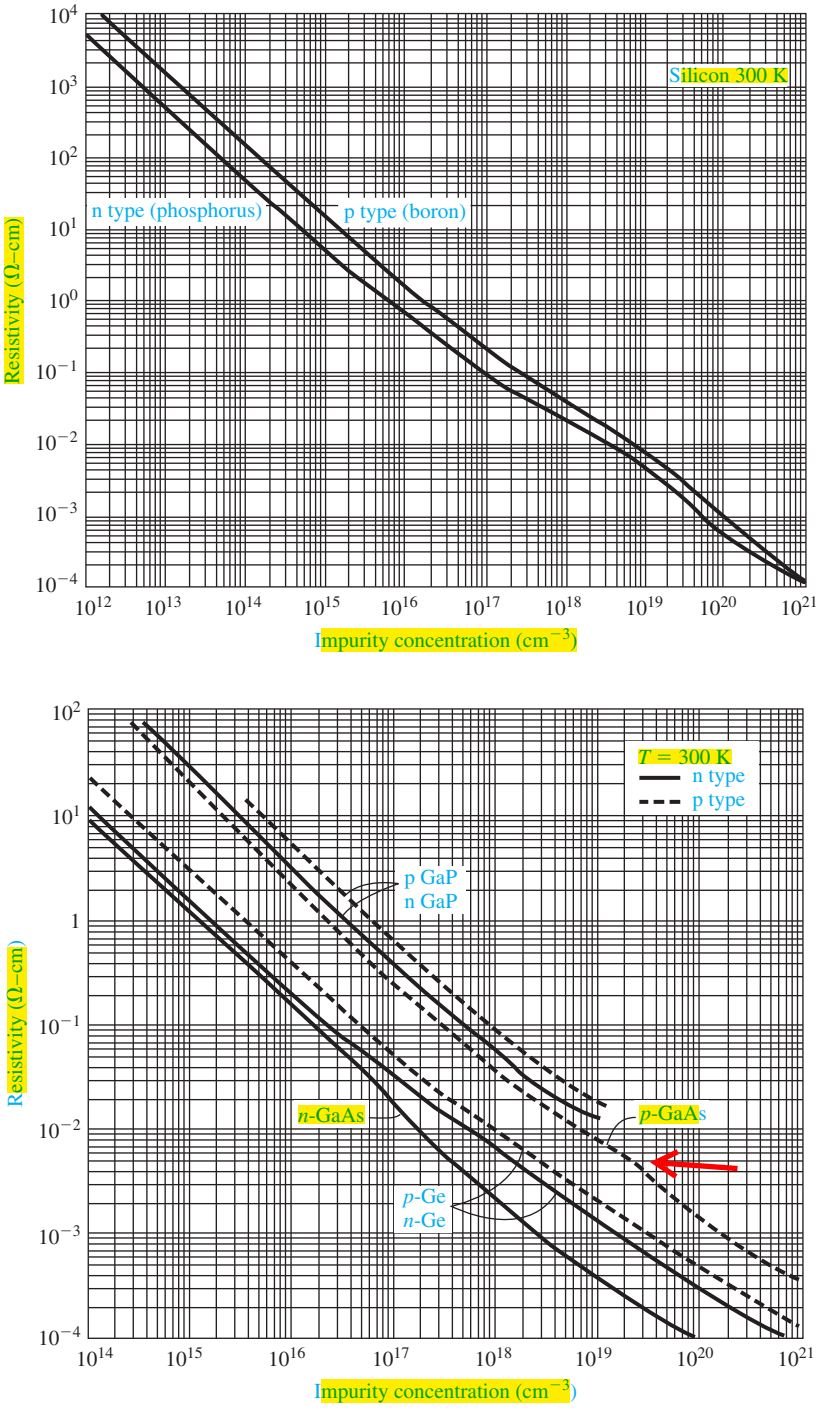


Figure 5.4 | Resistivity versus impurity concentration at $T = 300$ K in (a) silicon and (b) germanium, gallium arsenide, and gallium phosphide.
(From Sze [14].)

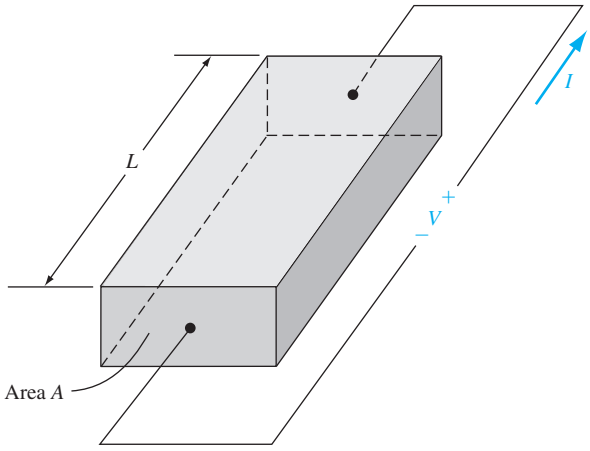


Figure 5.5 | Bar of semiconductor material as a resistor.

If we have a bar of semiconductor material as shown in Figure 5.5 with a voltage applied that produces a current I, then we can write

$$J = \frac{I}{A} \tag{5.21a}$$

and

$$\mathrm{E} = \frac{V}{L} \tag{5.21b}$$

We can now rewrite Equation (5.19) as

$$\frac{I}{A} = \sigma\left(\frac{V}{L}\right) \tag{5.22a}$$

or

$$V = \left(\frac{L}{\sigma A}\right) I = \left(\frac{\rho L}{A}\right) I = IR \tag{5.22b}$$

Equation (5.22b) is Ohm's law for a semiconductor. The resistance is a function of resistivity, or conductivity, as well as the geometry of the semiconductor.

If we consider, for example, a p-type semiconductor with an acceptor doping $N_a (N_d = 0)$ in which $N_a \gg n_i$, and if we assume that the electron and hole mobilities are of the same order of magnitude, then the conductivity becomes

$$\sigma = e(\mu_n n + \mu_p p) \approx e\mu_p p \tag{5.23}$$

If we also assume complete ionization, then Equation (5.23) becomes

$$\sigma \approx e\mu_p N_a \approx \frac{1}{\rho} \tag{5.24}$$

The conductivity and resistivity of an extrinsic semiconductor are a function primarily of the majority carrier parameters.

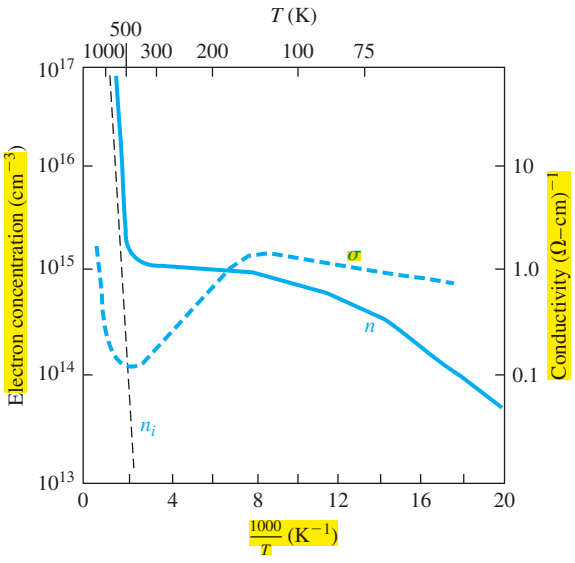


Figure 5.6 | Electron concentration and conductivity versus inverse temperature for silicon.
(After Sze [14].)

We may plot the carrier concentration and conductivity of a semiconductor as a function of temperature for a particular doping concentration. Figure 5.6 shows the electron concentration and conductivity of silicon as a function of inverse temperature for the case when $N_d = 10^{15}$ cm^{-3}. In the midtemperature range, or extrinsic range, as shown, we have complete ionization—the electron concentration remains essentially constant. However, the mobility is a function of temperature so the conductivity varies with temperature in this range. At higher temperatures, the intrinsic carrier concentration increases and begins to dominate the electron concentration as well as the conductivity. In the lower temperature range, freeze-out begins to occur; the electron concentration and conductivity decrease with decreasing temperature.

EXAMPLE 5.3

Objective: Determine the doping concentration and majority carrier mobility given the type and conductivity of a compensated semiconductor.

Consider compensated n-type silicon at $T = 300$ K, with a conductivity of $\sigma = 16\ (\Omega\text{-cm})^{-1}$ and an acceptor doping concentration of 10^{17} cm^{-3}. Determine the donor concentration and the electron mobility.

■ Solution

For n-type silicon at $T = 300$ K, we can assume complete ionization; therefore the conductivity, assuming $N_d - N_a \gg n_i$, is given by

$$\sigma \approx e\mu_n n = e\mu_n (N_d - N_a)$$

We have that

$$16 = (1.6 \times 10^{-19})\mu_n(N_d - 10^{17})$$

Since mobility is a function of the ionized impurity concentration, we can use Figure 5.3 along with trial and error to determine μ_n and N_d. For example, if we choose $N_d = 2 \times 10^{17}$, then $N_I = N_d^+ + N_a^- = 3 \times 10^{17}$ so that $\mu_n \approx 510$ cm^2/V-s which gives $\sigma = 8.16$ $(\Omega\text{-cm})^{-1}$. If we choose $N_d = 5 \times 10^{17}$, then $N_I = 6 \times 10^{17}$ so that $\mu_n \approx 325$ cm^2/V-s, which gives $\sigma = 20.8$ $(\Omega\text{-cm})^{-1}$. The doping is bounded between these two values. Further trial and error yields

$$N_d \approx 3.5 \times 10^{17} \text{ cm}^{-3} \qquad \text{and} \qquad \mu_n \approx 400 \text{ cm}^2\text{/V-s}$$

which gives

$$\sigma \approx 16 \text{ } (\Omega\text{-cm})^{-1}$$

Comment

We can see from this example that, in a high-conductivity semiconductor material, mobility is a strong function of carrier concentration.

EXERCISE PROBLEM

Ex 5.3 A compensated p-type silicon material at $T = 300$ K has impurity doping concentrations of $N_a = 2.8 \times 10^{17}$ cm^{-3} and $N_d = 8 \times 10^{16}$ cm^{-3}. Determine the (*a*) hole mobility, (*b*) conductivity, and (*c*) resistivity.

[Ans. (*a*) $\mu_p \cong 200$ cm^2/V-s; (*b*) $\sigma = 6.4$ $(\Omega\text{-cm})^{-1}$, (*c*) $\rho = 0.156$ $(\Omega\text{-cm})$]

DESIGN EXAMPLE 5.4

Objective: Design a semiconductor resistor with a specified resistance to handle a given current density.

A silicon semiconductor at $T = 300$ K is initially doped with donors at a concentration of $N_d = 5 \times 10^{15}$ cm^{-3}. Acceptors are to be added to form a compensated p-type material. The resistor is to have a resistance of 10 kΩ and handle a current density of 50 A/cm^2 when 5 V is applied.

Solution

For 5 V applied to a 10-kΩ resistor, the total current is

$$I = \frac{V}{R} = \frac{5}{10} = 0.5 \text{ mA}$$

If the current density is limited to 50 A/cm^2, then the cross-sectional area is

$$A = \frac{I}{J} = \frac{0.5 \times 10^{-3}}{50} = 10^{-5} \text{ cm}^2$$

If we, somewhat arbitrarily at this point, limit the electric field to E = 100 V/cm, then the length of the resistor is

$$L = \frac{V}{\text{E}} = \frac{5}{100} = 5 \times 10^{-2} \text{ cm}$$

From Equation (5.22b), the conductivity of the semiconductor is

$$\sigma = \frac{L}{RA} = \frac{5 \times 10^{-2}}{(10^4)(10^{-5})} = 0.50 \text{ } (\Omega\text{-cm})^{-1}$$

The conductivity of a compensated p-type semiconductor is

$$\sigma \approx e\mu_p p = e\mu_p (N_a - N_d)$$

where the mobility is a function of the total ionized impurity concentration $N_a + N_d$.

Using trial and error, if $N_a = 1.25 \times 10^{16}$ cm^{-3}, then $N_a + N_d = 1.75 \times 10^{16}$ cm^{-3}, and the hole mobility, from Figure 5.3, is approximately $\mu_p = 410$ cm^2/V-s. The conductivity is then

$$\sigma = e\mu_p(N_a - N_d) = (1.6 \times 10^{-19})(410)(1.25 \times 10^{16} - 5 \times 10^{15}) = 0.492$$

which is very close to the value we need.

■ Comment

Since the mobility is related to the total ionized impurity concentration, the determination of the impurity concentration to achieve a particular conductivity is not straightforward.

■ EXERCISE PROBLEM

Ex 5.4 A bar of p-type silicon, such as shown in Figure 5.5, has a cross-sectional area $A = 10^{-6}$ cm^2 and a length $L = 1.2 \times 10^{-3}$ cm. For an applied voltage of 5 V, a current of 2 mA is required. What is the required (*a*) resistance, (*b*) resistivity, and (*c*) impurity doping concentration? (*d*) What is the resulting hole mobility?

[Ans. (*a*) 2.5 kΩ; (*b*) 2.083 (Ω-cm); (*c*) $N_a \cong 7.3 \times 10^{15}$ cm^{-3}; (*d*) $\mu_p \cong 410$ cm^2/V-s]

For an intrinsic material, the conductivity can be written as

$$\sigma_i = e(\mu_n + \mu_p)\, n_i \tag{5.25}$$

The concentrations of electrons and holes are equal in an intrinsic semiconductor, so the intrinsic conductivity includes both the electron and hole mobility. Since, in general, the electron and hole mobilities are not equal, the intrinsic conductivity is not the minimum value possible at a given temperature.

5.1.4 Velocity Saturation

So far in our discussion of drift velocity, we have assumed that mobility is not a function of electric field, meaning that the drift velocity will increase linearly with applied electric field. The total velocity of a particle is the sum of the random thermal velocity and drift velocity. At $T = 300$ K, the average random thermal energy is given by

$$\frac{1}{2}mv_{th}^2 = \frac{3}{2}kT = \frac{3}{2}(0.0259) = 0.03885 \text{ eV} \tag{5.26}$$

This energy translates into a mean thermal velocity of approximately 10^7 cm/s for an electron in silicon. If we assume an electron mobility of $\mu_n = 1350$ cm^2/V-s in low-doped silicon, a drift velocity of 10^5 cm/s, or 1 percent of the thermal velocity, is achieved if the applied electric field is approximately 75 V/cm. This applied electric field does not appreciably alter the energy of the electron.

Figure 5.7 is a plot of average drift velocity as a function of applied electric field for electrons and holes in silicon, gallium arsenide, and germanium. At low electric fields, where there is a linear variation of velocity with electric field, the slope of

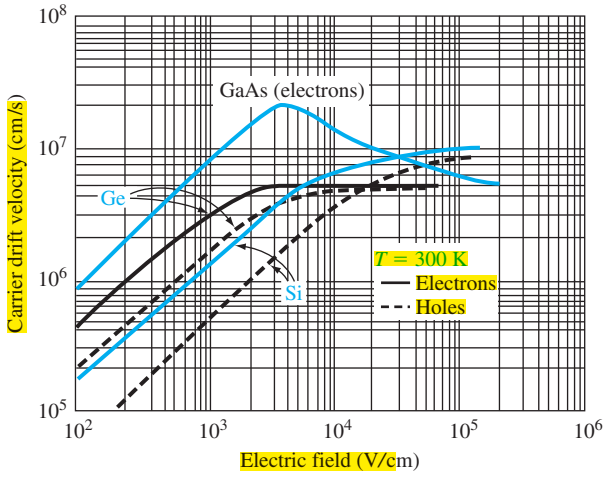


Figure 5.7 | Carrier drift velocity versus electric field for high-purity silicon, germanium, and gallium arsenide. (*From Sze [14].*)

the drift velocity versus electric field curve is the mobility. The behavior of the drift velocity of carriers at high electric fields deviates substantially from the linear relationship observed at low fields. The drift velocity of electrons in silicon, for example, saturates at approximately 10^7 cm/s at an electric field of approximately 30 kV/cm. If the drift velocity of a charge carrier saturates, then the drift current density also saturates and becomes independent of the applied electric field.

The experimental carrier drift velocity versus electric field in silicon can be approximated for electrons by [2]

$$v_n = \frac{v_s}{\left[1 + \left(\frac{E_{on}}{E}\right)^2\right]^{1/2}} \tag{5.27a}$$

and for holes by

$$v_p = \frac{v_s}{\left[1 + \left(\frac{E_{op}}{E}\right)\right]} \tag{5.27b}$$

faltam o quadrado e a raiz.

The variables are $v_s = 10^7$ cm/s at $T = 300$ K, $E_{on} = 7 \times 10^3$ V/cm, and $E_{op} = 2 \times 10^4$ V/cm.

We may note that for small electric fields, the drift velocities reduce to

$$v_n \cong \left(\frac{E}{E_{on}}\right) \cdot v_s \tag{5.28a}$$

and

$$v_p \cong \left(\frac{E}{E_{op}}\right) \cdot v_s \tag{5.28b}$$

At low electric fields, the drift velocities are linear functions of the electric field as we have discussed. However, for large electric fields, the drift velocities approach the saturation value.

The drift velocity versus electric field characteristic of gallium arsenide is more complicated than for silicon or germanium. At low fields, the slope of the drift velocity versus E-field is constant and is the low-field electron mobility, which is approximately 8500 cm^2/V-s for gallium arsenide. The low-field electron mobility in gallium arsenide is much larger than in silicon. As the field increases, the electron drift velocity in gallium arsenide reaches a peak and then decreases. A differential mobility is the slope of the v_d versus E curve at a particular point on the curve and the negative slope of the drift velocity versus electric field represents a negative differential mobility. The negative differential mobility produces a negative differential resistance; this characteristic is used in the design of oscillators.

The negative differential mobility can be understood by considering the E versus k diagram for gallium arsenide, which is shown again in Figure 5.8. The density of states effective mass of the electron in the lower valley is $m_n^* = 0.067\,m_0$. The small effective mass leads to a large mobility. As the E-field increases, the energy of the electron increases and the electron can be scattered into the upper valley, where the density of states effective mass is $0.55\,m_0$. The larger effective mass in the upper valley yields a smaller mobility. This intervalley transfer mechanism results in a decreasing average drift velocity of electrons with electric field, or the negative differential mobility characteristic.

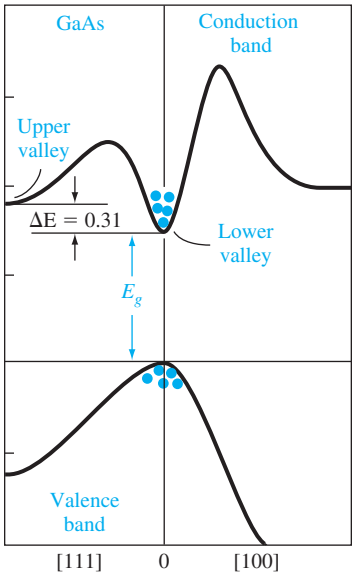


Figure 5.8 | Energy-band structure for gallium arsenide showing the upper valley and lower valley in the conduction band.
(*From Sze [15].*)

TEST YOUR UNDERSTANDING

TYU 5.1 Consider a sample of silicon at $T = 300$ K doped at an impurity concentration of $N_d = 10^{15}$ cm^{-3} and $N_a = 10^{14}$ cm^{-3}. Assume electron and hole mobilities given in Table 5.1. Calculate the drift current density if the applied electric field is E = 35 V/cm. (Ans. 6.80 A/cm^2)

TYU 5.2 Silicon at $T = 300$ K is doped with impurity concentrations of $N_d = 5 \times 10^{16}$ cm^{-3} and $N_a = 2 \times 10^{16}$ cm^{-3}. (*a*) What are the electron and hole mobilities? (*b*) Determine the conductivity and resistivity of the material.

[Ans. (*a*) $\mu_n = 1000$ cm^2/V-s, $\mu_p = 350$ cm^2/V-s; (*b*) $\sigma = 4.8\ (\Omega\text{-cm})^{-1}$, $\rho = 0.208\ \Omega$-cm]

TYU 5.3 For a particular silicon semiconductor device at $T = 300$ K, the required material is n type with a resistivity of 0.10 Ω-cm. (*a*) Determine the required impurity doping concentration and (*b*) the resulting electron mobility.

[Ans. (*a*) From Figure 5.4, $N_d \approx 9 \times 10^{16}$ cm^{-3}; (*b*) $\mu_n \approx 695$ cm^2/V-s]

5.2 | CARRIER DIFFUSION

There is a second mechanism, in addition to drift, that can induce a current in a semiconductor. We may consider a classic physics example in which a container, as shown in Figure 5.9, is divided into two parts by a membrane. The left side contains gas molecules at a particular temperature and the right side is initially empty. The gas molecules are in continual random thermal motion so that, when the membrane is broken, the gas molecules flow into the right side of the container. *Diffusion* is the process whereby particles flow from a region of high concentration toward a region of low concentration. If the gas molecules were electrically charged, the net flow of charge would result in a *diffusion current.*

5.2.1 Diffusion Current Density

To begin to understand the diffusion process in a semiconductor, we will consider a simplified analysis. Assume that an electron concentration varies in one dimension as shown in Figure 5.10. The temperature is assumed to be uniform so that the average thermal velocity of electrons is independent of x. To calculate the current, we will determine the net flow of electrons per unit time per unit area crossing the plane at $x = 0$. If the distance l shown in Figure 5.10 is less than the mean-free path of an electron, that is, the average

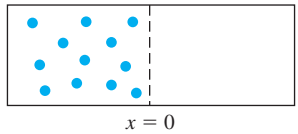


Figure 5.9 | Container divided by a membrane with gas molecules on one side.

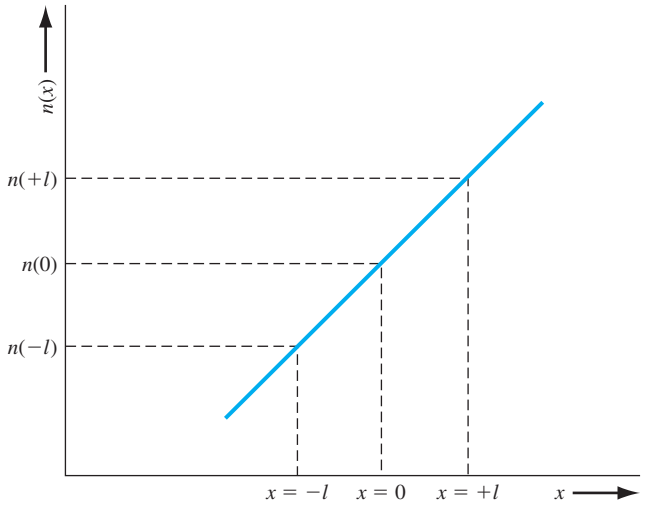


Figure 5.10 | Electron concentration versus distance.

distance an electron travels between collisions ($l < v_{th}\tau_{cn}$), then on the average, electrons moving to the right at $x = -l$ and electrons moving to the left at $x = +l$ will cross the $x = 0$ plane. One half of the electrons at $x = -l$ will be traveling to the right at any instant of time and one half of the electrons at $x = +l$ will be traveling to the left at any given time. The net rate of electron flow, F_n, in the $+x$ direction at $x = 0$ is given by

$$F_n = \frac{1}{2}n(-l)v_{th} - \frac{1}{2}n(+l)v_{th} = \frac{1}{2}v_{th}[n(-l) - n(+l)] \tag{5.29}$$

If we expand the electron concentration in a Taylor series about $x = 0$ keeping only the first two terms, then we can write Equation (5.29) as

$$F_n = \frac{1}{2}v_{th}\left\{\left[n(0) - l\frac{dn}{dx}\right] - \left[n(0) + l\frac{dn}{dx}\right]\right\} \tag{5.30}$$

which becomes

$$F_n = -v_{th}l\frac{dn}{dx} \tag{5.31}$$

Each electron has a charge $(-e)$, so the current is

$$J = -eF_n = +ev_{th}l\frac{dn}{dx} \tag{5.32}$$

The current described by Equation (5.32) is the electron diffusion current and is proportional to the spatial derivative, or density gradient, of the electron concentration.

The diffusion of electrons from a region of high concentration to a region of low concentration produces a flux of electrons flowing in the negative x direction for this example. Since electrons have a negative charge, the conventional current direction is in the positive x direction. Figure 5.11a shows these one-dimensional flux and

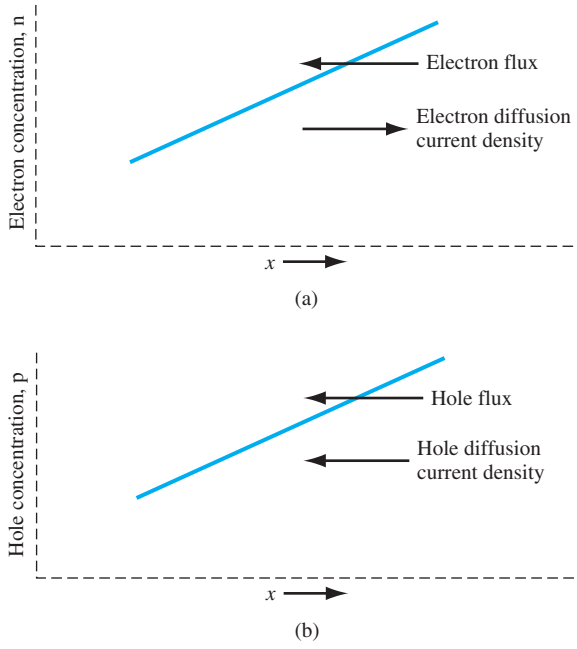


Figure 5.11 | (a) Diffusion of electrons due to a density gradient. (b) Diffusion of holes due to a density gradient.

current directions. We may write the electron diffusion current density for this one-dimensional case, in the form

$$J_{nx|dif} = eD_n \frac{dn}{dx} \tag{5.33}$$

where D_n is called the *electron diffusion coefficient,* has units of cm^2/s, and is a positive quantity. If the electron density gradient becomes negative, the electron diffusion current density will be in the negative x direction.

Figure 5.11b shows an example of a hole concentration as a function of distance in a semiconductor. The diffusion of holes, from a region of high concentration to a region of low concentration, produces a flux of holes in the negative x direction. Since holes are positively charged particles, the conventional diffusion current density is also in the negative x direction. The hole diffusion current density is proportional to the hole density gradient and to the electronic charge, so we may write

$$J_{px|dif} = -eD_p \frac{dp}{dx} \tag{5.34}$$

for the one-dimensional case. The parameter D_p is called the *hole diffusion coefficient,* has units of cm^2/s, and is a positive quantity. If the hole density gradient becomes negative, the hole diffusion current density will be in the positive x direction.

EXAMPLE 5.5

Objective: Calculate the diffusion current density given a density gradient.

Assume that, in an n-type gallium arsenide semiconductor at $T = 300$ K, the electron concentration varies linearly from 1×10^{18} to 7×10^{17} cm^{-3} over a distance of 0.10 cm. Calculate the diffusion current density if the electron diffusion coefficient is $D_n = 225$ cm^2/s.

■ Solution

The diffusion current density is given by

$$J_{n|dif} = eD_n \frac{dn}{dx} \approx eD_n \frac{\Delta n}{\Delta x}$$

$$= (1.6 \times 10^{-19})(225)\left(\frac{1 \times 10^{18} - 7 \times 10^{17}}{0.10}\right) = 108 \text{ A/cm}^2$$

■ Comment

A significant diffusion current density can be generated in a semiconductor material with only a modest density gradient.

■ EXERCISE PROBLEM

Ex 5.5 The hole density in silicon is given by $p(x) = 10^{16}\, e^{-(x/L_p)}$ $(x \geq 0)$ where $L_p = 2 \times 10^{-4}$ cm. Assume the hole diffusion coefficient is $D_p = 8$ cm^2/s. Determine the hole diffusion current density at (a) $x = 0$, (b) $x = 2 \times 10^{-4}$ cm, and (c) $x = 10^{-3}$ cm.

[Ans. (a) $J_p = 64$ A/cm^2; (b) $J_p = 23.54$ A/cm^2; (c) $J_p = 0.431$ A/cm^2]

5.2.2 Total Current Density

We now have four possible independent current mechanisms in a semiconductor. These components are electron drift and diffusion currents and hole drift and diffusion currents. The total current density is the sum of these four components, or, for the one-dimensional case,

$$J = en\mu_n \mathrm{E}_x + ep\mu_p \mathrm{E}_x + eD_n \frac{dn}{dx} - eD_p \frac{dp}{dx} \tag{5.35}$$

This equation may be generalized to three dimensions as

$$J = en\mu_n \mathrm{E} + ep\mu_p \mathrm{E} + eD_n \nabla n - eD_p \nabla p \tag{5.36}$$

The electron mobility gives an indication of how well an electron moves in a semiconductor as a result of the force of an electric field. The electron diffusion coefficient gives an indication of how well an electron moves in a semiconductor as a result of a density gradient. The electron mobility and diffusion coefficient are not independent parameters. Similarly, the hole mobility and diffusion coefficient are not independent parameters. The relationship between mobility and the diffusion coefficient is developed in the next section.

The expression for the total current in a semiconductor contains four terms. Fortunately in most situations, we will only need to consider one term at any one time at a particular point in a semiconductor.

TEST YOUR UNDERSTANDING

TYU 5.4 The electron concentration in silicon is given by $n(x) = 10^{15}\, e^{-(x/Ln)}$ cm^{-3} ($x \geq 0$) where $L_n = 10^{-4}$ cm. The electron diffusion coefficient is $D_n = 25$ cm^2/s. Determine the electron diffusion current density at (*a*) $x = 0$, (*b*) $x = 10^{-4}$ cm, and (*c*) $x \to \infty$. [(*a*) -40 A/cm^2; (*b*) -14.7 A/cm^2; (*c*) 0]

TYU 5.5 The hole concentration in silicon varies linearly from $x = 0$ to $x = 0.01$ cm. The hole diffusion coefficient is $D_p = 10$ cm^2/s, the hole diffusion current density is 20 A/cm^2, and the hole concentration at $x = 0$ is $p = 4 \times 10^{17}$ cm^{-3}. What is the value of the hole concentration at $x = 0.01$ cm? (2.75×10^{17} cm^{-3})

5.3 | GRADED IMPURITY DISTRIBUTION

In most cases so far, we have assumed that the semiconductor is uniformly doped. In many semiconductor devices, however, there may be regions that are nonuniformly doped. We will investigate how a nonuniformly doped semiconductor reaches thermal equilibrium and, from this analysis, we will derive the Einstein relation, which relates mobility and the diffusion coefficient.

5.3.1 Induced Electric Field

Consider a semiconductor that is nonuniformly doped with donor impurity atoms. If the semiconductor is in thermal equilibrium, the Fermi energy level is constant through the crystal so the energy-band diagram may qualitatively look like that shown in Figure 5.12. The doping concentration decreases as x increases in this case. There will be a diffusion of majority carrier electrons from the region of high concentration to the region of low concentration, which is in the $+x$ direction. The flow of negative electrons leaves behind positively charged donor ions. The separation of

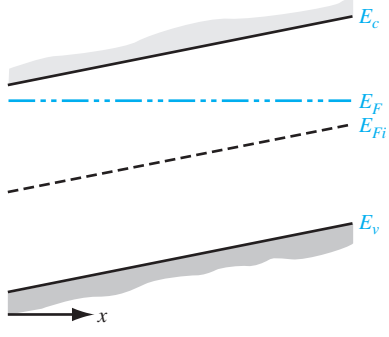


Figure 5.12 | Energy-band diagram for a semiconductor in thermal equilibrium with a nonuniform donor impurity concentration.

positive and negative charge induces an electric field that is in a direction to oppose the diffusion process. When equilibrium is reached, the mobile carrier concentration is not exactly equal to the fixed impurity concentration and the induced electric field prevents any further separation of charge. In most cases of interest, the space charge induced by this diffusion process is a small fraction of the impurity concentration, thus the mobile carrier concentration is not too different from the impurity dopant density.

The electric potential ϕ is related to electron potential energy by the charge $(-e)$, so we can write

$$\phi = +\frac{1}{e}(E_F - E_{Fi}) \tag{5.37}$$

The electric field for the one-dimensional situation is defined as

$$\mathrm{E}_x = -\frac{d\phi}{dx} = \frac{1}{e}\frac{dE_{Fi}}{dx} \tag{5.38}$$

If the intrinsic Fermi-level changes as a function of distance through a semiconductor in thermal equilibrium, an electric field exists in the semiconductor.

If we assume a quasi-neutrality condition in which the electron concentration is almost equal to the donor impurity concentration, then we can still write

$$n_0 = n_i \exp\left[\frac{E_F - E_{Fi}}{kT}\right] \approx N_d(x) \tag{5.39}$$

Solving for $E_F - E_{Fi}$, we obtain

$$E_F - E_{Fi} = kT \ln\left(\frac{N_d(x)}{n_i}\right) \tag{5.40}$$

The Fermi level is constant for thermal equilibrium so when we take the derivative with respect to x we obtain

$$-\frac{dE_{Fi}}{dx} = \frac{kT}{N_d(x)}\frac{dN_d(x)}{dx} \tag{5.41}$$

The electric field can then be written, combining Equations (5.41) and (5.38), as

$$\mathrm{E}_x = -\left(\frac{kT}{e}\right)\frac{1}{N_d(x)}\frac{dN_d(x)}{dx} \tag{5.42}$$

Since we have an electric field, there will be a potential difference through the semiconductor due to the nonuniform doping.

EXAMPLE 5.6

Objective: Determine the induced electric field in a semiconductor in thermal equilibrium, given a linear variation in doping concentration.

Assume that the donor concentration in an n-type semiconductor at $T = 300$ K is given by

$$N_d(x) = 10^{16} - 10^{19}x \qquad (\mathrm{cm}^{-3})$$

where x is given in cm and ranges between $0 \le x \le 1$ μm

■ Solution

Taking the derivative of the donor concentration, we have

$$\frac{dN_d(x)}{dx} = -10^{19} \qquad (\text{cm}^{-4})$$

The electric field is given by Equation (5.42), so we have

$$\mathrm{E}_x = \frac{-(0.0259)(-10^{19})}{(10^{16} - 10^{19}x)}$$

At $x = 0$, for example, we find

$$\mathrm{E}_x = 25.9 \text{ V/cm}$$

■ Comment

We may recall from our previous discussion of drift current that fairly small electric fields can produce significant drift current densities, so that an induced electric field from nonuniform doping can significantly influence semiconductor device characteristics.

■ EXERCISE PROBLEM

Ex 5.6 Assume the donor concentration in an n-type semiconductor at $T = 300$ K is given by $N_d(x) = 10^{16}\, e^{-x/L}$ where $L = 2 \times 10^{-2}$ cm. Determine the induced electric field in the semiconductor at (*a*) $x = 0$ and (*b*) $x = 10^{-4}$ cm.

[Ans. $\mathrm{E} = 1.295$ V/cm for (*a*) and (*b*)]

5.3.2 The Einstein Relation

If we consider the nonuniformly doped semiconductor represented by the energy-band diagram shown in Figure 5.12 and assume there are no electrical connections so that the semiconductor is in thermal equilibrium, then the individual electron and hole currents must be zero. We can write

$$J_n = 0 = en\mu_n \mathrm{E}_x + eD_n \frac{dn}{dx} \tag{5.43}$$

If we assume quasi-neutrality so that $n \approx N_d(x)$, then we can rewrite Equation (5.43) as

$$J_n = 0 = e\mu_n N_d(x)\mathrm{E}_x + eD_n \frac{dN_d(x)}{dx} \tag{5.44}$$

Substituting the expression for the electric field from Equation (5.42) into Equation (5.44), we obtain

$$0 = -e\mu_n N_d(x)\left(\frac{kT}{e}\right)\frac{1}{N_d(x)}\frac{dN_d(x)}{dx} + eD_n \frac{dN_d(x)}{dx} \tag{5.45}$$

Equation (5.45) is valid for the condition

$$\frac{D_n}{\mu_n} = \frac{kT}{e} \tag{5.46a}$$

The hole current must also be zero in the semiconductor. From this condition, we can show that

$$\frac{D_p}{\mu_p} = \frac{kT}{e} \tag{5.46b}$$

Combining Equations (5.46a) and (5.46b) gives

$$\frac{D_n}{\mu_n} = \frac{D_p}{\mu_p} = \frac{kT}{e} \tag{5.47}$$

The diffusion coefficient and mobility are not independent parameters. This relation between the mobility and diffusion coefficient, given by Equation (5.47), is known as the *Einstein relation.*

EXAMPLE 5.7

Objective: Determine the diffusion coefficient given the carrier mobility.

Assume that the mobility of a particular carrier is 1000 cm²/V-s at $T = 300$ K.

■ Solution

Using the Einstein relation, we have that

$$D = \left(\frac{kT}{e}\right)\mu = (0.0259)(1000) = 25.9 \text{ cm}^2/\text{s}$$

■ Comment

Although this example is fairly simple and straightforward, it is important to keep in mind the relative orders of magnitude of the mobility and diffusion coefficient. The diffusion coefficient is approximately 40 times smaller than the mobility at room temperature.

■ EXERCISE PROBLEM

Ex 5.7 Assume the electron diffusion coefficient of a semiconductor at $T = 300$ K is $D_n = 215$ cm²/s. Determine the electron mobility. (Ans. $\mu_n = 8301$ cm²/V-s)

Table 5.2 shows the diffusion coefficient values at $T = 300$ K corresponding to the mobilities listed in Table 5.1 for silicon, gallium arsenide, and germanium.

The relation between the mobility and diffusion coefficient given by Equation (5.47) contains temperature. It is important to keep in mind that the major temperature effects are a result of lattice scattering and ionized impurity scattering processes, as discussed in Section 5.1.2. As the mobilities are strong functions of temperature because of the scattering processes, the diffusion coefficients are also strong functions of temperature. The specific temperature dependence given in Equation (5.47) is a small fraction of the real temperature characteristic.

Table 5.2 | Typical mobility and diffusion coefficient values at $T = 300$ K (μ = cm²/V-s and D = cm²/s)

	μ_m	D_n	μ_p	D_p
Silicon	1350	35	480	12.4
Gallium arsenide	8500	220	400	10.4
Germanium	3900	101	1900	49.2

*5.4 | THE HALL EFFECT

The Hall effect is a consequence of the forces that are exerted on moving charges by electric and magnetic fields. The Hall effect is used to distinguish whether a semiconductor is n type or p type[3] and to measure the majority carrier concentration and majority carrier mobility. The Hall effect device, as discussed in this section, is used to experimentally measure semiconductor parameters. However, it is also used extensively in engineering applications as a magnetic probe and in other circuit applications.

The force on a particle having a charge q and moving in a magnetic field is given by

$$F = qv \times B \tag{5.48}$$

where the cross product is taken between velocity and magnetic field so that the force vector is perpendicular to both the velocity and magnetic field.

Figure 5.13 illustrates the Hall effect. A semiconductor with a current I_x is placed in a magnetic field perpendicular to the current. In this case, the magnetic field is in the z direction. Electrons and holes flowing in the semiconductor will experience a force as indicated in the figure. The force on both electrons and holes is in the $(-y)$ direction. In a p-type semiconductor ($p_0 > n_0$), there will be a buildup of positive charge on the $y = 0$ surface of the semiconductor and, in an n-type semiconductor ($n_0 > p_0$), there will be a buildup of negative charge on the $y = 0$ surface. This net charge induces an electric field in the y direction as shown in the figure. In steady

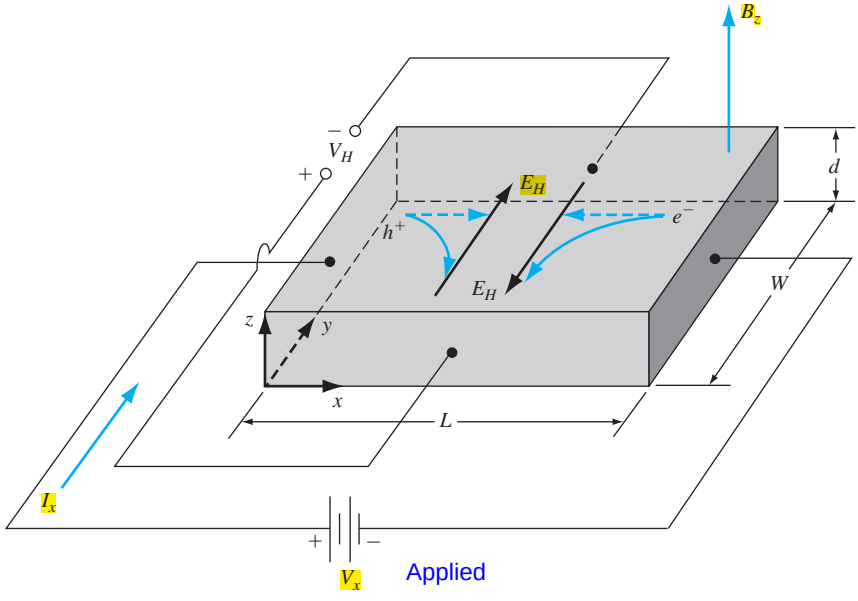


Figure 5.13 | Geometry for measuring the Hall effect.

*Indicates sections that will aid in the total summation of understanding of semiconductor devices, but may be skipped the first time through the text without loss of continuity.

[3]We will assume an extrinsic semiconductor material in which the majority carrier concentration is much larger than the minority carrier concentration.

state, the magnetic field force will be exactly balanced by the induced electric field force. This balance may be written as

$$F = q[\mathrm{E} + v \times B] = 0 \tag{5.49a}$$

which becomes

$$q\mathrm{E}_y = qv_xB_z \tag{5.49b}$$

The induced electric field in the y direction is called the *Hall field*. The Hall field produces a voltage across the semiconductor which is called the *Hall voltage*. We can write

$$V_H = +\mathrm{E}_H W \tag{5.50}$$

where E_H is assumed positive in the $+y$ direction and V_H is positive with the polarity shown.

In a p-type semiconductor, in which holes are the majority carrier, the Hall voltage will be positive as defined in Figure 5.13. In an n-type semiconductor, in which electrons are the majority carrier, the Hall voltage will have the opposite polarity. The polarity of the Hall voltage is used to determine whether an extrinsic semiconductor is n type or p type.

Substituting Equation (5.50) into Equation (5.49) gives

$$V_H = v_x WB_z \tag{5.51}$$

For a p-type semiconductor, the drift velocity of holes can be written as

$$v_{dx} = \frac{J_x}{ep} = \frac{I_x}{(ep)(Wd)} \tag{5.52}$$

where e is the magnitude of the electronic charge. Combining Equations (5.52) and (~~5.50~~ 5.51), we have

$$V_H = \frac{I_x B_z}{epd} \tag{5.53}$$

or, solving for the hole concentration, we obtain

$$p = \frac{I_x B_z}{edV_H} \tag{5.54}$$

The majority carrier hole concentration is determined from the current, magnetic field, and Hall voltage.

For an n-type semiconductor, the Hall voltage is given by

$$V_H = -\frac{I_x B_z}{ned} \tag{5.55}$$

so that the electron concentration is

$$n = -\frac{I_x B_z}{edV_H} \tag{5.56}$$

Note that the Hall voltage is negative for the n-type semiconductor; therefore, the electron concentration determined from Equation (5.56) is actually a positive quantity.

Once the majority carrier concentration has been determined, we can calculate the low-field majority carrier mobility. For a p-type semiconductor, we can write

$$J_x = ep\mu_p E_x \tag{5.57}$$

The current density and electric field can be converted to current and voltage so that Equation (5.57) becomes

$$\frac{I_x}{Wd} = \frac{ep\mu_p V_x}{L} \tag{5.58}$$

The hole mobility is then given by

$$\mu_p = \frac{I_x L}{epV_x Wd} \tag{5.59}$$

Similarly for an n-type semiconductor, the low-field electron mobility is determined from

$$\mu_n = \frac{I_x L}{enV_x Wd} \tag{5.60}$$

EXAMPLE 5.8

Objective: Determine the majority carrier concentration and mobility, given Hall effect parameters.

Consider the geometry shown in Figure 5.13. Let $L = 10^{-1}$ cm, $W = 10^{-2}$ cm, and $d = 10^{-3}$ cm. Also assume that $I_x = 1.0$ mA, $V_x = 12.5$ V, $B_z = 500$ gauss $= 5 \times 10^{-2}$ tesla, and $V_H = -6.25$ mV.

■ Solution

A negative Hall voltage for this geometry implies that we have an n-type semiconductor. Using Equation (5.56), we can calculate the electron concentration as

$$n = \frac{-(10^{-3})(5 \times 10^{-2})}{(1.6 \times 10^{-19})(10^{-5})(-6.25 \times 10^{-3})} = 5 \times 10^{21}\ \text{m}^{-3} = 5 \times 10^{15}\ \text{cm}^{-3}$$

The electron mobility is then determined from Equation (5.60) as

$$\mu_n = \frac{(10^{-3})(10^{-3})}{(1.6 \times 10^{-19})(5 \times 10^{21})(12.5)(10^{-4})(10^{-5})} = 0.10\ \text{m}^2/\text{V-s}$$

or

$$\mu_n = 1000\ \text{cm}^2/\text{V-s}$$

■ Comment

It is important to note that the MKS units must be used consistently in the Hall effect equations to yield correct results.

■ EXERCISE PROBLEM

Ex 5.8 A p-type silicon sample with the geometry shown in Figure 5.13 has parameters $L = 0.2$ cm, $W = 10^{-2}$ cm, and $d = 8 \times 10^{-4}$ cm. The semiconductor parameters are $p = 10^{16}$ cm^{-3} and $\mu_p = 320$ cm^2/V-s. For $V_x = 10$ V and $B_z = 500$ gauss $= 5 \times 10^{-2}$ tesla, determine I_x and V_H. (Ans. $I_x = 0.2048$ mA, $V_H = 0.80$ mV)

5.5 | SUMMARY

- The two basic transport mechanisms are drift, due to an applied electric field, and diffusion, due to a density gradient.
- Carriers reach an average drift velocity in the presence of an applied electric field, due to scattering events. Two scattering processes within a semiconductor are lattice scattering and impurity scattering.
- The average drift velocity is a linear function of the applied electric field for small values of electric field, but the drift velocity reaches a saturation limit that is on the order of 10^7 cm/s at high electric fields.
- Carrier mobility is the ratio of the average drift velocity and applied electric field. The electron and hole mobilities are functions of temperature and of the ionized impurity concentration.
- The drift current density is the product of conductivity and electric field (a form of Ohm's law). Conductivity is a function of the carrier concentrations and mobilities. Resistivity is the inverse of conductivity.
- The diffusion current density is proportional to the carrier diffusion coefficient and the carrier density gradient.
- The diffusion coefficient and mobility are related through the Einstein relation.
- The Hall effect is a consequence of a charged carrier moving in the presence of perpendicular electric and magnetic fields. The charged carrier is deflected, inducing a Hall voltage. The polarity of the Hall voltage is a function of the semiconductor conductivity type. The majority carrier concentration and mobility can be determined from the Hall voltage.

GLOSSARY OF IMPORTANT TERMS

conductivity A material parameter related to carrier drift; quantitatively, the ratio of drift current density to electric field.

diffusion The process whereby particles flow from a region of high concentration to a region of low concentration.

diffusion coefficient The parameter relating particle flux to the particle density gradient.

diffusion current The current that results from the diffusion of charged particles.

drift The process whereby charged particles move while under the influence of an electric field.

drift current The current that results from the drift of charged particles.

drift velocity The average velocity of charged particles in the presence of an electric field.

Einstein relation The relation between the mobility and the diffusion coefficient.

Hall voltage The voltage induced across a semiconductor in a Hall effect measurement.

ionized impurity scattering The interaction between a charged carrier and an ionized impurity center.

lattice scattering The interaction between a charged carrier and a thermally vibrating lattice atom.

mobility The parameter relating carrier drift velocity and electric field.

resistivity The reciprocal of conductivity; a material parameter that is a measure of the resistance to current.

velocity saturation The saturation of carrier drift velocity with increasing electric field.

CHECKPOINT

After studying this chapter, the reader should have the ability to:

- Discuss carrier drift current density.
- Explain why carriers reach an average drift velocity in the presence of an applied electric field.
- Discuss the mechanisms of lattice scattering and impurity scattering.
- Define mobility and discuss the temperature and ionized impurity concentration dependence on mobility.
- Define conductivity and resistivity.
- Discuss velocity saturation.
- Discuss carrier diffusion current density.
- State the Einstein relation.
- Describe the Hall effect.

REVIEW QUESTIONS

1. Write the equation for the total drift current density. Is the linear relationship between drift current density and electric field always valid? Why or why not.
2. Define electron and hole mobility. What is the unit of mobility?
3. Explain the temperature dependence of mobility. Why is the carrier mobility a function of the ionized impurity concentrations?
4. Define conductivity. Define resistivity. What are the units of conductivity and resistivity?
5. Sketch the drift velocity of electrons in silicon versus electric field. Repeat for GaAs.
6. Write the equations for the diffusion current densities of electrons and holes.
7. What is the Einstein relation?
8. What is the direction of the induce electric field in a semiconductor with a graded donor impurity concentration? Repeat for a graded acceptor impurity concentration.
9. Describe the Hall effect.
10. Explain why the polarity of the Hall voltage changes depending on the conductivity type (n type or p type) of the semiconductor.

PROBLEMS

(*Note:* Use the semiconductor parameters given in Appendix B if the parameters are not specifically given in a problem.)

Section 5. 1 Carrier Drift

5.1 The concentration of donor impurity atoms in silicon is $N_d = 10^{15}$ cm^{-3}. Assume an electron mobility of $\mu_n = 1300$ cm^2/V-s and a hole mobility of $\mu_p = 450$ cm^2/V-s. (*a*) Calculate the resistivity of the material. (*b*) What is the conductivity of the material?

5.2 A p-type silicon material is to have a conductivity of $\sigma = 1.80$ $(\Omega\text{-cm})^{-1}$. If the mobility values are $\mu_n = 1250$ cm^2/V-s and $\mu_p = 380$ cm^2/V-s, what must be the acceptor impurity concentration in the material?

5.3 (a) The required conductivity of an n-type silicon sample at $T = 300$ K is to be $\sigma = 10\,(\Omega\text{-cm})^{-1}$. What donor impurity concentration is required? What is the electron mobility corresponding to this impurity concentration? (b) A p-type silicon material is required to have a resistivity of $\rho = 0.20\,(\Omega\text{-cm})$. What acceptor impurity concentration is required and what is the corresponding hole mobility?

5.4 (a) The resistivity of a p-type GaAs material at $T = 300$ K is required to be $\rho = 0.35\,(\Omega\text{-cm})$. Determine the acceptor impurity concentration that is required. What is the hole mobility corresponding to this impurity concentration? (b) An n-type GaAs material is required to have a conductivity of $\sigma = 120\,(\Omega\text{-cm})^{-1}$. What donor impurity concentration is required and what is the corresponding electron mobility?

5.5 A silicon sample is 2.5 cm long and has a cross-sectional area of 0.1 cm^2. The silicon is n type with a donor impurity concentration of $N_d = 2 \times 10^{15}$ cm^{-3}. The resistance of the sample is measured and found to be $70\,\Omega$. What is the electron mobility?

5.6 Consider a homogeneous gallium arsenide semiconductor at $T = 300$ K with $N_d = 10^{16}$ cm^{-3} and $N_a = 0$. (a) Calculate the thermal-equilibrium values of electron and hole concentrations. (b) For an applied E-field of 10 V/cm, calculate the drift current density. (c) Repeat parts (a) and (b) if $N_d = 0$ and $N_a = 10^{16}$ cm^{-3}.

5.7 A silicon crystal having a cross-sectional area of 0.001 cm^2 and a length of 10^{-3} cm is connected at its ends to a 10-V battery. At $T = 300$ K, we want a current of 100 mA in the silicon. Calculate (a) the required resistance R, (b) the required conductivity, (c) the density of donor atoms to be added to achieve this conductivity, and (d) the concentration of acceptor atoms to be added to form a compensated p-type material with the conductivity given from part (b) if the initial concentration of donor atoms is $N_d = 10^{15}$ cm^{-3}.

5.8 (a) A silicon semiconductor resistor is in the shape of a rectangular bar with a cross-sectional area of 8.5×10^{-4} cm^2, a length of 0.075 cm, and is doped with a concentration of 2×10^{16} cm^{-3} boron atoms. Let $T = 300$ K. A bias of 2 volts is applied across the length of the silicon device. Calculate the current in the resistor. (b) Repeat part (a) if the length is increased by a factor of three. (c) Determine the average drift velocity of holes in parts (a) and (b).

5.9 (a) A GaAs semiconductor resistor is doped with donor impurities at a concentration of $N_d = 2 \times 10^{15}$ cm^{-3} and has a cross-sectional area of 5×10^{-5} cm^2. A current of $I = 25$ mA is induced in the resistor with an applied bias of 5 V. Determine the length of the resistor. (b) Using the results of part (a), calculate the drift velocity of the electrons. (c) If the bias applied to the resistor in part (a) increases to 20 V, determine the resulting current if the electrons are traveling at their saturation velocity of 5×10^6 cm/s.

5.10 (a) Three volts is applied across a 1-cm-long semiconductor bar. The average electron drift velocity is 10^4 cm/s. Find the electron mobility. (b) If the electron mobility in part (a) were 800 cm^2/V-s, what is the average electron drift velocity?

5.11 Use the velocity–field relations for silicon and gallium arsenide shown in Figure 5.7 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.

5.12 A perfectly compensated semiconductor is one in which the donor and acceptor impurity concentrations are exactly equal. Assuming complete ionization, determine the resistivity of silicon at $T = 300$ K in which the impurity concentrations are (a) $N_a = N_d = 10^{14}$ cm^{-3}, (b) $N_a = N_d = 10^{16}$ cm^{-3}, and (c) $N_a = N_d = 10^{18}$ cm^{-3}.

5.13 (*a*) In a p-type gallium arsenide semiconductor, the conductivity is $\sigma = 5\ (\Omega\text{-cm})^{-1}$ at $T = 300$ K. Calculate the thermal-equilibrium values of the electron and hole concentrations. (*b*) Repeat part (*a*) for n-type silicon if the resistivity is $\rho = 8\ \Omega$-cm.

5.14 In a particular semiconductor material, $\mu_n = 1000\ \text{cm}^2/\text{V-s}$, $\mu_p = 600\ \text{cm}^2/\text{V-s}$, and $N_C = N_V = 10^{19}\ \text{cm}^{-3}$. These parameters are independent of temperature. The measured conductivity of the intrinsic material is $\sigma = 10^{-6}\ (\Omega\text{-cm})^{-1}$ at $T = 300$ K. Find the conductivity at $T = 500$ K.

5.15 (*a*) Calculate the resistivity at $T = 300$ K of intrinsic (*i*) silicon, (*ii*) germanium, and (*iii*) gallium arsenide. (*b*) If rectangular semiconductor bars are fabricated using the materials in part (*a*), determine the resistance of each bar if its cross-sectional area is 85 μm^2 and length is 200 μm.

5.16 An n-type silicon material at $T = 300$ K has a conductivity of $0.25\ (\Omega\text{-cm})^{-1}$. (*a*) What is the donor impurity concentration and the corresponding electron mobility? (*b*) Determine the expected conductivity of the material at (*i*) $T = 250$ K and (*ii*) $T = 400$ K.

5.17 The conductivity of a semiconductor layer varies with depth as $\sigma(x) = \sigma_o \exp(-x/d)$ where $\sigma_o = 20\,(\Omega\text{-cm})^{-1}$ and $d = 0.3\ \mu$m. If the thickness of the semiconductor layer is $t = 1.5\ \mu$m, determine the average conductivity of this layer.

5.18 An n-type silicon resistor has a length $L = 150\ \mu$m, width $W = 7.5\ \mu$m, and thickness $T = 1\ \mu$m. A voltage of 2 V is applied across the length of the resistor. The donor impurity concentration varies linearly through the thickness of the resistor with $N_d = 2 \times 10^{16}\ \text{cm}^{-3}$ at the top surface and $N_d = 2 \times 10^{15}\ \text{cm}^{-3}$ at the bottom surface. Assume an average carrier mobility of $\mu_n = 750\ \text{cm}^2/\text{V-s}$. (*a*) What is the electric field in the resistor? (*b*) Determine the average conductivity of the silicon. (*c*) Calculate the current in the resistor. (*d*) Determine the current density near the top surface and the current density near the bottom surface.

5.19 Consider silicon doped at impurity concentrations of $N_d = 2 \times 10^{16}\ \text{cm}^{-3}$ and $N_a = 0$. An empirical expression relating electron drift velocity to electric field is given by

$$v_d = \frac{\mu_{n0}\text{E}}{\sqrt{1 + \left(\dfrac{\mu_{n0}\text{E}}{v_{sat}}\right)^2}}$$

where $\mu_{n0} = 1350\ \text{cm}^2/\text{V-s}$, $v_{sat} = 1.8 \times 10^7\ \text{cm/s}$, and E is given in V/cm. Plot electron drift current density (magnitude) versus electric field (log–log scale) over the range $0 \leq \text{E} \leq 10^6$ V/cm.

5.20 Consider silicon at $T = 300$ K. Assume the electron mobility is $\mu_n = 1350\ \text{cm}^2/\text{V-s}$. The kinetic energy of an electron in the conduction band is $(1/2)m_n^* v_d^2$, where m_n^* is the effective mass and v_d is the drift velocity. Determine the kinetic energy of an electron in the conduction band if the applied electric field is (*a*) 10 V/cm and (*b*) 1 kV/cm.

5.21 Consider a semiconductor that is uniformly doped with $N_d = 10^{14}\ \text{cm}^{-3}$ and $N_a = 0$, with an applied electric field of E = 100 V/cm. Assume that $\mu_n = 1000\ \text{cm}^2/\text{V-s}$ and $\mu_p = 0$. Also assume the following parameters:

$$N_c = 2 \times 10^{19}\,(T/300)^{3/2}\ \text{cm}^{-3}$$
$$N_v = 1 \times 10^{19}\,(T/300)^{3/2}\ \text{cm}^{-3}$$
$$E_g = 1.10\ \text{eV}$$

(*a*) Calculate the electric-current density at $T = 300$ K. (*b*) At what temperature will this current increase by 5 percent? (Assume the mobilities are independent of temperature.)

5.22 A semiconductor material has electron and hole mobilities μ_n and μ_p, respectively. When the conductivity is considered as a function of the hole concentration p_0, (*a*) show that the minimum value of conductivity, σ_{min}, can be written as

$$\sigma_{min} = \frac{2\sigma_i(\mu_n \mu_p)^{1/2}}{(\mu_n + \mu_p)}$$

where σ_i is the intrinsic conductivity, and (b) show that the corresponding hole concentration is $p_0 = n_i(\mu_n/\mu_p)^{1/2}$.

5.23 Consider three samples of silicon at $T = 300$ K. The n-type sample is doped with arsenic atoms to a concentration of $N_d = 5 \times 10^{16}$ cm^{-3}. The p-type sample is doped with boron atoms to a concentration of $N_a = 2 \times 10^{16}$ cm^{-3}. The compensated sample is doped with both the donors and acceptors described in the n-type and p-type samples. (*a*) Find the equilibrium electron and hole concentrations in each sample, (*b*) determine the majority carrier mobility in each sample, (*c*) calculate the conductivity of each sample, (*d*) and determine the electric field required in each sample to induce a drift current density of $J = 120$ A/cm^2.

5.24 Three scattering mechanisms are present in a particular semiconductor material. If only the first scattering mechanism were present, the mobility would be $\mu_1 = 2000$ cm^2/V-s, if only the second mechanism were present, the mobility would be $\mu_2 = 1500$ cm^2/V-s, and if only the third mechanism were present, the mobility would be $\mu_3 = 500$ cm^2/V-s. What is the net mobility?

5.25 Assume that the mobility of electrons in silicon at $T = 300$ K is $\mu_n = 1300$ cm^2/V-s. Also assume that the mobility is limited by lattice scattering and varies as $T^{-3/2}$. Determine the electron mobility at (*a*) $T = 200$ K and (*b*) $T = 400$ K.

5.26 Two scattering mechanisms exist in a semiconductor. If only the first mechanism were present, the mobility would be 250 cm^2/V-s. If only the second mechanism were present, the mobility would be 500 cm^2/V-s. Determine the mobility when both scattering mechanisms exist at the same time.

5.27 The effective density of states functions in silicon can be written in the form

$$N_c = 2.8 \times 10^{19}\left(\frac{T}{300}\right)^{3/2} \qquad N_v = 1.04 \times 10^{19}\left(\frac{T}{300}\right)^{3/2}$$

Assume the mobilities are given by

$$\mu_n = 1350\left(\frac{T}{300}\right)^{-3/2} \qquad \mu_p = 480\left(\frac{T}{300}\right)^{-3/2}$$

Assume the bandgap energy is $E_g = 1.12$ eV and independent of temperature. Plot the intrinsic conductivity as a function of T over the range $200 \leq T \leq 600$ K.

5.28 (*a*) Assume that the electron mobility in an n-type semiconductor is given by

$$\mu_n = \frac{1350}{\left(1 + \dfrac{N_d}{5 \times 10^{16}}\right)^{1/2}} \text{ cm}^2\text{/V-s}$$

where N_d is the donor concentration in cm^{-3}. Assuming complete ionization, plot the conductivity as a function of N_d over the range $10^{15} \leq N_d \leq 10^{18}$ cm^{-3}. (*b*) Compare the results of part (*a*) to that if the mobility were assumed to be a constant equal to 1350 cm^2/V-s. (*c*) If an electric field of $E = 10$ V/cm is applied to the semiconductor, plot the electron drift current density of parts (*a*) and (*b*).

Section 5.2 Carrier Diffusion

5.29 Consider a sample of silicon at $T = 300$ K. Assume that the electron concentration varies linearly with distance, as shown in Figure P5.29. The diffusion current density is found to be $J_n = 0.19$ A/cm^2. If the electron diffusion coefficient is $D_n = 25$ cm^2/s, determine the electron concentration at $x = 0$.

5.30 The steady-state electron distribution in silicon can be approximated by a linear function of x. The maximum electron concentration occurs at $x = 0$ and is $n(0) = 2 \times 10^{16}$ cm^{-3}. At $x = 0.012$ cm, the electron concentration is 5×10^{15} cm^{-3}. If the electron diffusion coefficient is $D_n = 27$ cm^2/s, determine the electron diffusion current density.

5.31 The electron diffusion current density in a semiconductor is a constant and is given by $J_n = -2$ A/cm^2. The electron concentration at $x = 0$ is $n(0) = 10^{15}$ cm^{-3}. (*a*) Calculate the electron concentration at $x = 20$ μm if the material is silicon with $D_n = 30$ cm^2/s. (*b*) Repeat part (*a*) if the material is GaAs with $D_n = 230$ cm^2/s.

5.32 The hole concentration in p-type GaAs is given by $p(x) = 10^{16}(1 + x/L)^2$ cm^{-3} for $-L \le x \le 0$ where $L = 12$ μm. The hole diffusion coefficient is $D_p = 10$ cm^2/s. Calculate the hole diffusion current density at (*a*) $x = 0$, (*b*) $x = -6$ μm, and (*c*) $x = -12$ μm.

5.33 In silicon, the electron concentration is given by $n(x) = 10^{15}e^{-x/L_n}$ cm^{-3} for $x \ge 0$ and the hole concentration is given by $p(x) = 5 \times 10^{15}e^{+x/L_p}$ cm^{-3} for $x \le 0$. The parameter values are $L_n = 2 \times 10^{-3}$ cm and $L_P = 5 \times 10^{-4}$ cm. The electron and hole diffusion coefficients are $D_n = 25$ cm^2/s and $D_P = 10$ cm^2/s, respectively. The total current density is defined as the sum of the electron and hole diffusion current densities at $x = 0$. Calculate the total current density.

5.34 The concentration of holes in a semiconductor is given by $p(x) = 5 \times 10^{15}e^{-x/L_p}$ cm^{-3} for $x \ge 0$. Determine the hole diffusion current density at (*a*) $x = 0$ and (*b*) $x = L_p$ if the material is (*i*) silicon with $D_p = 10$ cm^2/s and $L_p = 50$ μm, and (*ii*) germanium with $D_p = 48$ cm^2/s and $L_p = 22.5$ μm.

5.35 The electron concentration in silicon at $T = 300$ K is given by

$$n(x) = 10^{16}\exp\left(\frac{-x}{18}\right) \text{ cm}^{-3}$$

where x is measured in μm and is limited to $0 \le x \le 25$ μm. The electron diffusion coefficient is $D_n = 25$ cm^2/s and the electron mobility is $\mu_n = 960$ cm^2/V-s. The total electron current density through the semiconductor is constant and equal to

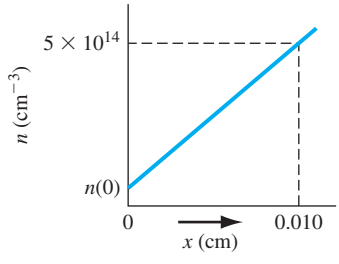


Figure P5.29 | Figure for Problem 5.29.

$J_n = -40$ A/cm^2. The electron current has both diffusion and drift current components. Determine the electric field as a function of x which must exist in the semiconductor.

5.36 The total current in a semiconductor is constant and equal to $J = -10$ A/cm^2. The total current is composed of a hole drift current and electron diffusion current. Assume that the hole concentration is a constant and equal to 10^{16} cm^{-3} and assume that the electron concentration is given by $n(x) = 2 \times 10^{15} e^{-x/L}$ cm^{-3} where $L = 15$ μm. The electron diffusion coefficient is $D_n = 27$ cm^2/s and the hole mobility is $\mu_p = 420$ cm^2/V-s. Calculate (*a*) the electron diffusion current density for $x > 0$, (*b*) the hole drift current density for $x > 0$, and (*c*) the required electric field for $x > 0$.

***5.37** A constant electric field, E = 12 V/cm, exists in the $+x$ direction of an n-type gallium arsenide semiconductor for $0 \le x \le 50$ μm. The total current density is a constant and is $J = 100$ A/cm^2. At $x = 0$, the drift and diffusion currents are equal. Let $T = 300$ K and $\mu_n = 8000$ cm^2/V-s. (*a*) Determine the expression for the electron concentration $n(x)$. (*b*) Calculate the electron concentration at $x = 0$ and at $x = 50$ μm. (*c*) Calculate the drift and diffusion current densities at $x = 50$ μm.

***5.38** In n-type silicon, the Fermi energy level varies linearly with distance over a short range. At $x = 0$, $E_F - E_{Fi} = 0.4$ eV and, at $x = 10^{-3}$ cm, $E_F - E_{Fi} = 0.15$ eV. (*a*) Write the expression for the electron concentration over the distance. (*b*) If the electron diffusion coefficient is $D_n = 25$ cm^2/s, calculate the electron diffusion current density at (*i*) $x = 0$ and (*ii*) $x = 5 \times 10^{-4}$ cm.

***5.39** (*a*) The electron concentration in a semiconductor is given by $n = 10^{16}(1 - x/L)$ cm^{-3} for $0 \le x \le L$, where $L = 10$ μm. The electron mobility and diffusion coefficient are $\mu_n = 1000$ cm^2/V-s and $D_n = 25.9$ cm^2/s. An electric field is applied such that the total electron current density is a constant over the given range of x and is $J_n = -80$ A/cm^2. Determine the required electric field versus distance function. (*b*) Repeat part (*a*) if $J_n = -20$ A/cm^2.

Section 5.3 Graded Impurity Distribution

5.40 Consider an n-type semiconductor at $T = 300$ K in thermal equilibrium (no current). Assume that the donor concentration varies as $N_d(x) = N_{d0} e^{-x/L}$ over the range $0 \le x \le L$ where $N_{d0} = 10^{16}$ cm^{-3} and $L = 10$ μm. (*a*) Determine the electric field as a function of x for $0 \le x \le L$. (*b*) Calculate the potential difference between $x = 0$ and $x = L$ (with the potential at $x = 0$ being positive with respect to that at $x = L$).

5.41 Using the data in Example 5.6, calculate the potential difference between $x = 0$ and $x = 1$ μm.

5.42 Determine the doping profile in a semiconductor at $T = 300$ K that will induce a constant electric field of 500 V/cm over a length of 0.1 cm.

***5.43** In GaAs, the donor impurity concentration varies as $N_{d0} \exp(-x/L)$ for $0 \le x \le L$, where $L = 0.1$ μm and $N_{d0} = 5 \times 10^{16}$ cm^{-3}. Assume $\mu_n = 6000$ cm^2/V-s and $T = 300$ K. (*a*) Derive the expression for the electron diffusion current density versus distance over the given range of x. (*b*) Determine the induced electric field that generates a drift current density that compensates the diffusion current density.

5.44 (*a*) Consider the electron mobility in silicon for $N_d = 10^{17}$cm^{-3} from Figure 5.2a. Calculate and plot the electron diffusion coefficient versus temperature over the range $-50 \le T \le 200$°C. (*b*) Repeat part (*a*) if the electron diffusion coefficient is given by $D_n = (0.0259)\mu_n$ for all temperatures. What conclusion can be made about the temperature dependence of the diffusion coefficient?

*Asterisks next to problems indicate problems that are more difficult.

5.45 Consider a semiconductor at $T = 300$ K. (*a*) (*i*) Determine the electron diffusion coefficient if the electron mobility is $\mu_n = 1150$ cm^2/V-s. (*ii*) Repeat (*i*) of part (*a*) if the electron mobility is $\mu_n = 6200$ cm^2/V-s. (*b*) (*i*) Determine the hole mobility if the hole diffusion coefficient is $D_p = 8$ cm^2/s. (*ii*) Repeat (*i*) of part (*b*) if the hole diffusion coefficient is $D_p = 35$ cm^2/s.

Section 5.4 The Hall Effect

(*Note:* Refer to Figure 5.13 for the geometry of the Hall effect.)

5.46 Silicon, at $T = 300$ K, is uniformly doped with phosphorus atoms at a concentration of 2×10^{16} cm^{-3}. A Hall device has the same geometrical dimensions as given in Example 5.8. The current is $I_x = 1.2$ mA and the magnetic field is $B_Z = 500$ gauss $= 5 \times 10^{-2}$ tesla. Determine (*a*) the Hall voltage and (*b*) the Hall field.

5.47 Germanium is doped with 5×10^{15} donor atoms per cm^3 at $T = 300$ K. The dimensions of the Hall device are $d = 5 \times 10^{-3}$ cm, $W = 2 \times 10^{-2}$ cm, and $L = 10^{-1}$ cm. The current is $I_x = 250\ \mu A$, the applied voltage is $V_x = 100$ mV, and the magnetic flux density is $B_z = 500$ gauss $= 5 \times 10^{-2}$ tesla. Calculate: (*a*) the Hall voltage, (*b*) the Hall field, and (*c*) the carrier mobility.

5.48 A semiconductor Hall device at $T = 300$ K has the following geometry: $d = 10^{-3}$cm, $W = 10^{-2}$ cm, and $L = 10^{-1}$cm. The following parameters are measured: $I_X = 0.50$ mA, $V_X = 15$ V, $V_H = -5.2$ mV, and $B_Z = 0.10$ tesla. Determine the (*a*) conductivity type, (*b*) majority carrier concentration, and (*c*) majority carrier mobility.

5.49 Consider silicon at $T = 300$ K. A Hall effect device is fabricated with the following geometry: $d = 5 \times 10^{-3}$ cm, $W = 5 \times 10^{-2}$ cm, and $L = 0.50$ cm. The electrical parameters measured are: $I_x = 0.50$ mA, $V_x = 1.25$ V, and $B_z = 650$ gauss $= 6.5 \times 10^{-2}$ tesla. The Hall field is $E_H = -16.5$ mV/cm. Determine (*a*) the Hall voltage, (*b*) the conductivity type, (*c*) the majority carrier concentration, and (*d*) the majority carrier mobility.

5.50 Consider a gallium arsenide sample at $T = 300$ K. A Hall effect device has been fabricated with the following geometry: $d = 0.0\ 1$ cm, $W = 0.05$ cm, and $L = 0.5$ cm. The electrical parameters are: $I_x = 2.5$ mA, $V_x = 2.2$ V, and $B_z = 2.5 \times 10^{-2}$ tesla. The Hall voltage is $V_H = -4.5$ mV. Find: (*a*) the conductivity type, (*b*) the majority carrier concentration, (*c*) the mobility, and (*d*) the resistivity.

Summary and Review

5.51 An n-type silicon semiconductor resistor is to be designed so that it carries a current of 5 mA with an applied voltage of 5 V. (*a*) If $N_d = 3 \times 10^{14}$ cm^{-3} and $N_a = 0$, design a resistor to meet the required specifications. (*b*) If $N_d = 3 \times 10^{16}$ cm^{-3} and $N_a = 2.5 \times 10^{16}$ cm^{-3}, redesign the resistor. (*c*) Discuss the relative lengths of the two designs compared to the doping concentration. Is there a linear relationship?

5.52 In fabricating a Hall effect device, the two points at which the Hall voltage is measured may not be lined up exactly perpendicular to the current I_x (see Figure 5.13). Discuss the effect this misalignment will have on the Hall voltage. Show that a valid Hall voltage can be obtained from two measurements: first with the magnetic field in the $+z$ direction, and then in the $-z$ direction.

5.53 Another technique for determining the conductivity type of a semiconductor is called the hot probe method. It consists of two probes and an ammeter that indicates the

direction of current. One probe is heated and the other is at room temperature. No voltage is applied, but a current will exist when the probes touch the semiconductor. Explain the operation of this hot probe technique and sketch a diagram indicating the direction of current for p- and n-type semiconductor samples.

READING LIST

*1. Bube, R. H. *Electrons in Solids: An Introductory Survey,* 3rd ed. San Diego, CA: Academic Press, 1992.

2. Caughey, D. M., and R. E. Thomas. "Carrier Mobilities in Silicon Empirically, Related to Doping and Field." *Proc. IEEE* 55 (1967), p. 2192.

3. Dimitrijev, S. *Principles of Semiconductor Devices.* New York: Oxford University, 2006.

4. Kano, K. *Semiconductor Devices*. Upper Saddle River, NJ: Prentice Hall, 1998.

*5. Lundstrom, M. *Fundamentals of Carrier Transport*. Vol. X of *Modular Series on Solid State Devices.* Reading, MA: Addison-Wesley, 1990.

6. Muller, R. S., and T. I. Kamins. *Device Electronics for Integrated Circuits,* 2nd ed. New York: Wiley, 1986.

7. Navon, D. H. *Semiconductor Microdevices and Materials*. New York: Holt, Rinehart & Winston, 1986.

8. Pierret, R. F. *Semiconductor Device Fundamentals*. Reading, MA: Addison-Wesley, 1996.

9. Shur, M. *Introduction to Electronic Devices*. New York: John Wiley and Sons, 1996.

*10. Shur, M. *Physics of Semiconductor Devices.* Englewood Cliffs, NJ: Prentice Hall, 1990.

11. Singh, J. *Semiconductor Devices: An Introduction*. New York: McGraw-Hill, 1994.

12. Singh, J. *Semiconductor Devices: Basic Principles*. New York: John Wiley and Sons, 2001.

13. Streetman, B. G., and S. K. Banerjee. *Solid State Electronic Devices*, 6th ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2006.

14. Sze, S. M. and K. K. Ng. *Physics of Semiconductor Devices,* 3rd ed. Hoboken, NJ: John Wiley and Sons, 2007.

15. Sze, S. M. *Semiconductor Devices: Physics and Technology,* 2nd ed. New York: John Wiley and Sons, 2001.

*16. van der Ziel, A. *Solid State Physical Electronics,* 2nd ed. Englewood Cliffs, NJ: Prentice Hall, 1968.

17. Wang, S. *Fundamentals of Semiconductor Theory and Device Physics*. Englewood Cliffs, NJ: Prentice Hall, 1989.

18. Yang, E. S. *Microelectronic Devices*. New York: McGraw-Hill, 1988.

*Indicates references that are at an advanced level compared to this text.

CHAPTER 6

Nonequilibrium Excess Carriers in Semiconductors

Our discussion of the physics of semiconductors in Chapter 4 was based on thermal equilibrium. When a voltage is applied or a current exists in a semiconductor device, the semiconductor is operating under nonequilibrium conditions. In our discussion of current transport in Chapter 5, we did not address nonequilibrium conditions but implicitly assumed that equilibrium was not significantly disturbed. Excess electrons in the conduction band and excess holes in the valence band may exist in addition to the thermal-equilibrium concentrations if an external excitation is applied to the semiconductor. In this chapter, we discuss the behavior of nonequilibrium electron and hole concentrations as functions of time and space coordinates.

Excess electrons and excess holes do not move independently of each other. These excess carriers diffuse, drift, and recombine with the same effective diffusion coefficient, drift mobility, and lifetime. This phenomenon is called ambipolar transport. We develop the ambipolar transport equation that describes the behavior of excess electrons and holes. Excess carriers dominate the electrical properties of a semiconductor material, and the behavior of excess carriers is fundamental to the operation of semiconductor devices. ■

6.0 | PREVIEW

In this chapter, we will:

- Describe the process of generation and recombination of excess carriers in a semiconductor.
- Define the recombination rate and generation rate of excess carriers, and define the excess carrier lifetime.
- Discuss why excess electrons and excess holes do not move independently of each other. The movement of excess carriers is called *ambipolar transport*, and the ambipolar transport equation is derived.

- Apply the ambipolar transport equation to various situations to determine the time behavior and spatial behavior of excess carriers.
- Define the quasi-Fermi energy level.
- Analyze the effect of defects in a semiconductor on the excess carrier lifetime.
- Analyze the effect of defects at a semiconductor surface on the excess carrier concentration.

6.1 | CARRIER GENERATION AND RECOMBINATION

In this chapter, we discuss carrier generation and recombination, which we can define as follows: *generation* is the process whereby electrons and holes are created, and *recombination* is the process whereby electrons and holes are annihilated.

Any deviation from thermal equilibrium will tend to change the electron and hole concentrations in a semiconductor. A sudden increase in temperature, for example, will increase the rate at which electrons and holes are thermally generated so that their concentrations will change with time until new equilibrium values are reached. An external excitation, such as light (a flux of photons), can also generate electrons and holes, creating a nonequilibrium condition. To understand the generation and recombination processes, we first consider direct band-to-band generation and recombination, and then, later, the effect of allowed electronic energy states within the bandgap, referred to as traps or recombination centers.

6.1.1 The Semiconductor in Equilibrium

We have determined the thermal-equilibrium concentration of electrons and holes in the conduction and valence bands, respectively. In thermal equilibrium, these concentrations are independent of time. However, electrons are continually being thermally excited from the valence band into the conduction band by the random nature of the thermal process. At the same time, electrons moving randomly through the crystal in the conduction band may come in close proximity to holes and "fall" into the empty states in the valence band. This recombination process annihilates both the electron and hole. Since the net carrier concentrations are independent of time in thermal equilibrium, the rate at which electrons and holes are generated and the rate at which they recombine must be equal. The generation and recombination processes are schematically shown in Figure 6.1.

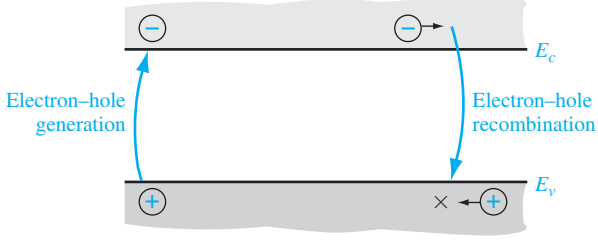


Figure 6.1 | Electron–hole generation and recombination.

Let G_{n0} and G_{p0} be the thermal-generation rates of electrons and holes, respectively, given in units of #/cm^3-s. For the direct band-to-band generation, the electrons and holes are created in pairs, so we must have that

$$G_{n0} = G_{p0} \tag{6.1}$$

Let R_{n0} and R_{p0} be the recombination rates of electrons and holes, respectively, for a semiconductor in thermal equilibrium, again given in units of #/cm^3-s. In direct band-to-band recombination, electrons and holes recombine in pairs, so that

$$R_{n0} = R_{p0} \tag{6.2}$$

In thermal equilibrium, the concentrations of electrons and holes are independent of time; therefore, the generation and recombination rates are equal, so we have

$$G_{n0} = G_{p0} = R_{n0} = R_{p0} \tag{6.3}$$

6.1.2 Excess Carrier Generation and Recombination

Additional notation is introduced in this chapter. Table 6.1 lists some of the more pertinent symbols used throughout the chapter. Other symbols will be defined as we advance through the chapter.

Electrons in the valence band may be excited into the conduction band when, for example, high-energy photons are incident on a semiconductor. When this happens, not only is an electron created in the conduction band, but a hole is created in the valence band; thus, an electron–hole pair is generated. The additional electrons and holes created are called *excess electrons* and *excess holes.*

The excess electrons and holes are generated by an external force at a particular rate. Let g'_n be the generation rate of excess electrons and g'_p be that of excess holes. These generation rates also have units of #/cm^3-s. For the direct band-to-band generation, the excess electrons and holes are also created in pairs, so we must have

$$g'_n = g'_p \tag{6.4}$$

Table 6.1 | Relevant notation used in Chapter 6

Symbol	Definition
n_0, p_0	Thermal-equilibrium electron and hole concentrations (independent of time and also usually position)
n, p	Total electron and hole concentrations (may be functions of time and/or position)
$\delta n = n - n_0$ $\delta p = p - p_0$	Excess electron and hole concentrations (may be functions of time and/or position)
g'_n, g'_p	Excess electron and hole generation rates
R'_n, R'_p	Excess electron and hole recombination rates
τ_{n0}, τ_{p0}	Excess minority carrier electron and hole lifetimes

When excess electrons and holes are created, the concentration of electrons in the conduction band and of holes in the valence band increase above their thermal-equilibrium value. We may write

$$n = n_0 + \delta n \tag{6.5a}$$

and

$$p = p_0 + \delta p \tag{6.5b}$$

where n_0 and p_0 are the thermal-equilibrium concentrations, and δn and δp are the excess electron and hole concentrations. Figure 6.2 shows the excess electron–hole generation process and the resulting carrier concentrations. The external force has perturbed the equilibrium condition so that the semiconductor is no longer in thermal equilibrium. We may note from Equations (6.5a) and (6.5b) that, in a nonequilibrium condition, $np \neq n_0 p_0 = n_i^2$.

A steady-state generation of excess electrons and holes will not cause a continual buildup of the carrier concentrations. As in the case of thermal equilibrium, an electron in the conduction band may "fall down" into the valence band, leading to the process of excess electron–hole recombination. Figure 6.3 shows this process. The

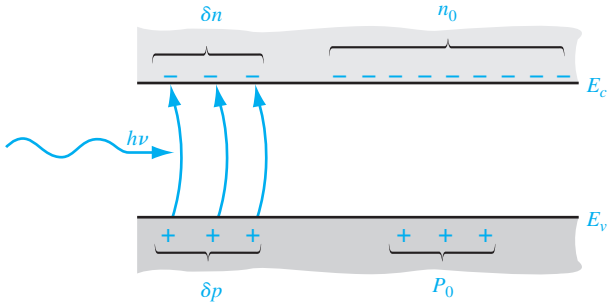


Figure 6.2 | Creation of excess electron and hole densities by photons.

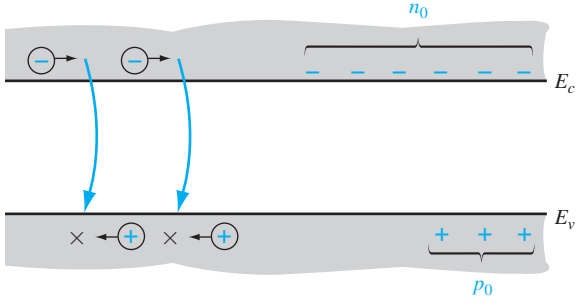


Figure 6.3 | Recombination of excess carriers reestablishing thermal equilibrium.

recombination rate for excess electrons is denoted by R'_n and for excess holes by R'_p. Both parameters have units of #/cm^3-s. The excess electrons and holes recombine in pairs, so the recombination rates must be equal. We can then write

$$R'_n = R'_p \tag{6.6}$$

In the direct band-to-band recombination that we are considering, the recombination occurs spontaneously; thus, the probability of an electron and hole recombining is constant with time. The rate at which electrons recombine must be proportional to the electron concentration and must also be proportional to the hole concentration. If there are no electrons or holes, there can be no recombination.

The net rate of change in the electron concentration can be written as

$$\frac{dn(t)}{dt} = \alpha_r\left[n_i^2 - n(t)p(t)\right] \tag{6.7}$$

where

$$n(t) = n_0 + \delta n(t) \tag{6.8a}$$

and

$$p(t) = p_0 + \delta p(t) \tag{6.8b}$$

The first term, $\alpha_r n_i^2$, in Equation (6.7) is the thermal-equilibrium generation rate. Since excess electrons and holes are created and recombine in pairs, we have that $\delta n(t) = \delta p(t)$. (Excess electron and hole concentrations are equal so we can simply use the phrase *excess carriers* to mean either.) The thermal-equilibrium parameters, n_0 and p_0, are independent of time; therefore, Equation (6.7) becomes

$$\begin{aligned}\frac{d(\delta n(t))}{dt} &= \alpha_r\left[n_i^2 - (n_0 + \delta n(t))(p_0 + \delta p(t))\right] \\ &= -\alpha_r \delta n(t)[(n_0 + p_0) + \delta n(t)]\end{aligned} \tag{6.9}$$

Equation (6.9) can easily be solved if we impose the condition of *low-level injection.* Low-level injection puts limits on the magnitude of the excess carrier concentration compared with the thermal-equilibrium carrier concentrations. In an extrinsic n-type material, we generally have $n_0 \gg p_0$ and, in an extrinsic p-type material, we generally have $p_0 \gg n_0$. Low-level injection means that the excess carrier concentration is much less than the thermal-equilibrium majority carrier concentration. Conversely, high-level injection occurs when the excess carrier concentration becomes comparable to or greater than the thermal-equilibrium majority carrier concentrations.

If we consider a p-type material ($p_0 \gg n_0$) under low-level injection ($\delta n(t) \ll p_0$), then Equation (6.9) becomes

$$\frac{d(\delta n(t))}{dt} = -\alpha_r p_0 \delta n(t) \tag{6.10}$$

The solution to the equation is an exponential decay from the initial excess concentration, or

$$\delta n(t) = \delta n(0) e^{-\alpha_r p_0 t} = \delta n(0) e^{-t/\tau_{n0}} \tag{6.11}$$

where $\tau_{n0} = (\alpha_r p_0)^{-1}$ and is a constant for the low-level injection. Equation (6.11) describes the decay of excess minority carrier electrons so that τ_{n0} is often referred to as the *excess minority carrier lifetime.*[1]

The recombination rate—which is defined as a positive quantity—of excess minority carrier electrons can be written, using Equation (6.10), as

$$R'_n = \frac{-d(\delta n(t))}{dt} = +\alpha_r p_0 \delta n(t) = \frac{\delta n(t)}{\tau_{n0}} \tag{6.12}$$

For the direct band-to-band recombination, the excess majority carrier holes recombine at the same rate, so that for the p-type material

$$\boxed{R'_n = R'_p = \frac{\delta n(t)}{\tau_{n0}}} \tag{6.13}$$

In the case of an n-type material ($n_0 \gg p_0$) under low-level injection ($\delta n(t) \ll n_0$), the decay of minority carrier holes occurs with a time constant $\tau_{p0} = (\alpha_r n_0)^{-1}$, where τ_{p0} is also referred to as the excess minority carrier lifetime. The recombination rate of the majority carrier electrons will be the same as that of the minority carrier holes, so we have

$$\boxed{R'_n = R'_p = \frac{\delta n(t)}{\tau_{p0}}} \tag{6.14}$$

The generation rates of excess carriers are not functions of electron or hole concentrations. In general, the generation and recombination rates may be functions of the space coordinates and time.

EXAMPLE 6.1

Objective: Determine the behavior of excess carriers as a function of time.

Assume that excess carriers have been generated uniformly in a semiconductor to a concentration of $\delta n(0) = 10^{15}$ cm^{-3}. The forcing function generating the excess carriers turns off at time $t = 0$. Assuming the excess carrier lifetime is $\tau_{n0} = 10^{-6}$ s, determine $\delta n(t)$ for $t > 0$.

■ Solution

From Equation (6.11), we have

$$\delta n(t) = \delta n(0) e^{-t/\tau_{n0}} = 10^{15}\, e^{-t/10^{-6}} \text{ cm}^{-3}$$

For example, at: $t = 0$, $\quad \delta n = 10^{15}$ cm^{-3}

$t = 1\mu$s, $\quad \delta n = 10^{15}\, e^{-1/1} = 3.68 \times 10^{14}$ cm^{-3}

$t = 4\mu$s, $\quad \delta n = 10^{15}\, e^{-4/1} = 1.83 \times 10^{13}$ cm^{-3}

$t = 10\mu$s, $\quad \delta n = 10^{15}\, e^{-10/1} = 4.54 \times 10^{10}$ cm^{-3}

■ Comment

These results simply demonstrate the exponential decay of excess carriers with time after an excitation source is removed.

[1] In Chapter 5 we defined τ as a mean time between collisions. We define τ here as the mean time before a recombination event occurs. The two parameters are not related.

EXERCISE PROBLEM

Ex 6.1 Using the parameters in Example 6.1, calculate the recombination rate of the excess carriers for (*a*) $t = 0$, (*b*) $t = 1\mu$s, (*c*) $t = 4\mu$s, and (*d*) $t = 10\mu$s.

[Ans. (*a*) 10^{21} cm^{-3}s^{-1}; (*b*) 3.68×10^{20} cm^{-3}s^{-1}; (*c*) 1.83×10^{19} cm^{-3}s^{-1}; (*d*) 4.54×10^{16} cm^{-3}s^{-1}]

6.2 | CHARACTERISTICS OF EXCESS CARRIERS

The generation and recombination rates of excess carriers are important parameters, but how the excess carriers behave with time and in space in the presence of electric fields and density gradients is of equal importance. As mentioned at the beginning of this chapter, the excess electrons and holes do not move independently of each other, but they diffuse and drift with the same effective diffusion coefficient and with the same effective mobility. This phenomenon is called ambipolar transport. The question that must be answered is what is the effective diffusion coefficient and what is the effective mobility that characterizes the behavior of these excess carriers? To answer these questions, we must develop the continuity equations for the carriers and then develop the ambipolar transport equations.

The final results show that, for an extrinsic semiconductor under low injection (this concept will be defined in the analysis), the effective diffusion coefficient and mobility parameters are those of the minority carrier. This result is thoroughly developed in the following derivations. As will be seen in the following chapters, the behavior of the excess carriers has a profound impact on the characteristics of semiconductor devices.

6.2.1 Continuity Equations

The continuity equations for electrons and holes are developed in this section. Figure 6.4 shows a differential volume element in which a one-dimensional hole-particle flux is entering the differential element at x and is leaving the element at $x + dx$. The parameter F_{px}^{+} is the hole-particle flux, or flow, and has units of number of holes/cm^2-s. For the x component of the particle current density shown, we may write

$$F_{px}^{+}(x + dx) = F_{px}^{+}(x) + \frac{\partial F_{px}^{+}}{\partial x} \cdot dx \tag{6.15}$$

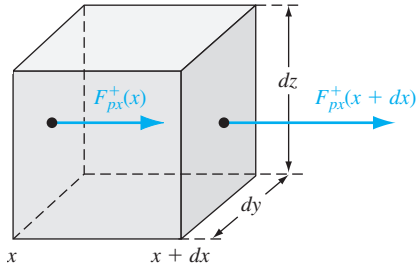


Figure 6.4 | Differential volume showing x component of the hole-particle flux.

This equation is a Taylor expansion of $F_{px}^{+}(x + dx)$, where the differential length dx is small, so that only the first two terms in the expansion are significant. The net increase in the number of holes per unit time within the differential volume element due to the x-component of hole flux is given by

$$\frac{\partial p}{\partial t}\,dx\,dy\,dz = [F_{px}^{+}(x) - F_{px}^{+}(x + dx)]\,dy\,dz = -\frac{\partial F_{px}^{+}}{\partial x}\,dx\,dy\,dz \tag{6.16}$$

If $F_{px}^{+}(x) > F_{px}^{+}(x + dx)$, for example, there will be a net increase in the number of holes in the differential volume element with time. If we generalize to a three-dimensional hole flux, then the right side of Equation (6.16) may be written as $-\nabla \cdot F_{p}^{+}\,dx\,dy\,dz$, where $\nabla \cdot F_{p}^{+}$ is the divergence of the flux vector. We will limit ourselves to a one-dimensional analysis.

The generation rate and recombination rate of holes will also affect the hole concentration in the differential volume. The net increase in the number of holes per unit time in the differential volume element is then given by

$$\frac{\partial p}{\partial t}\,dx\,dy\,dz = -\frac{\partial F_{p}^{+}}{\partial x}\,dx\,dy\,dz + g_p\,dx\,dy\,dz - \frac{p}{\tau_{pt}}\,dx\,dy\,dz \tag{6.17}$$

where p is the density of holes. The first term on the right side of Equation (6.17) is the increase in the number of holes per unit time due to the hole flux, the second term is the increase in the number of holes per unit time due to the generation of holes, and the last term is the decrease in the number of holes per unit time due to the recombination of holes. The recombination rate for holes is given by p/τ_{pt} where τ_{pt} includes the thermal-equilibrium carrier lifetime and the excess carrier lifetime.

If we divide both sides of Equation (6.17) by the differential volume $dx\,dy\,dz$, the net increase in the hole concentration per unit time is

$$\frac{\partial p}{\partial t} = -\frac{\partial F_{p}^{+}}{\partial x} + g_p - \frac{p}{\tau_{pt}} \tag{6.18}$$

Equation (6.18) is known as the continuity equation for holes.

Similarly, the one-dimensional continuity equation for electrons is given by

$$\frac{\partial n}{\partial t} = -\frac{\partial F_{n}^{-}}{\partial x} + g_n - \frac{n}{\tau_{nt}} \tag{6.19}$$

where F_{n}^{-} is the electron-particle flow, or flux, also given in units of number of electrons/cm^2-s.

6.2.2 Time-Dependent Diffusion Equations

In Chapter 5, we derived the hole and electron current densities, which are given, in one dimension, by

$$J_p = e\mu_p pE - eD_p\frac{\partial p}{\partial x} \tag{6.20}$$

and

$$J_n = e\mu_n nE + eD_n\frac{\partial n}{\partial x} \tag{6.21}$$

If we divide the hole current density by $(+e)$ and the electron current density by $(-e)$, we obtain each particle flux. These equations become

$$\frac{J_p}{(+e)} = F_p^+ = \mu_p p\mathrm{E} - D_p\frac{\partial p}{\partial x} \tag{6.22}$$

and

$$\frac{J_n}{(-e)} = F_n^- = -\mu_n n\mathrm{E} - D_n\frac{\partial n}{\partial x} \tag{6.23}$$

Taking the divergence of Equations (6.22) and (6.23), and substituting back into the continuity equations of (6.18) and (6.19), we obtain

$$\frac{\partial p}{\partial t} = -\mu_p\frac{\partial(p\mathrm{E})}{\partial x} + D_p\frac{\partial^2 p}{\partial x^2} + g_p - \frac{p}{\tau_{pt}} \tag{6.24}$$

and

$$\frac{\partial n}{\partial t} = +\mu_n\frac{\partial(n\mathrm{E})}{\partial x} + D_n\frac{\partial^2 n}{\partial x^2} + g_n - \frac{n}{\tau_{nt}} \tag{6.25}$$

Keeping in mind that we are limiting ourselves to a one-dimensional analysis, we can expand the derivative of the product as

$$\frac{\partial(p\mathrm{E})}{\partial x} = \mathrm{E}\frac{\partial p}{\partial x} + p\frac{\partial \mathrm{E}}{\partial x} \tag{6.26}$$

In a more generalized three-dimensional analysis, Equation (6.26) would have to be replaced by a vector identity. Equations (6.24) and (6.25) can be written in the form

$$D_p\frac{\partial^2 p}{\partial x^2} - \mu_p\left(\mathrm{E}\frac{\partial p}{\partial x} + p\frac{\partial \mathrm{E}}{\partial x}\right) + g_p - \frac{p}{\tau_{pt}} = \frac{\partial p}{\partial t} \tag{6.27}$$

and

$$D_n\frac{\partial^2 n}{\partial x^2} + \mu_n\left(\mathrm{E}\frac{\partial n}{\partial x} + n\frac{\partial \mathrm{E}}{\partial x}\right) + g_n - \frac{n}{\tau_{nt}} = \frac{\partial n}{\partial t} \tag{6.28}$$

Equations (6.27) and (6.28) are the time-dependent diffusion equations for holes and electrons, respectively. Since both the hole concentration p and the electron concentration n contain the excess concentrations, Equations (6.27) and (6.28) describe the space and time behavior of the excess carriers.

The hole and electron concentrations are functions of both the thermal equilibrium and the excess values, which are given in Equations (6.5a) and (6.5b). The thermal-equilibrium concentrations, n_0 and p_0, are not functions of time. For the special case of a homogeneous semiconductor, n_0 and p_0 are also independent of the space coordinates. Equations (6.27) and (6.28) may then be written in the form

$$D_p\frac{\partial^2(\delta p)}{\partial x^2} - \mu_p\left(\mathrm{E}\frac{\partial(\delta p)}{\partial x} + p\frac{\partial \mathrm{E}}{\partial x}\right) + g_p - \frac{p}{\tau_{pt}} = \frac{\partial(\delta p)}{\partial t} \tag{6.29}$$

and

$$D_n\frac{\partial^2(\delta n)}{\partial x^2} + \mu_n\left(\mathrm{E}\frac{\partial(\delta n)}{\partial x} + n\frac{\partial \mathrm{E}}{\partial x}\right) + g_n - \frac{n}{\tau_{nt}} = \frac{\partial(\delta n)}{\partial t} \tag{6.30}$$

Note that Equations (6.29) and (6.30) contain terms involving the total concentrations, p and n, and terms involving only the excess concentrations, δp and δn.

6.3 | AMBIPOLAR TRANSPORT

Originally, we assumed that the electric field in the current Equations (6.20) and (6.21) was an applied electric field. This electric field term appears in the time-dependent diffusion equations given by Equations (6.29) and (6.30). If a pulse of excess electrons and a pulse of excess holes are created at a particular point in a semiconductor with an applied electric field, the excess holes and electrons *will tend* to drift in opposite directions. However, because the electrons and holes are charged particles, any separation will induce an internal electric field between the two sets of particles. This internal electric field will create a force attracting the electrons and holes back toward each other. This effect is shown in Figure 6.5. The electric field term in Equations (6.29) and (6.30) is then composed of the externally applied field plus the induced internal field. This E-field may be written as

$$\mathrm{E} = \mathrm{E}_{app} + \mathrm{E}_{int} \tag{6.31}$$

where E_{app} is the applied electric field and E_{int} is the induced internal electric field.

Since the internal E-field creates a force attracting the electrons and holes, this E-field will hold the pulses of excess electrons and excess holes together. The negatively charged electrons and positively charged holes then will drift or diffuse together with a single effective mobility or diffusion coefficient. This phenomenon is called *ambipolar diffusion* or *ambipolar transport.*

6.3.1 Derivation of the Ambipolar Transport Equation

The time-dependent diffusion Equations (6.29) and (6.30) describe the behavior of the excess carriers. However, a third equation is required to relate the excess electron and hole concentrations to the internal electric field. This relation is Poisson's equation, which may be written as

$$\nabla \cdot \mathrm{E}_{int} = \frac{e(\delta p - \delta n)}{\epsilon_s} = \frac{\partial \mathrm{E}_{int}}{\partial x} \tag{6.32}$$

where ϵ_s is the permittivity of the semiconductor material.

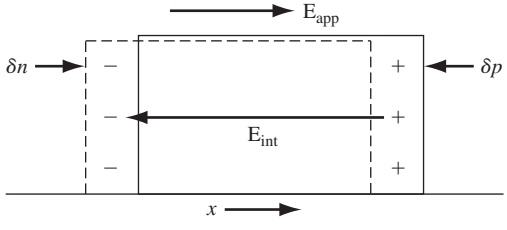


Figure 6.5 | The creation of an internal electric field as excess electrons and holes *tend* to separate.

To make the solution of Equations (6.29), (6.30), and (6.32) more tractable, we need to make some approximations. We can show that only a relatively small internal electric field is sufficient to keep the excess electrons and holes drifting and diffusing together. Hence, we can assume that

$$|\mathrm{E}_{\mathrm{int}}| \ll |\mathrm{E}_{\mathrm{app}}| \tag{6.33}$$

However, the $\nabla \cdot \mathrm{E}_{\mathrm{int}}$ term may not be negligible. We will impose the condition of charge neutrality: We will assume that the excess electron concentration is just balanced by an equal excess hole concentration at any point in space and time. If this assumption were exactly true, there would be no induced internal electric field to keep the two sets of particles together. However, only a very small difference in the excess electron concentration and excess hole concentration will set up an internal E-field sufficient to keep the particles diffusing and drifting together. We can show that a 1 percent difference in δp and δn, for example, will result in non-negligible values of the $\nabla \cdot \mathrm{E} = \nabla \cdot \mathrm{E}_{\mathrm{int}}$ term in Equations (6.29) and (6.30).

We can combine Equations (6.29) and (6.30) to eliminate the $\nabla \cdot \mathrm{E}$ term. Considering Equations (6.1) and (6.4), we can define

$$g_n = g_p \equiv g \tag{6.34}$$

and considering Equations (6.2) and (6.6), we can define

$$R_n = \frac{n}{\tau_{nt}} = R_p = \frac{p}{\tau_{pt}} \equiv R \tag{6.35}$$

The lifetimes in Equation (6.35) include the thermal-equilibrium carrier lifetimes and the excess carrier lifetimes. If we impose the charge neutrality condition, then $\delta n \approx \delta p$. We will denote both the excess electron and excess hole concentrations in Equations (6.29) and (6.30) by δn. We may then rewrite Equations (6.29) and (6.30) as

$$D_p \frac{\partial^2(\delta n)}{\partial x^2} - \mu_p\left(\mathrm{E}\frac{\partial(\delta n)}{\partial x} + p\frac{\partial \mathrm{E}}{\partial x}\right) + g - R = \frac{\partial(\delta n)}{\partial t} \tag{6.36}$$

and

$$D_n \frac{\partial^2(\delta n)}{\partial x^2} + \mu_n\left(\mathrm{E}\frac{\partial(\delta n)}{\partial x} + n\frac{\partial \mathrm{E}}{\partial x}\right) + g - R = \frac{\partial(\delta n)}{\partial t} \tag{6.37}$$

If we multiply Equation (6.36) by $\mu_n n$, multiply Equation (6.37) by $\mu_p p$, and add the two equations, the $\nabla \cdot \mathrm{E} = \partial \mathrm{E}/\partial x$ term will be eliminated. The result of this addition gives

$$(\mu_n n D_p + \mu_p p D_n)\frac{\partial^2(\delta n)}{\partial x^2} + (\mu_n \mu_p)(p - n)\mathrm{E}\frac{\partial(\delta n)}{\partial x} + (\mu_n n + \mu_p p)(g - R) = (\mu_n n + \mu_p p)\frac{\partial(\delta n)}{\partial t} \tag{6.38}$$

If we divide Equation (6.38) by the term $(\mu_n n + \mu_p p)$, this equation becomes

$$\boxed{D'\frac{\partial^2(\delta n)}{\partial x^2} + \mu'\mathrm{E}\frac{\partial(\delta n)}{\partial x} + g - R = \frac{\partial(\delta n)}{\partial t}} \tag{6.39}$$

where

$$D' = \frac{\mu_n n D_p + \mu_p p D_n}{\mu_n n + \mu_p p} \tag{6.40}$$

and

$$\boxed{\mu' = \frac{\mu_n \mu_p (p - n)}{\mu_n n + \mu_p p}} \tag{6.41}$$

Equation (6.39) is called the *ambipolar transport equation* and describes the behavior of the excess electrons and holes in time and space. The parameter D' is called the *ambipolar diffusion coefficient* and μ' is called the *ambipolar mobility.*

The Einstein relation relates the mobility and diffusion coefficient by

$$\frac{\mu_n}{D_n} = \frac{\mu_p}{D_p} = \frac{e}{kT} \tag{6.42}$$

Using these relations, the ambipolar diffusion coefficient may be written in the form

$$\boxed{D' = \frac{D_n D_p (n + p)}{D_n n + D_p p}} \tag{6.43}$$

The ambipolar diffusion coefficient, D', and the ambipolar mobility, μ', are functions of the electron and hole concentrations, n and p, respectively. Since both n and p contain the excess carrier concentration δn, the coefficient in the ambipolar transport equation are not constants. The ambipolar transport equation, given by Equation (6.39), then, is a nonlinear differential equation.

6.3.2 Limits of Extrinsic Doping and Low Injection

The ambipolar transport equation may be simplified and linearized by considering an extrinsic semiconductor and by considering low-level injection. The ambipolar diffusion coefficient, from Equation (6.43), may be written as

$$D' = \frac{D_n D_p [(n_0 + \delta n) + (p_0 + \delta n)]}{D_n (n_0 + \delta n) + D_p (p_0 + \delta n)} \tag{6.44}$$

where n_0 and p_0 are the thermal-equilibrium electron and hole concentrations, respectively, and δn is the excess carrier concentration. If we consider a p-type semiconductor, we can assume that $p_0 \gg n_0$. The condition of low-level injection, or just low injection, means that the excess carrier concentration is much smaller than the thermal-equilibrium majority carrier concentration. For the p-type semiconductor, then, low injection implies that $\delta n \ll p_0$. Assuming that $n_0 \ll p_0$ and $\delta n \ll p_0$, and assuming that D_n and D_p are on the same order of magnitude, the ambipolar diffusion coefficient from Equation (6.44) reduces to

$$D' = D_n \tag{6.45}$$

If we apply the conditions of an extrinsic p-type semiconductor and low injection to the ambipolar mobility, Equation (6.41) reduces to

$$\mu' = \mu_n \tag{6.46}$$

It is important to note that for an extrinsic p-type semiconductor under low injection, the ambipolar diffusion coefficient and the ambipolar mobility coefficient reduce to the minority carrier electron parameter values, which are constants. The ambipolar transport equation reduces to a linear differential equation with constant coefficients.

If we now consider an extrinsic n-type semiconductor under low injection, we may assume that $p_0 \ll n_0$ and $\delta n \ll n_0$. The ambipolar diffusion coefficient from Equation (6.43) reduces to

$$D' = D_p \tag{6.47}$$

and the ambipolar mobility from Equation (6.41) reduces to

$$\mu' = -\mu_p \tag{6.48}$$

The ambipolar parameters again reduce to the minority-carrier values, which are constants. Note that, for the n-type semiconductor, the ambipolar mobility is a negative value. The ambipolar mobility term is associated with carrier drift; therefore, the sign of the drift term depends on the charge of the particle. The equivalent ambipolar particle is negatively charged, as one can see by comparing Equations (6.30) and (6.39). If the ambipolar mobility reduces to that of a positively charged hole, a negative sign is introduced as shown in Equation (6.48).

The remaining terms we need to consider in the ambipolar transport equation are the generation rate and the recombination rate. Recall that the electron and hole recombination rates are equal and are given by Equation (6.35) as $R_n = R_p = n/\tau_{nt} = p/\tau_{pt} \equiv R$, where τ_{nt} and τ_{pt} are the mean electron and hole lifetimes, respectively. If we consider the inverse lifetime functions, then $1/\tau_{nt}$ is the probability per unit time that an electron will encounter a hole and recombine. Likewise, $1/\tau_{pt}$ is the probability per unit time that a hole will encounter an electron and recombine. If we again consider an extrinsic p-type semiconductor under low injection, the concentration of majority carrier holes will be essentially constant, even when excess carriers are present. Then, the probability per unit time of a minority carrier electron encountering a majority carrier hole will be essentially constant. Hence, the minority carrier electron lifetime, $\tau_{nt} \equiv \tau_n$, will remain a constant for the extrinsic p-type semiconductor under low injection.

Similarly, if we consider an extrinsic n-type semiconductor under low injection, the minority carrier hole lifetime, $\tau_{pt} \equiv \tau_p$, will remain constant. Even under the condition of low injection, the minority carrier hole concentration may increase by several orders of magnitude. The probability per unit time of a majority carrier electron encountering a hole may change drastically. The majority carrier lifetime, then, may change substantially when excess carriers are present.

Consider, again, the generation and recombination terms in the ambipolar transport equation. For electrons we may write

$$g - R = g_n - R_n = (G_{n0} + g'_n) - (R_{n0} + R'_n) \tag{6.49}$$

where G_{n0} and g'_n are the thermal-equilibrium electron and excess electron generation rates, respectively. The terms R_{n0} and R'_n are the thermal-equilibrium electron and excess electron recombination rates, respectively. For thermal equilibrium, we have that

$$G_{n0} = R_{n0} \tag{6.50}$$

so Equation (6.49) reduces to

$$g - R = g'_n - R'_n = g'_n - \frac{\delta n}{\tau_n} \tag{6.51}$$

where τ_n is the excess minority carrier electron lifetime.

For the case of holes, we may write

$$g - R = g_p - R_p = (G_{p0} + g'_p) - (R_{p0} + R'_p) \tag{6.52}$$

where G_{p0} and g'_p are the thermal-equilibrium hole and excess hole generation rates, respectively. The terms R_{p0} and R'_p are the thermal-equilibrium hole and excess hole recombination rates, respectively. Again, for thermal equilibrium, we have that

$$G_{p0} = R_{p0} \tag{6.53}$$

so that Equation (6.52) reduces to

$$g - R = g'_p - R'_p = g'_p - \frac{\delta p}{\tau_p} \tag{6.54}$$

where τ_p is the excess minority carrier hole lifetime.

The generation rate for excess electrons must equal the generation rate for excess holes. We may then define a generation rate for excess carriers as g', so that $g'_n = g'_p \equiv g'$. We also determined that the minority carrier lifetime is essentially a constant for low injection. Then, the term $g - R$ in the ambipolar transport equation may be written in terms of the minority carrier parameters.

The ambipolar transport equation, given by Equation (6.39), for a p-type semiconductor under low injection then becomes

$$\boxed{D_n \frac{\partial^2(\delta n)}{\partial x^2} + \mu_n \mathrm{E} \frac{\partial(\delta n)}{\partial x} + g' - \frac{\delta n}{\tau_{n0}} = \frac{\partial(\delta n)}{\partial t}} \tag{6.55}$$

The parameter δn is the excess minority carrier electron concentration, the parameter τ_{n0} is the minority carrier lifetime under low injection, and the other parameters are the usual minority carrier electron parameters.

Similarly, for an extrinsic n-type semiconductor under low injection, the ambipolar transport equation becomes

$$\boxed{D_p \frac{\partial^2(\delta p)}{\partial x^2} - \mu_p \mathrm{E} \frac{\partial(\delta p)}{\partial x} + g' - \frac{\delta p}{\tau_{p0}} = \frac{\partial(\delta p)}{\partial t}} \tag{6.56}$$

The parameter δp is the excess minority carrier hole concentration, the parameter τ_{p0} is the minority carrier hole lifetime under low injection, and the other parameters are the usual minority carrier hole parameters.

It is extremely important to note that the transport and recombination parameters in Equations (6.55) and (6.56) are those of the minority carrier. *Equations (6.55) and (6.56) describe the drift, diffusion, and recombination of excess minority carriers as a function of spatial coordinates and as a function of time.* Recall that we had imposed the condition of charge neutrality; the excess minority carrier concentration is equal to the excess majority carrier concentration. The excess majority carriers, then, diffuse and drift with the excess minority carriers; thus, the behavior of the excess majority carrier is determined by the minority carrier parameters. This ambipolar phenomenon is extremely important in semiconductor physics, and is the basis for describing the characteristics and behavior of semiconductor devices.

6.3.3 Applications of the Ambipolar Transport Equation

We solve the ambipolar transport equation for several problems. These examples help illustrate the behavior of excess carriers in a semiconductor material, and the results are used later in the discussion of the pn junction and the other semiconductor devices.

The following examples use several common simplifications in the solution of the ambipolar transport equation. Table 6.2 summarizes these simplifications and their effects.

EXAMPLE 6.2

Objective: Determine the time behavior of excess carriers as a semiconductor returns to thermal equilibrium.

Consider an infinitely large, homogeneous n-type semiconductor with zero applied electric field. Assume that at time $t = 0$, a uniform concentration of excess carriers exists in the crystal, but assume that $g' = 0$ for $t > 0$. If we assume that the concentration of excess carriers is much smaller than the thermal-equilibrium electron concentration, then the low-injection condition applies. Calculate the excess carrier concentration as a function of time for $t \geq 0$.

Table 6.2 | Common ambipolar transport equation simplifications

Specification	Effect
Steady state	$\frac{\partial(\delta n)}{\partial t} = 0, \quad \frac{\partial(\delta p)}{\partial t} = 0$
Uniform distribution of excess carriers (uniform generation rate)	$D_n \frac{\partial^2(\delta n)}{\partial x^2} = 0, \quad D_p \frac{\partial^2(\delta n)}{\partial x^2} = 0$
Zero electric field	$E \frac{\partial(\delta n)}{\partial x} = 0, \quad E \frac{\partial(\delta p)}{\partial x} = 0$
No excess carrier generation	$g' = 0$
No excess carrier recombination (infinite lifetime)	$\frac{\delta n}{\tau_{n0}} = 0, \quad \frac{\delta p}{\tau_{p0}} = 0$

■ Solution

For the n-type semiconductor, we need to consider the ambipolar transport equation for the minority carrier holes, which is given by Equation (6.56). The equation is

$$D_p \frac{\partial^2(\delta p)}{\partial x^2} - \mu_p \mathrm{E}\frac{\partial(\delta p)}{\partial x} + g' - \frac{\delta p}{\tau_{p0}} = \frac{\partial(\delta p)}{\partial t}$$

We are assuming a uniform concentration of excess holes so that $\partial^2(\delta p)/\partial x^2 = \partial(\delta p)/\partial x = 0$. For $t > 0$, we are also assuming that $g' = 0$. Equation (6.56) reduces to

$$\frac{d(\delta p)}{dt} = -\frac{\delta p}{\tau_{p0}} \tag{6.57}$$

Since there is no spatial variation, the total time derivative may be used. At low injection, the minority carrier hole lifetime, τ_{p0}, is a constant. The left-side of Equation (6.57) is the time rate of change of δp and the right-side of the equation is the recombination rate. The solution to Equation (6.57) is

$$\boxed{\delta p(t) = \delta p(0)e^{-t/\tau_{p0}}} \tag{6.58}$$

where $\delta p(0)$ is the uniform concentration of excess carriers that exists at time $t = 0$. The concentration of excess holes decays exponentially with time, with a time constant equal to the minority carrier hole lifetime.

From the charge-neutrality condition, we have that $\delta n = \delta p$, so the excess electron concentration is given by

$$\delta n(t) = \delta p(t) = \delta p(0)e^{-t/\tau_{p0}} \tag{6.59}$$

■ Comment

The excess electrons and holes recombine at the rate determined by the excess minority carrier hole lifetime in the n-type semiconductor.

■ EXERCISE PROBLEM

Ex 6.2 Consider n-type GaAs doped at $N_d = 10^{16}$ cm^{-3}. Assume that 10^{14} electron–hole pairs have been uniformly created per cm^3 at $t = 0$, and assume the minority carrier hole lifetime is $\tau_{p0} = 50$ ns. Determine the time at which the minority carrier hole concentration reaches (*a*) 1/e of its initial value and (*b*) 10% of its initial value.

[Ans. (*a*) $t = 50$ ns; (*b*) $t = 115$ ns]

EXAMPLE 6.3

Objective: Determine the time dependence of excess carriers in reaching a steady-state condition.

Again consider an infinitely large, homogeneous n-type semiconductor with a zero applied electric field. Assume that, for $t < 0$, the semiconductor is in thermal equilibrium and that, for $t \geq 0$, a uniform generation rate exists in the crystal. Calculate the excess carrier concentration as a function of time assuming the condition of low injection.

Solution

The condition of a uniform generation rate and a homogeneous semiconductor again implies that $\partial^2(\delta p)/\partial x^2 = \partial(\delta p)/\partial x = 0$ in Equation (6.56). The equation, for this case, reduces to

$$g' - \frac{\delta p}{\tau_{p0}} = \frac{d(\delta p)}{dt} \tag{6.60}$$

The solution to this differential equation is

$$\boxed{\delta p(t) = g'\tau_{p0}(1 - e^{-t/\tau_{p0}})} \tag{6.61}$$

Comment

We may note that, as $t \to \infty$, a steady-state excess hole and electron concentration of $g'\tau_{p0}$ is reached. Equation (6.60) contains both a generation rate term and a recombination rate term for the excess carriers.

EXERCISE PROBLEM

Ex 6.3 In Example 6.3, consider n-type silicon at $T = 300$ K doped to $N_d = 5 \times 10^{16}$ cm^{-3}. Assume that $g' = 5 \times 10^{21}$ cm^{-3} s^{-1} and let $\tau_{p0} = 10^{-7}$ s. (*a*) Determine $\delta p(t)$ at (*i*) $t = 0$, (*ii*) $t = 10^{-7}$ s, (*iii*) $t = 5 \times 10^{-7}$ s, and (*iv*) $t \to \infty$. (*b*) Considering the results of part (*a*), is the low-injection condition maintained?

[Ans. (*a*) (*i*) 0; (*ii*) 3.16×10^{14} cm^{-3}; (*iii*) 4.966×10^{14} cm^{-3}; (*iv*) 5×10^{14} cm^{-3}; (*b*) $\delta p(\max) = 0.01N_d$, yes]

The excess minority carrier hole concentration increases with time with the same time constant τ_{p0}, which is the excess minority carrier lifetime. The excess carrier concentration reaches a steady-state value as time goes to infinity, even though a steady-state generation of excess electrons and holes exists. This steady-state effect can be seen from Equation (6.60) by setting $d(\delta p)/dt = 0$. The remaining terms simply state that, in steady state, the generation rate is equal to the recombination rate.

TEST YOUR UNDERSTANDING

TYU 6.1 Silicon at T=300 K has been doped with boron atoms to a concentration of $N_a = 5 \times 10^{16}$ cm^{-3}. Excess carriers have been generated in the uniformly doped material to a concentration of 10^{15} cm^{-3}. The minority carrier lifetime is 5 μs. (*a*) What carrier type is the minority carrier? (*b*) Assuming $g' = E = 0$ for $t > 0$, determine the minority carrier concentration for $t > 0$.

[Ans. (*a*) electrons; (*b*) $10^{15}\, e^{-t/5\times10^{-6}}$ cm^{-3}]

TYU 6.2 Consider silicon with the same parameters as given in TYU 6.1. The material is in thermal equilibrium for $t < 0$. At $t = 0$, a source generating excess carriers is turned on, producing a generation rate of $g' = 10^{20}$ cm^{-3}-s^{-1}. (*a*) What carrier type is the minority carrier? (*b*) Determine the minority carrier concentration for $t > 0$. (*c*) What is the minority carrier concentration as $t \to \infty$?

[Ans. (*a*) electrons; (*b*) $5 \times 10^{14}\,[1 - e^{-t/5\times10^{-6}}]$ cm^{-3} (*c*) 5×10^{14} cm^{-3}]

EXAMPLE 6.4

Objective: Determine the steady-state spatial dependence of the excess carrier concentration.

Consider a p-type semiconductor that is homogeneous and infinite in extent. Assume a zero applied electric field. For a one-dimensional crystal, assume that excess carriers are being generated at $x = 0$ only, as indicated in Figure 6.6. The excess carriers being generated at $x = 0$ will begin diffusing in both the $+x$ and $-x$ directions. Calculate the steady-state excess carrier concentration as a function of x.

■ Solution

The ambipolar transport equation for excess minority carrier electrons is given by Equation (6.55), and is written as

$$D_n \frac{\partial^2(\delta n)}{\partial x^2} + \mu_n \mathrm{E} \frac{\partial(\delta n)}{\partial x} + g' - \frac{\delta n}{\tau_{n0}} = \frac{\partial(\delta n)}{\partial t}$$

From our assumptions, we have $\mathrm{E} = 0$, $g' = 0$ for $x \neq 0$, and $\partial(\delta n)/\partial t = 0$ for steady state. Assuming a one-dimensional crystal, Equation (6.55) reduces to

$$D_n \frac{d^2(\delta n)}{dx^2} - \frac{\delta n}{\tau_{n0}} = 0 \tag{6.62}$$

Dividing by the diffusion coefficient, Equation (6.62) may be written as

$$\frac{d^2(\delta n)}{dx^2} - \frac{\delta n}{D_n \tau_{n0}} = \frac{d^2(\delta n)}{dx^2} - \frac{\delta n}{L_n^2} = 0 \tag{6.63}$$

where we have defined $L_n^2 = D_n \tau_{n0}$. The parameter L_n has the unit of length and is called the minority carrier electron diffusion length. The general solution to Equation (6.63) is

$$\delta n(x) = Ae^{-x/L_n} + Be^{x/L_n} \tag{6.64}$$

As the minority carrier electrons diffuse away from $x = 0$, they will recombine with the majority carrier holes. The minority carrier electron concentration will then decay toward zero at both $x = +\infty$ and $x = -\infty$. These boundary conditions mean that $B \equiv 0$ for $x > 0$ and $A \equiv 0$ for $x < 0$. The solution to Equation (6.63) may then be written as

$$\boxed{\delta n(x) = \delta n(0) e^{-x/L_n} \qquad x \geq 0} \tag{6.65a}$$

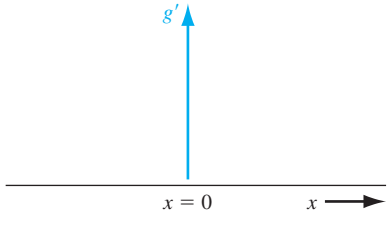


Figure 6.6 | Steady-state generation rate at $x = 0$.

and

$$\boxed{\delta n(x) = \delta n(0)e^{+x/L_n} \quad x \leq 0} \tag{6.65b}$$

where $\delta n(0)$ is the value of the excess electron concentration at $x = 0$. The steady-state excess electron concentration decays exponentially with distance away from the source at $x = 0$.

Comment

We may note that the steady-state excess concentration decays to $1/e$ of its value at $x = L_n$.

EXERCISE PROBLEM

Ex 6.4 In Example 6.4, consider p-type silicon at $T = 300$ K doped to $N_a = 5 \times 10^{16}$ cm^{-3}. Assume that $\tau_{n0} = 5 \times 10^{-7}$ s, $D_n = 25$ cm^2/s, and $\delta n(0) = 10^{15}$ cm^{-3}. (*a*) Calculate the value of diffusion length L_n. (*b*) Determine δn at (*i*) $x = 0$, (*ii*) $x = +30$ μm, (*iii*) $x = -50$ μm, (*iv*) $x = +85$ μm, and (*v*) $x = -120$ μm.

[Ans. (*a*) $L_n = 35.36$ μm; (*b*) (*i*) 10^{15} cm^{-3}; (*ii*) 4.28×10^{14} cm^{-3}, (*iii*) 2.43×10^{14} cm^{-3}, (*iv*) 9.04×10^{13} cm^{-3}, (*v*) 3.36×10^{13} cm^{-3}]

As before, we will assume charge neutrality; thus, the steady-state excess majority carrier hole concentration also decays exponentially with distance with the same characteristic minority carrier electron diffusion length L_n. Figure 6.7 is a plot of the total electron and hole concentrations as a function of distance. We are assuming low injection, that is, $\delta n(0) \ll p_0$ in the p-type semiconductor. The total concentration of

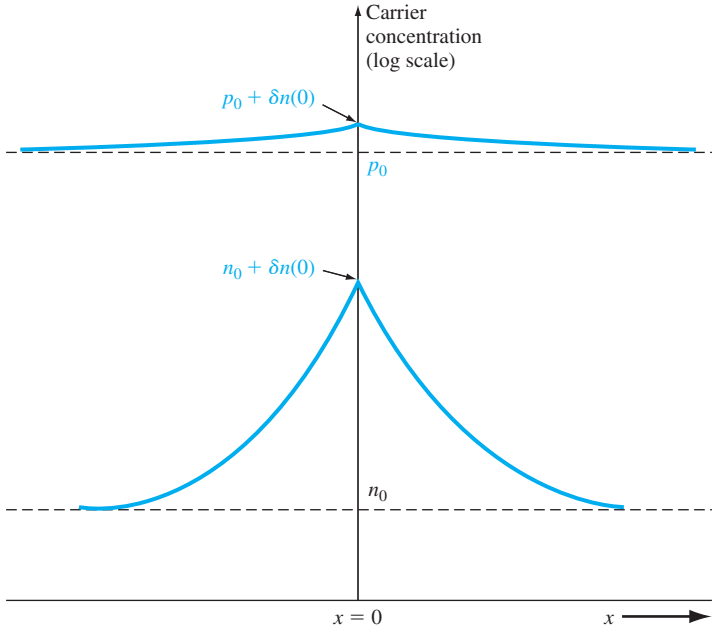


Figure 6.7 | Steady-state electron and hole concentrations for the case when excess electrons and holes are generated at $x = 0$.

majority carrier holes barely changes. However, we may have $\delta n(0) \gg n_0$ and still satisfy the low-injection condition. The minority carrier concentration may change by many orders of magnitude.

TEST YOUR UNDERSTANDING

TYU 6.3 Excess electrons and holes are generated at the end of a silicon bar ($x = 0$). The silicon is doped with phosphorus atoms to a concentration of $N_d = 10^{17}$ cm^{-3}. The minority carrier lifetime is 1 μs, the electron diffusion coefficient is $D_n = 25$ cm^2/s, and the hole diffusion coefficient is $D_p = 10$ cm^2/s. If $\delta n(0) = \delta p(0) = 10^{15}$ cm^{-3}, determine the steady-state electron and hole concentrations in the silicon for $x > 0$.

[Ans. $\delta n(x) = \delta p(x) = 10^{15} e^{-x/3.16\times10^{-3}}$ cm^{-3}, where x is in cm]

TYU 6.4 Using the parameters given in TYU 6.3, calculate the electron and hole diffusion current densities at $x = 10$ μm.

[Ans. $J_p = +0.369$ A/cm^2, $J_n = -0.369$ A/cm^2]

The three previous examples, which applied the ambipolar transport equation to specific situations, assumed either a homogeneous or a steady-state condition; only the time variation or the spatial variation was considered. Now consider an example in which both the time and spatial dependence are considered in the same problem.

EXAMPLE 6.5

Objective: Determine both the time dependence and spatial dependence of the excess carrier concentration.

Assume that a finite number of electron–hole pairs is generated instantaneously at time $t = 0$ and at $x = 0$, but assume $g' = 0$ for $t > 0$. Assume we have an n-type semiconductor with a constant applied electric field equal to E_0, which is applied in the $+x$ direction. Calculate the excess carrier concentration as a function of x and t.

■ Solution

The one-dimensional ambipolar transport equation for the minority carrier holes can be written from Equation (6.56) as

$$D_p \frac{\partial^2(\delta p)}{\partial x^2} - \mu_p E_0 \frac{\partial(\delta p)}{\partial x} - \frac{\delta p}{\tau_{p0}} = \frac{\partial(\delta p)}{\partial t} \tag{6.66}$$

The solution to this partial differential equation is of the form

$$\delta p(x, t) = p'(x, t)e^{-t/\tau_{p0}} \tag{6.67}$$

By substituting Equation (6.67) into Equation (6.66), we are left with the partial differential equation

$$D_p \frac{\partial^2 p'(x, t)}{\partial x^2} - \mu_p E_0 \frac{\partial p'(x, t)}{\partial x} = \frac{\partial p'(x, t)}{\partial t} \tag{6.68}$$

Equation (6.68) is normally solved using Laplace transform techniques. The solution, without going through the mathematical details, is

$$p'(x, t) = \frac{1}{(4\pi D_p t)^{1/2}} \exp\left[\frac{-(x - \mu_p E_0 t)^2}{4D_p t}\right] \tag{6.69}$$

The total solution, from Equations (6.67) and (6.69), for the excess minority carrier hole concentration is

$$\delta p(x, t) = \frac{e^{-t/\tau_{p0}}}{(4\pi D_p t)^{1/2}} \exp\left[\frac{-(x - \mu_p E_0 t)^2}{4D_p t}\right] \tag{6.70}$$

Comment

We could show that Equation (6.70) is a solution to the partial differential equation, Equation (6.66), by direct substitution. We may also note that Equation (6.70) is not normalized.

EXERCISE PROBLEM

Ex 6.5 Consider the result of Example 6.5. Let $D_p = 10$ cm^2/s, $\tau_{p0} = 10^{-7}$ s, $\mu_p = 400$ cm^2/V-s, and $E_0 = 100$ V/cm. (*a*) Determine δp for $t = 10^{-7}$ s at (*i*) $x = 20$ μm, (*ii*) $x = 40$ μm, and (*iii*) $x = 60$ μm. (*b*) Determine δp for $x = 40$ μm at (*i*) $t = 5 \times 10^{-8}$ s, (*ii*) $t = 10^{-7}$ s, and (*iii*) $t = 2 \times 10^{-7}$ s. Compare the results to the curves shown in Figure 6.9.

[Ans. (*a*) (*i*) 38.18, (*ii*) 103.8, (*iii*) 38.18; (*b*) (*i*) 32.75, (*ii*) 103.8, (*iii*) 3.65]

Equation (6.70) can be plotted as a function of distance x, for various times. Figure 6.8 shows such a plot for the case when the applied electric field is zero. For $t > 0$, the excess minority carrier holes diffuse in both the $+x$ and $-x$ directions. During this time, the excess majority carrier electrons, which were generated, diffuse at exactly the same rate as the holes. As time proceeds, the excess holes recombine with the excess electrons so that at $t = \infty$ the excess hole concentration is zero. In this particular example, both diffusion and recombination processes are occurring at the same time.

Figure 6.9 shows a plot of Equation (6.70) as a function of distance x at various times for the case when the applied electric field is not zero. In this case, the pulse of excess minority carrier holes is drifting in the $+x$ direction, which is the direction of the electric field. We still have the same diffusion and recombination processes as we had before. An important point to consider is that, with charge neutrality, $\delta n = \delta p$ at any instant of time and at any point in space. The excess electron concentration is equal to the excess hole concentration. In this case, then, the excess electron pulse is moving in the same direction as the applied electric field even though the electrons have a negative charge. In the ambipolar transport process, the excess carriers are characterized by the minority carrier parameters. In this example, the excess carriers behave according to the minority carrier hole parameters, which include D_p, μ_p, and τ_{p0}. The excess majority carrier electrons are being pulled along by the excess minority carrier holes.

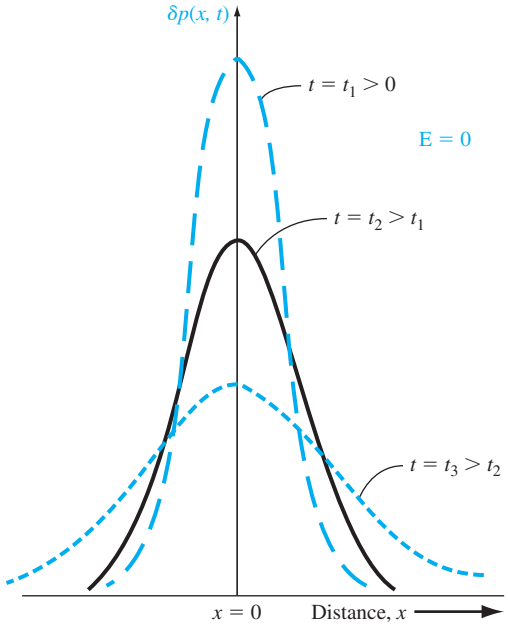


Figure 6.8 | Excess hole concentration versus distance at various times for zero applied electric field.

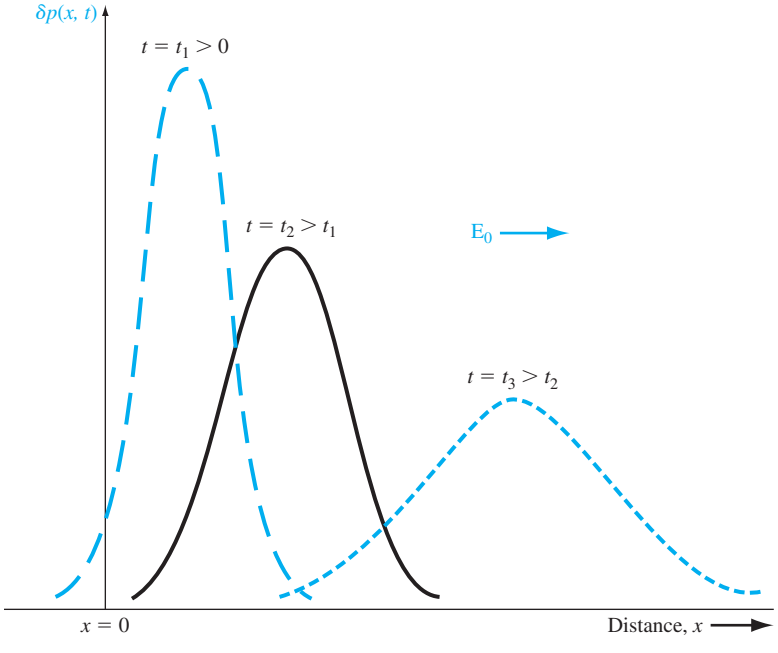


Figure 6.9 | Excess hole concentration versus distance at various times for a constant applied electric field.

TEST YOUR UNDERSTANDING

TYU 6.5 As a good approximation, the peak value of a normalized excess carrier concentration, given by Equation (6.70), occurs at $x = \mu_p E_0 t$. Assume the following parameters: $\tau_{p0} = 5\ \mu$s, $D_p = 10\ \text{cm}^2/\text{s}$, $\mu_p = 386\ \text{cm}^2/V\text{-s}$, and $E_0 = 10\ \text{V/cm}$. Calculate the peak value at times of (*a*) $t = 1\ \mu$s, (*b*) $t = 5\ \mu$s, (*c*) $t = 15\ \mu$s, and (*d*) $t = 25\ \mu$s. What are the corresponding values of x for parts (*a*) to (*d*)?

[Ans. (*a*) 73.0, $x = 38.6\ \mu$m; (*b*) 14.7, $x = 193\ \mu$m; (*c*) 1.15, $x = 579\ \mu$m; (*d*) 0.120, $x = 965\ \mu$m]

TYU 6.6 The excess carrier concentration, given by Equation (6.70), is to be calculated at distances of one diffusion length away from the peak value. Using the parameters given in TYU 6.5, calculate the values of δp for (*a*) $t = 1\ \mu$s at (*i*) 1.093×10^{-2} cm and (*ii*) $x = -3.21 \times 10^{-3}$ cm; (*b*) $t = 5\ \mu$s at (*i*) $x = 2.64 \times 10^{-2}$ cm and (*ii*) $x = 1.22 \times 10^{-2}$ cm; (*c*) $t = 15\ \mu$s at (*i*) $x = 6.50 \times 10^{-2}$ cm and (*ii*) $x = 5.08 \times 10^{-2}$ cm.

[Ans. (*a*) (*i*) 20.9, (*ii*) 20.9; (*b*) (*i*) 11.4, (*ii*) 11.4; (*c*) (*i*) 1.05, (*ii*) 1.05]

6.3.4 Dielectric Relaxation Time Constant

We have assumed in the previous analysis that a quasi-neutrality condition exists—that is, the concentration of excess holes is balanced by an equal concentration of excess electrons. Suppose that we have a situation as shown in Figure 6.10, in which a uniform concentration of holes δp is suddenly injected into a portion of the surface of a semiconductor. We will instantly have a concentration of excess holes and a net positive charge density that is not balanced by a concentration of excess electrons. How is charge neutrality achieved and how fast?

There are three defining equations to be considered. Poisson's equation is

$$\nabla \cdot \mathrm{E} = \frac{\rho}{\epsilon} \tag{6.71}$$

The current equation, Ohm's law, is

$$J = \sigma \mathrm{E} \tag{6.72}$$

The continuity equation, neglecting the effects of generation and recombination, is

$$\nabla \cdot J = -\frac{\partial \rho}{\partial t} \tag{6.73}$$

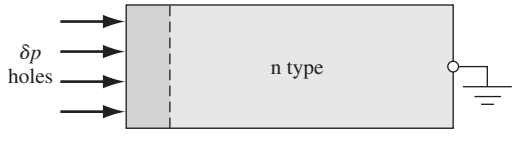


Figure 6.10 | The injection of a concentration of holes into a small region at the surface of an n-type semiconductor.

The parameter ρ is the net charge density and the initial value is given by $e(\delta p)$. We will assume that δp is uniform over a short distance at the surface. The parameter ϵ is the permittivity of the semiconductor.

Taking the divergence of Ohm's law and using Poisson's equation, we find

$$\nabla \cdot J = \sigma \nabla \cdot E = \frac{\sigma \rho}{\epsilon} \tag{6.74}$$

Substituting Equation (6.74) into the continuity equation, we have

$$\frac{\sigma \rho}{\epsilon} = -\frac{\partial \rho}{\partial t} = -\frac{d\rho}{dt} \tag{6.75}$$

Since Equation (6.75) is a function of time only, we can write the equation as a total derivative. Equation (6.75) can be rearranged as

$$\frac{d\rho}{dt} + \left(\frac{\sigma}{\epsilon}\right)\rho = 0 \tag{6.76}$$

Equation (6.76) is a first-order differential equation whose solution is

$$\rho(t) = \rho(0)e^{-(t/\tau_d)} \tag{6.77}$$

where

$$\tau_d = \frac{\epsilon}{\sigma} \tag{6.78}$$

and is called the dielectric relaxation time constant.

EXAMPLE 6.6

Objective: Calculate the dielectric relaxation time constant for a particular semiconductor.

Consider n-type silicon with a donor impurity concentration of $N_d = 10^{16}$ cm^{-3}.

■ **Solution**

The conductivity is found as

$$\sigma \approx e\mu_n N_d = (1.6 \times 10^{-19})(1200)(10^{16}) = 1.92\ (\Omega\text{-cm})^{-1}$$

where the value of mobility is the approximate value found from Figure 5.3. The permittivity of silicon is

$$\epsilon = \epsilon_r \epsilon_0 = (11.7)(8.85 \times 10^{-14})\ \text{F/cm}$$

The dielectric relaxation time constant is then

$$\tau_d = \frac{\epsilon}{\sigma} = \frac{(11.7)(8.85 \times 10^{-14})}{1.92} = 5.39 \times 10^{-13}\ \text{s}$$

or

$$\tau_d = 0.539\ \text{ps}$$

Comment

Equation (6.77) then predicts that in approximately four time constants, or in approximately 2 ps, the net charge density is essentially zero; that is, quasi-neutrality has been achieved. Since the continuity equation, Equation (6.73), does not contain any generation or recombination terms, the initial positive charge is then neutralized by pulling in electrons from the bulk n-type material to create excess electrons. This process occurs very quickly compared to the normal excess carrier lifetimes of approximately 0.1 μs. The condition of quasi-charge-neutrality is then justified.

EXERCISE PROBLEM

Ex 6.6 (*a*) Consider n-type GaAs doped to $N_d = 5 \times 10^{15}$ cm^{-3}. Determine the dielectric relaxation time. (*b*) Repeat part (*a*) for p-type silicon doped to $N_a = 2 \times 10^{16}$ cm^{-3}.

[Ans. (*a*) $\tau_d = 0.193$ ps; (*b*) $\tau_d = 0.809$ ps]

*6.3.5 Haynes–Shockley Experiment

We have derived the mathematics describing the behavior of excess carriers in a semiconductor. The Haynes–Shockley experiment was one of the first experiments to actually measure excess carrier behavior.

Figure 6.11 shows the basic experimental arrangement. The voltage source V_1 establishes an applied electric field E_0 in the $+x$ direction in the n-type semiconductor sample. Excess carriers are effectively injected into the semiconductor at contact A. Contact B is a rectifying contact that is under reverse bias by the voltage source V_2. The contact B will collect a fraction of the excess carriers as they drift through the semiconductor. The collected carriers will generate an output voltage, V_0.

This experiment corresponds to the problem we discussed in Example 6.5. Figure 6.12 shows the excess carrier concentrations at contacts A and B for two conditions. Figure 6.12a shows the idealized excess carrier pulse at contact A at time $t = 0$. For a given electric field E_{01}, the excess carriers will drift along the

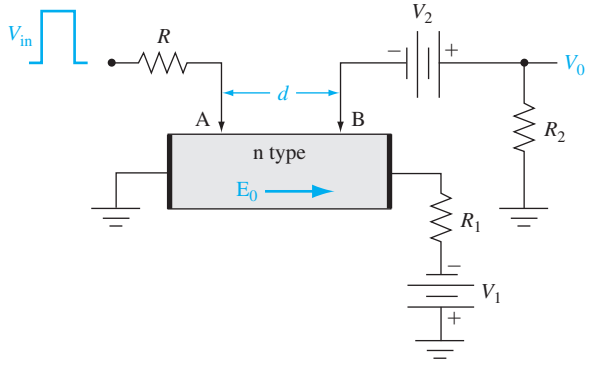


Figure 6.11 | The basic Haynes–Shockley experimental arrangement.

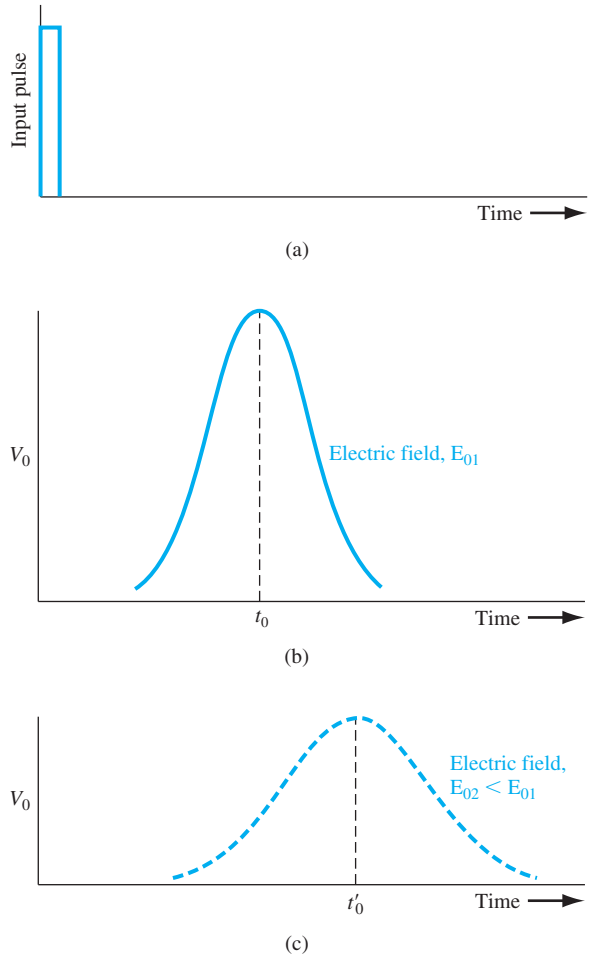


Figure 6.12 | (a) The idealized excess carrier pulse at terminal A at $t = 0$. (b) The excess carrier pulse versus time at terminal B for a given applied electric field. (c) The excess carrier pulse versus time at terminal B for a smaller applied electric field.

semiconductor producing an output voltage as a function of time given in Figure 6.12b. The peak of the pulse will arrive at contact B at time t_0. If the applied electric field is reduced to a value E_{02}, $E_{02} < E_{01}$, the output voltage response at contact B will look approximately as shown in Figure 6.12c. For the smaller electric field, the drift velocity of the pulse of excess carriers is smaller, and so it will take a longer time for the pulse to reach the contact B. During this longer time period, there is more diffusion and more recombination. The excess carrier pulse shapes shown in Figure 6.12b, c are different for the two electric field conditions.

The minority carrier mobility, lifetime, and diffusion coefficient can be determined from this single experiment. As a good first approximation, the peak of the minority carrier pulse will arrive at contact B when the exponent involving distance and time in Equation (6.70) is zero, or

$$x - \mu_p \mathrm{E}_0 t = 0 \tag{6.79a}$$

In this case $x = d$, where d is the distance between contacts A and B, and $t = t_0$, where t_0 is the time at which the peak of the pulse reaches contact B. The mobility may be calculated as

$$\mu_p = \frac{d}{E_0 t_0} \tag{6.79b}$$

Figure 6.13 again shows the output response as a function of time. At times t_1 and t_2, the magnitude of the excess concentration is e^{-1} of its peak value. If the time difference between t_1 and t_2 is not too large, $e^{-t/\tau_{p0}}$ and $(4\pi D_p t)^{1/2}$ do not change appreciably during this time; then the equation

$$(d - \mu_p \mathrm{E}_0 t)^2 = 4 D_p t \tag{6.80}$$

is satisfied at both $t = t_1$ and $t = t_2$. If we set $t = t_1$ and $t = t_2$ in Equation (6.80) and add the two resulting equations, we may show that the diffusion coefficient is given by

$$D_p = \frac{(\mu_p \mathrm{E}_0)^2 (\Delta t)^2}{16 t_0} \tag{6.81}$$

where

$$\Delta t = t_2 - t_1 \tag{6.82}$$

The area S under the curve shown in Figure 6.13 is proportional to the number of excess holes that have not recombined with majority carrier electrons. We may write

$$S = K \exp\left(\frac{-t_0}{\tau_{p0}}\right) = K \exp\left(\frac{-d}{\mu_p \mathrm{E}_0 \tau_{p0}}\right) \tag{6.83}$$

where K is a constant. By varying the electric field, the area under the curve will change. A plot of ln (S) as a function of $(d/\mu_p \mathrm{E}_0)$ will yield a straight line whose slope is $(1/\tau_{p0})$, so the minority carrier lifetime can also be determined from this experiment.

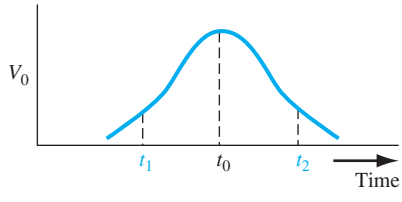


Figure 6.13 | The output excess carrier pulse versus time to determine the diffusion coefficient.

The Haynes–Shockley experiment is elegant in the sense that the three basic processes of drift, diffusion, and recombination are all observed in a single experiment. The determination of mobility is straightforward and can yield accurate values. The determination of the diffusion coefficient and lifetime is more complicated and may lead to some inaccuracies.

6.4 | QUASI-FERMI ENERGY LEVELS

The thermal-equilibrium electron and hole concentrations are functions of the Fermi energy level. We can write

$$n_0 = n_i \exp\left(\frac{E_F - E_{Fi}}{kT}\right) \tag{6.84a}$$

and

$$p_0 = n_i \exp\left(\frac{E_{Fi} - E_F}{kT}\right) \tag{6.84b}$$

where E_F and E_{Fi} are the Fermi energy and intrinsic Fermi energy, respectively, and n_i is the intrinsic carrier concentration. Figure 6.14a shows the energy-band diagram for an n-type semiconductor in which $E_F > E_{Fi}$. For this case, we may note from Equations (6.84a) and (6.84b) that $n_0 > n_i$ and $p_0 < n_i$, as we would expect. Similarly, Figure 6.14b shows the energy-band diagram for a p-type semiconductor in which $E_F < E_{Fi}$. Again we may note from Equations (6.84a) and (6.84b) that $n_0 < n_i$ and $p_0 > n_i$, as we would expect for the p-type material. These results are for thermal equilibrium.

If excess carriers are created in a semiconductor, we are no longer in thermal equilibrium and the Fermi energy is strictly no longer defined. However, we may define a quasi-Fermi level for electrons and a quasi-Fermi level for holes that apply for nonequilibrium. If δn and δp are the excess electron and hole concentrations, respectively, we may write

$$n_0 + \delta n = n_i \exp\left(\frac{E_{Fn} - E_{Fi}}{kT}\right) \tag{6.85a}$$

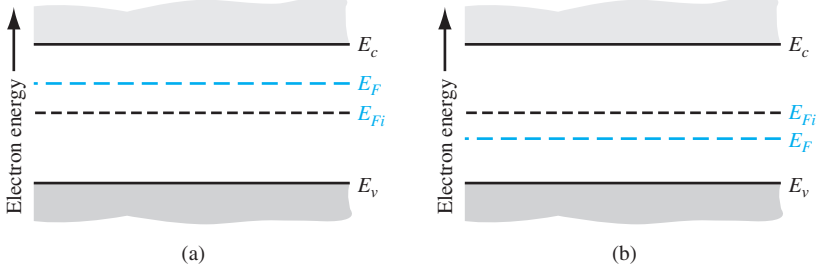


Figure 6.14 | Thermal-equilibrium energy-band diagrams for (a) n-type semiconductor and (b) p-type semiconductor.

and

$$p_0 + \delta p = n_i \exp\left(\frac{E_{Fi} - E_{Fp}}{kT}\right) \tag{6.85b}$$

where E_{Fn} and E_{Fp} are the quasi-Fermi energy levels for electrons and holes, respectively. The total electron concentration and the total hole concentration are functions of the quasi-Fermi levels.

EXAMPLE 6.7

Objective: Calculate the quasi-Fermi energy levels.

Consider an n-type semiconductor at $T = 300$ K with carrier concentrations of $n_0 = 10^{15}$ cm^{-3}, $n_i = 10^{10}$ cm^{-3}, and $p_0 = 10^5$ cm^{-3}. In nonequilibrium, assume that the excess carrier concentrations are $\delta n = \delta p = 10^{13}$ cm^{-3}.

■ Solution

The Fermi level for thermal equilibrium can be determined from Equation (6.84a). We have

$$E_F - E_{Fi} = kT \ln\left(\frac{n_0}{n_i}\right) = 0.2982 \text{ eV}$$

We can use Equation (6.85a) to determine the quasi-Fermi level for electrons in nonequilibrium. We can write

$$E_{Fn} - E_{Fi} = kT \ln\left(\frac{n_0 + \delta n}{n_i}\right) = 0.2984 \text{ eV}$$

Equation (6.85b) can be used to calculate the quasi-Fermi level for holes in nonequilibrium. We can write

$$E_{Fi} - E_{Fp} = kT \ln\left(\frac{p_0 + \delta p}{n_i}\right) = 0.179 \text{ eV}$$

■ Comment

We may note that the quasi-Fermi level for electrons is above E_{Fi} while the quasi-Fermi level for holes is below E_{Fi}.

■ EXERCISE PROBLEM

Ex 6.7 Impurity concentrations of $N_d = 3 \times 10^{15}$ cm^{-3} and $N_a = 10^{16}$ cm^{-3} are added to silicon at $T = 300$ K. Excess carriers are generated in the semiconductor such that the steady-state excess carrier concentrations are $\delta n = \delta p = 4 \times 10^{14}$ cm^{-3}. (*a*) Determine the thermal-equilibrium Fermi level with respect to the intrinsic Fermi level. (*b*) Find E_{Fn} and E_{Fp} with respect to E_{Fi}.

[Ans. (*a*) $E_{Fi} - E_F = 0.33808$ eV; (*b*) $E_{Fi} - E_{Fp} = 0.33952$ eV, $E_{Fn} - E_{Fi} = 0.26395$ eV]

Figure 6.15a shows the energy-band diagram with the Fermi energy level corresponding to thermal equilibrium. Figure 6.15b now shows the energy-band diagram under the nonequilibrium condition. Since the majority carrier electron concentration does not change significantly for this low-injection condition, the quasi-Fermi level for electrons is not much different from the thermal-equilibrium Fermi level.

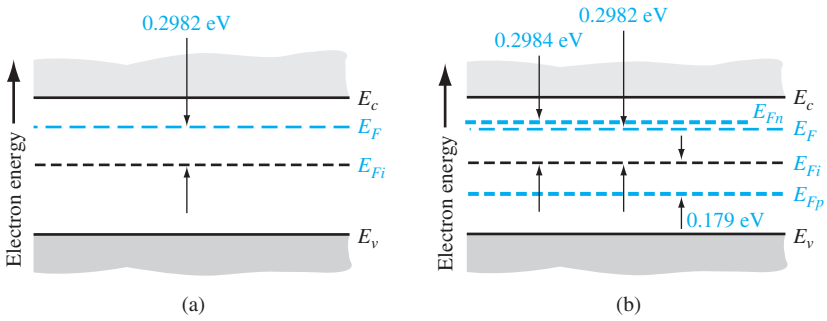


Figure 6.15 | (a) Thermal-equilibrium energy-band diagram for $N_d = 10^{15}$ cm^{-3} and $n_i = 10^{10}$ cm^{-3}. (b) Quasi-Fermi levels for electrons and holes if 10^{13} cm^{-3} excess carriers are present.

The quasi-Fermi energy level for the minority carrier holes is significantly different from the Fermi level and illustrates the fact that we have deviated from thermal equilibrium significantly. Since the electron concentration has increased, the quasi-Fermi level for electrons has moved slightly closer to the conduction band. The hole concentration has increased significantly so that the quasi-Fermi level for holes has moved much closer to the valence band. We will consider the quasi-Fermi energy levels again when we discuss forward-biased pn junctions.

*6.5 | EXCESS CARRIER LIFETIME

The rate at which excess electrons and holes recombine is an important characteristic of the semiconductor and influences many of the device characteristics, as we will see in later chapters. We considered recombination briefly at the beginning of this chapter and argued that the recombination rate is inversely proportional to the mean carrier lifetime. We have assumed up to this point that the mean carrier lifetime is simply a parameter of the semiconductor material.

We have been considering an ideal semiconductor in which electronic energy states do not exist within the forbidden-energy bandgap. This ideal effect is present in a perfect single-crystal material with an ideal periodic-potential function. In a real semiconductor material, defects occur within the crystal and disrupt the perfect periodic-potential function. If the density of these defects is not too great, the defects will create discrete electronic energy states within the forbidden-energy band. These allowed energy states may be the dominant effect in determining the mean carrier lifetime. The mean carrier lifetime may be determined from the Shockley–Read–Hall theory of recombination.

6.5.1 Shockley–Read–Hall Theory of Recombination

An allowed energy state, also called a *trap,* within the forbidden bandgap may act as a *recombination center,* capturing both electrons and holes with almost equal probability. This equal probability of capture means that the capture cross sections for

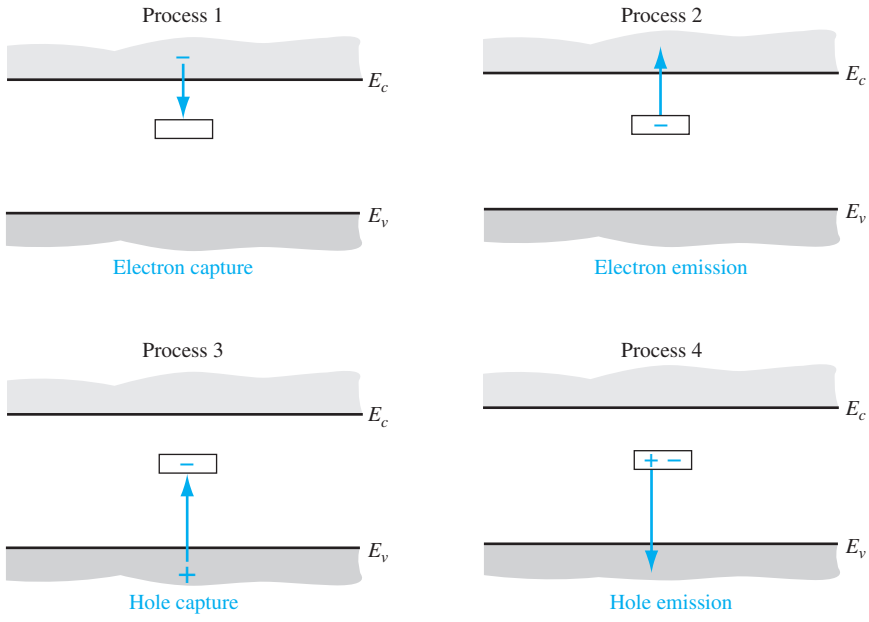


Figure 6.16 | The four basic trapping and emission processes for the case of an acceptor-type trap.

electrons and holes are approximately equal. The Shockley–Read–Hall theory of recombination assumes that a single recombination center, or trap, exists at an energy E_t within the bandgap. There are four basic processes, shown in Figure 6.16, that may occur at this single trap. We will assume that the trap is an acceptor-type trap; that is, it is negatively charged when it contains an electron and is neutral when it does not contain an electron.

The four basic processes are as follows:

Process 1: The capture of an electron from the conduction band by an initially neutral empty trap.

Process 2: The inverse of process 1—the emission of an electron that is initially occupying a trap level back into the conduction band.

Process 3: The capture of a hole from the valence band by a trap containing an electron. (Or we may consider the process to be the emission of an electron from the trap into the valence band.)

Process 4: The inverse of process 3—the emission of a hole from a neutral trap into the valence band. (Or we may consider this process to be the capture of an electron from the valence band.)

In process 1, the rate at which electrons from the conduction band are captured by the traps is proportional to the density of electrons in the conduction band and

proportional to the density of empty trap states. We can then write the electron capture rate as

$$R_{cn} = C_n N_t [1 - f_F(E_t)] n \tag{6.86}$$

where

R_{cn} = capture rate (#/cm^3-s)
C_n = constant proportional to electron-capture cross section
N_t = total concentration of trapping centers
n = electron concentration in the conduction band
$f_F(E_t)$ = Fermi function at the trap energy

The Fermi function at the trap energy is given by

$$f_F(E_t) = \frac{1}{1 + \exp\left(\dfrac{E_t - E_F}{kT}\right)} \tag{6.87}$$

which is the probability that a trap will contain an electron. The function $[1 - f_F(E_t)]$ is then the probability that the trap is empty. In Equation (6.87), we have assumed that the degeneracy factor is 1, which is the usual approximation made in this analysis. However, if a degeneracy factor is included, it will eventually be absorbed in other constants later in the analysis.

For process 2, the rate at which electrons are emitted from filled traps back into the conduction band is proportional to the number of filled traps, so that

$$R_{en} = E_n N_t f_F(E_t) \tag{6.88}$$

where

R_{en} = emission rate (#/cm^3-s)
E_n = constant
$f_F(E_t)$ = probability that the trap is occupied

In thermal equilibrium, the rate of electron capture from the conduction band and the rate of electron emission back into the conduction band must be equal. Then

$$R_{en} = R_{cn} \tag{6.89}$$

so that

$$E_n N_t f_{F0}(E_t) = C_n N_t [1 - f_{F0}(E_t)] n_0 \tag{6.90}$$

where f_{F0} denotes the thermal-equilibrium Fermi function. Note that, in thermal equilibrium, the value of the electron concentration in the capture rate term is the equilibrium value n_0. Using the Boltzmann approximation for the Fermi function, we can find E_n in terms of C_n as

$$E_n = n' C_n \tag{6.91}$$

where n' is defined as

$$n' = N_c \exp\left[\frac{-(E_c - E_t)}{kT}\right] \tag{6.92}$$

The parameter n' is equivalent to an electron concentration that would exist in the conduction band if the trap energy E_t coincided with the Fermi energy E_F.

In nonequilibrium, excess electrons exist, so that the net rate at which electrons are captured from the conduction band is given by

$$R_n = R_{cn} - R_{en} \tag{6.93}$$

which is just the difference between the capture rate and the emission rate. Combining Equations (6.86) and (6.88) with (6.93) gives

$$R_n = [C_n N_t (1 - f_F(E_t))n] - [E_n N_t f_F(E_t)] \tag{6.94}$$

We may note that, in this equation, the electron concentration n is the total concentration, which includes the excess electron concentration. The remaining constants and terms in Equation (6.94) are the same as defined previously and the Fermi energy in the Fermi probability function needs to be replaced by the quasi-Fermi energy for electrons. The constants E_n and C_n are related by Equation (6.91), so the net recombination rate can be written as

$$R_n = C_n N_t [n(1 - f_F(E_t)) - n' f_F(E_t)] \tag{6.95}$$

If we consider processes 3 and 4 in the recombination theory, the net rate at which holes are captured from the valence band is given by

$$R_p = C_p N_t [p f_F(E_t) - p'(1 - f_F(E_t))] \tag{6.96}$$

where C_p is a constant proportional to the hole capture rate, and p' is given by

$$p' = N_v \exp\left[\frac{-(E_t - E_v)}{kT}\right] \tag{6.97}$$

In a semiconductor in which the trap density is not too large, the excess electron and hole concentrations are equal and the recombination rates of electrons and holes are equal. If we set Equation (6.95) equal to Equation (6.96) and solve for the Fermi function, we obtain

$$f_F(E_t) = \frac{C_n n + C_p p'}{C_n(n + n') + C_p(p + p')} \tag{6.98}$$

We may note that $n'p' = n_i^2$. Then, substituting Equation (6.98) back into either Equation (6.95) or (6.96) gives

$$\boxed{R_n = R_p = \frac{C_n C_p N_t (np - n_i^2)}{C_n(n + n') + C_p(p + p')} \equiv R} \tag{6.99}$$

Equation (6.99) is the recombination rate of electrons and holes due to the recombination center at $E = E_t$. If we consider thermal equilibrium, then $np = n_0 p_0 = n_i^2$, so

that $R_n = R_p = 0$. Equation (6.99), then, is the recombination rate of excess electrons and holes.

Since R in Equation (6.99) is the recombination rate of the excess carriers, we may write

$$R = \frac{\delta n}{\tau} \tag{6.100}$$

where δn is the excess carrier concentration and τ is the lifetime of the excess carriers.

6.5.2 Limits of Extrinsic Doping and Low Injection

We simplified the ambipolar transport equation, Equation (6.39), from a nonlinear differential equation to a linear differential equation by applying limits of extrinsic doping and low injection. We may apply these same limits to the recombination rate equation.

Consider an n-type semiconductor under low injection. Then

$$n_0 \gg p_0, \quad n_0 \gg \delta p, \quad n_0 \gg n', \quad n_0 \gg p'$$

where δp is the excess minority carrier hole concentration. The assumptions of $n_0 \gg n'$ and $n_0 \gg p'$ imply that the trap level energy is near midgap so that n' and p' are not too different from the intrinsic carrier concentration. With these assumptions, Equation (6.99) reduces to

$$R = C_p N_t \delta p \tag{6.101}$$

The recombination rate of excess carriers in the n-type semiconductor is a function of the parameter C_p, which is related to the minority carrier hole capture cross section. The recombination rate, then, is a function of the minority carrier parameter in the same way that the ambipolar transport parameters reduced to their minority carrier values.

The recombination rate is related to the mean carrier lifetime. Comparing Equations (6.100) and (6.101), we may write

$$R = \frac{\delta n}{\tau} = C_p N_t \delta p \equiv \frac{\delta p}{\tau_{p0}} \tag{6.102}$$

where

$$\tau_{p0} = \frac{1}{C_p N_t} \tag{6.103}$$

and where τ_{p0} is defined as the excess minority carrier hole lifetime. If the trap concentration increases, the probability of excess carrier recombination increases; thus, the excess minority carrier lifetime decreases.

Similarly, if we have a strongly extrinsic p-type material under low injection, we can assume that

$$p_0 \gg n_0, \quad p_0 \gg \delta n, \quad p_0 \gg n', \quad p_0 \gg p'$$

The lifetime then becomes that of the excess minority carrier electron lifetime, or

$$\tau_{n0} = \frac{1}{C_n N_t} \tag{6.104}$$

Again note that for the n-type material, the lifetime is a function of C_p, which is related to the capture rate of the minority carrier hole. And for the p-type material, the lifetime is a function of C_n, which is related to the capture rate of the minority carrier electron. The excess carrier lifetime for an extrinsic material under low injection reduces to that of the minority carrier.

EXAMPLE 6.8

Objective: Determine the excess carrier lifetime in an intrinsic semiconductor.

If we substitute the definitions of excess carrier lifetimes from Equations (6.103) and (6.104) into Equation (6.99), the recombination rate can be written as

$$R = \frac{\left(np - n_i^2\right)}{\tau_{p0}(n + n') + \tau_{n0}(p + p')} \tag{6.105}$$

Consider an intrinsic semiconductor containing excess carriers. Then $n = n_i + \delta n$ and $p = n_i + \delta n$. Also assume that $n' = p' = n_i$.

■ Solution

Equation (6.105) now becomes

$$R = \frac{2n_i\delta n + (\delta n)^2}{(2n_i + \delta n)(\tau_{p0} + \tau_{n0})}$$

If we also assume very low injection, so that $\delta n \ll 2n_i$, then we can write

$$R = \frac{\delta n}{\tau_{p0} + \tau_{n0}} = \frac{\delta n}{\tau}$$

where τ is the excess carrier lifetime. We see that $\tau = \tau_{p0} + \tau_{n0}$ in the intrinsic material.

■ Comment

The excess carrier lifetime increases as we change from an extrinsic to an intrinsic semiconductor.

■ EXERCISE PROBLEM

Ex 6.8 Consider silicon at $T = 300$ K doped at concentrations of $N_d = 10^{15}$ cm^{-3} and $N_a = 0$. Assume that $n' = p' = n_i$ in the excess carrier recombination rate equation and assume parameter values of $\tau_{n0} = \tau_{p0} = 5 \times 10^{-7}$ s. Calculate the recombination rate of excess carriers if $\delta n = \delta p = 10^{14}$ cm^{-3}.

(Ans. 1.83×10^{20} cm^{-3} s^{-1})

Intuitively, we can see that the number of majority carriers that are available for recombining with excess minority carriers decreases as the extrinsic semiconductor becomes intrinsic. Since there are fewer carriers available for recombining in the intrinsic material, the mean lifetime of an excess carrier increases.

*6.6 | SURFACE EFFECTS

In all previous discussions, we have implicitly assumed the semiconductors were infinite in extent; thus, we were not concerned with any boundary conditions at a semiconductor surface. In any real application of semiconductors, the material is not infinitely large and therefore surfaces do exist between the semiconductor and an adjacent medium.

6.6.1 Surface States

When a semiconductor is abruptly terminated, the perfect periodic nature of the idealized single-crystal lattice ends abruptly at the surface. The disruption of the periodic-potential function results in allowed electronic energy states within the energy bandgap. In the previous section, we argued that simple defects in the semiconductor would create discrete energy states within the bandgap. The abrupt termination of the periodic potential at the surface results in a distribution of allowed energy states within the bandgap, shown schematically in Figure 6.17 along with the discrete energy states in the bulk semiconductor.

The Shockley–Read–Hall recombination theory shows that the excess minority carrier lifetime is inversely proportional to the density of trap states. We may argue that since the density of traps at the surface is larger than in the bulk, the excess minority carrier lifetime at the surface will be smaller than the corresponding lifetime in the bulk material. If we consider an extrinsic n-type semiconductor, for example, the recombination rate of excess carriers in the bulk, given by Equation (6.102), is

$$R = \frac{\delta p}{\tau_{p0}} \equiv \frac{\delta p_B}{\tau_{p0}} \tag{6.106}$$

where δp_B is the concentration of excess minority carrier holes in the bulk material. We may write a similar expression for the recombination rate of excess carriers at the surface as

$$R_s = \frac{\delta p_s}{\tau_{p0s}} \tag{6.107}$$

where δp_s is the excess minority carrier hole concentration at the surface and τ_{p0s} is the excess minority carrier hole lifetime at the surface.

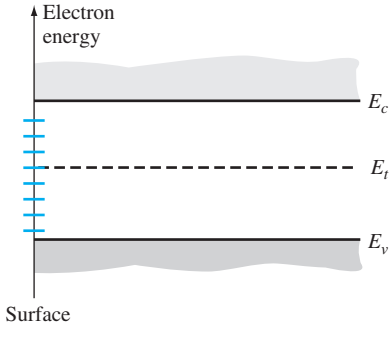


Figure 6.17 | Distribution of surface states within the forbidden bandgap.

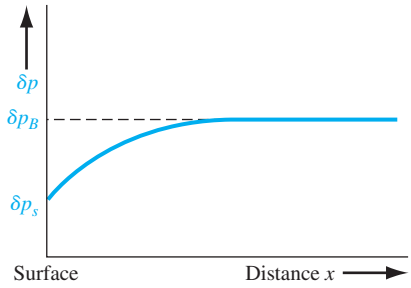


Figure 6.18 | Steady-state excess hole concentration versus distance from a semiconductor surface.

Assume that excess carriers are being generated at a constant rate throughout the entire semiconductor material. We showed that, in steady state, the generation rate is equal to the recombination rate for the case of a homogeneous, infinite semiconductor. Using this argument, the recombination rates at the surface and in the bulk material must be equal. Since $\tau_{p0s} < \tau_{p0}$, then the excess minority carrier concentration at the surface is smaller than the excess minority carrier concentration in the bulk region, or $\delta p_s < \delta p_B$. Figure 6.18 shows an example of the excess carrier concentration plotted as a function of distance from the semiconductor surface.

EXAMPLE 6.9

Objective: Determine the steady-state excess carrier concentration as a function of distance from the surface of a semiconductor.

Consider Figure 6.18, in which the surface is at $x = 0$. Assume that in the n-type semiconductor $\delta p_B = 10^{14}$ cm^{-3} and $\tau_{p0} = 10^{-6}$ s in the bulk, and $\tau_{p0s} = 10^{-7}$ s at the surface. Assume zero applied electric field and let $D_p = 10$ cm^2/s.

■ Solution

From Equations (6.106) and (6.107), we have

$$\frac{\delta p_B}{\tau_{p0}} = \frac{\delta p_s}{\tau_{p0s}}$$

so that

$$\delta p_s = \delta p_B\left(\frac{\tau_{p0s}}{\tau_{p0}}\right) = (10^{14})\left(\frac{10^{-7}}{10^{-6}}\right) = 10^{13}\ \text{cm}^{-3}$$

From Equation (6.56), we can write

$$D_p\frac{d^2(\delta p)}{dx^2} + g' - \frac{\delta p}{\tau_{p0}} = 0 \tag{6.108}$$

The generation rate can be determined from the steady-state conditions in the bulk, or

$$g' = \frac{\delta p_B}{\tau_{p0}} = \frac{10^{14}}{10^{-6}} = 10^{20}\ \text{cm}^{-3}\text{-s}^{-1}$$

The solution to Equation (6.107) is of the form

$$\delta p(x) = g'\tau_{p0} + Ae^{x/L_p} + Be^{-x/L_p} \tag{6.109}$$

As $x \to +\infty$, $\delta p(x) = \delta p_B = g'\tau_{p0} = 10^{14}$ cm^{-3}, which implies that $A = 0$. At $x = 0$, we have

$$\delta p(0) = \delta p_s = 10^{14} + B = 10^{13} \text{ cm}^{-3}$$

so that $B = -9 \times 10^{13}$. The entire solution for the minority carrier hole concentration as a function of distance from the surface is

$$\delta p(x) = 10^{14}\,(1 - 0.9e^{-x/L_p})$$

where

$$L_p = \sqrt{D_p\tau_{p0}} = \sqrt{(10)(10^{-6})} = 31.6\ \mu\text{m}$$

■ Comment

The excess carrier concentration is smaller at the surface than in the bulk.

■ EXERCISE PROBLEM

Ex 6.9 (*a*) Repeat Example 6.9 for the case when $\tau_{p0s} = 0$. (*b*) What is the excess hole concentration at $x = 0$? (*c*) For this particular case, what is the recombination rate of excess carriers at the surface?

[Ans. (a) $\delta p(x) = g'\tau_{p0}(1 - e^{-x/L_p})$; (b) $\delta p(0) = 0$; (c) $R' = \infty$]

6.6.2 Surface Recombination Velocity

A gradient in the excess carrier concentration exists near the surface as shown in Figure 6.18; excess carriers from the bulk region diffuse toward the surface where they recombine. This diffusion toward the surface can be described by

$$-D_p\left[\hat{n} \cdot \frac{d(\delta p)}{dx}\right]\Bigg|_{\text{surf}} = s\delta p|_{\text{surf}} \tag{6.110}$$

where each side of the equation is evaluated at the surface. The parameter $\hat{n}$ is the unit outward vector normal to the surface. Using the geometry of Figure 6.18, $d(\delta p)/dx$ is a positive quantity and $\hat{n}$ is negative, so that the parameter s is a positive quantity.

A dimensional analysis of Equation (6.110) shows that the parameter s has units of cm/s, or velocity. The parameter s is called the *surface recombination velocity*. If the excess concentrations at the surface and in the bulk region were equal, then the gradient term would be zero and the surface recombination velocity would be zero. As the excess concentration at the surface becomes smaller, the gradient term becomes larger, and the surface recombination velocity increases. The surface recombination velocity gives some indication of the surface characteristics as compared with the bulk region.

Equation (6.110) may be used as a boundary condition to the general solution given by Equation (6.109) in Example 6.8. Using Figure 6.18, we have that $\hat{n} = -1$, and Equation (6.110) becomes

$$D_p \frac{d(\delta p)}{dx}\bigg|_{\text{surf}} = s\delta p|_{\text{surf}} \tag{6.111}$$

We have argued that the coefficient A is zero in Equation (6.109). Then, from Equation (6.109), we can write that

$$\delta p_{\text{surf}} = \delta p(0) = g'\tau_{p0} + B \tag{6.112a}$$

and

$$\frac{d(\delta p)}{dx}\bigg|_{\text{surf}} = \frac{d(\delta p)}{dx}\bigg|_{x=0} = -\frac{B}{L_p} \tag{6.112b}$$

Substituting Equations (6.112a) and (6.112b) into Equation (6.111) and solving for the coefficient B, we obtain

$$B = \frac{-sg'\tau_{p0}}{(D_p/L_p) + s} \tag{6.113}$$

The excess minority carrier hole concentration can then be written as

$$\boxed{\delta p(x) = g'\tau_{p0}\left(1 - \frac{sL_p e^{-x/L_p}}{D_p + sL_p}\right)} \tag{6.114}$$

EXAMPLE 6.10

Objective: Determine the value of surface recombination velocity corresponding to the parameters given in Example 6.9.

From Example 6.9, we have that $g'\tau_{p0} = 10^{14}\ \text{cm}^{-3}$, $D_p = 10\ \text{cm}^2/\text{s}$, $L_p = 31.6\ \mu\text{m}$, and $\delta p(0) = 10^{13}\,\text{cm}^{-3}$.

■ Solution

Writing Equation (6.114) at the surface, we have

$$\delta p(0) = g'\tau_{p0}\left[1 - \frac{s}{(D_p/L_p) + s}\right]$$

Solving for the surface recombination velocity, we find that

$$s = \frac{D_p}{L_p}\left(\frac{g'\tau_{p0}}{\delta p(0)} - 1\right)$$

which becomes

$$s = \frac{10}{31.6 \times 10^{-4}}\left[\frac{10^{14}}{10^{13}} - 1\right] = 2.85 \times 10^4\ \text{cm/s}$$

■ Comment

This example shows that a surface recombination velocity of approximately $s = 3 \times 10^4$ cm/s could seriously degrade the performance of semiconductor devices, such as solar cells, since these devices tend to be fabricated close to a surface.

■ EXERCISE PROBLEM

Ex 6.10 (*a*) Using Equation (6.114), determine $\delta p(x)$ for (*i*) $s = \infty$ and (*ii*) $s = 0$. (*b*) What does (*i*) an infinite surface recombination velocity ($s = \infty$) and (*ii*) a zero surface recombination velocity ($s = 0$) imply?

[Ans. (*a*) (*i*) $\delta p(x) = g'\tau_{p0}(1 - e^{-x/L_p})$, (*ii*) $\delta p(x) = g'\tau_{p0}$; (*b*) (*i*) $\delta p(0) = 0$, (*ii*) $\delta p(0) = \text{constant}$ and $\delta p(x) = g'\tau_{p0}$]

In the above example, the surface influences the excess carrier concentration to the extent that, even at a distance of $L_p = 31.6\ \mu$m from the surface, the excess carrier concentration is only two-thirds of the value in the bulk. We will see in later chapters that device performance is dependent in large part on the properties of excess carriers.

6.7 | SUMMARY

- The processes of excess electron and hole generation and recombination were discussed. The excess carrier generation rate and recombination rate were defined.
- Excess electrons and holes do not move independently of each other, but move together. This common movement is called ambipolar transport.
- The ambipolar transport equation was derived and limits of low injection and extrinsic doping were applied to the coefficients. Under these conditions, the excess electrons and holes diffuse and drift together with the characteristics of the minority carrier, a result that is fundamental to the behavior of semiconductor devices.
- The concept of excess carrier lifetime was developed.
- Examples of excess carrier behavior as a function of time, as a function of space, and as a function of both time and space were examined.
- The quasi-Fermi level for electrons and the quasi-Fermi level for holes were defined. The degree of quasi-Fermi level splitting is a measure of departure from thermal equilibrium.
- The Shockley–Read–Hall theory of recombination was considered. Expressions for the excess minority carrier lifetime were developed. Generation and recombination of excess carriers increase as a result of traps in a semiconductor.
- The effect of a semiconductor surface influences the behavior of excess electrons and holes. The surface recombination velocity was defined.

GLOSSARY OF IMPORTANT TERMS

ambipolar diffusion coefficient The effective diffusion coefficient of excess carriers.

ambipolar mobility The effective mobility of excess carriers.

ambipolar transport The process whereby excess electrons and holes diffuse, drift, and recombine with the same effective diffusion coefficient, mobility, and lifetime.

ambipolar transport equation The equation describing the behavior of excess carriers as a function of time and space coordinates.

carrier generation The process of elevating electrons from the valence band into the conduction band, creating an electron–hole pair.

carrier recombination The process whereby an electron "falls" into an empty state in the valence band (a hole) so that an electron–hole pair is annihilated.

excess carriers The term describing both excess electrons and excess holes.

excess electrons The concentration of electrons in the conduction band over and above the thermal-equilibrium concentration.

excess holes The concentration of holes in the valence band over and above the thermal-equilibrium concentration.

excess minority carrier lifetime The average time that an excess minority carrier exists before it recombines.

generation rate The rate (#/cm^3-s) at which electron–hole pairs are created.

low-level injection The condition in which the excess carrier concentration is much smaller than the thermal-equilibrium majority carrier concentration.

minority carrier diffusion length The average distance a minority carrier diffuses before recombining: a parameter equal to $\sqrt{D\tau}$ where D and τ are the minority carrier diffusion coefficient and lifetime, respectively.

quasi-Fermi level The quasi-Fermi level for electrons and the quasi-Fermi level for holes relate the nonequilibrium electron and hole concentrations, respectively, to the intrinsic carrier concentration and the intrinsic Fermi level.

recombination rate The rate (#/cm^3-s) at which electron–hole pairs recombine.

surface recombination velocity A parameter that relates the gradient of the excess carrier concentration at a surface to the surface concentration of excess carriers.

surface states The electronic energy states that exist within the bandgap at a semiconductor surface.

CHECKPOINT

After studying this chapter, the reader should have the ability to:

- Describe the concept of excess carrier generation and recombination.
- Describe the concept of an excess carrier lifetime.
- Describe how the time-dependent diffusion equations for holes and electrons are derived.
- Describe how the ambipolar transport equation is derived.
- Understand the consequence of the coefficients in the ambipolar transport equation reducing to the minority carrier values under low injection and extrinsic semiconductors.
- Apply the ambipolar transport equation to various problems.
- Understand the concept of the dielectric relaxation time constant and what it means.
- Calculate the quasi-Fermi levels for electrons and holes.
- Calculate the excess carrier recombination rate for a given concentration of excess carriers.
- Understand the effect of a surface on the excess carrier concentrations.

REVIEW QUESTIONS

1. Why are the electron generation rate and recombination rate equal in thermal equilibrium?
2. Define the excess carrier recombination rate in terms of excess carrier concentration and lifetime.
3. Explain how the density of holes, for example, can change as a result of a change in the flux of particles.
4. Why is the general ambipolar transport equation nonlinear?
5. Explain qualitatively why a pulse of excess electrons and holes would move together in the presence of an applied electric field.
6. Explain qualitatively why the excess carrier lifetime reduces to that of the minority carrier under low injection.
7. What is the time dependence of the density of excess carriers when the generation rate becomes zero?
8. In the presence of an external force, why doesn't the density of excess carriers continue to increase with time?
9. When a concentration of one type of excess carrier is suddenly created in a semiconductor, what is the mechanism by which the net charge density quickly becomes zero?
10. State the definition of the quasi-Fermi level for electrons. Repeat for holes.
11. Explain why the presence of traps in a semiconductor increases the recombination rate of excess carriers.
12. Why, in general, is the concentration of excess carriers less at the surface of a semiconductor than in the bulk?

PROBLEMS

(*Note:* Use the semiconductor parameters listed in Appendix B if they are not specifically given in a problem. Assume $T = 300$ K.)

Section 6.1 Carrier Generation and Recombination

6.1 Consider silicon at $T = 300$ K that is doped with donor impurity atoms to a concentration of $N_d = 5 \times 10^{15}$ cm^{-3}. The excess carrier lifetime is 2×10^{-7}s. (*a*) Determine the thermal equilibrium recombination rate of holes. (*b*) Excess carriers are generated such that $\delta n = \delta p = 10^{14}$ cm^{-3}. What is the recombination rate of holes for this condition?

6.2 GaAs, at $T = 300$ K, is uniformly doped with acceptor impurity atoms to a concentration of $N_a = 2 \times 10^{16}$ cm^{-3}. Assume an excess carrier lifetime of 5×10^{-7}s. (*a*) Determine the electron–hole recombination rate if the excess electron concentration is $\delta n = 5 \times 10^{14}$ cm^{-3}. (*b*) Using the results of part (*a*), what is the lifetime of holes?

6.3 An n-type silicon sample contains a donor concentration of $N_d = 10^{16}$ cm^{-3}. The minority carrier hole lifetime is found to be $\tau_{p0} = 20$ μs. (*a*) What is the lifetime of the majority carrier electrons? (*b*) Determine the thermal-equilibrium generation rate for electrons and holes in this material. (*c*) Determine the thermal-equilibrium recombination rate for electrons and holes in this material.

6.4 (*a*) A sample of semiconductor has a cross-sectional area of 1 cm^2 and a thickness of 0.1 cm. Determine the number of electron–hole pairs that are generated per unit volume per unit time by the uniform absorption of 1 watt of light at a wavelength of 6300 Å. Assume each photon creates one electron–hole pair. (*b*) If the excess minority carrier lifetime is 10 μs, what is the steady-state excess carrier concentration?

Section 6.2 Mathematical Analysis of Excess Carriers

6.5 Derive Equation (6.27) from Equations (6.18) and (6.20).

6.6 Consider a one-dimensional hole flux as shown in Figure 6.4. If the generation rate of holes in this differential volume is $g_p = 10^{20}$ cm^{-3}-s^{-1} and the recombination rate is 2×10^{19} cm^{-3}-s^{-1}, what must be the gradient in the particle current density to maintain a steady-state hole concentration?

6.7 Repeat Problem 6.6 if the generation rate becomes zero.

Section 6.3 Ambipolar Transport

6.8 Starting with the continuity equations given by Equations (6.29) and (6.30), derive the ambipolar transport equation given by Equation (6.39).

6.9 A silicon sample at $T = 300$ K has a uniform acceptor concentration of 7×10^{15} cm^{-3}. The excess carrier lifetime is $\tau_{n0} = 10^{-7}$ s. (*a*) Determine the ambipolar mobility. (*b*) Find the ambipolar diffusion coefficient. (*c*) What are the electron and hole lifetimes?

6.10 Germanium at $T = 300$ K is uniformly doped with donor impurity atoms to a concentration of 4×10^{13} cm^{-3}. The excess carrier lifetime is found to be $\tau_{p0} = 2 \times 10^{-6}$ s. (*a*) Determine the ambipolar (*i*) diffusion coefficient and (*ii*) mobility. (*b*) Find the electron and hole lifetimes.

6.11 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

$$\Delta\sigma = e(\mu_n + \mu_p)\tau_{p0}g'$$

6.12 Consider a silicon sample at $T = 300$ K that is uniformly doped with acceptor impurity atoms at a concentration of $N_a = 10^{16}$ cm^{-3}. At $t = 0$, a light source is turned on generating excess carriers uniformly throughout the sample at a rate of $g' = 8 \times 10^{20}$ cm^{-3} s^{-1}. Assume the minority carrier lifetime is $\tau_{n0} = 5 \times 10^{-7}$ s, and assume mobility values of $\mu_n = 900$ cm^2/V-s and $\mu_p = 380$ cm^2/V-s. (*a*) Determine the conductivity of the silicon as a function of time for $t \geq 0$. (*b*) What is the value of conductivity at (*i*) $t = 0$ and (*ii*) $t = \infty$?

6.13 An n-type GaAs semiconductor at $T = 300$ K is uniformly doped at $N_d = 5 \times 10^{15}$ cm^{-3}. The minority carrier lifetime is $\tau_{p0} = 5 \times 10^{-8}$ s. Assume mobility values of $\mu_n = 7500$ cm^2/V-s and $\mu_p = 310$ cm^2/V-s. A light source is turned on at $t = 0$ generating excess carriers uniformly at a rate of $g' = 4 \times 10^{21}$ cm^{-3} s^{-1} and turns off at $t = 10^{-6}$ s. (*a*) Determine the excess carrier concentrations versus time over the range $0 \leq t \leq \infty$. (*b*) Calculate the conductivity of the semiconductor versus time over the same time period as part (*a*).

6.14 A bar of silicon at $T = 300$ K has a length of $L = 0.05$ cm and a cross-sectional area of $A = 10^{-5}$ cm^2. The semiconductor is uniformly doped with $N_d = 8 \times 10^{15}$ cm^{-3} and $N_a = 2 \times 10^{15}$ cm^{-3}. A voltage of 10 V is applied across the length of the material. For

$t < 0$, the semiconductor has been uniformly illuminated with light, producing an excess carrier generation rate of $g' = 8 \times 10^{20}$ cm^{-3} s^{-1}. The minority carrier lifetime is $\tau_{p0} = 5 \times 10^{-7}$ s. At $t = 0$, the light source is turned off. Determine the current in the semiconductor as a function of time for $t \geq 0$.

6.15 Silicon at $T = 300$ K is uniformly doped with impurity atoms at concentrations of $N_a = 2 \times 10^{16}$ cm^{-3} and $N_d = 6 \times 10^{15}$ cm^{-3} and is in thermal equilibrium for $t < 0$. A light source turns on at $t = 0$ producing excess carriers with a uniform generation rate of $g' = 2 \times 10^{21}$ cm^{-3} s^{-1}. The electric field is zero. (*a*) If the maximum, steady-state excess carrier concentrations are $\delta n = \delta p = 5 \times 10^{14}$ cm^{-3}, determine the excess minority carrier lifetime. (*b*) Derive the expressions for the excess carrier concentration and excess carrier recombination rate as a function of time. (*c*) Determine the times at which the excess carrier concentration is equal to (*i*) one-fourth, (*ii*) one-half, (*iii*) three-fourths, and (*iv*) 95% of the steady-state value.

6.16 In a GaAs material at $T = 300$ K, the doping concentrations are $N_d = 8 \times 10^{15}$ cm^{-3} and $N_a = 2 \times 10^{15}$ cm^{-3}. The thermal equilibrium recombination rate is $R_o = 4 \times 10^4$ cm^{-3} s^{-1}. (*a*) What is the minority carrier lifetime? (*b*) A uniform generation rate for excess carriers results in an excess carrier recombination rate of $R' = 2 \times 10^{21}$ cm^{-3} s^{-1}. What is the steady-state excess carrier concentration? (*c*) What is the excess carrier lifetime?

6.17 (*a*) Consider a silicon sample at $T = 300$ K doped with 10^{16} cm^{-3} donor atoms. Let $\tau_{p0} = 5 \times 10^{-7}$ s. A light source turns on at $t = 0$ producing excess carriers with a uniform generation rate of $g' = 5 \times 10^{20}$ cm^{-3} s^{-1}. At $t = 5 \times 10^{-7}$ s, the light source turns off. (*i*) Derive the expression(s) for the excess carrier concentration as a function of time over the range $0 \leq t \leq \infty$. (*ii*) What is the value of the excess concentration when the light source turns off. (*b*) Repeat part (*a*) for the case when the light source turns off at $t = 2 \times 10^{-6}$ s. (*c*) Sketch the excess minority carrier concentrations versus time for parts (*a*) and (*b*).

6.18 A semiconductor is uniformly doped with 10^{17} cm^{-3} acceptor atoms and has the following properties: $D_n = 27$ cm^2/s, $D_p = 12$ cm^2/s, $\tau_{n0} = 5 \times 10^{-7}$ s, and $\tau_{p0} = 10^{-7}$ s. An external source has been turned on for $t < 0$ producing a uniform concentration of excess carriers at a generation rate of $g' = 10^{21}$ cm^{-3} s^{-1}. The source turns off at time $t = 0$ and back on at time $t = 2 \times 10^{-6}$ s. (*a*) Derive the expressions for the excess carrier concentration as a function of time for $0 \leq t \leq \infty$. (*b*) Determine the value of excess carrier concentration at (*i*) $t = 0$, (*ii*) $t = 2 \times 10^{-6}$ s, and (*iii*) $t = \infty$. (*c*) Plot the excess carrier concentration as a function of time.

6.19 Consider a bar of p-type silicon that is uniformly doped to a value of $N_a = 2 \times 10^{16}$ cm^{-3} at $T = 300$ K. The applied electric field is zero. A light source is incident on the end of the semiconductor as shown in Figure P6.19. The steady-state concentration of excess carriers generated at $x = 0$ is $\delta p(0) = \delta n(0) = 2 \times 10^{14}$ cm^{-3}. Assume the following

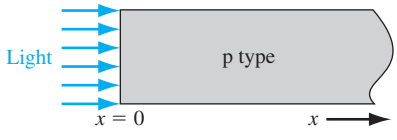


Figure P6.19 | Figure for Problems 6.19 and 6.21.

parameters: $\mu_n = 1200$ cm^2/V-s, $\mu_p = 400$ cm^2/V-s, $\tau_{n0} = 10^{-6}$ s, and $\tau_{p0} = 5 \times 10^{-7}$ s. Neglecting surface effects, (*a*) determine the steady-state excess electron and hole concentrations as a function of distance into the semiconductor, and (*b*) calculate the steady-state electron and hole diffusion current densities as a function of distance into the semiconductor.

6.20 The $x = 0$ end of an $N_a = 1 \times 10^{14}$ cm^{-3} doped semi-infinite ($x \geq 0$) bar of silicon maintained at $T = 300$ K is attached to a "minority carrier digester" which makes $n_p = 0$ at $x = 0$ (n_p is the minority carrier electron concentration in a p-type semiconductor). The electric field is zero. (*a*) Determine the thermal-equilibrium values of n_{p0} and p_{p0}. (*b*) What is the excess minority carrier concentration at $x = 0$? (*c*) Derive the expression for the steady-state excess minority carrier concentration as a function of x.

6.21 In a p-type silicon semiconductor, excess carriers are being generated at the end of the bar at $x = 0$ as shown in Figure P6.19. The uniform doping concentrations are $N_a = 7 \times 10^{16}$ cm^{-3} and $N_d = 2 \times 10^{16}$ cm^{-3}. The steady-state excess carrier concentrations at $x = 0$ are $\delta p(0) = \delta n(0) = 5 \times 10^{14}$ cm^{-3}. (Neglect surface effects.) The electric field is zero. Assume semiconductor parameters of $\tau_{n0} = \tau_{p0} = 10^{-6}$ s, $D_n = 25$ cm^2/s, and $D_p = 10$ cm^2/s. (*a*) Calculate δn and the electron and hole diffusion current densities at $x = 0$. (*b*) Repeat part (*a*) for $x = 5 \times 10^{-3}$ cm. (*c*) Repeat part (*a*) for $x = 15 \times 10^{-3}$ cm.

6.22 Consider an n-type silicon sample. Excess carriers are generated at $x = 0$ such as shown in Figure 6.6. A constant electric field E_0 is applied in the $+x$ direction. Show that the steady-state excess carrier concentration is given by

$$\delta p(x) = A \exp(s_- x) \qquad x > 0 \qquad \text{and} \qquad \delta p(x) = A \exp(s_+ x) \qquad x < 0$$

where

$$s_{\mp} = \frac{1}{L_p}\left[\beta \mp \sqrt{1+\beta^2}\right]$$

and

$$\beta = \frac{\mu_p L_p E_0}{2D_n}$$

6.23 Plot the excess carrier concentration $\delta p(x)$ versus x from Problem 6.22 for (*a*) $E_0 = 0$ and (*b*) $E_0 = 10$ V/cm.

***6.24** Consider the semiconductor described in Problem 6.19. Assume a constant electric field E_0 is applied in the $+x$ direction. (*a*) Derive the expression for the steady-state excess electron concentration. (Assume the solution is of the form $e^{-\alpha x}$.) (*b*) Plot δn versus x for (*i*) $E_0 = 0$ and (*ii*) $E_0 = 12$ V/cm. (*c*) Explain the general characteristics of the two curves plotted in part (*b*).

6.25 Assume that a p-type semiconductor is in thermal equilibrium for $t < 0$ and has an infinite minority carrier lifetime. Also assume that the semiconductor is uniformly illuminated, resulting in a uniform generation rate, $g'(t)$, which is given by

$$g'(t) = G_0' \qquad \text{for } 0 < t < T$$
$$g'(t) = 0 \qquad \text{for } t < 0 \text{ and } t > T$$

where G_0' is a constant. Find the excess minority carrier concentration as a function of time.

***6.26** Consider the n-type semiconductor shown in Figure P6.26. Illumination produces a constant excess carrier generation rate, G_0', in the region $-L < x < +L$. Assume that the

*Asterisks next to problems indicate problems that are more difficult.

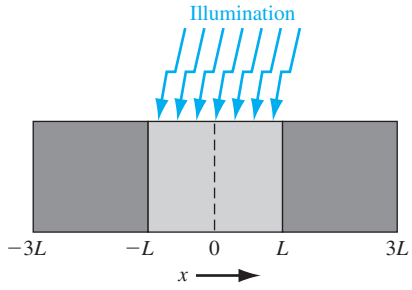


Figure P6.26 | Figure for Problem 6.26.

minority carrier lifetime is infinite and assume that the excess minority carrier hole concentration is zero at $x = -3L$ and at $x = +3L$. Find the steady-state excess minority carrier concentration versus x, for the case of low injection and for zero applied electric field.

6.27 An n-type semiconductor at $T = 300$ K is used in the Haynes–Shockley experiment. The length of the sample is 0.4 cm and the applied voltage is $V_1 = 8$V. The contacts A and B are separated by 0.25 cm. The peak of the pulse arrives at contact B 32 μs after carrier injection at contact A. The width of the pulse is $\Delta t = 9.35$ μs. Determine the hole mobility and diffusion coefficient. Compare the results with the Einstein relation.

6.28 Consider the function $f(x, t) = (4\pi Dt)^{-1/2} \exp(-x^2/4Dt)$. (*a*) Show that this function is a solution to the differential equation $D(\partial^2 f/\partial x^2) = \partial f/\partial t$. (*b*) Show that the integral of the function $f(x, t)$ over x from $-\infty$ to $+\infty$ is unity for all values of time. (*c*) Show that this function approaches a δ function as t approaches zero.

6.29 The basic equation in the Haynes–Shockley experiment is given by Equation (6.70). (*a*) Plot $\delta p(x, t)$ versus x for various values of t and for $E_0 = 0$ as well as for $E_0 \neq 0$. (*b*) Plot $\delta p(x, t)$ versus t for various values of x and for $E_0 = 0$ as well as for $E_0 \neq 0$.

Section 6.4 Quasi-Fermi Energy Levels

6.30 An n-type silicon semiconductor, doped at $N_d = 4 \times 10^{16}$ cm^{-3}, is steadily illuminated such that $g' = 2 \times 10^{21}$ cm^{-3} s^{-1}. Assume $\tau_{n0} = 10^{-6}$ s and $\tau_{p0} = 5 \times 10^{-7}$ s. (*a*) Determine the thermal-equilibrium value of $E_F - E_{Fi}$. (*b*) Calculate the quasi-Fermi levels for electrons and holes with respect to E_{Fi}. (*c*) What is the difference (in eV) between E_{Fn} and E_F ?

6.31 Consider a p-type silicon semiconductor at $T = 300$ K doped at $N_a = 5 \times 10^{15}$ cm^{-3}. (*a*) Determine the position of the Fermi level with respect to the intrinsic Fermi level. (*b*) Excess carriers are generated such that the excess carrier concentration is 10 percent of the thermal-equilibrium majority carrier concentration. Determine the quasi-Fermi levels with respect to the intrinsic Fermi level. (*c*) Plot the Fermi level and quasi-Fermi levels with respect to the intrinsic level.

6.32 Consider n-type silicon doped at $N_d = 5 \times 10^{15}$ cm^{-3}. It is found that $E_{Fn} - E_F = 1.02 \times 10^{-3}$ eV. (*a*) What is the excess carrier concentration? (*b*) Determine $E_{Fn} - E_{Fi}$. (*c*) Calculate $E_{Fi} - E_{Fp}$.

6.33 A p-type silicon sample is doped at $N_a = 6 \times 10^{15}$ cm^{-3}. It is determined that $E_{Fn} - E_{Fi} = 0.270$ eV. (*a*) Determine the excess carrier concentration. (*b*) Find $E_{Fi} - E_{Fp}$. (*c*) (*i*) Derive the expression for $E_F - E_{Fp}$. (*ii*) Find $E_F - E_{Fp}$.

6.34 Consider n-type GaAs doped at $N_d = 10^{16}$ cm^{-3}. (*a*) For an excess concentration of $\delta p = (0.02)N_d$, determine (*i*) $E_{Fn} - E_{Fi}$ and (*ii*) $E_{Fi} - E_{Fp}$. (*b*) Repeat part (*a*) if $\delta p = (0.1)N_d$.

6.35 A p-type gallium arsenide semiconductor at $T = 300$ K is doped at $N_a = 10^{16}$ cm^{-3}. The excess carrier concentration varies linearly from 10^{14} cm^{-3} to zero over a distance of 50 μm. Plot the position of the quasi-Fermi levels with respect to the intrinsic Fermi level versus distance.

6.36 Consider p-type silicon at $T = 300$ K doped to $N_a = 5 \times 10^{14}$ cm^{-3}. Assume excess carriers are present and assume that $E_F - E_{Fp} = (0.01)kT$. (*a*) Does this condition correspond to low injection? Why or why not? (*b*) Determine $E_{Fn} - E_{Fi}$.

6.37 An n-type silicon sample is doped with donors at a concentration of $N_d = 10^{16}$ cm^{-3}. Excess carriers are generated such that the excess hole concentration is given by $\delta p(x) = 10^{14} \exp(-x/10^{-4})$cm^{-3}. Plot the function $E_{Fi} - E_{Fp}$ versus x over the range $0 \le x \le 4 \times 10^{-4}$.

6.38 An n-type silicon semiconductor is doped at $N_d = 2 \times 10^{16}$ cm^{-3}. (*a*) Plot $E_{Fi} - E_{Fp}$ as a function of δp over the range $10^{11} \le \delta p \le 10^{15}$ cm^{-3}. (Use a log scale for δp.) (*b*) Repeat part (*a*) for $E_{Fn} - E_{Fi}$.

Section 6.5 Excess Carrier Lifetime

6.39 Consider Equation (6.99) and the definitions of τ_{p0} and τ_{n0} by Equations (6.103) and (6.104). Let $n' = p' = n_i$. Assume that in a particular region of a semiconductor, $n = p = 0$. (*a*) Determine the recombination rate R. (*b*) Explain what this result means physically.

6.40 Again consider Equation (6.99) and the definitions of τ_{p0} and τ_{n0} given by Equations (6.103) and (6.104). Let $\tau_{p0} = 10^{-7}$ s and $\tau_{n0} = 5 \times 10^{-7}$ s. Also let $n' = p' = n_i = 10^{10}$ cm^{-3}. Assume very low injection so that $\delta n \ll n_i$. Calculate $R/\delta n$ for a semiconductor which is (*a*) n-type ($n_0 \gg p_0$), (*b*) intrinsic ($n_0 = p_0 = n_i$), and (*c*) p-type ($p_0 \gg n_0$).

Section 6.6 Surface Effects

***6.41** Consider an n-type semiconductor as shown in Figure P6.41, doped at $N_d = 10^{16}$ cm^{-3} and with a uniform excess carrier generation rate equal to $g' = 10^{21}$ cm^{-3}-s^{-1}. Assume that $D_p = 10$ cm^2/s and $\tau_{p0} = 10^{-7}$ s. The electric field is zero. (*a*) Determine the steady-state excess minority carrier concentration versus x if the surface recombination velocity at $x = 0$ is (*i*) $s = 0$, (*ii*) $s = 2000$ cm/s, and (*iii*) $s = \infty$. (*b*) Calculate the excess minority carrier concentration at $x = 0$ for (*i*) $s = 0$, (*ii*) $s = 2000$ cm/s, and (*iii*) $s = \infty$.

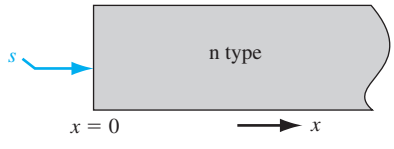


Figure P6.41 | Figure for Problem 6.41.

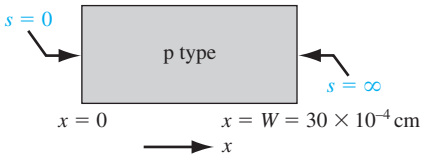


Figure P6.42 | Figure for Problem 6.42.

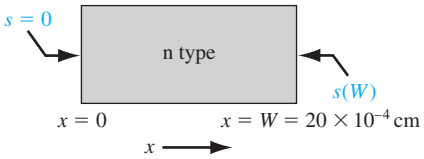


Figure P6.43 | Figure for Problem 6.43.

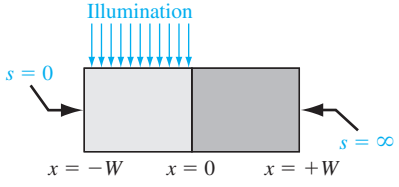


Figure P6.44 | Figure for Problem 6.44.

***6.42** (*a*) Consider the p-type semiconductor shown in Figure P6.42 with the following parameters: $N_a = 5 \times 10^{16}$ cm^{-3}, $D_n = 25$ cm^2/s, and $\tau_{n0} = 5 \times 10^{-7}$ s. The surface recombination velocities at the two surfaces are shown. The electric field is zero. The semiconductor is illuminated at $x = 0$ with an excess carrier generation rate equal to $g' = 2 \times 10^{21}$ cm^{-3}-s^{-1}. Determine the excess minority carrier electron concentration versus x in steady state. (*b*) Repeat part (*a*) for $\tau_{n0} = \infty$.

***6.43** Consider the n-type semiconductor shown in Figure P6.43. Assume that $D_p = 10$ cm^2/s and $\tau_{p0} = \infty$. The electric field is zero. Assume that a flux of excess electrons and holes is incident at $x = 0$. Let the flux of each carrier type be 10^{19} carriers/cm^2-s. Determine the minority carrier hole current versus x if the surface recombination velocity is (*a*) $s(W) = \infty$ and (*b*) $s(W) = 2000$ cm/s.

***6.44** A p-type semiconductor is shown in Figure P6.44. The surface recombination velocities are shown. The semiconductor is uniformly illuminated for $-W < x < 0$ producing a constant excess carrier generation rate G'_0. Determine the steady-state excess carrier concentration versus x if the minority carrier lifetime is infinite and if the electric field is zero.

6.45 Plot $\delta p(x)$ versus x for various values of s using Equation (6.113). Choose reasonable parameter values.

Summary and Review

***6.46** Consider an n-type semiconductor as shown in Figure P6.41. The material is doped at $N_d = 3 \times 10^{16}$ cm^{-3} and $N_a = 0$. Assume that $D_p = 12$ cm^2/s and $\tau_{p0} = 2 \times 10^{-7}$ s. The electric field is zero. "Design" the surface recombination velocity so that the minority carrier diffusion current density at the surface is no greater than $J_p = -0.18$ A/cm^2 with a uniform excess carrier generation rate equal to $g' = 3 \times 10^{21}$ cm^{-3}-s^{-1}.

6.47 Consider a semiconductor with excess carriers present. From the definition of carrier lifetimes and recombination rates, determine the average time that an electron

stays in the conduction band and the average time that a hole stays in the valence band. Discuss these relations for (*a*) an intrinsic semiconductor and (*b*) an n-type semiconductor.

***6.48** (*a*) Design a GaAs photoconductor that is 4 μm thick. Assume that the material is doped at $N_d = 10^{16}$ cm^{-3} and has lifetime values of $\tau_{n0} = 10^{-7}$ s and $\tau_{p0} = 5 \times 10^{-8}$ s. With an excitation of $g' = 10^{21}$ cm^{-3} s^{-1}, a photocurrent of at least 2 μA is required with an applied voltage of 2 V. (*b*) Repeat the design for a silicon photoconductor that has the same parameters as given in part (*a*).

READING LIST

1. Bube, R. H. *Photoelectronic Properties of Semiconductors,* New York: Cambridge University Press, 1992.

***2.** deCogan, D. *Solid State Devices: A Quantum Physics Approach.* New York: Springer-Verlag, 1987.

3. Dimitrijev, S. *Principles of Semiconductor Devices.* New York: Oxford University Press, 2006.

4. Hall, R. H. "Electron–Hole Recombination." *Physical Review* 87, no. 2 (July 15, 1952), p. 387.

5. Haynes, J. R., and W. Shockley. "The Mobility and Life of Injected Holes and Electrons in Germanium." *Physical Review* 81, no. 5 (March 1, 1951), pp. 835–843.

***6.** Hess, K. *Advanced Theory of Semiconductor Devices.* Englewood Cliffs, NJ: Prentice Hall, 1988.

7. Kano, K. *Semiconductor Devices.* Upper Saddle River, NJ: Prentice Hall, 1998.

8. Kingston, R. H. *Semiconductor Surface Physics.* Philadelphia, PA: University of Pennsylvania Press, 1957.

9. McKelvey, J. P. *Solid State Physics for Engineering and Materials Science.* Malabar, FL: Krieger Publishing, 1993.

10. Pierret, R. F. *Semiconductor Device Fundamentals.* Reading, MA: Addison-Wesley, 1996.

11. Shockley, W., and W. T. Read, Jr. "Statistics of the Recombinations of Holes and Electrons." *Physical Review* 87, no. 5 (September 1, 1952), pp. 835–842.

12. Singh, J. *Semiconductor Devices: An Introduction.* New York: McGraw-Hill, 1994.

13. Singh, J. *Semiconductor Devices: Basic Principles.* New York: John Wiley and Sons, 2001.

14. Streetman, B. G., and S. K. Banerjee. *Solid State Electronic Devices,* 6th ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2006.

***15.** Wang, S. *Fundamentals of Semiconductor Theory and Device Physics.* Englewood Cliffs, NJ: Prentice Hall, 1989.

16. Wolfe, C. M., N. Holonyak, Jr., and G. E. Stillman. *Physical Properties of Semiconductors*. Englewood Cliffs, NJ: Prentice Hall, 1989.

*Indicates references that are at an advanced level compared to this text.

CHAPTER 7

The pn Junction

Up to this point in the text, we have been considering the properties of the semiconductor material. We calculated electron and hole concentrations in thermal equilibrium and determined the position of the Fermi level. We then considered the nonequilibrium condition in which excess electrons and holes are present in the semiconductor. We now wish to consider the situation in which a p-type and an n-type semiconductor are brought into contact with one another to form a pn junction.

Most semiconductor devices contain at least one junction between p-type and n-type semiconductor regions. Semiconductor device characteristics and operation are intimately connected to these pn junctions, so considerable attention is devoted initially to this basic device.

The electrostatics of the zero-biased and reverse-biased pn junction is considered in this chapter. The current–voltage characteristics of the pn junction diode are developed in the next chapter. ■

7.0 | PREVIEW

In this chapter, we will:

- Consider a uniformly doped pn junction, in which one region of the semiconductor is uniformly doped with acceptor atoms and the adjacent region is uniformly doped with donor atoms.
- Determine the energy-band diagram of a pn junction in thermal equilibrium.
- Discuss the creation of a space charge region between the p and n regions.
- Apply Poisson's equation to determine the electric field in the space charge region and calculate the built-in potential barrier.
- Analyze the changes that occur in the pn junction when a reverse-biased voltage is applied. Derive expressions for space charge width and depletion capacitance.
- Analyze the voltage breakdown characteristics of a pn junction.
- Consider the properties of a nonuniformly doped pn junction. Specific doping profiles can lead to desirable properties of the pn junction.

7.1 | BASIC STRUCTURE OF THE pn JUNCTION

Figure 7.1a schematically shows the pn junction. It is important to realize that the entire semiconductor is a single-crystal material in which one region is doped with acceptor impurity atoms to form the p region and the adjacent region is doped with donor atoms to form the n region. The interface separating the n and p regions is referred to as the *metallurgical junction.*

The impurity doping concentrations in the p and n regions are shown in Figure 7.1b. For simplicity, we will consider a *step junction* in which the doping concentration is uniform in each region and there is an abrupt change in doping at the junction. Initially, at the metallurgical junction, there is a very large density gradient in both electron and hole concentrations. Majority carrier electrons in the n region will begin diffusing into the p region, and majority carrier holes in the p region will begin diffusing into the n region. If we assume there are no external connections to the semiconductor, then this diffusion process cannot continue indefinitely. As electrons diffuse from the n region, positively charged donor atoms are left behind. Similarly, as holes diffuse from the p region, they uncover negatively charged acceptor atoms. The net positive and negative charges in the n and p regions induce an electric field in the region near the metallurgical junction, in the direction from the positive to the negative charge, or from the n to the p region.

The net positively and negatively charged regions are shown in Figure 7.2. These two regions are referred to as the *space charge region.* Essentially all electrons and holes are swept out of the space charge region by the electric field. Since the space charge region is depleted of any mobile charge, this region is also referred to as the

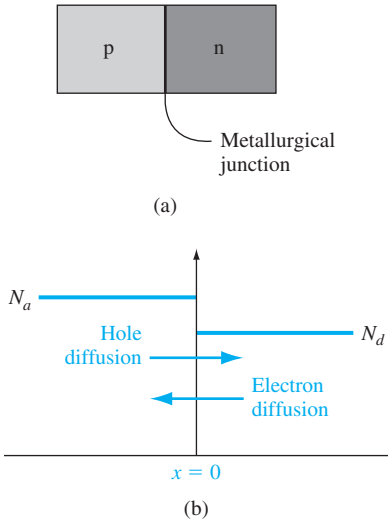


Figure 7.1 | (a) Simplified geometry of a pn junction; (b) doping profile of an ideal uniformly doped pn junction.

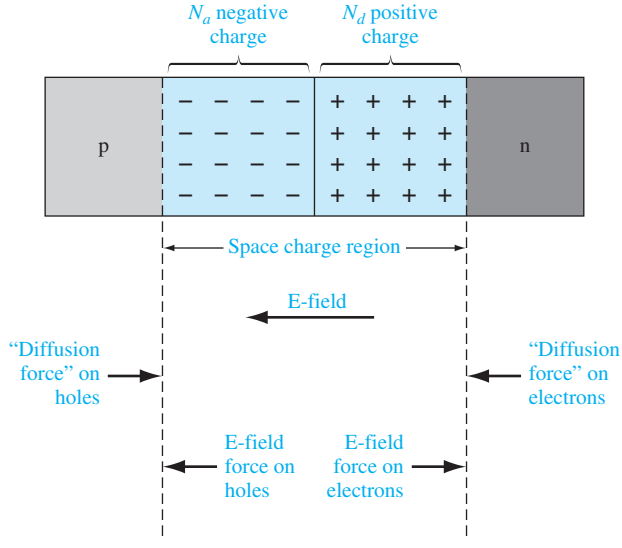


Figure 7.2 | The space charge region, the electric field, and the forces acting on the charged carriers.

depletion region; these two terms will be used interchangeably. Density gradients still exist in the majority carrier concentrations at each edge of the space charge region. We can think of a density gradient as producing a "diffusion force" that acts on the majority carriers. These diffusion forces, acting on the electrons and holes at the edges of the space charge region, are shown in the figure. The electric field in the space charge region produces another force on the electrons and holes, which is in the opposite direction to the diffusion force for each type of particle. In thermal equilibrium, the diffusion force and the E-field force exactly balance each other.

7.2 | ZERO APPLIED BIAS

We have considered the basic pn junction structure and discussed briefly how the space charge region is formed. In this section we will examine the properties of the step junction in thermal equilibrium, where no currents exist and no external excitation is applied. We will determine the space charge region width, electric field, and potential through the depletion region.

The analysis in this chapter is based on two assumptions that we have considered in previous chapters. The first assumption is that the Boltzmann approximation is valid, which means that each semiconductor region is nondegenerately doped. The second assumption is that complete ionization exists, which means that the temperature of the pn junction is not "too low."

7.2.1 Built-in Potential Barrier

If we assume that no voltage is applied across the pn junction, then the junction is in thermal equilibrium—the Fermi energy level is constant throughout the entire system. Figure 7.3 shows the energy-band diagram for the pn junction in thermal equilibrium. The conduction and valance band energies must bend as we go through the space charge region, since the relative position of the conduction and valence bands with respect to the Fermi energy changes between p and n regions.

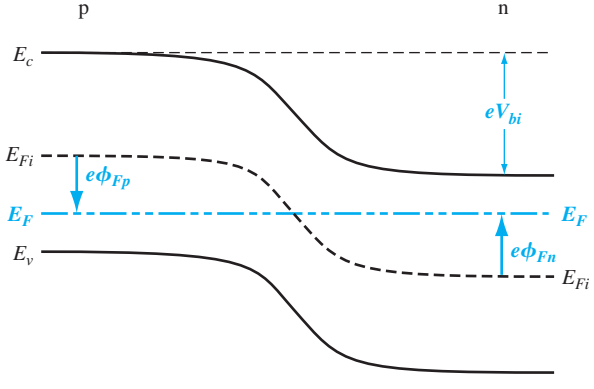


Figure 7.3 | Energy-band diagram of a pn junction in thermal equilibrium.

Electrons in the conduction band of the n region see a potential barrier in trying to move into the conduction band of the p region. This potential barrier is referred to as the *built-in potential barrier* and is denoted by V_{bi}. The built-in potential barrier maintains equilibrium between majority carrier electrons in the n region and minority carrier electrons in the p region, and also between majority carrier holes in the p region and minority carrier holes in the n region. This potential difference across the junction cannot be measured with a voltmeter because new potential barriers will be formed between the probes and the semiconductor that will cancel V_{bi}. The potential V_{bi} maintains equilibrium, so no current is produced by this voltage.

The intrinsic Fermi level is equidistant from the conduction band edge through the junction; thus, the built-in potential barrier can be determined as the difference between the intrinsic Fermi levels in the p and n regions. We can define the potentials ϕ_{Fn} and ϕ_{Fp} as shown in Figure 7.3, so we have

$$V_{bi} = |\phi_{Fn}| + |\phi_{Fp}| \tag{7.1}$$

In the n region, the electron concentration in the conduction band is given by

$$n_0 = N_c \exp\left[\frac{-(E_c - E_F)}{kT}\right] \tag{7.2}$$

which can also be written in the form

$$n_0 = n_i \exp\left[\frac{E_F - E_{Fi}}{kT}\right] \tag{7.3}$$

where n_i and E_{Fi} are the intrinsic carrier concentration and the intrinsic Fermi energy, respectively. We may define the potential ϕ_{Fn} in the n region as

$$e\phi_{Fn} = E_{Fi} - E_F \tag{7.4}$$

Equation (7.3) may then be written as

$$n_0 = n_i \exp\left[\frac{-(e\phi_{Fn})}{kT}\right] \tag{7.5}$$

Taking the natural log of both sides of Equation (7.5), setting $n_0 = N_d$, and solving for the potential, we obtain

$$\phi_{Fn} = \frac{-kT}{e} \ln\left(\frac{N_d}{n_i}\right) \tag{7.6}$$

Similarly, in the p region, the hole concentration is given by

$$p_0 = N_a = n_i \exp\left[\frac{E_{Fi} - E_F}{kT}\right] \tag{7.7}$$

where N_a is the acceptor concentration. We can define the potential ϕ_{Fp} in the p region as

$$e\phi_{Fp} = E_{Fi} - E_F \tag{7.8}$$

Combining Equations (7.7) and (7.8), we find that

$$\phi_{Fp} = +\frac{kT}{e}\ln\left(\frac{N_a}{n_i}\right) \tag{7.9}$$

Finally, the built-in potential barrier for the step junction is found by substituting Equations (7.6) and (7.9) into Equation (7.1), which yields

$$V_{bi} = \frac{kT}{e}\ln\left(\frac{N_a N_d}{n_i^2}\right) = V_t \ln\left(\frac{N_a N_d}{n_i^2}\right) \tag{7.10}$$

where $V_t = kT/e$ and is defined as the thermal voltage.

At this time, we should note a subtle but important point concerning notation. Previously in the discussion of a semiconductor material, N_d and N_a denoted donor and acceptor impurity concentrations in the same region, thereby forming a compensated semiconductor. From this point on in the text, N_d and N_a will denote the net donor and acceptor concentrations in the individual n and p regions, respectively. If the p region, for example, is a compensated material, then N_a will represent the difference between the actual acceptor and donor impurity concentrations. The parameter N_d is defined in a similar manner for the n region.

EXAMPLE 7.1

Objective: Calculate the built-in potential barrier in a pn junction.

Consider a silicon pn junction at $T = 300$ K with doping concentrations of $N_a = 2 \times 10^{17}$ cm^{-3} and $N_d = 10^{15}$ cm^{-3}.

■ Solution

The built-in potential barrier is determined from Equation (7.10) as

$$V_{bi} = V_t \ln\left(\frac{N_a N_d}{n_i^2}\right) = (0.0259)\ln\left[\frac{(2\times 10^{17})(10^{15})}{(1.5\times 10^{10})^2}\right] = 0.713 \text{ V}$$

If we change the doping concentration in the p region of the pn junction such that the doping concentrations become $N_a = 10^{16}$ cm^{-3} and $N_d = 10^{15}$ cm^{-3}, then the built-in potential barrier becomes $V_{bi} = 0.635$ V.

■ Comment

The built-in potential barrier changes only slightly as the doping concentrations change by orders of magnitude because of the logarithmic dependence.

■ EXERCISE PROBLEM

Ex 7.1 (*a*) Calculate the built-in potential barrier in a silicon pn junction at $T = 300$ K for (*i*) $N_a = 5 \times 10^{15}$ cm^{-3}, $N_d = 10^{17}$ cm^{-3} and (*ii*) $N_a = 2 \times 10^{16}$ cm^{-3}, $N_d = 2 \times 10^{15}$ cm^{-3}. (*b*) Repeat part (*a*) for a GaAs pn junction.

[Ans. (*a*) (*i*) 0.736 V, (*ii*) 0.671 V; (*b*) (*i*) 1.20 V, (*ii*) 1.14 V]

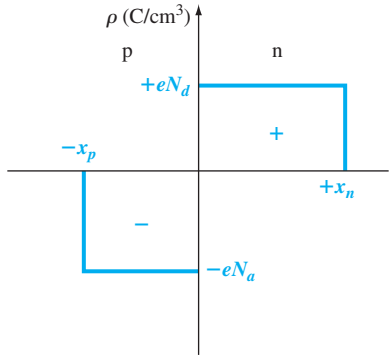


Figure 7.4 | The space charge density in a uniformly doped pn junction assuming the abrupt junction approximation.

7.2.2 Electric Field

An electric field is created in the depletion region by the separation of positive and negative space charge densities. Figure 7.4 shows the volume charge density distribution in the pn junction assuming uniform doping and assuming an abrupt junction approximation. We will assume that the space charge region abruptly ends in the n region at $x = +x_n$ and abruptly ends in the p region at $x = -x_p$ (x_p is a positive quantity).

The electric field is determined from Poisson's equation, which, for a one-dimensional analysis, is

$$\frac{d^2\phi(x)}{dx^2} = \frac{-\rho(x)}{\epsilon_s} = -\frac{d\mathrm{E}(x)}{dx} \tag{7.11}$$

where $\phi(x)$ is the electric potential, $\mathrm{E}(x)$ is the electric field, $\rho(x)$ is the volume charge density, and ϵ_s is the permittivity of the semiconductor. From Figure 7.4, the charge densities are

$$\rho(x) = -eN_a \qquad -x_p < x < 0 \tag{7.12a}$$

and

$$\rho(x) = eN_d \qquad 0 < x < x_n \tag{7.12b}$$

The electric field in the p region is found by integrating Equation (7.11). We have

$$\mathrm{E} = \int \frac{\rho(x)}{\epsilon_s}\,dx = -\int \frac{eN_a}{\epsilon_s}\,dx = \frac{-eN_a}{\epsilon_s}\,x + C_1 \tag{7.13}$$

where C_1 is a constant of integration. The electric field is assumed to be zero in the neutral p region for $x < -x_p$ since the currents are zero in thermal equilibrium. Since there are no surface charge densities within the pn junction structure, the electric

field is a continuous function. The constant of integration is determined by setting E = 0 at $x = -x_p$. The electric field in the p region is then given by

$$\mathrm{E} = \frac{-eN_a}{\epsilon_s}(x + x_p) \qquad -x_p \leq x \leq 0 \tag{7.14}$$

In the n region, the electric field is determined from

$$\mathrm{E} = \int \frac{(eN_d)}{\epsilon_s}\,dx = \frac{eN_d}{\epsilon_s}x + C_2 \tag{7.15}$$

where C_2 is again a constant of integration and is determined by setting E = 0 at $x = x_n$, since the E-field is assumed to be zero in the n region and is a continuous function. Then

$$\mathrm{E} = \frac{-eN_d}{\epsilon_s}(x_n - x) \qquad 0 \leq x \leq x_n \tag{7.16}$$

The electric field is also continuous at the metallurgical junction, or at $x = 0$. Setting Equations (7.14) and (7.16) equal to each other at $x = 0$ gives

$$N_a x_p = N_d x_n \tag{7.17}$$

Equation (7.17) states that the number of negative charges per unit area in the p region is equal to the number of positive charges per unit area in the n region.

Figure 7.5 is a plot of the electric field in the depletion region. The electric field direction is from the n to the p region, or in the negative x direction for this geometry. For the uniformly doped pn junction, the E-field is a linear function of distance through the junction, and the maximum (magnitude) electric field occurs at the metallurgical junction. An electric field exists in the depletion region even when no voltage is applied between the p and n regions.

The potential in the junction is found by integrating the electric field. In the p region then, we have

$$\phi(x) = -\int \mathrm{E}(x)dx = \int \frac{eN_a}{\epsilon_s}(x + x_p)dx \tag{7.18}$$

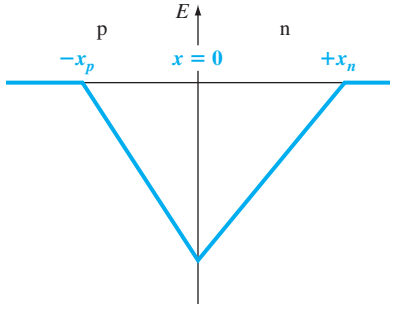


Figure 7.5 | Electric field in the space charge region of a uniformly doped pn junction.

or

$$\phi(x) = \frac{eN_a}{\epsilon_s}\left(\frac{x^2}{2} + x_p \cdot x\right) + C_1' \tag{7.19}$$

where C_1' is again a constant of integration. The potential difference through the pn junction is the important parameter, rather than the absolute potential, so we may arbitrarily set the potential equal to zero at $x = -x_p$. The constant of integration is then found as

$$C_1' = \frac{eN_a}{2\epsilon_s} x_p^2 \tag{7.20}$$

so that the potential in the p region can now be written as

$$\phi(x) = \frac{eN_a}{2\epsilon_s}(x + x_p)^2 \qquad (-x_p \le x \le 0) \tag{7.21}$$

The potential in the n region is determined by integrating the electric field in the n region, or

$$\phi(x) = \int \frac{eN_d}{\epsilon_s}(x_n - x)dx \tag{7.22}$$

Then

$$\phi(x) = \frac{eN_d}{\epsilon_s}\left(x_n \cdot x - \frac{x^2}{2}\right) + C_2' \tag{7.23}$$

where C_2' is another constant of integration. The potential is a continuous function, so setting Equation (7.21) equal to Equation (7.23) at the metallurgical junction, or at $x = 0$, gives

$$C_2' = \frac{eN_a}{2\epsilon_s} x_p^2 \tag{7.24}$$

The potential in the n region can thus be written as

$$\phi(x) = \frac{eN_d}{\epsilon_s}\left(x_n \cdot x - \frac{x^2}{2}\right) + \frac{eN_a}{2\epsilon_s} x_p^2 \qquad (0 \le x \le x_n) \tag{7.25}$$

Figure 7.6 is a plot of the potential through the junction and shows the quadratic dependence on distance. The magnitude of the potential at $x = x_n$ is equal to the built-in potential barrier. Then from Equation (7.25), we have

$$V_{bi} = |\phi(x = x_n)| = \frac{e}{2\epsilon_s}\left(N_d x_n^2 + N_a x_p^2\right) \tag{7.26}$$

The potential energy of an electron is given by $E = -e\phi$, which means that the electron potential energy also varies as a quadratic function of distance through the space charge region. The quadratic dependence on distance was shown in the energy-band diagram of Figure 7.3, although we did not explicitly know the shape of the curve at that time.

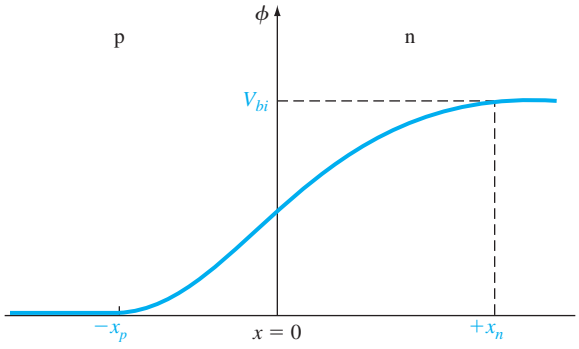


Figure 7.6 | Electric potential through the space charge region of a uniformly doped pn junction.

7.2.3 Space Charge Width

We can determine the distance that the space charge region extends into the p and n regions from the metallurgical junction. This distance is known as the space charge width. From Equation (7.17), we may write, for example,

$$x_p = \frac{N_d x_n}{N_a} \tag{7.27}$$

Then, substituting Equation (7.27) into Equation (7.26) and solving for x_n, we obtain

$$\boxed{x_n = \left\{\frac{2\epsilon_s V_{bi}}{e}\left[\frac{N_a}{N_d}\right]\left[\frac{1}{N_a + N_d}\right]\right\}^{1/2}} \tag{7.28}$$

Equation (7.28) gives the space charge width, or the width of the depletion region, x_n extending into the n-type region for the case of zero applied voltage.

Similarly, if we solve for x_n from Equation (7.17) and substitute into Equation (7.26), we find

$$\boxed{x_p = \left\{\frac{2\epsilon_s V_{bi}}{e}\left[\frac{N_d}{N_a}\right]\left[\frac{1}{N_a + N_d}\right]\right\}^{1/2}} \tag{7.29}$$

where x_p is the width of the depletion region extending into the p region for the case of zero applied voltage.

The total depletion or space charge width W is the sum of the two components, or

$$W = x_n + x_p \tag{7.30}$$

Using Equations (7.28) and (7.29), we obtain

$$W = \left\{\frac{2\epsilon_s V_{bi}}{e}\left[\frac{N_a + N_d}{N_a N_d}\right]\right\}^{1/2} \tag{7.31}$$

The built-in potential barrier can be determined from Equation (7.10), and then the total space charge region width is obtained using Equation (7.31).

EXAMPLE 7.2

Objective: Calculate the space charge width and electric field in a pn junction for zero bias.

Consider a silicon pn junction at $T = 300$ K with doping concentrations of $N_a = 10^{16}$ cm^{-3} and $N_d = 10^{15}$ cm^{-3}.

Solution

In Example 7.1, we determined the built-in potential barrier as $V_{bi} = 0.635$ V. From Equation (7.31), the space charge width is

$$W = \left\{\frac{2\epsilon_s V_{bi}}{e}\left[\frac{N_a + N_d}{N_a N_d}\right]\right\}^{1/2}$$

$$= \left\{\frac{2(11.7)(8.85 \times 10^{-14})(0.635)}{1.6 \times 10^{-19}}\left[\frac{10^{16} + 10^{15}}{(10^{16})(10^{15})}\right]\right\}^{1/2}$$

$$= 0.951 \times 10^{-4} \text{ cm} = 0.951\ \mu\text{m}$$

Using Equations (7.28) and (7.29), we can find $x_n = 0.8644\ \mu$m, and $x_p = 0.0864\ \mu$m.

The peak electric field at the metallurgical junction, using Equation (7.16) for example, is

$$\mathrm{E}_{\max} = -\frac{eN_d x_n}{\epsilon_s} = -\frac{(1.6 \times 10^{-19})(10^{15})(0.8644 \times 10^{-4})}{(11.7)(8.85 \times 10^{-14})} = -1.34 \times 10^4 \text{ V/cm}$$

Comment

The peak electric field in the space charge region of a pn junction is quite large. We must keep in mind, however, that there is no mobile charge in this region; hence there will be no drift current. We may also note, from this example, that the width of each space charge region is a reciprocal function of the doping concentration: The depletion region will extend further into the lower-doped region.

EXERCISE PROBLEM

Ex 7.2 A silicon pn junction at $T = 300$ K with zero applied bias has doping concentrations of $N_d = 5 \times 10^{16}$ cm^{-3} and $N_a = 5 \times 10^{15}$ cm^{-3}. Determine x_n, x_p, W, and $|\mathrm{E}_{\max}|$.

(Ans. $x_n = 4.11 \times 10^{-6}$ cm, $x_p = 4.11 \times 10^{-5}$ cm, $W = 4.52 \times 10^{-5}$ cm, $|\mathrm{E}_{\max}| = 3.18 \times 10^4$ V/cm)

TEST YOUR UNDERSTANDING

TYU 7.1 Calculate V_{bi}, x_n, x_p, W, and $|\mathrm{E}_{\max}|$ for a silicon pn junction at zero bias and $T = 300$ K for doping concentrations of (*a*) $N_a = 2 \times 10^{17}$ cm^{-3}, $N_d = 10^{16}$ cm^{-3} and (*b*) $N_a = 4 \times 10^{15}$ cm^{-3}, $N_d = 3 \times 10^{16}$ cm^{-3}.

[Ans. (*a*) $V_{bi} = 0.772$ V, $x_n = 0.3085\ \mu$m, $x_p = 0.0154\ \mu$m, $W = 0.3240\ \mu$m, $|\mathrm{E}_{\max}| = 4.77 \times 10^4$ V/cm; (*b*) $V_{bi} = 0.699$ V, $x_n = 0.0596\ \mu$m, $x_p = 0.4469\ \mu$m, $W = 0.5064\ \mu$m, $|\mathrm{E}_{\max}| = 2.76 \times 10^4$ V/cm]

TYU 7.2 Repeat Exercise Problem Ex 7.2 for a GaAs pn junction. $(|\mathrm{E}_{\max}| = 3.86 \times 10^4$ V/cm, $W = 0.6149\ \mu$m, $x_p = 0.5590\ \mu$m, $x_n = 0.05590\ \mu$m, $V_{bi} = 1.186$ V. Ans.)

7.3 | REVERSE APPLIED BIAS

If we apply a potential between the p and n regions, we will no longer be in an equilibrium condition—the Fermi energy level will no longer be constant through the system. Figure 7.7 shows the energy-band diagram of the pn junction for the case when a positive voltage is applied to the n region with respect to the p region. As the positive potential is downward, the Fermi level on the n side is below the Fermi level on the p side. The difference between the two is equal to the applied voltage in units of energy.

The total potential barrier, indicated by V_{total}, has increased. The applied potential is the reverse-biased condition. The total potential barrier is now given by

$$V_{\text{total}} = |\phi_{Fn}| + |\phi_{Fp}| + V_R \tag{7.32}$$

where V_R is the magnitude of the applied reverse-biased voltage. Equation (7.32) can be rewritten as

$$V_{\text{total}} = V_{bi} + V_R \tag{7.33}$$

where V_{bi} is the same built-in potential barrier we had defined in thermal equilibrium.

7.3.1 Space Charge Width and Electric Field

Figure 7.8 shows a pn junction with an applied reverse-biased voltage V_R. Also indicated in the figure are the electric field in the space charge region and the electric field E_{app}, induced by the applied voltage. The electric fields in the neutral p and n regions are essentially zero, or at least very small, which means that the magnitude of the electric field in the space charge region must increase above the thermal-equilibrium value due to the applied voltage. The electric field originates on positive charge and terminates on negative charge; this means that the number of positive and negative

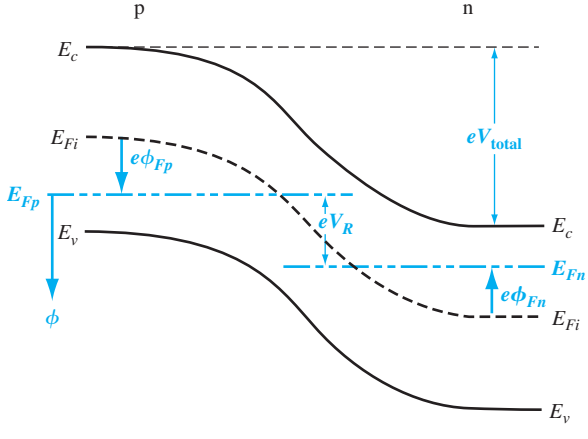


Figure 7.7 | Energy-band diagram of a pn junction under reverse bias.

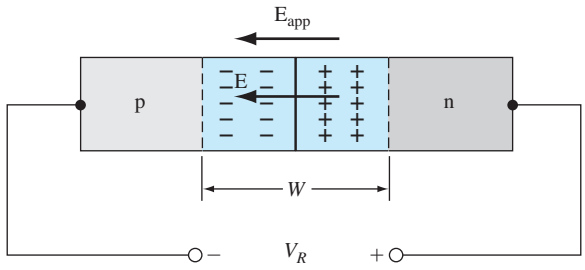


Figure 7.8 | A pn junction, with an applied reverse-biased voltage, showing the directions of the electric field induced by V_R and the space charge electric field.

charges must increase if the electric field increases. For given impurity doping concentrations, the number of positive and negative charges in the depletion region can be increased only if the space charge width W increases. The space charge width W increases, therefore, with an increasing reverse-biased voltage V_R. We are assuming that the electric field in the bulk n and p regions is zero. This assumption will become clearer in the next chapter when we discuss the current–voltage characteristics.

In all of the previous equations, the built-in potential barrier can be replaced by the total potential barrier. The total space charge width can be written from Equation (7.31) as

$$W = \left\{\frac{2\epsilon_s(V_{bi} + V_R)}{e}\left[\frac{N_a + N_d}{N_a N_d}\right]\right\}^{1/2} \tag{7.34}$$

showing that the total space charge width increases as we apply a reverse-biased voltage. By substituting the total potential barrier V_{total} into Equations (7.28) and (7.29), the space charge widths in the n and p regions, respectively, can be found as a function of applied reverse-biased voltage.

EXAMPLE 7.3

Objective: Calculate the width of the space charge region in a pn junction when a reverse-biased voltage is applied.

Again consider a silicon pn junction at $T = 300$ K with doping concentrations of $N_a = 10^{16}$ cm^{-3} and $N_d = 10^{15}$ cm^{-3}. Assume that $n_i = 1.5 \times 10^{10}$ cm^{-3} and $V_R = 5$ V.

■ Solution

The built-in potential barrier was calculated in Example 7.1 for this case and is $V_{bi} = 0.635$ V. The space charge width is determined from Equation (7.34). We have

$$W = \left\{\frac{2(11.7)(8.85 \times 10^{-14})(0.635 + 5)}{1.6 \times 10^{-19}}\left[\frac{10^{16} + 10^{15}}{(10^{16})(10^{15})}\right]\right\}^{1/2}$$

so that

$$W = 2.83 \times 10^{-4} \text{ cm} = 2.83\ \mu\text{m}$$

■ Comment

The space charge width has increased from 0.951 μm at zero bias to 2.83 μm at a reverse bias of 5 V.

■ EXERCISE PROBLEM

Ex 7.3 (*a*) A silicon pn junction at $T = 300$ K has doping concentrations of $N_a = 5 \times 10^{15}$ cm^{-3} and $N_d = 5 \times 10^{16}$ cm^{-3}. A reverse-biased voltage of $V_R = 4$ V is applied. Determine V_{bi}, x_n, x_p, and W. (*b*) Repeat part (*a*) for $V_R = 8$ V.

[Ans. (*a*) $V_{bi} = 0.718$ V, $x_n = 0.1054$ μm, $x_p = 1.054$ μm, $W = 1.159$ μm; (*b*) $V_{bi} = 0.718$, $x_n = 0.1432$ μm, $x_p = 1.432$ μm, $W = 1.576$ μm]

The magnitude of the electric field in the depletion region increases with an applied reverse-biased voltage. The electric field is still given by Equations (7.14) and (7.16) and is still a linear function of distance through the space charge region. Since x_n and x_p increase with reverse-biased voltage, the magnitude of the electric field also increases. The maximum electric field still occurs at the metallurgical junction.

The maximum electric field at the metallurgical junction, from Equations (7.14) and (7.16), is

$$\mathrm{E}_{\max} = \frac{-eN_d x_n}{\epsilon_s} = \frac{-eN_a x_p}{\epsilon_s} \tag{7.35}$$

If we use either Equation (7.28) or (7.29) in conjunction with the total potential barrier, $V_{bi} + V_R$, then

$$\boxed{\mathrm{E}_{\max} = -\left\{\frac{2e(V_{bi} + V_R)}{\epsilon_s}\left(\frac{N_a N_d}{N_a + N_d}\right)\right\}^{1/2}} \tag{7.36}$$

We can show that the maximum electric field in the pn junction can also be written as

$$\boxed{\mathrm{E}_{\max} = \frac{-2(V_{bi} + V_R)}{W}} \tag{7.37}$$

where W is the total space charge width.

DESIGN EXAMPLE 7.4

Objective: Design a pn junction to meet maximum electric field and voltage specifications.

Consider a silicon pn junction at $T = 300$ K with a p-type doping concentration of $N_a = 2 \times 10^{17}$ cm^{-3}. Determine the n-type doping concentration such that the maximum electric field is $|\mathrm{E}_{\max}| = 2.5 \times 10^5$ V/cm at a reverse-biased voltage of $V_R = 25$ V.

■ Solution

The maximum electric field is given by Equation (7.36). Neglecting V_{bi} compared to V_R, we can write

$$|\mathrm{E}_{\max}| \cong \left\{\frac{2eV_R}{\epsilon_s}\left(\frac{N_a N_d}{N_a + N_d}\right)\right\}^{1/2}$$

or

$$2.5 \times 10^5 = \left\{\frac{2(1.6 \times 10^{-19})(25)}{(11.7)(8.85 \times 10^{-14})}\left[\frac{(2 \times 10^{17})N_d}{2 \times 10^{17} + N_d}\right]\right\}^{1/2}$$

which yields

$$N_d = 8.43 \times 10^{15}\ \text{cm}^{-3}$$

Conclusion

A smaller value of N_d results in a smaller value of $|E_{max}|$ for a given reverse-biased voltage. The value of N_d determined in this example, then, is the maximum value that will meet the specifications.

EXERCISE PROBLEM

Ex 7.4 The maximum electric field in a reverse-biased GaAs pn junction at $T = 300$ K is to be limited to $|E_{max}| = 7.2 \times 10^4$ V/cm. The doping concentrations are $N_d = 5 \times 10^{15}$ cm^{-3} and $N_a = 3 \times 10^{16}$ cm^{-3}. Determine the maximum reverse-biased voltage that can be applied. (Ans. $V_R = 12.3$ V)

7.3.2 Junction Capacitance

Since we have a separation of positive and negative charges in the depletion region, a capacitance is associated with the pn junction. Figure 7.9 shows the charge densities in the depletion region for applied reverse-biased voltages of V_R and $V_R + dV_R$. An increase

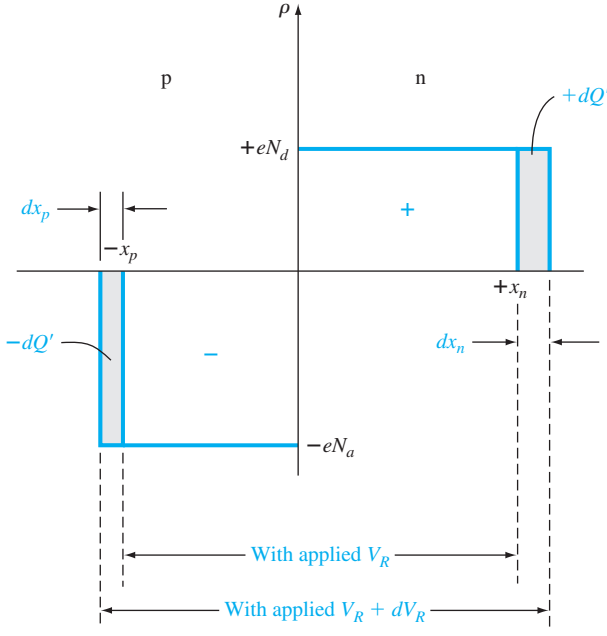


Figure 7.9 | Differential change in the space charge width with a differential change in reverse-biased voltage for a uniformly doped pn junction.

in the reverse-biased voltage dV_R will uncover additional positive charges in the n region and additional negative charges in the p region. The junction capacitance is defined as

$$C' = \frac{dQ'}{dV_R} \tag{7.38}$$

where

$$dQ' = eN_d\,dx_n = eN_a\,dx_p \tag{7.39}$$

The differential charge dQ' is in units of C/cm^2 so that the capacitance C' is in units of farads per square centimeter F/cm^2), or capacitance per unit area.

For the total potential barrier, Equation (7.28) may be written as

$$x_n = \left\{\frac{2\epsilon_s(V_{bi} + V_R)}{e}\left[\frac{N_a}{N_d}\right]\left[\frac{1}{N_a + N_d}\right]\right\}^{1/2} \tag{7.40}$$

The junction capacitance can be written as

$$C' = \frac{dQ'}{dV_R} = eN_d\frac{dx_n}{dV_R} \tag{7.41}$$

so that

$$\boxed{C' = \left\{\frac{e\epsilon_s N_a N_d}{2(V_{bi} + V_R)(N_a + N_d)}\right\}^{1/2}} \tag{7.42}$$

Exactly the same capacitance expression is obtained by considering the space charge region extending into the p region x_p. The junction capacitance is also referred to as the *depletion layer capacitance.*

EXAMPLE 7.5

Objective: Calculate the junction capacitance of a pn junction.

Consider the same pn junction as that in Example 7.3. Again assume that $V_R = 5$ V.

■ Solution

The junction capacitance is found from Equation (7.42) as

$$C' = \left\{\frac{(1.6 \times 10^{-19})(11.7)(8.85 \times 10^{-14})(10^{16})(10^{15})}{2(0.635 + 5)(10^{16} + 10^{15})}\right\}^{1/2}$$

or

$$C' = 3.66 \times 10^{-9}\ \text{F/cm}^2$$

If the cross-sectional area of the pn junction is, for example, $A = 10^{-4}$ cm^2, then the total junction capacitance is

$$C = C' \cdot A = 0.366 \times 10^{-12}\ \text{F} = 0.366\ \text{pF}$$

■ Comment

The value of junction capacitance is usually in the pF, or smaller, range.

EXERCISE PROBLEM

Ex 7.5 Consider a GaAs pn junction at $T = 300$ K doped to $N_a = 5 \times 10^{15}$ cm^{-3} and $N_d = 2 \times 10^{16}$ cm^{-3}. (*a*) Calculate V_{bi}. (*b*) Determine the junction capacitance C' for $V_R = 4$ V. (*c*) Repeat part (*b*) for $V_R = 8$ V.

[Ans. (*a*) $V_{bi} = 1.16$ V; (*b*) $C' = 8.48 \times 10^{-9}$ F/cm^2; (*c*) $C' = 6.36 \times 10^{-9}$ F/cm^2]

If we compare Equation (7.34) for the total depletion width W of the space charge region under reverse bias and Equation (7.42) for the junction capacitance C', we find that we can write

$$C' = \frac{\epsilon_s}{W} \tag{7.43}$$

Equation (7.43) is the same as the capacitance per unit area of a parallel plate capacitor. Considering Figure 7.9, we may have come to this same conclusion earlier. Keep in mind that the space charge width is a function of the reverse-biased voltage so that the junction capacitance is also a function of the reverse-biased voltage applied to the pn junction.

7.3.3 One-Sided Junctions

Consider a special pn junction called the one-sided junction. If, for example, $N_a \gg N_d$, this junction is referred to as a p^+n junction. The total space charge width, from Equation (7.34), reduces to

$$W \approx \left\{\frac{2\epsilon_s(V_{bi} + V_R)}{eN_d}\right\}^{1/2} \tag{7.44}$$

Considering the expressions for x_n and x_p, we have for the p^+n junction

$$x_p \ll x_n \tag{7.45}$$

and

$$W \approx x_n \tag{7.46}$$

Almost the entire space charge layer extends into the low-doped region of the junction. This effect can be seen in Figure 7.10.

The junction capacitance of the p^+n junction reduces to

$$C' \approx \left\{\frac{e\epsilon_s N_d}{2(V_{bi} + V_R)}\right\}^{1/2} \tag{7.47}$$

The depletion layer capacitance of a one-sided junction is a function of the doping concentration in the low-doped region. Equation (7.47) may be manipulated to give

$$\left(\frac{1}{C'}\right)^2 = \frac{2(V_{bi} + V_R)}{e\epsilon_s N_d} \tag{7.48}$$

which shows that the inverse capacitance squared is a linear function of applied reverse-biased voltage.

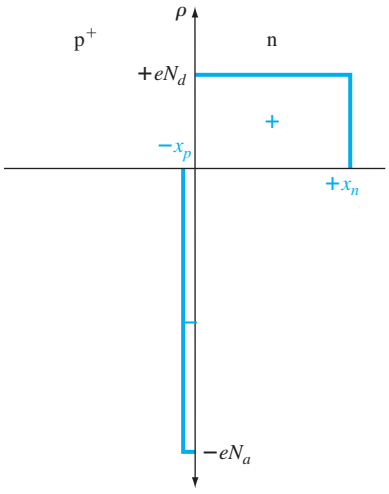


Figure 7.10 | Space charge density of a one-sided p⁻n junction.

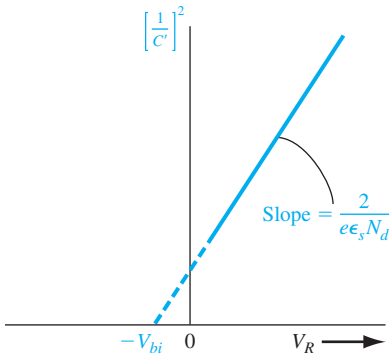


Figure 7.11 | $(1/C')^2$ versus V_R of a uniformly doped pn junction.

Figure 7.11 shows a plot of Equation (7.48). The built-in potential of the junction can be determined by extrapolating the curve to the point where $(1/C')^2 = 0$. The slope of the curve is inversely proportional to the doping concentration of the low-doped region in the junction; thus, this doping concentration can be experimentally determined. The assumptions used in the derivation of this capacitance include uniform doping in both semiconductor regions, the abrupt junction approximation, and a planar junction.

EXAMPLE 7.6

Objective: Determine the impurity doping concentrations in a p^+n junction given the parameters from Figure 7.11.

Assume that the intercept and the slope of the curve in Figure 7.11 are $V_{bi} = 0.725$ V and 6.15×10^{15} $(F/cm^2)^{-2}$ $(V)^{-1}$, respectively, for a silicon p^+n junction at $T = 300$ K.

■ Solution

The slope of the curve in Figure 7.11 is given by $2/e\,\epsilon_s\,N_d$, so we may write

$$N_d = \frac{2}{e\,\epsilon_s} \cdot \frac{1}{slope} = \frac{2}{(1.6 \times 10^{-19})(11.7)(8.85 \times 10^{-14})(6.15 \times 10^{15})}$$

or

$$N_d = 1.96 \times 10^{15}\ cm^{-3}$$

From the expression for V_{bi}, which is

$$V_{bi} = V_t \ln\left(\frac{N_a N_d}{n_i^2}\right)$$

we can solve for N_a as

$$N_a = \frac{n_i^2}{N_d} \exp\left(\frac{V_{bi}}{V_t}\right) = \frac{(1.5 \times 10^{10})^2}{1.963 \times 10^{15}} \exp\left(\frac{0.725}{0.0259}\right)$$

which yields

$$N_a = 1.64 \times 10^{17}\ \text{cm}^{-3}$$

Comment

The results of this example show that $N_a \gg N_d$; therefore the assumption of a one-sided junction was valid.

EXERCISE PROBLEM

Ex 7.6 The experimentally measured junction capacitance of a one-sided silicon n^+p junction biased at $V_R = 3$ V and at $T = 300$ K is $C = 0.105$ pF. The built-in potential barrier is found to be $V_{bi} = 0.765$ V. The cross-sectional area is $A = 10^{-5}$ cm^2. Find the doping concentrations. (Ans. $N_d = 5.01 \times 10^{15}$ cm^{-3}, $N_d = 3.02 \times 10^{17}$ cm^{-3})

A one-sided pn junction is useful for experimentally determining the doping concentrations and built-in potential.

TEST YOUR UNDERSTANDING

TYU 7.3 (*a*) A silicon pn junction at $T = 300$ K is reverse biased at $V_R = 8$ V. The doping concentrations are $N_a = 5 \times 10^{16}$ cm^{-3} and $N_d = 5 \times 10^{15}$ cm^{-3}. Determine x_n, x_p, W, and $|E_{max}|$. (*b*) Repeat part (*a*) for a reverse-biased voltage of $V_R = 12$ V.
[Ans. (*a*) $x_n = 1.43 \times 10^{-4}$ cm, $x_p = 1.43 \times 10^{-5}$ cm, $W = 1.57 \times 10^{-4}$ cm, $|E_{max}| = 1.11 \times 10^5$ V/cm; (*b*) $x_n = 1.73 \times 10^{-4}$ cm, $x_p = 1.73 \times 10^{-5}$ cm, $W = 1.90 \times 10^{-4}$ cm, $|E_{max}| = 1.34 \times 10^5$ V/cm]

TYU 7.4 A silicon pn junction at $T = 300$ K has doping concentrations of $N_d = 3 \times 10^{16}$ cm^{-3} and $N_a = 8 \times 10^{15}$ cm^{-3}, and has a cross-sectional area of $A = 5 \times 10^{-5}$ cm^2. Determine the junction capacitance at (*a*) $V_R = 2$ V and (*b*) $V_R = 5$ V.
[Ans. (*a*) 0.694 pF; (*b*) 0.478 pF]

7.4 | JUNCTION BREAKDOWN

In the last section, we determined the effects of applying a reverse-biased voltage across the pn junction. However, the reverse-biased voltage may not increase without limit; at some particular voltage, the reverse-biased current will increase rapidly. The applied voltage at this point is called the *breakdown voltage.*

Two physical mechanisms give rise to the reverse-biased breakdown in a pn junction: the *Zener effect* and the *avalanche effect.* Zener breakdown occurs in highly doped pn junctions through a tunneling mechanism. In a highly doped junction, the conduction and valence bands on opposite sides of the junction are sufficiently close during reverse bias that electrons may tunnel directly from the valence band on the p side into the conduction band on the n side. This tunneling process is schematically shown in Figure 7.12a.

The avalanche breakdown process occurs when electrons and/or holes, moving across the space charge region, acquire sufficient energy from the electric field to create electron–hole pairs by colliding with atomic electrons within the depletion region. The avalanche process is schematically shown in Figure 7.12b. The newly

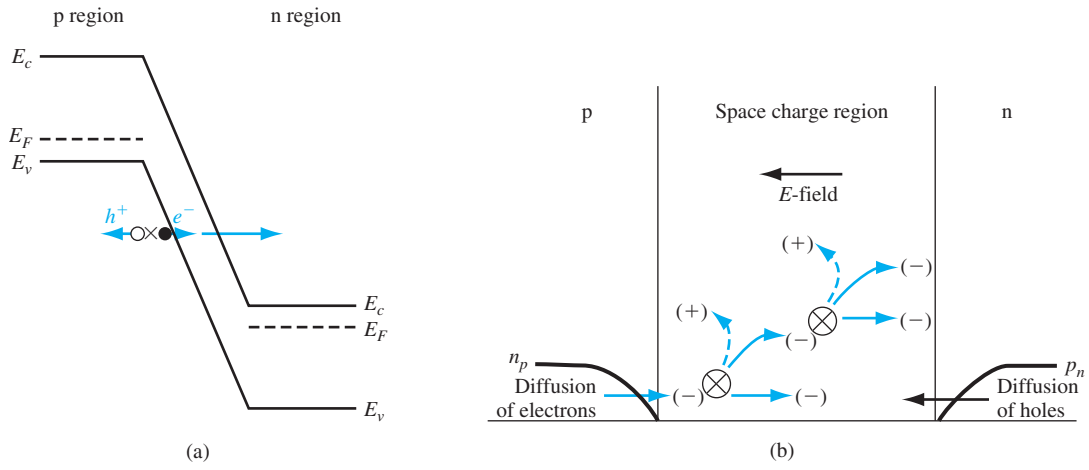


Figure 7.12 | (a) Zener breakdown mechanism in a reverse-biased pn junction; (b) avalanche breakdown process in a reverse-biased pn junction.

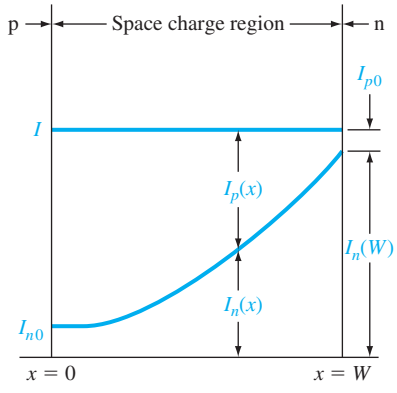


Figure 7.13 | Electron and hole current components through the space charge region during avalanche multiplication.

created electrons and holes move in opposite directions due to the electric field and thereby create a reverse-biased current. In addition, the newly generated electrons and/or holes may acquire sufficient energy to ionize other atoms, leading to the avalanche process. For most pn junctions, the predominant breakdown mechanism will be the avalanche effect.

If we assume that a reverse-biased electron current I_{n0} enters the depletion region at $x = 0$ as shown in Figure 7.13, the electron current I_n will increase with distance through the depletion region due to the avalanche process. At $x = W$, the electron current may be written as

$$I_n(W) = M_n I_{n0} \tag{7.49}$$

where M_n is a multiplication factor. The hole current is increasing through the depletion region from the n to p region and reaches a maximum value at $x = 0$. The total current is constant through the pn junction in steady state.

We can write an expression for the incremental electron current at some point x as

$$dI_n(x) = I_n(x)\alpha_n\, dx + I_p(x)\alpha_p\, dx \tag{7.50}$$

where α_n and α_p are the electron and hole ionization rates, respectively. The ionization rates are the number of electron–hole pairs generated per unit length by an electron (α_n) or by a hole (α_p). Equation (7.50) may be written as

$$\frac{dI_n(x)}{dx} = I_n(x)\alpha_n + I_p(x)\alpha_p \tag{7.51}$$

The total current I is given by

$$I = I_n(x) + I_p(x) \tag{7.52}$$

which is a constant. Solving for $I_p(x)$ from Equation (7.52) and substituting into Equation (7.51), we obtain

$$\frac{dI_n(x)}{dx} + (\alpha_p - \alpha_n)I_n(x) = \alpha_p I \tag{7.53}$$

If we make the assumption that the electron and hole ionization rates are equal so that

$$\alpha_n = \alpha_p \equiv \alpha \tag{7.54}$$

then Equation (7.53) may be simplified and integrated through the space charge region. We will obtain

$$I_n(W) - I_n(0) = I\int_0^W \alpha\, dx \tag{7.55}$$

Using Equation (7.49), Equation (7.55) may be written as

$$\frac{M_n I_{n0} - I_n(0)}{I} = \int_0^W \alpha\, dx \tag{7.56}$$

Since $M_n\, I_{n0} \approx I$ and since $I_n(0) = I_{n0}$, Equation (7.56) becomes

$$1 - \frac{1}{M_n} = \int_0^W \alpha\, dx \tag{7.57}$$

The avalanche breakdown voltage is defined to be the voltage at which M_n approaches infinity. The avalanche breakdown condition is then given by

$$\int_0^W \alpha dx = 1 \tag{7.58}$$

The ionization rates are strong functions of electric field and, since the electric field is not constant through the space charge region, Equation (7.58) is not easy to evaluate.

If we consider, for example, a one-sided p^+n junction, the maximum electric field is given by

$$\mathrm{E_{max}} = \frac{eN_d\, x_n}{\epsilon_s} \tag{7.59}$$

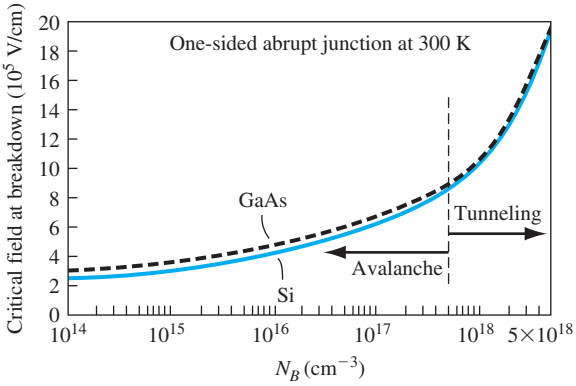


Figure 7.14 | Critical electric field at breakdown in a one-sided junction as a function of impurity doping concentrations. *(From Sze and N$_g$[14].)*

The depletion width x_n is given approximately as

$$x_n \approx \left\{\frac{2\epsilon_s V_R}{e} \cdot \frac{1}{N_d}\right\}^{1/2} \tag{7.60}$$

where V_R is the magnitude of the applied reverse-biased voltage. We have neglected the built-in potential V_{bi}.

If we now define V_R to be the breakdown voltage V_B, the maximum electric field, E_{max}, will be defined as a critical electric field, E_{crit}, at breakdown. Combining Equations (7.59) and (7.60), we may write

$$\boxed{V_B = \frac{\epsilon_s E_{crit}^2}{2eN_B}} \tag{7.61}$$

where N_B is the semiconductor doping in the low-doped region of the one-sided junction. The critical electric field, plotted in Figure 7.14, is a slight function of doping.

We have been considering a uniformly doped planar junction. The breakdown voltage will decrease for a linearly graded junction. (See Section 7.5.) Figure 7.15 shows a plot of the breakdown voltage for a one-sided abrupt junction and a linearly graded junction. If we take into account the curvature of a diffused junction as well, the breakdown voltage will be further degraded.

DESIGN EXAMPLE 7.7

Objective: Design an ideal one-sided n^+p junction diode to meet a breakdown voltage specification.

Consider a silicon pn junction diode at $T = 300$ K. Assume that $N_d = 3 \times 10^{18}$ cm^{-3}. Design the diode such that the breakdown voltage is $V_B = 100$ V.

■ Solution

From Figure 7.15, we find that the doping concentration in the low-doped side of a one-sided abrupt junction should be approximately 4×10^{15} cm^{-3} for a breakdown voltage of 100 V.

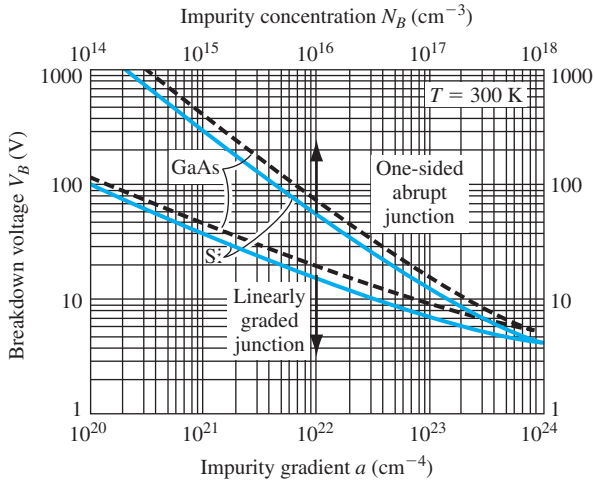


Figure 7.15 | Breakdown voltage versus impurity concentration in uniformly doped and linearly graded junctions. *(From Sze [14].)*

For a doping concentration of 4×10^{15} cm^{-3}, the critical electric field, from Figure 7.14, is approximately 3.7×10^5 V/cm. Then, using Equation (7.61), we find the breakdown voltage as

$$V_B = \frac{\epsilon_s E_{crit}^2}{2eN_B} = \frac{(11.7)(8.85 \times 10^{-14})(3.7 \times 10^5)^2}{2(1.6 \times 10^{-19})(4 \times 10^{15})} = 110 \text{ V}$$

which correlates very well with the results from Figure 7.15.

Conclusion

As Figure 7.15 shows, the breakdown voltage increases as the doping concentration decreases in the low-doped region.

EXERCISE PROBLEM

Ex 7.7 A one-sided, planar, uniformly doped silicon pn junction diode is required to have a reverse-biased breakdown voltage of $V_B = 60$ V. What is the maximum doping concentration in the low-doped region such that this specification is met?

(Ans. $N_B \approx 8 \times 10^{15}$ cm^{-3})

*7.5 | NONUNIFORMLY DOPED JUNCTIONS

In the pn junctions considered so far, we have assumed that each semiconductor region has been uniformly doped. In actual pn junction structures, this is rarely true. In some electronic applications, specific nonuniform doping profiles are used to obtain special pn junction capacitance characteristics.

7.5.1 Linearly Graded Junctions

If we start with a uniformly doped n-type semiconductor, for example, and diffuse acceptor atoms through the surface, the impurity concentrations will tend to be like those shown in Figure 7.16. The point $x = x'$ on the figure corresponds to the metallurgical junction. The depletion region extends into the p and n regions from the metallurgical junction as we have discussed previously. The net p-type doping concentration near the metallurgical junction may be approximated as a linear function of distance from the metallurgical junction. Likewise, as a first approximation, the net n-type doping concentration is also a linear function of distance extending into the n region from the metallurgical junction. This effective doping profile is referred to as a linearly graded junction.

Figure 7.17 shows the space charge density in the depletion region of the linearly graded junction. For convenience, the metallurgical junction is placed at $x = 0$. The space charge density can be written as

$$\rho(x) = eax \tag{7.62}$$

where a is the gradient of the net impurity concentration.

The electric field and potential in the space charge region can be determined from Poisson's equation. We can write

$$\frac{d\mathrm{E}}{dx} = \frac{\rho(x)}{\epsilon_s} = \frac{eax}{\epsilon_s} \tag{7.63}$$

so that the electric field can be found by integration as

$$\mathrm{E} = \int \frac{eax}{\epsilon_s}\,dx = \frac{ea}{2\epsilon_s}\left(x^2 - x_0^2\right) \tag{7.64}$$

The electric field in the linearly graded junction is a quadratic function of distance rather than the linear function found in the uniformly doped junction. The maximum electric field again occurs at the metallurgical junction. We may note that the electric field is zero at both $x = +x_0$ and at $x = -x_0$. The electric field in a nonuniformly

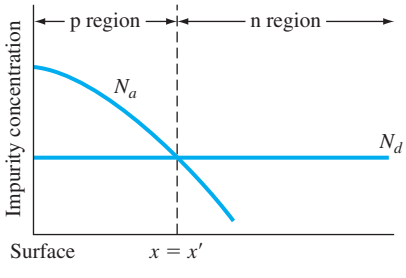


Figure 7.16 | Impurity concentrations of a pn junction with a nonuniformly doped p region.

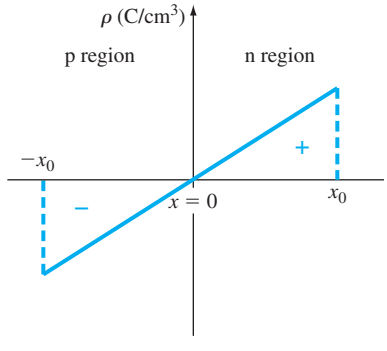


Figure 7.17 | Space charge density in a linearly graded pn junction.

doped semiconductor is not exactly zero, but the magnitude of this field is small, so setting E = 0 in the bulk regions is still a good approximation.

The potential is again found by integrating the electric field as

$$\phi(x) = -\int \mathrm{E}\, dx \tag{7.65}$$

If we arbitrarily set $\phi = 0$ at $x = -x_0$, then the potential through the junction is

$$\phi(x) = \frac{-ea}{2\epsilon_s}\left(\frac{x^3}{3} - x_0^2 x\right) + \frac{ea}{3\epsilon_s}x_0^3 \tag{7.66}$$

The magnitude of the potential at $x = +x_0$ will equal the built-in potential barrier for this function. We then have that

$$\phi(x_0) = \frac{2}{3}\cdot\frac{eax_0^3}{\epsilon_s} = V_{bi} \tag{7.67}$$

Another expression for the built-in potential barrier for a linearly graded junction can be approximated from the expression used for a uniformly doped junction. We can write

$$V_{bi} = V_t \ln\left[\frac{N_d(x_0)N_a(-x_0)}{n_i^2}\right] \tag{7.68}$$

where $N_d(x_0)$ and $N_a(-x_0)$ are the doping concentrations at the edges of the space charge region. We can relate these doping concentrations to the gradient, so that

$$N_d(x_0) = ax_0 \tag{7.69a}$$

and

$$N_a(-x_0) = ax_0 \tag{7.69b}$$

Then the built-in potential barrier for the linearly graded junction becomes

$$V_{bi} = V_t \ln\left(\frac{ax_0}{n_i}\right)^2 \tag{7.70}$$

There may be situations in which the doping gradient is not the same on either side of the junction, but we will not consider that condition here.

If a reverse-biased voltage is applied to the junction, the potential barrier increases. The built-in potential barrier V_{bi} in the above equations is then replaced by the total potential barrier $V_{bi} + V_R$. Solving for x_0 from Equation (7.67) and using the total potential barrier, we obtain

$$x_0 = \left\{\frac{3}{2}\cdot\frac{\epsilon_s}{ea}(V_{bi} + V_R)\right\}^{1/3} \tag{7.71}$$

The junction capacitance per unit area can be determined by the same method that we used for the uniformly doped junction. Figure 7.18 shows the differential charge

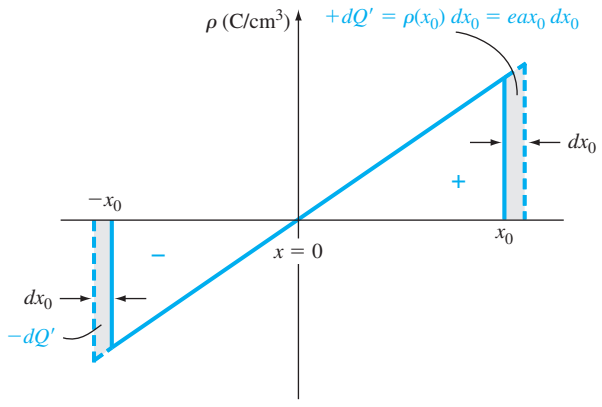


Figure 7.18 | Differential change in space charge width with a differential change in reverse-biased voltage for a linearly graded pn junction.

dQ', which is uncovered as a differential voltage dV_R is applied. The junction capacitance is then

$$C' = \frac{dQ'}{dV_R} = (eax_0)\frac{dx_0}{dV_R} \tag{7.72}$$

Using Equation (7.71), we obtain[1]

$$C' = \left\{\frac{ea\epsilon_s^2}{12(V_{bi} + V_R)}\right\}^{1/3} \tag{7.73}$$

We may note that C' is proportional to $(V_{bi} + V_R)^{-1/3}$ for the linearly graded junction as compared to $C'\alpha(V_{bi} + V_R)^{-1/2}$ for the uniformly doped junction. In the linearly graded junction, the capacitance is less dependent on reverse-biased voltage than in the uniformly doped junction.

7.5.2 Hyperabrupt Junctions

The uniformly doped junction and linearly graded junction are not the only possible doping profiles. Figure 7.19 shows a generalized one-sided p^+n junction where the generalized n-type doping concentration for $x > 0$ is given by

$$N = Bx^m \tag{7.74}$$

The case of $m = 0$ corresponds to the uniformly doped junction, and $m = +1$ corresponds to the linearly graded junction just discussed. The cases of $m = +2$ and $m = +3$ shown would approximate a fairly low-doped epitaxial n-type layer grown

[1]In a more exact analysis, V_{bi} in Equation (7.73) is replaced by a gradient voltage. However, this analysis is beyond the scope of this text.

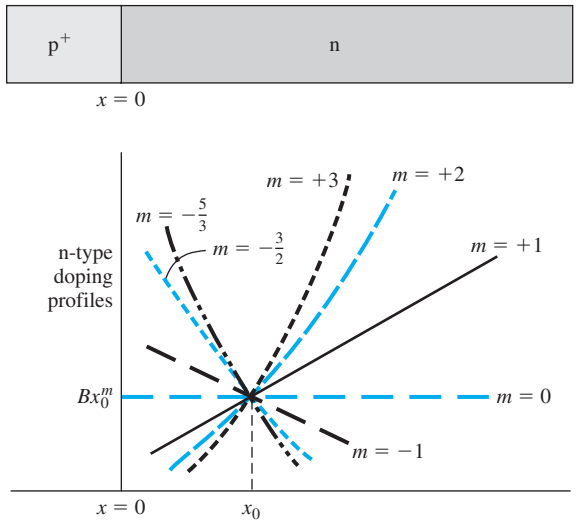


Figure 7.19 | Generalized doping profiles of a one-sided p^+n junction.
(From Sze [14].)

on a much more heavily doped n^+ substrate layer. When the value of m is negative, we have what is referred to as a *hyperabrupt junction.* In this case, the n-type doping is larger near the metallurgical junction than in the bulk semiconductor. Equation (7.74) is used to approximate the n-type doping over a small region near $x = x_0$ and does not hold at $x = 0$ when m is negative.

The junction capacitance can be derived using the same analysis method as before and is given by

$$C' = \left\{\frac{eB\epsilon_s^{(m+1)}}{(m+2)(V_{bi}+V_R)}\right\}^{1/(m+2)} \tag{7.75}$$

When m is negative, the capacitance becomes a very strong function of reverse-biased voltage, a desired characteristic in *varactor diodes.* The term *varactor* comes from the words *var*iable re*actor* and means a device whose reactance can be varied in a controlled manner with bias voltage.

If a varactor diode and an inductance are in parallel, the resonant frequency of the *LC* circuit is

$$f_r = \frac{1}{2\pi\sqrt{LC}} \tag{7.76}$$

The capacitance of the diode, from Equation (7.75), can be written in the form

$$C = C_0(V_{bi}+V_R)^{-1/(m+2)} \tag{7.77}$$

In a circuit application, we would, in general, like to have the resonant frequency be a linear function of reverse-biased voltage V_R, so we need

$$C \alpha V^{-2} \tag{7.78}$$

From Equation (7.77), the parameter m required is found from

$$\frac{1}{m+2} = 2 \tag{7.79a}$$

or

$$m = -\frac{3}{2} \tag{7.79b}$$

A specific doping profile will yield the desired capacitance characteristic.

7.6 | SUMMARY

- A uniformly doped pn junction is initially considered, in which one region of a semiconductor is uniformly doped with acceptor impurities and the adjacent region is uniformly doped with donor impurities.
- A space charge region, or depletion region, is formed on either side of the metallurgical junction separating the n and p regions. This region is essentially depleted of any mobile electrons or holes. A net positive charge density, due to the positively charged donor impurity ions, exists in the n region and a net negative charge density, due to the negatively charged acceptor impurity ions, exists in the p region.
- An electric field exists in the depletion region due to the net space charge density. The direction of the electric field is from the n region to the p region.
- A potential difference exists across the space charge region. Under zero applied bias, this potential difference, known as the built-in potential barrier, maintains thermal equilibrium and holds back majority carrier electrons in the n region and majority carrier holes in the p region.
- An applied reverse-biased voltage (n region positive with respect to the p region) increases the potential barrier, the space charge width, and the magnitude of the electric field.
- As the reverse-biased voltage changes, the amount of charge in the depletion region changes. This change in charge with voltage defines the junction capacitance.
- Avalanche breakdown occurs when a sufficiently large reverse-biased voltage is applied to the pn junction. A large reverse-biased current may then be induced in the pn junction. The breakdown voltage, as a function of the doping concentrations in the pn junction, is derived. In a one-sided pn junction, the breakdown voltage is a function of the doping concentration in the low-doped region.
- The linearly graded junction represents a nonuniformly doped pn junction. Expressions for the electric field, built-in potential barrier, and junction capacitance are derived. The functional relationships differ from those of the uniformly doped junction.
- Specific doping profiles can be used to obtain specific capacitance characteristics. A hyperabrupt junction is one in which the doping decreases away from the metallurgical junction. This type of junction is advantageous in varactor diodes that are used in resonant circuits.

GLOSSARY OF IMPORTANT TERMS

abrupt junction approximation The assumption that there is an abrupt discontinuity in space charge density between the space charge region and the neutral semiconductor region.

avalanche breakdown The process whereby a large reverse-biased pn junction current is created due to the generation of electron–hole pairs by the collision of electrons and/or holes with atomic electrons within the space charge region.

built-in potential barrier The electrostatic potential difference between the p and n regions of a pn junction in thermal equilibrium.

critical electric field The peak electric field in the space charge region at breakdown.

depletion layer capacitance Another term for junction capacitance.

depletion region Another term for space charge region.

hyperabrupt junction A pn junction in which the doping concentration on one side decreases away from the metallurgical junction to achieve a specific capacitance–voltage characteristic.

junction capacitance The capacitance of the pn junction under reverse bias.

linearly graded junction A pn junction in which the doping concentrations on either side of the metallurgical junction are approximated by a linear distribution.

metallurgical junction The interface between the p- and n-doped regions of a pn junction.

one-sided junction A pn junction in which one side of the junction is much more heavily doped than the adjacent side.

reverse bias The condition in which a positive voltage is applied to the n region with respect to the p region of a pn junction so that the potential barrier between the two regions increases above the thermal-equilibrium built-in potential barrier.

space charge region The region on either side of the metallurgical junction in which there is a net charge density due to ionized donors in the n region and ionized acceptors in the p region.

space charge width The width of the space charge region, a function of doping concentrations and applied voltage.

varactor diode A diode whose reactance can be varied in a controlled manner with bias voltage.

CHECKPOINT

After studying this chapter, the reader should have the ability to:

- Describe why and how the space charge region is formed.
- Draw the energy-band diagram of a zero-biased and reverse-biased pn junction.
- Define and derive the expression of the built-in potential barrier voltage.
- Derive the expression for the electric field in space charge region of the pn junction.
- Describe what happens to the parameters of the space charge region when a reverse-biased voltage is applied.
- Define and explain the junction capacitance.
- Describe the characteristics and properties of a one-sided pn junction.
- Describe the avalanche breakdown mechanism in a reverse-biased pn junction.
- Describe how a linearly graded junction is formed.
- Define a hyperabrupt junction.

REVIEW QUESTIONS

1. Define the built-in potential voltage and describe how it maintains thermal equilibrium.
2. Why is an electric field formed in the space charge region? Why is the electric field a linear function of distance in a uniformly doped pn junction?
3. Where does the maximum electric field occur in the space charge region?
4. Why is the space charge width larger in the lower doped side of a pn junction?
5. What is the functional dependence of the space charge width on reverse-biased voltage?
6. Why does the space charge width increase with reverse-biased voltage?
7. Why does a capacitance exist in a reverse-biased pn junction? Why does the capacitance decrease with increasing reverse-biased voltage?
8. What is a one-sided pn junction? What parameters can be determined in a one-sided pn junction?
9. Why does the breakdown voltage of a pn junction decrease as the doping concentration increases?
10. What is a linearly graded junction?
11. What is a hyperabrupt junction and what is one advantage or characteristic of such a junction?

PROBLEMS

Section 7.2 Zero Applied Bias

7.1 (*a*) Calculate V_{bi} in a silicon pn junction at $T = 300$ K for (*a*) $N_a = 2 \times 10^{15}$ cm^{-3} and $N_d =$ (*i*) 2×10^{15}, (*ii*) 2×10^{16}, and (*iii*) 2×10^{17} cm^{-3}. (*b*) Repeat part (*a*) for $N_a = 2 \times 10^{17}$ cm^{-3}.

7.2 Calculate the built-in potential barrier, V_{bi}, for Si, Ge, and GaAs pn junctions if they each have the following dopant concentrations at $T = 300$ K:

(*a*) $N_d = 10^{14}$ cm^{-3} $\quad N_a = 10^{17}$ cm^{-3}

(*b*) $N_d = 5 \times 10^{16}$ $\quad N_a = 5 \times 10^{16}$

(*c*) $N_d = 10^{17}$ $\quad N_a = 10^{17}$

7.3 (*a*) Plot the built-in potential barrier for a symmetrical ($N_a = N_d$) silicon pn junction at $T = 300$ K over the range $10^{14} \leq N_a = N_d \leq 10^{17}$ cm^{-3}. (*b*) Repeat part (*a*) for a GaAs pn junction. (*c*) Repeat parts (*a*) and (*b*) for $T = 400$ K.

7.4 An abrupt silicon pn junction at zero bias has dopant concentrations of $N_a = 10^{17}$ cm^{-3} and $N_d = 5 \times 10^{15}$ cm^{-3}. $T = 300$ K. (*a*) Calculate the Fermi level on each side of the junction with respect to the intrinsic Fermi level. (*b*) Sketch the equilibrium energy-band diagram for the junction and determine V_{bi} from the diagram and the results of part (*a*). (*c*) Calculate V_{bi} using Equation (7.10), and compare the results to part (*b*). (*d*) Determine x_n, x_p, and the peak electric field for this junction.

7.5 Repeat problem 7.4 for the case when the doping concentrations are $N_a = N_d = 2 \times 10^{16}$ cm^{-3}.

7.6 A silicon pn junction in thermal equilibrium at $T = 300$ K is doped such that $E_F - E_{Fi} = 0.365$ eV in the n region and $E_{Fi} - E_F = 0.330$ eV in the p region.

(*a*) Sketch the energy-band diagram for the pn junction. (*b*) Find the impurity doping concentration in each region. (*c*) Determine V_{bi}.

7.7 Consider a uniformly doped GaAs pn junction with doping concentrations of $N_a = 2 \times 10^{15}$ cm^{-3} and $N_d = 4 \times 10^{16}$ cm^{-3}. Plot the built-in potential barrier V_{bi} versus temperature over the range $200 \leq T \leq 400$ K.

7.8 (*a*) Consider a uniformly doped silicon pn junction at $T = 300$ K. At zero bias, 25 percent of the total space charge region is in the n-region. The built-in potential barrier is $V_{bi} = 0.710$ V. Determine (*i*) N_a, (*ii*) N_d, (*iii*) x_n, (*iv*) x_p, and (*v*) $|E_{max}|$. (*b*) Repeat part (*a*) for a GaAs pn junction with $V_{bi} = 1.180$ V.

7.9 Consider the impurity doping profile shown in Figure P7.9 in a silicon pn junction. For zero applied voltage, (*a*) determine V_{bi}, (*b*) calculate x_n and x_p, (*c*) sketch the thermal equilibrium energy-band diagram, and (*d*) plot the electric field versus distance through the junction.

7.10 Consider a uniformly doped silicon pn junction with doping concentrations $N_a = 2 \times 10^{17}$ cm^{-3} and $N_d = 4 \times 10^{16}$ cm^{-3}. (*a*) Determine V_{bi} at $T = 300$ K. (*b*) Determine the temperature at which V_{bi} increases by 2 percent. (Trial and error may have to be used.)

7.11 The doping concentrations in a uniformly doped silicon pn junction are $N_a = 4 \times 10^{16}$ cm^{-3} and $N_d = 2 \times 10^{15}$ cm^{-3}. The measured built-in potential barrier is $V_{bi} = 0.550$ V. Determine the temperature at which this result occurs.

7.12 An "isotype" step junction is one in which the same impurity type doping changes from one concentration value to another value. An n-n isotype doping profile is shown in Figure P7.12. (*a*) Sketch the thermal equilibrium energy-band diagram of the isotype junction. (*b*) Using the energy-band diagram, determine the built-in potential barrier. (*c*) Discuss the charge distribution through the junction.

7.13 A particular type of junction is an n region adjacent to an intrinsic region. This junction can be modeled as an n-type region to a lightly doped p-type region. Assume the doping concentrations in silicon at $T = 300$ K are $N_d = 10^{16}$ cm^{-3} and $N_a = 10^{12}$ cm^{-3}. For zero applied bias, determine (*a*) V_{bi}, (*b*) x_n, (*c*) x_p, and (*d*) $|E_{max}|$. Sketch the electric field versus distance through the junction.

7.14 We are assuming an abrupt depletion approximation for the space charge region. That is, no free carriers exist within the depletion region and the semiconductor abruptly changes to a neutral region outside the space charge region. This approximation is

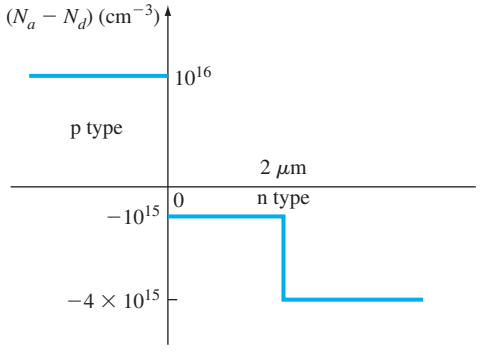


Figure P7.9 | Figure for Problem 7.9.

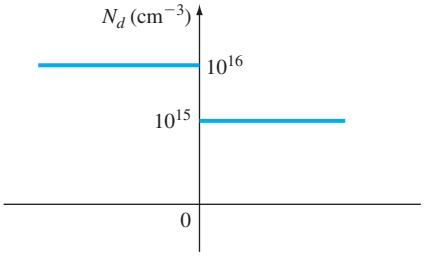


Figure P7.12 | Figure for Problem 7.12.

adequate for most applications, but the abrupt transition does not exist. The space charge region changes over a distance of a few Debye lengths, where the Debye length in the n region is given by

$$L_D = \left[\frac{\epsilon_s kT}{e^2 N_d}\right]^{1/2}$$

Calculate L_D and find the ratio of L_D/x_n for the following conditions. The p-type doping concentration is $N_a = 8 \times 10^{17}$ cm^{-3} and the n-type doping concentration is (*a*) $N_d = 8 \times 10^{14}$ cm^{-3}, (*b*) $N_d = 2.2 \times 10^{16}$ cm^{-3}, and (*c*) $N_d = 8 \times 10^{17}$ cm^{-3}.

7.15 Examine the electric field versus distance through a uniformly doped silicon pn junction at $T = 300$ K as a function of doping concentrations. Assume zero applied bias. Sketch the electric field versus distance through the space charge region and calculate $|E_{max}|$ for: (*a*) $N_a = 10^{17}$ cm^{-3} and $10^{14} \leq N_d \leq 10^{17}$ cm^{-3} and (*b*) $N_a = 10^{14}$ cm^{-3} and $10^{14} \leq N_d \leq 10^{17}$ cm^{-3}. (*c*) What can be said about the results for $N_d \geq 100\, N_a$ or $N_a \geq 100\, N_d$?

Section 7.3 Reverse Applied Bias

7.16 An abrupt silicon pn junction at $T = 300$ K has impurity doping concentrations of $N_a = 5 \times 10^{16}$ cm^{-3} and $N_d = 10^{15}$ cm^{-3}. Calculate (*a*) V_{bi}, (*b*) W at (*i*) $V_R = 0$ and (*ii*) $V_R = 5$ V, and (*c*) $|E_{max}|$ at (*i*) $V_R = 0$ and (*ii*) $V_R = 5$.

7.17 Consider the pn junction described in Problem 7.10 for $T = 300$ K. The cross-sectional area of the junction is 2×10^{-4} cm^2 and the applied reverse-biased voltage is $V_R = 2.5$ V. Calculate (*a*) V_{bi}, (*b*) x_n, x_p, W, (*c*) $|E_{max}|$, and (*d*) the junction capacitance.

7.18 An ideal one-sided silicon p$^+$n junction at $T = 300$ K is uniformly doped on both sides of the metallurgical junction. It is found that the doping relation is $N_a = 80\, N_d$ and the built-in potential barrier is $V_{bi} = 0.740$ V. A reverse-biased voltage of $V_R = 10$ V is applied. Determine (*a*) N_a, N_d; (*b*) x_p, x_n; (*c*) $|E_{max}|$; and (*d*) C_j'.

7.19 A silicon n$^+$p junction is biased at $V_R = 5$ V. (*a*) Determine the change in built-in potential barrier if the doping concentration in the p region increases by a factor of 3. (*b*) Determine the ratio of junction capacitance when the acceptor doping is $3N_a$ compared to that when the acceptor doping is N_a. (*c*) Why does the junction capacitance increase when the doping concentration increases?

7.20 (*a*) The peak electric field in a reverse-biased silicon pn junction is $|E_{max}| = 3 \times 10^5$ V/cm. The doping concentrations are $N_d = 4 \times 10^{15}$ cm^{-3} and $N_a = 4 \times 10^{17}$ cm^{-3}. Find the magnitude of the reverse-biased voltage. (*b*) Repeat part (*a*) for $N_d = 4 \times 10^{16}$ cm^{-3} and $N_a = 4 \times 10^{17}$ cm^{-3}. (*c*) Repeat part (*a*) for $N_d = N_a = 4 \times 10^{17}$ cm^{-3}.

7.21 Consider two p$^+$n silicon junctions at $T = 300$ K reverse biased at $V_R = 5$ V. The impurity doping concentrations in junction *A* are $N_a = 10^{18}$ cm^{-3} and $N_d = 10^{15}$ cm^{-3}, and those in junction *B* are $N_a = 10^{18}$ cm^{-3} and $N_d = 10^{16}$ cm^{-3}. Calculate the ratio of the following parameters for junction *A* to junction *B*: (*a*) W, (*b*) $|E_{max}|$, and (*c*) C_j'.

7.22 Consider a uniformly doped GaAs pn junction at $T = 300$ K. The junction capacitance at zero bias is $C_j(0)$ and the junction capacitance with a 10-V reverse-biased voltage is $C_j(10)$. The ratio of the capacitances is

$$\frac{C_j(0)}{C_j(10)} = 3.13$$

Also under reverse bias, the space charge width into the p region is 0.2 of the total space charge width. Determine (*a*) V_{bi} and (*b*) N_a, N_d.

7.23 A GaAs pn junction at $T = 300$ K has impurity doping concentrations of $N_a = 2 \times 10^{16}$ cm^{-3} and $N_d = 5 \times 10^{15}$ cm^{-3}. It is determined that the ratio of junction capacitance at two reverse-biased voltages is $C_j'(V_{R1})/C_j'(V_{R2}) = 1.5$, where $V_{R1} = 0.5$ V. Determine V_{R2}.

7.24 (*a*) The impurity doping concentrations in a silicon pn junction at $T = 300$ K are $N_a = 2 \times 10^{15}$ cm^{-3} and $N_d = 4 \times 10^{16}$ cm^{-3}. The cross-sectional area of the pn junction is 5×10^{-4} cm^2. Determine the junction capacitance at (*i*) $V_R = 0$ and (*ii*) $V_R = 5$ V. (*b*) Repeat part (*a*) for a GaAs pn junction.

7.25 An abrupt silicon pn junction at $T = 300$ K is uniformly doped with $N_a = 2 \times 10^{17}$ cm^{-3} and $N_d = 5 \times 10^{15}$ cm^{-3}. The cross-sectional area of the pn junction is 8×10^{-4} cm^2. An inductance is placed in parallel with the pn junction. (*a*) With a reverse-biased voltage of $V_R = 10$ V applied to the pn junction, the resonant frequency of the circuit is $f = 1.25$ MHz. What is the value of the inductance? (*b*) Using the results of part (*a*), what is the resonant frequency if the reverse-biased voltage is (*i*) $V_R = 1$ V and (*ii*) $V_R = 5$ V?

7.26 (*a*) A uniformly doped silicon p$^+$n junction at $T = 300$ K is to be designed such that, at a reverse-biased of $V_R = 10$ V, the maximum electric field is limited to $|E_{max}| = 2.5 \times 10^5$ V/cm. Determine the maximum doping concentration in the n region. (Use an approximate value for V_{bi}.) (*b*) Repeat part (*a*) if the maximum electric field is limited to $|E_{max}| = 10^5$ V/cm.

7.27 (*a*) A GaAs pn junction at $T = 300$ K, with a cross-sectional area of 10^{-4} cm^2, is to be designed that meets the following specifications. At a reverse-biased voltage of $V_R = 2$ V, 20 percent of the total space charge width is to be in the p region and the total junction capacitance is to be 0.6 pF. Determine N_a, N_d, and W. (*b*) Repeat part (*a*) if $V_R = 5$ V.

7.28 A silicon pn junction at $T = 300$ K has the doping profile shown in Figure P7.28. Calculate (*a*) V_{bi}, (*b*) x_n and x_p at zero bias, and (*c*) the applied bias required so that $x_n = 30$ μm.

7.29 Consider a silicon pn junction with the doping profile shown in Figure P7.29. $T = 300$ K. (*a*) Calculate the applied reverse-biased voltage required so that the space charge region extends entirely through the p region. (*b*) Determine the space charge width into the n$^+$ region with the reverse-biased voltage calculated in part (*a*). (*c*) Calculate the peak electric field for this applied voltage.

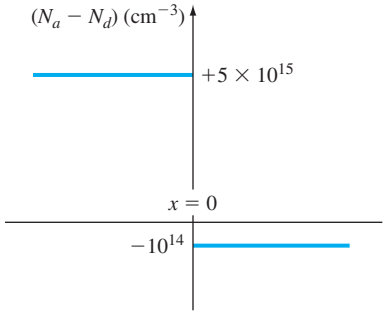


Figure P7.28 | Figure for Problem 7.28.

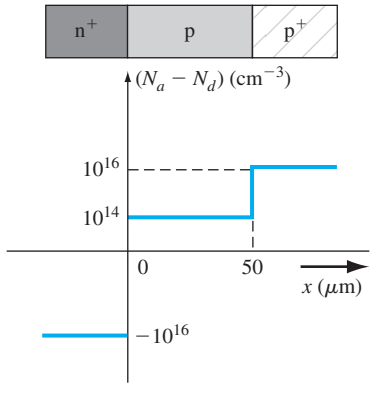


Figure P7.29 | Figure for Problem 7.29.

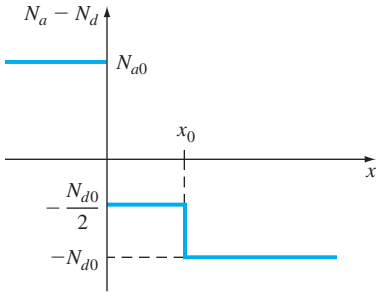


Figure P7.33 | Figure for Problem 7.33.

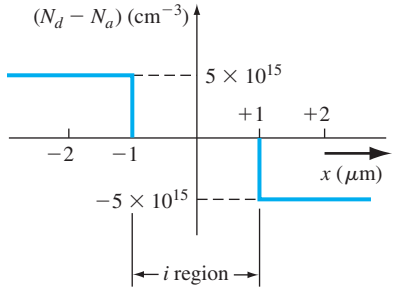


Figure P7.34 | Figure for Problem 7.34.

7.30 A silicon p$^+$n junction has doping concentrations of $N_a = 2 \times 10^{17}$ cm^{-3} and $N_d = 2 \times 10^{15}$ cm^{-3}. The cross-sectional area is 10^{-5} cm^2. Calculate (*a*) V_{bi} and (*b*) the junction capacitance at (*i*) $V_R = 1$ V, (*ii*) $V_R = 3$ V, and (*iii*) $V_R = 5$ V. (*c*) Plot $1/C^2$ versus V_R and show that the slope can be used to find N_d and the intercept at the voltage axis yields V_{bi}.

7.31 The total junction capacitance of a GaAs pn junction at $T = 300$ K is found to be 1.10 pF at $V_R = 1$ V. The doping concentration in one region is measured and found to be 8×10^{16} cm^{-3}, and the built-in potential is found to be $V_{bi} = 1.20$ V. Determine (*a*) the doping in the other region of the pn junction and (*b*) the cross-sectional area. (*c*) The reverse-biased voltage is changed and the capacitance is found to be 0.80 pF. What is the value of V_R?

7.32 Examine how the capacitance C' and the function $(1/C')^2$ vary with reverse-biased voltage V_R as the doping concentrations change. In particular, consider these plots versus N_a for $N_a \geq 100\, N_d$ and versus N_d for $N_d \geq 100\, N_a$.

***7.33** A pn junction has the doping profile shown in Figure P7.33. Assume that $x_n > x_0$ for all reverse-biased voltages. (*a*) What is the built-in potential across the junction? (*b*) For the abrupt junction approximation, sketch the charge density through the junction. (*c*) Derive the expression for the electric field through the space charge region.

***7.34** A silicon PIN junction has the doping profile shown in Figure P7.34. The "I" corresponds to an ideal intrinsic region in which there is no impurity doping concentration. A reverse-biased voltage is applied to the PIN junction so that the total depletion width extends from -2 μm to $+2$ μm. (*a*) Using Poisson's equation, calculate the magnitude of the electric field at $x = 0$. (*b*) Sketch the electric field through the PIN junction. (*c*) Calculate the reverse-biased voltage that must be applied.

Section 7.4 Junction Breakdown

7.35 Consider a silicon n$^+$p junction diode. The critical electric field for breakdown in silicon is approximately $E_{crit} = 4 \times 10^5$ V/cm. Determine the maximum p-type doping concentration such that the breakdown voltage is (*a*) 40 V and (*b*) 20 V.

7.36 Design an abrupt silicon n$^+$p junction diode that has a reverse breakdown voltage of 80 V.

7.37 (*a*) The n-type doping concentration in an abrupt p$^+$n GaAs junction diode is $N_d = 10^{16}$ cm^{-3}. Determine the breakdown voltage. (*b*) Repeat part (*a*) for $N_d = 10^{15}$ cm^{-3}.

*Asterisks next to problems indicate problems that are more difficult.

7.38 (*a*) A symmetrically doped silicon pn junction diode has doping concentrations of $N_a = N_d = 2 \times 10^{16}$ cm^{-3}. Assuming the critical electric field is $E_{crit} = 4 \times 10^5$ V/cm, determine the breakdown voltage. (*b*) Repeat part (*a*) if the doping concentrations are $N_a = N_d = 5 \times 10^{15}$ cm^{-3}.

7.39 An abrupt silicon p$^+$n junction has an n-region doping concentration of $N_d = 5 \times 10^{15}$ cm^{-3}. What must be the minimum n-region width such that avalanche breakdown occurs before the depletion region reaches an ohmic contact (punchthrough)?

7.40 A silicon pn junction diode is doped with $N_a = N_d = 10^{18}$ cm^{-3}. Zener breakdown occurs when the peak electric field reaches 10^6 V/cm. Determine the reverse-biased breakdown voltage.

7.41 A diode will very often have the doping profile shown in Figure P7.29, which is known as an n$^+$pp$^+$ diode. Under reverse bias, the depletion region must remain within the p region to avoid premature breakdown. Assume the p-region doping is 10^{15} cm^{-3}. Determine the reverse-biased voltage such that the depletion region remains within the p region and does not reach breakdown if the p-region width is (*a*) 75 μm and (*b*) 150 μm. For each case, state whether the maximum depletion width or the breakdown voltage is reached first.

7.42 Consider a silicon pn junction at $T = 300$ K whose doping profile varies linearly from $N_a = 10^{18}$ cm^{-3} to $N_d = 10^{18}$ cm^{-3} over a distance of 2 μm. Estimate the breakdown voltage.

Section 7.5 Nonuniformly Doped Junctions

7.43 Consider a linearly graded junction. (*a*) Starting with Equation (7.62), derive the expression for the electric field given in Equation (7.64). (*b*) Derive the expression for the potential through the space charge region given by Equation (7.66).

7.44 The built-in potential barrier of a linearly graded silicon pn junction at $T = 300$ K is $V_{bi} = 0.70$ V. The junction capacitance measured at $V_R = 3.5$ V is $C' = 7.2 \times 10^{-9}$ F/cm^2. Find the gradient, a, of the net impurity concentration.

Summary and Review

7.45 (*a*) A one-sided silicon n$^+$p junction at $T = 300$ K is doped at $N_d = 3 \times 10^{17}$ cm^{-3}. Design the junction such that $C_j = 0.45$ pF at $V_R = 5$ V. (*b*) Calculate the junction capacitance at (*i*) $V_R = 2.5$ V and (*ii*) $V_R = 0$ V.

7.46 A one-sided p$^+$n junction with a cross-sectional area of 10^{-5} cm^2 has a measured built-in potential of $V_{bi} = 0.8$ V at $T = 300$ K. A plot of $(1/C_j)^2$ versus V_R is linear for $V_R < 1$ V and is essentially constant for $V_R > 1$ V. The capacitance is $C_j = 0.082$ pF at $V_R = 1$ V. Determine the doping concentrations on either side of the metallurgical junction that will produce this capacitance characteristic.

***7.47** Silicon, at $T = 300$ K, is doped at $N_{d1} = 10^{15}$ cm^{-3} for $x < 0$ and $N_{d2} = 5 \times 10^{16}$ cm^{-3} for $x > 0$ to form an $n - n$ step junction. (*a*) Sketch the energy-band diagram. (*b*) Derive an expression for V_{bi}. (*c*) Sketch the charge density, electric field, and potential through the junction. (*d*) Explain where the charge density came from and is located.

***7.48** A diffused silicon pn junction has a linearly graded junction on the p side with $a = 2 \times 10^{19}$ cm^{-4}, and a uniform doping of 10^{15} cm^{-3} on the n side. (*a*) If the depletion width on the p side is 0.7 μm at zero bias, find the total depletion width, built-in potential, and maximum electric field at zero bias. (*b*) Plot the potential function through the junction.

READING LIST

1. Dimitrijev, S. *Principles of Semiconductor Devices.* New York: Oxford University Press, 2006.
2. Kano, K. *Semiconductor Devices.* Upper Saddle River, NJ: Prentice Hall, 1998.
*3. Li, S. S. *Semiconductor Physical Electronics*. New York: Plenum Press, 1993.
4. Muller, R. S., and T. I. Kamins. *Device Electronics for Integrated Circuits,* 2nd ed. New York: John Wiley and Sons, 1986.
5. Navon, D. H. S*emiconductor Microdevices and Materials*. New York: Holt, Rinehart & Winston, 1986.
6. Neudeck, G. W. *The PN Junction Diode*. Vol. 2 of the *Modular Series on Solid State Devices,* 2nd ed. Reading, MA: Addison-Wesley, 1989.
*7. Ng, K. K. *Complete Guide to Semiconductor Devices*. New York: McGraw-Hill, 1995.
8. Pierret, R. F. *Semiconductor Device Fundamentals. Reading,* MA: Addison-Wesley, 1996.
*9. Roulston, D. J. *An Introduction to the Physics of Semiconductor Devices.* New York: Oxford University Press, 1999.
10. Shur, M. *Introduction to Electronic Devices*. New York: John Wiley and Sons, 1996.
*11. Shur, M. *Physics of Semiconductor Devices*. Englewood Cliffs, NJ: Prentice Hall, 1990.
12. Singh, J. *Semiconductor Devices: Basic Principles*. New York: John Wiley and Sons, 2001.
13. Streetman, B. G., and S. K. Banerjee. *Solid State Electronic Devices,* 6th ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2006.
14. Sze, S. M., and K. K. Ng. *Physics of Semiconductor Devices,* 3rd ed. Hoboken, NJ: John Wiley and Sons, 2007.
15. Sze, S. M. *Semiconductor Devices: Physics and Technology,* 2nd ed. New York: John Wiley and Sons, 2001.
*16. Wang, S. *Fundamentals of Semiconductor Theory and Device Physics*. Englewood Cliffs, NJ: Prentice Hall, 1989.
17. Yang, E. S. *Microelectronic Devices*. New York: McGraw-Hill, 1988.

*Indicates references that are at an advanced level compared to this text.

CHAPTER 8

The pn Junction Diode

In the last chapter, we discussed the electrostatics of the pn junction in thermal equilibrium and under reverse bias. We determined the built-in potential barrier at thermal equilibrium and calculated the electric field in the space charge region. We also considered the junction capacitance.

In this chapter, we consider the pn junction with a forward-bias voltage applied and determine the current–voltage characteristics. The potential barrier of the pn junction is lowered when a forward-bias voltage is applied, allowing electrons and holes to flow across the space charge region. When holes flow from the p region across the space charge region into the n region, they become excess minority carrier holes and are subject to the excess minority carrier diffusion, drift, and recombination processes discussed in Chapter 6. Likewise, when electrons from the n region flow across the space charge region into the p region, they become excess minority carrier electrons and are subject to these same processes. ■

8.0 | PREVIEW

In this chapter, we will:

- Consider the process by which the potential barrier of a pn junction is lowered when a forward-bias voltage is applied, so holes and electrons can flow across the junction generating a diode current.
- Derive the boundary conditions for excess holes in the n region and excess-electrons in the p region, and analyze the behavior of these excess carriers under a forward bias.
- Derive the ideal current–voltage relation of the forward-biased pn junction diode.
- Describe and analyze nonideal effects in the pn junction diode such as high-level injection, and generation and recombination currents.
- Develop a small-signal equivalent circuit of the pn junction diode. This equivalent circuit is used to relate small time-varying currents and voltages in the pn junction.

- Discuss large signal diode switching characteristics.
- Describe a specialized pn junction called a tunnel diode.

8.1 | pn JUNCTION CURRENT

When a forward-bias voltage is applied to a pn junction, a current will be induced in the device. We initially consider a qualitative discussion of how charges flow in the pn junction and then consider the mathematical derivation of the current–voltage relationship.

8.1.1 Qualitative Description of Charge Flow in a pn Junction

We can qualitatively understand the mechanism of the current in a pn junction by again considering the energy-band diagrams. Figure 8.1a shows the energy-band diagram of a pn junction in thermal equilibrium that was developed in the last chapter. We argued that the potential barrier seen by the electrons, for example, holds back the large concentration of electrons in the n region and keeps them from flowing into the p region. Similarly, the potential barrier seen by the holes holds back the large concentration of holes in the p region and keeps them from flowing into the n region. The potential barrier, then, maintains thermal equilibrium.

Figure 8.1b shows the energy-band diagram of a reverse-biased pn junction. The potential of the n region is positive with respect to the p region so the Fermi energy

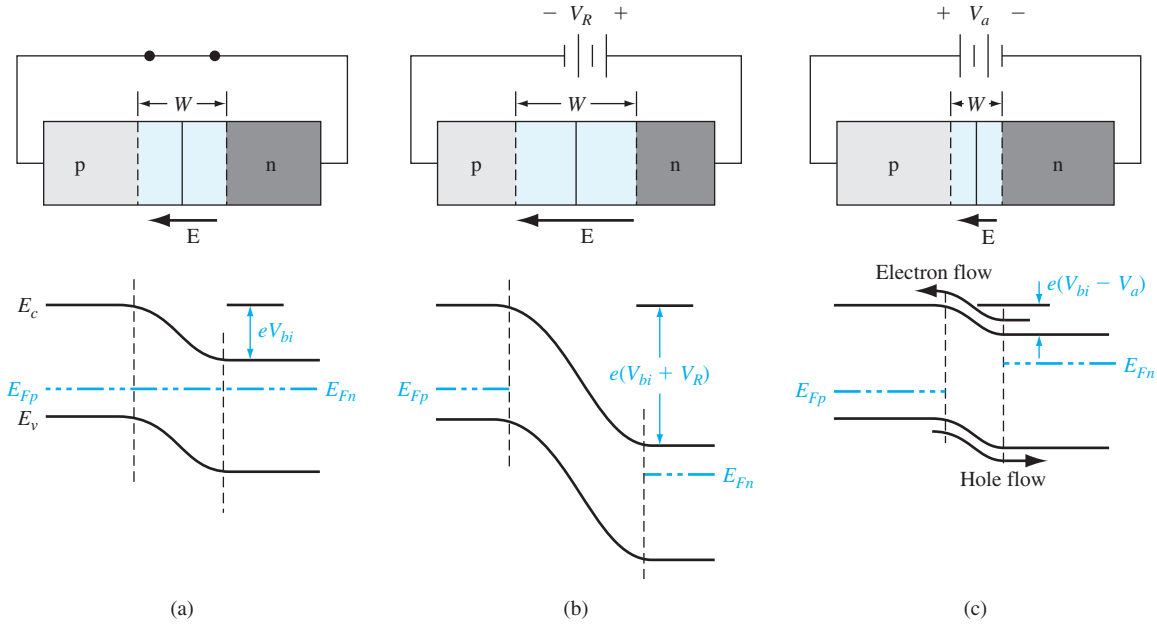


Figure 8.1 | A pn junction and its associated energy-band diagram for (a) zero bias, (b) reverse bias, and (c) forward bias.

in the n region is lower than that in the p region. The total potential barrier is now larger than that for the zero-bias case. We argued in the last chapter that the increased potential barrier continues to hold back the electrons and holes so that there is still essentially no charge flow and hence essentially no current.

Figure 8.1c shows the energy-band diagram for the case when a positive voltage is applied to the p region with respect to the n region. The Fermi level in the p region is now lower than that in the n region. The total potential barrier is now reduced. The smaller potential barrier means that the electric field in the depletion region is also reduced. The smaller electric field means that the electrons and holes are no longer held back in the n and p regions, respectively. There will be a diffusion of holes from the p region across the space charge region where they will flow into the n region. Similarly, there will be a diffusion of electrons from the n region across the space charge region where they will flow into the p region. The flow of charge generates a current through the pn junction.

The injection of holes into the n region means that these holes are minority carriers. Likewise, the injection of electrons into the p region means that these electrons are minority carriers. The behavior of these minority carriers is described by the ambipolar transport equations that were discussed in Chapter 6. There will be diffusion as well as recombination of excess carriers in these regions. The diffusion of carriers implies that there will be diffusion currents. The mathematical derivation of the pn junction current–voltage relationship is considered in the next section.

8.1.2 Ideal Current–Voltage Relationship

The ideal current–voltage relationship of a pn junction is derived on the basis of four assumptions. (The last assumption has three parts, but each part deals with current.) They are:

1. The abrupt depletion layer approximation applies. The space charge regions have abrupt boundaries, and the semiconductor is neutral outside of the depletion region.

2. The Maxwell–Boltzmann approximation applies to carrier statistics.

3. The concepts of low injection and complete ionization apply.

4a. The total current is a constant throughout the entire pn structure.

4b. The individual electron and hole currents are continuous functions through the pn structure.

4c. The individual electron and hole currents are constant throughout the depletion region.

Notation can sometimes appear to be overwhelming in the equations in this chapter. Table 8.1 lists some of the various electron and hole concentration terms that appear. Many terms have already been used in previous chapters but are repeated here for convenience.

Table 8.1 | Commonly used terms and notation for this chapter

Term	Meaning
N_a	Acceptor concentration in the p region of the pn junction
N_d	Donor concentration in the n region of the pn junction
$n_{n0} = N_d$	Thermal-equilibrium majority carrier electron concentration in the n region
$p_{p0} = N_a$	Thermal-equilibrium majority carrier hole concentration in the p region
$n_{p0} = n_i^2/N_a$	Thermal-equilibrium minority carrier electron concentration in the p region
$p_{n0} = n_i^2/N_d$	Thermal-equilibrium minority carrier hole concentration in the n region
n_p	Total minority carrier electron concentration in the p region
p_n	Total minority carrier hole concentration in the n region
$n_p(-x_p)$	Minority carrier electron concentration in the p region at the space charge edge
$p_n(x_n)$	Minority carrier hole concentration in the n region at the space charge edge
$\delta n_p = n_p - n_{p0}$	Excess minority carrier electron concentration in the p region
$\delta p_n = p_n - p_{n0}$	Excess minority carrier hole concentration in the n region

8.1.3 Boundary Conditions

Figure 8.2 shows the conduction-band energy through the pn junction in thermal equilibrium. The n region contains many more electrons in the conduction band than the p region; the built-in potential barrier prevents this large density of electrons from flowing into the p region. The built-in potential barrier maintains equilibrium between the carrier distributions on either side of the junction.

An expression for the built-in potential barrier was derived in the last chapter and was given by Equation (7.10) as

$$V_{bi} = V_t \ln\left(\frac{N_a N_d}{n_i^2}\right)$$

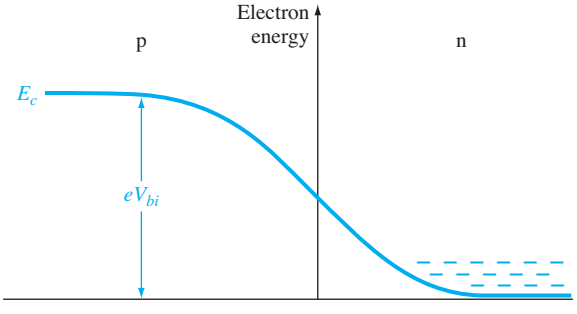


Figure 8.2 | Conduction-band energy through a pn junction.

If we divide the equation by $V_t = kT/e$, take the exponential of both sides, and then take the reciprocal, we obtain

$$\frac{n_i^2}{N_a N_d} = \exp\left(\frac{-eV_{bi}}{kT}\right) \tag{8.1}$$

If we assume complete ionization, we can write

$$n_{n0} \approx N_d \tag{8.2}$$

where n_{n0} is the thermal-equilibrium concentration of majority carrier electrons in the n region. In the p region, we can write

$$n_{p0} \approx \frac{n_i^2}{N_a} \tag{8.3}$$

where n_{p0} is the thermal-equilibrium concentration of minority carrier electrons. Substituting Equations (8.2) and (8.3) into Equation (8.1), we obtain

$$n_{p0} = n_{n0} \exp\left(\frac{-eV_{bi}}{kT}\right) \tag{8.4}$$

This equation relates the minority carrier electron concentration on the p side of the junction to the majority carrier electron concentration on the n side of the junction in thermal equilibrium.

If a positive voltage is applied to the p region with respect to the n region, the potential barrier is reduced. Figure 8.3a shows a pn junction with an applied voltage V_a. The electric field in the bulk p and n regions is normally very small. Essentially all of the applied voltage is across the junction region. The electric field E_{app} induced by the applied voltage is in the opposite direction to the thermal-equilibrium space charge electric field, so the net electric field in the space charge region is reduced below the equilibrium value. The delicate balance between diffusion and the E-field

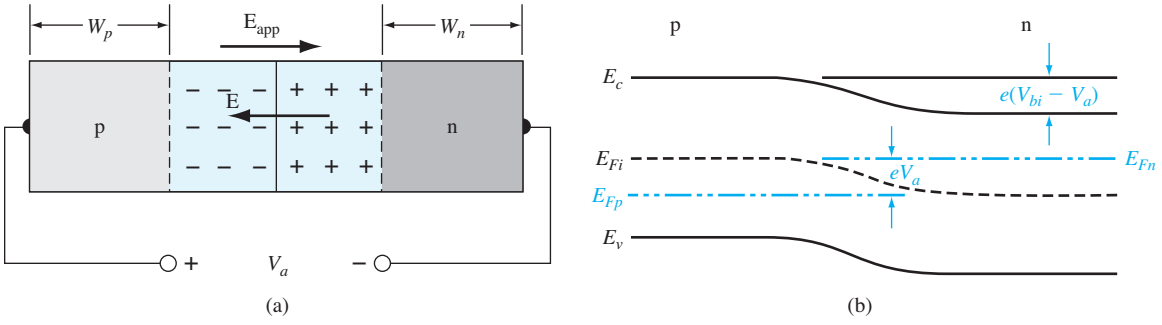


Figure 8.3 | (a) A pn junction with an applied forward-bias voltage showing the directions of the electric field induced by V_a and the space charge electric field. (b) Energy-band diagram of the forward-biased pn junction.

force achieved at thermal equilibrium is upset. The electric field force that prevented majority carriers from crossing the space charge region is reduced; majority carrier electrons from the n side are now injected across the depletion region into the p material, and majority carrier holes from the p side are injected across the depletion region into the n material. As long as the bias V_a is applied, the injection of carriers across the space charge region continues and a current is created in the pn junction. This bias condition is known as forward bias; the energy-band diagram of the forward-biased pn junction is shown in Figure 8.3b.

The potential barrier V_{bi} in Equation (8.4) can be replaced by $(V_{bi} - V_a)$ when the junction is forward biased. Equation (8.4) becomes

$$n_p = n_{n0} \exp\left(\frac{-e\,(V_{bi} - V_a)}{kT}\right) = n_{n0} \exp\left(\frac{-eV_{bi}}{kT}\right) \exp\left(\frac{+eV_a}{kT}\right) \tag{8.5}$$

If we assume low injection, the majority carrier electron concentration n_{n0}, for example, does not change significantly. However, the minority carrier concentration, n_p, can deviate from its thermal-equilibrium value n_{p0} by orders of magnitude. Using Equation (8.4), we can write Equation (8.5) as

$$\boxed{n_p = n_{p0} \exp\left(\frac{eV_a}{kT}\right)} \tag{8.6}$$

When a forward-bias voltage is applied to the pn junction, the junction is no longer in thermal equilibrium. The left side of Equation (8.6) is the total minority carrier electron concentration in the p region, which is now greater than the thermal equilibrium value. The forward-bias voltage lowers the potential barrier so that majority carrier electrons from the n region are injected across the junction into the p region, thereby increasing the minority carrier electron concentration. We have produced excess minority carrier electrons in the p region.

When the electrons are injected into the p region, these excess carriers are subject to the diffusion and recombination processes we discussed in Chapter 6. Equation (8.6), then, is the expression for the minority carrier electron concentration at the edge of the space charge region in the p region.

Exactly the same process occurs for majority carrier holes in the p region, which are injected across the space charge region into the n region under a forward-bias voltage. We can write that

$$\boxed{p_n = p_{n0} \exp\left(\frac{eV_a}{kT}\right)} \tag{8.7}$$

where p_n is the concentration of minority carrier holes at the edge of the space charge region in the n region. Figure 8.4 shows these results. By applying a forward-bias voltage, we create excess minority carriers in each region of the pn junction.

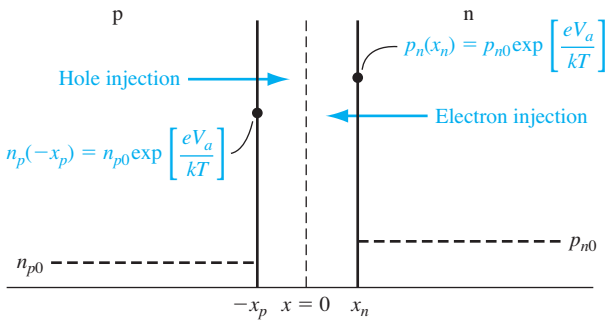


Figure 8.4 | Excess minority carrier concentrations at the space charge edges generated by the forward-bias voltage.

EXAMPLE 8.1

Objective: Calculate the minority carrier concentrations at the edge of the space charge regions in a forward-biased pn junction.

Consider a silicon pn junction at $T = 300$ K. Assume the doping concentration in the n region is $N_d = 10^{16}$ cm^{-3} and the doping concentration in the p region is $N_a = 6 \times 10^{15}$ cm^{-3}, and assume that a forward bias of 0.60 V is applied to the pn junction.

■ Solution

From Equations (8.6) and (8.7) and from Figure 8.4, we have

$$n_p(-x_p) = n_{po}\exp\left(\frac{eV_a}{kT}\right) \quad \text{and} \quad p_n(x_n) = p_{no}\exp\left(\frac{eV_a}{kT}\right)$$

The thermal-equilibrium minority carrier concentrations are

$$n_{po} = \frac{n_i^2}{N_a} = \frac{(1.5 \times 10^{10})^2}{6 \times 10^{15}} = 3.75 \times 10^4 \text{ cm}^{-3}$$

and

$$p_{no} = \frac{n_i^2}{N_d} = \frac{(1.5 \times 10^{10})^2}{10^{16}} = 2.25 \times 10^4 \text{ cm}^{-3}$$

We then have

$$n_p(-x_p) = 3.75 \times 10^4 \exp\left(\frac{0.60}{0.0259}\right) = 4.31 \times 10^{14} \text{ cm}^{-3}$$

and

$$p_n(x_n) = 2.25 \times 10^4 \exp\left(\frac{0.60}{0.0259}\right) = 2.59 \times 10^{14} \text{ cm}^{-3}$$

■ Comment

The minority carrier concentrations can increase by many orders of magnitude when a relatively small forward-bias voltage is applied. Low injection still applies, however, since the excess minority carrier concentrations at the space-charge edges are much less than the thermal-equilibrium majority carrier concentrations.

EXERCISE PROBLEM

Ex 8.1 A silicon pn junction at $T = 300$ K is doped with impurity concentrations of $N_d = 2 \times 10^{16}$ cm^{-3} and $N_a = 5 \times 10^{16}$ cm^{-3}. The junction is forward biased at $V_a = 0.650$ V. Determine the minority carrier concentrations at the space charge edges. Does low injection still apply?

[Ans. $(n_p)_{-x_p} = 3.57 \times 10^{14}$ cm^{-3}, $(p_n)_{x_n} = 8.92 \times 10^{14}$ cm^{-3}, yes]

The minority carrier concentrations at the space charge edges, given by Equations (8.6) and (8.7), were derived assuming that a forward-bias voltage ($V_a > 0$) was applied across the pn junction. However, nothing in the derivation prevents V_a from being negative (reverse bias). If a reverse-biased voltage greater than a few tenths of a volt is applied to the pn junction, then we see from Equations (8.6) and (8.7) that the minority carrier concentrations at the space charge edge are essentially zero. The minority carrier concentrations for the reverse-biased condition drop below the thermal-equilibrium values.

8.1.4 Minority Carrier Distribution

We developed, in Chapter 6, the ambipolar transport equation for excess minority carrier holes in an n region. This equation, in one dimension, is

$$D_p \frac{\partial^2(\delta p_n)}{\partial x^2} - \mu_p \mathrm{E} \frac{\partial(\delta p_n)}{\partial x} + g' - \frac{\delta p_n}{\tau_{p0}} = \frac{\partial(\delta p_n)}{\partial t} \tag{8.8}$$

where $\delta p_n = p_n - p_{n0}$ is the excess minority carrier hole concentration and is the difference between the total and thermal equilibrium minority carrier concentrations. The ambipolar transport equation describes the behavior of excess carriers as a function of time and spatial coordinates.

In Chapter 5, we calculated drift current densities in a semiconductor. We determined that relatively large currents could be created with fairly small electric fields. As a first approximation, we assume that the electric field is zero in both the neutral p and n regions. In the n region for $x > x_n$, we have that $\mathrm{E} = 0$ and $g' = 0$. If we also assume steady state so $\partial(\delta p_n)/\partial t = 0$, then Equation (8.8) reduces to

$$\frac{d^2(\delta p_n)}{dx^2} - \frac{\delta p_n}{L_p^2} = 0 \qquad (x > x_n) \tag{8.9}$$

where $L_p^2 = D_p \tau_{p0}$. For the same set of conditions, the excess minority carrier electron concentration in the p region is determined from

$$\frac{d^2(\delta n_p)}{dx^2} - \frac{\delta n_p}{L_n^2} = 0 \qquad (x < x_p) \tag{8.10}$$

where $L_n^2 = D_n \tau_{n0}$.

The boundary conditions for the total minority carrier concentrations are

$$p_n(x_n) = p_{n0}\exp\left(\frac{eV_a}{kT}\right) \tag{8.11a}$$

$$n_p(-x_p) = n_{p0}\exp\left(\frac{eV_a}{kT}\right) \tag{8.11b}$$

$$p_n(x \to +\infty) = p_{n0} \tag{8.11c}$$

$$n_p(x \to -\infty) = n_{p0} \tag{8.11d}$$

As minority carriers diffuse from the space charge edge into the neutral semiconductor regions, they recombine with majority carriers. We assume that the lengths W_n and W_p shown in Figure 8.3a are very long, meaning in particular that $W_n \gg L_p$ and $W_p \gg L_n$. The excess minority carrier concentrations must approach zero at distances far from the space charge region. The structure is referred to as a long pn junction.

The general solution to Equation (8.9) is

$$\delta p_n(x) = p_n(x) - p_{n0} = Ae^{x/L_p} + Be^{-x/L_p} \qquad (x \geq x_n) \tag{8.12}$$

and the general solution to Equation (8.10) is

$$\delta n_p(x) = n_p(x) - n_{p0} = Ce^{x/L_n} + De^{-x/L_n} \qquad (x \leq -x_p) \tag{8.13}$$

Applying the boundary conditions from Equations (8.11c) and (8.11d), the coefficients A and D must be zero. The coefficients B and C may be determined from the boundary conditions given by Equations (8.11a) and (8.11b). The excess carrier concentrations are then found to be, for $(x \geq x_n)$,

$$\boxed{\delta p_n(x) = p_n(x) - p_{n0} = p_{n0}\left[\exp\left(\frac{eV_a}{kT}\right) - 1\right]\exp\left(\frac{x_n - x}{L_p}\right)} \tag{8.14}$$

and, for $(x \leq -x_p)$,

$$\boxed{\delta n_p(x) = n_p(x) - n_{p0} = n_{p0}\left[\exp\left(\frac{eV_a}{kT}\right) - 1\right]\exp\left(\frac{x_p + x}{L_n}\right)} \tag{8.15}$$

The minority carrier concentrations decay exponentially with distance away from the junction to their thermal-equilibrium values. Figure 8.5 shows these results. Again, we have assumed that both the n-region and the p-region lengths are long compared to the minority carrier diffusion lengths.

In Chapter 6, we discussed the concept of quasi-Fermi levels, which apply to excess carriers in a nonequilibium condition. Since excess electrons exist in the neutral p region and excess holes exist in the neutral n region, we can apply quasi-Fermi levels to these regions. We had defined quasi-Fermi levels in terms of carrier concentrations as

$$p = p_o + \delta p = n_i\exp\left(\frac{E_{Fi} - E_{Fp}}{kT}\right) \tag{8.16}$$

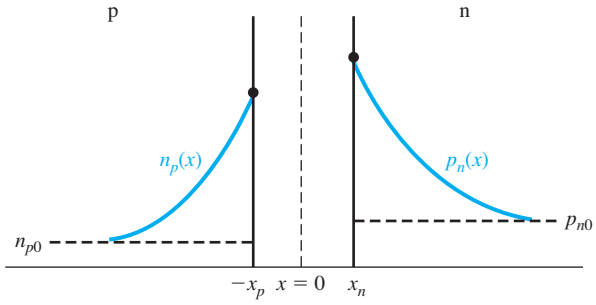


Figure 8.5 | Steady-state minority carrier concentrations in a pn junction under forward bias.

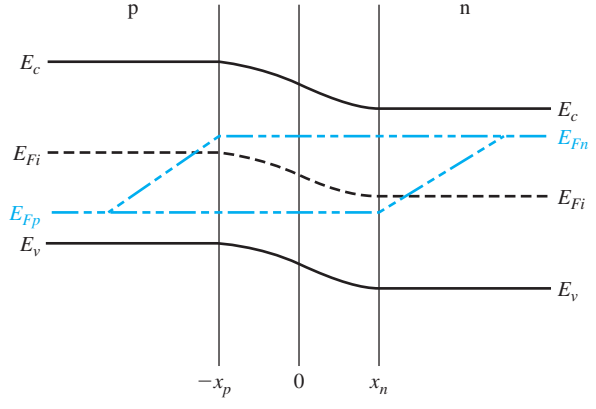


Figure 8.6 | Quasi-Fermi levels through a forward-biased pn junction.

and

$$n = n_o + \delta n = n_i \exp\left(\frac{E_{Fn} - E_{Fi}}{kT}\right) \tag{8.17}$$

Figure 8.6 shows the quasi-Fermi levels through the pn junction. From Equations (8.14) and (8.15), the carrier concentrations are exponential functions of distance, and from Equations (8.16) and (8.17), the carrier concentrations are exponential functions of the quasi-Fermi levels. The quasi-Fermi levels are then linear functions of distance in the neutral p and n regions as shown in Figure 8.6.

We may note that close to the space charge edge in the p region, $E_{Fn} - E_{Fi} > 0$ which means that $\delta n > n_i$. Further from the space charge edge, $E_{Fn} - E_{Fi} < 0$ which means that $\delta n < n_i$ and the excess electron concentration is approaching zero. The same discussion applies to the excess hole concentration in the n region.

At the space charge edge at $x = x_n$, we can write, for low injection

$$n_o p_n(x_n) = n_o p_{no} \exp\left(\frac{V_a}{V_t}\right) = n_i^2 \exp\left(\frac{V_a}{V_t}\right) \tag{8.18}$$

Combining Equations (8.16) and (8.17), we can write

$$np = n_i^2 \exp\left(\frac{E_{Fn} - E_{Fp}}{kT}\right) \tag{8.19}$$

Comparing Equations (8.18) and (8.19), the difference in quasi-Fermi levels is related to the applied bias V_a and represents the deviation from thermal equilibrium. The difference between E_{Fn} and E_{Fp} is nearly constant through the depletion region.

To review, a forward-bias voltage lowers the built-in potential barrier of a pn junction so that electrons from the n region are injected across the space charge region, creating excess minority carriers in the p region. These excess electrons begin diffusing into the bulk p region where they can recombine with majority carrier holes. The excess minority carrier electron concentration then decreases with distance from the junction. The same discussion applies to holes injected across the space charge region into the n region.

8.1.5 Ideal pn Junction Current

The approach we use to determine the current in a pn junction is based on the three parts of the fourth assumption stated earlier in this section. The total current in the junction is the sum of the individual electron and hole currents that are constant through the depletion region. Since the electron and hole currents are continuous functions through the pn junction, the total pn junction current will be the minority carrier hole diffusion current at $x = x_n$ plus the minority carrier electron diffusion current at $x = -x_p$. The gradients in the minority carrier concentrations, as shown in Figure 8.5, produce diffusion currents, and since we are assuming the electric field to be zero at the space charge edges, we can neglect any minority carrier drift current component. This approach in determining the pn junction current is shown in Figure 8.7.

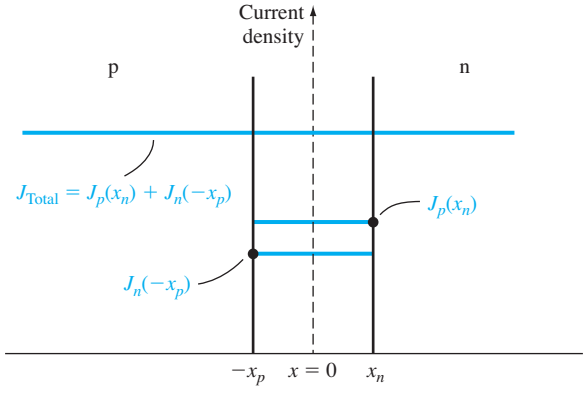


Figure 8.7 | Electron and hole current densities through the space charge region of a pn junction.

We can calculate the minority carrier hole diffusion current density at $x = x_n$ from the relation

$$J_p(x_n) = -eD_p \left.\frac{dp_n(x)}{dx}\right|_{x=x_n} \tag{8.20}$$

Since we are assuming uniformly doped regions, the thermal-equilibrium carrier concentration is constant, so the hole diffusion current density may be written as

$$J_p(x_n) = -eD_p \left.\frac{d(\delta p_n(x))}{dx}\right|_{x=x_n} \tag{8.21}$$

Taking the derivative of Equation (8.14) and substituting into Equation (8.21), we obtain

$$\boxed{J_p(x_n) = \frac{eD_p p_{n0}}{L_p}\left[\exp\left(\frac{eV_a}{kT}\right) - 1\right]} \tag{8.22}$$

The hole current density for this forward-bias condition is in the $+x$ direction, which is from the p to the n region.

Similarly, we may calculate the electron diffusion current density at $x = -x_p$. This may be written as

$$J_n(-x_p) = eD_n \left.\frac{d(\delta n_p(x))}{dx}\right|_{x=-x_p} \tag{8.23}$$

Using Equation (8.15), we obtain

$$\boxed{J_n(-x_p) = \frac{eD_n n_{p0}}{L_n}\left[\exp\left(\frac{eV_a}{kT}\right) - 1\right]} \tag{8.24}$$

The electron current density is also in the $+x$ direction.

An assumption we made at the beginning was that the individual electron and hole currents were continuous functions and constant through the space charge region. The total current is the sum of the electron and hole currents and is constant through the entire junction. Figure 8.7 again shows a plot of the magnitudes of these currents.

The total current density in the pn junction is then

$$J = J_p(x_n) + J_n(-x_p) = \left[\frac{eD_p p_{n0}}{L_p} + \frac{eD_n n_{p0}}{L_n}\right]\left[\exp\left(\frac{eV_a}{kT}\right) - 1\right] \tag{8.25}$$

Equation (8.25) is the ideal current–voltage relationship of a pn junction.

We may define a parameter J_s as

$$\boxed{J_s = \left[\frac{eD_p p_{n0}}{L_p} + \frac{eD_n n_{p0}}{L_n}\right]} \tag{8.26}$$

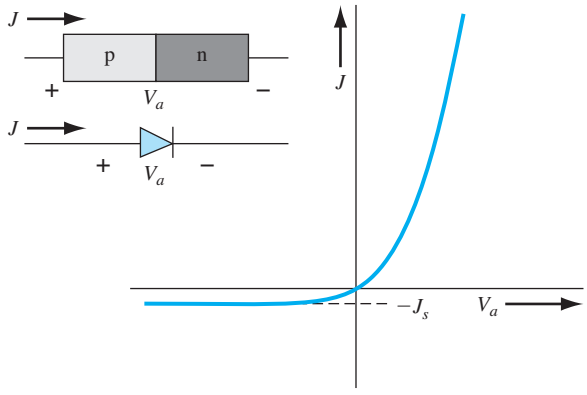


Figure 8.8 | Ideal *I–V* characteristic of a pn junction diode.

so that Equation (8.25) may be written as

$$J = J_s\left[\exp\left(\frac{eV_a}{kT}\right) - 1\right] \tag{8.27}$$

Equation (8.27), known as the ideal-diode equation, gives a good description of the current–voltage characteristics of the pn junction over a wide range of currents and voltages. Although Equation (8.27) was derived assuming a forward-bias voltage ($V_a > 0$), there is nothing to prevent V_a from being negative (reverse bias). Equation (8.27) is plotted in Figure 8.8 as a function of forward-bias voltage V_a. If the voltage V_a becomes negative (reverse bias) by a few kT/eV, then the reverse-biased current density becomes independent of the reverse-biased voltage. The parameter J_s is then referred to as the reverse-saturation current density. The current–voltage characteristics of the pn junction diode are obviously not bilateral.

EXAMPLE 8.2

Objective: Determine the ideal reverse-saturation current density in a silicon pn junction at $T = 300$ K.

Consider the following parameters in a silicon pn junction:

$$N_a = N_d = 10^{16}\ \text{cm}^{-3} \qquad n_i = 1.5 \times 10^{10}\ \text{cm}^{-3}$$
$$D_n = 25\ \text{cm}^2/\text{s} \qquad \tau_{p0} = \tau_{n0} = 5 \times 10^{-7}\ \text{s}$$
$$D_p = 10\ \text{cm}^2/\text{s} \qquad \epsilon_r = 11.7$$

■ Solution

The ideal reverse-saturation current density is given by

$$J_s = \frac{eD_n n_{p0}}{L_n} + \frac{eD_p p_{n0}}{L_p}$$

which may be rewritten as

$$J_s = en_i^2\left(\frac{1}{N_a}\sqrt{\frac{D_n}{\tau_{n0}}} + \frac{1}{N_d}\sqrt{\frac{D_p}{\tau_{p0}}}\right)$$

Then

$$J_s = (1.6 \times 10^{-19})(1.5 \times 10^{10})^2\left(\frac{1}{10^{16}}\sqrt{\frac{25}{5 \times 10^{-7}}} + \frac{1}{10^{16}}\sqrt{\frac{10}{5 \times 10^{-7}}}\right)$$

or $J_s = 4.16 \times 10^{-11}$ A/cm^2

Comment

The ideal reverse-biased saturation current density is very small. If the pn junction cross-sectional area were A $= 10^{-4}$ cm^2, for example, then the ideal reverse-biased diode current would be $I_s = 4.15 \times 10^{-15}$ A.

EXERCISE PROBLEM

Ex 8.2 Consider a GaAs pn junction diode at $T = 300$ K. The parameters of the device are $N_d = 2 \times 10^{16}$ cm^{-3}, $N_a = 8 \times 10^{15}$ cm^{-3}, $D_n = 210$ cm^2/s, $D_p = 8$ cm^2/s, $\tau_{no} = 10^{-7}$ s, and $\tau_{po} = 5 \times 10^{-8}$ s. Determine the ideal reverse-saturation current density.

(Ans. $J_s = 3.30 \times 10^{-18}$ A/cm^2)

If the forward-bias voltage in Equation (8.27) is positive by more than a few kT/eV, then the (-1) term in Equation (8.27) becomes negligible. Figure 8.9 shows the forward-bias current–voltage characteristic when the current is plotted on a log scale. Ideally, this plot yields a straight line when V_a is greater than a few kT/eV. The forward-bias current is an exponential function of the forward-bias voltage.

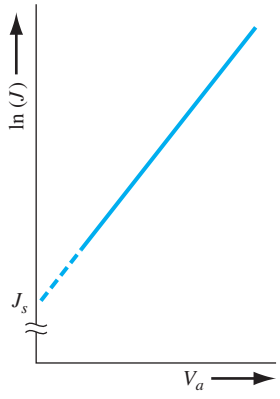


Figure 8.9 | Ideal I–V characteristic of a pn junction diode with the current plotted on a log scale.

EXAMPLE 8.3

Objective: Design a pn junction diode to produce particular electron and hole current densities at a given forward-bias voltage.

Consider a silicon pn junction diode at $T = 300$ K. Design the diode such that $J_n = 20$ A/cm^2 and $J_p = 5$ A/cm^2 at $V_a = 0.65$ V. Assume the remaining semiconductor parameters are as given in Example 8.2.

■ Solution

The electron diffusion current density is given by Equation (8.24) as

$$J_n = \frac{eD_n n_{p0}}{L_n}\left[\exp\left(\frac{eV_a}{kT}\right) - 1\right] = e\sqrt{\frac{D_n}{\tau_{n0}}} \cdot \frac{n_i^2}{N_a}\left[\exp\left(\frac{eV_a}{kT}\right) - 1\right]$$

Substituting the numbers, we have

$$20 = (1.6 \times 10^{-19})\sqrt{\frac{25}{5 \times 10^{-7}}} \cdot \frac{(1.5 \times 10^{10})^2}{N_a}\left[\exp\left(\frac{0.65}{0.0259}\right) - 1\right]$$

which yields

$$N_a = 1.01 \times 10^{15}\ \text{cm}^{-3}$$

The hole diffusion current density is given by Equation (8.22) as

$$J_p = \frac{eD_p p_{n0}}{L_p}\left[\exp\left(\frac{eV_a}{kT}\right) - 1\right] = e\sqrt{\frac{D_p}{\tau_{p0}}} \cdot \frac{n_i^2}{N_d}\left[\exp\left(\frac{eV_a}{kT}\right) - 1\right]$$

Substituting the numbers, we have

$$5 = (1.6 \times 10^{-19})\sqrt{\frac{10}{5 \times 10^{-7}}} \cdot \frac{(1.5 \times 10^{10})^2}{N_d}\left[\exp\left(\frac{0.65}{0.0259}\right) - 1\right]$$

which yields

$$N_d = 2.55 \times 10^{15}\ \text{cm}^{-3}$$

■ Comment

The relative magnitude of the electron and hole current densities through a diode can be varied by changing the doping concentrations in the device.

■ EXERCISE PROBLEM

Ex 8.3 Using the parameters given in Ex 8.2 for the GaAs diode, determine the electron and hole current densities at the space charge edges, and determine the total current density in the diode for a forward-bias voltage of $V_a = 1.05$ V.

[Ans. $J_n(-x_p) = 1.20$ A/cm^2, $J_p(x_n) = 0.1325$ A/cm^2, $J_T = 1.33$ A/cm^2]

8.1.6 Summary of Physics

We have been considering the case of a forward-bias voltage being applied to a pn junction. The forward-bias voltage lowers the potential barrier so that electrons and holes are injected across the space charge region. The injected carriers become minority carriers which then diffuse from the junction and recombine with majority carriers.

We calculated the minority carrier diffusion current densities at the edge of the space charge region. We can reconsider Equations (8.14) and (8.15) and determine

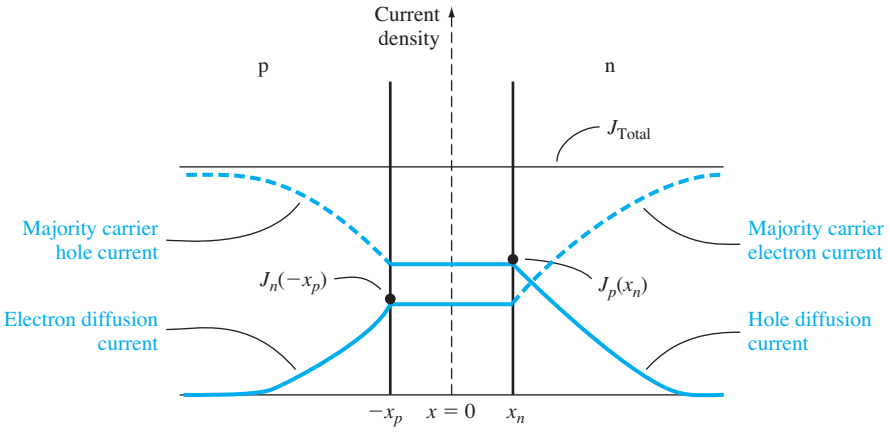


Figure 8.10 | Ideal electron and hole current components through a pn junction under forward bias.

the minority carrier diffusion current densities as a function of distance through the p and n regions. These results are

$$J_p(x) = \frac{eD_p p_{n0}}{L_p}\left[\exp\left(\frac{eV_a}{kT}\right) - 1\right]\exp\left(\frac{x_n - x}{L_p}\right) \qquad (x \geq x_n) \qquad (8.28)$$

and

$$J_n(x) = \frac{eD_n n_{p0}}{L_n}\left[\exp\left(\frac{eV_a}{kT}\right) - 1\right]\exp\left(\frac{x_p + x}{L_n}\right) \qquad (x \leq -x_p) \qquad (8.29)$$

The minority carrier diffusion current densities decay exponentially in each region. However, the total current through the pn junction is constant. The difference between total current and minority carrier diffusion current is a majority carrier current. Figure 8.10 shows the various current components through the pn structure. The drift of majority carrier holes in the p region far from the junction, for example, is to supply holes that are being injected across the space charge region into the n region and also to supply holes that are lost by recombination with excess minority carrier electrons. The same discussion applies to the drift of electrons in the n region.

We have seen that excess carriers are created in a forward-biased pn junction. From the results of the ambipolar transport theory derived in Chapter 6, the behavior of the excess carriers is determined by the minority carrier parameters for low injection. In determining the current–voltage relationship of the pn junction, we consider the flow of minority carriers since we know the behavior and characteristics of these particles. It may seem strange, at times, that we concern ourselves so much with minority carriers rather than with the vast number of majority carriers, but the reason for this can be found in the results derived from the ambipolar transport theory.

The fact that we now have drift current densities in the p and n regions implies that the electric field in these regions is not zero as we had originally assumed. We can calculate the electric field in the neutral regions and determine the validity of our zero-field approximation.

EXAMPLE 8.4

Objective: Calculate the electric field in a neutral region of a silicon diode to produce a given majority carrier drift current density.

Consider a silicon pn junction at $T = 300$ K with the parameters given in Example 8.2 and with an applied forward-bias voltage $V_a = 0.65$ V.

■ Solution

The total forward-bias current density is given by

$$J = J_s\left[\exp\left(\frac{eV}{kT}\right) - 1\right]$$

We determined the reverse-saturation current density in Example 8.2, so we can write

$$J = (4.155 \times 10^{-11})\left[\exp\left(\frac{0.65}{0.0259}\right) - 1\right] = 3.295 \text{ A/cm}^2$$

The total current far from the junction in the n region will be majority carrier electron drift current, so we can write

$$J = J_n \approx e\mu_n N_d \text{ E}$$

The doping concentration is $N_d = 10^{16}$ cm^{-3}, and, if we assume $\mu_n = 1350$ cm^2/V-s, then the electric field must be

$$\text{E} = \frac{J_n}{e\mu_n N_d} = \frac{3.295}{(1.6 \times 10^{-19})(1350)(10^{16})} = 1.525 \text{ V/cm}$$

■ Comment

We assumed, in the derivation of the current–voltage equation, that the electric field in the neutral p and n regions was zero. Although the electric field is not zero, this example shows that the magnitude is very small—thus the approximation of zero electric field is very good.

■ EXERCISE PROBLEM

Ex 8.4 Determine the electric field in the neutral n region and neutral p region for the GaAs pn junction diode described in Ex 8.3.

(Ans. $E_n = 0.0694$ V/cm, $E_p = 3.25$ V/cm)

8.1.7 Temperature Effects

The ideal reverse-saturation current density J_s, given by Equation (8.26), is a function of the thermal-equilibrium minority carrier concentrations n_{p0} and p_{n0}. These minority carrier concentrations are proportional to n_i^2, which is a very strong function of temperature. For a silicon pn junction, the ideal reverse-saturation current density will increase by approximately a factor of 4 for every 10°C increase in temperature.

The forward-bias current–voltage relation is given by Equation (8.27). This relation includes J_s as well as the exp (eV_a/kT) factor, making the forward-bias current–voltage relation a function of temperature also. As temperature increases, less forward-bias voltage is required to obtain the same diode current. If the voltage is held constant, the diode current will increase as temperature increases. The change in forward-bias current with temperature is less sensitive than the reverse-saturation current.

EXAMPLE 8.5

Objective: Determine the change in the forward-bias voltage on a pn junction with a change in temperature to maintain a constant diode current.

Consider a silicon pn junction initially biased at 0.60 V at $T = 300$ K. Assume the temperature increases to $T = 310$ K. Calculate the change in the forward-bias voltage required to maintain a constant current through the junction.

■ Solution

The forward-bias current can be written as follows:

$$J \propto \exp\left(\frac{-E_g}{kT}\right) \exp\left(\frac{eV_a}{kT}\right)$$

If the temperature changes, we may take the ratio of the diode currents at the two temperatures. This ratio is

$$\frac{J_2}{J_1} = \frac{\exp(-E_g/kT_2)\exp(eV_{a2}/kT_2)}{\exp(-E_g/kT_1)\exp(eV_{a1}/kT_1)}$$

If current is to be held constant, then $J_1 = J_2$, and we must have

$$\frac{E_g - eV_{a2}}{kT_2} = \frac{E_g - eV_{a1}}{kT_1}$$

For $T_1 = 300$ K, $T_2 = 310$ K, $E_g = 1.12$ eV, and $V_{a1} = 0.60$ V. We then find

$$\frac{1.12 - V_{a2}}{310} = \frac{1.12 - 0.60}{300}$$

which yields

$$V_{a2} = 0.5827 \text{ V}$$

■ Comment

The change in the forward-bias voltage is −17.3 mV for a 10°C temperature change.

■ EXERCISE PROBLEM

Ex 8.5 Repeat Example 8.5 for a GaAs pn junction diode biased at $V_a = 1.050$ V for $T = 300$ K.

(Ans. −12.3 mV)

8.1.8 The "Short" Diode

We assumed in the previous analysis that both p and n regions were long compared with the minority carrier diffusion lengths. In many pn junction structures, one region may, in fact, be short compared with the minority carrier diffusion length. Figure 8.11 shows one such example: the length W_n is assumed to be much smaller than the minority carrier hole diffusion length, L_p.

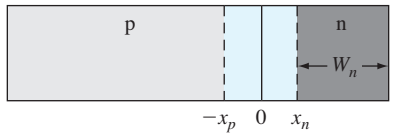


Figure 8.11 | Geometry of a "short" diode.

The steady-state excess minority carrier hole concentration in the n region is determined from Equation (8.9), which is given as

$$\frac{d^2(\delta p_n)}{dx^2} - \frac{\delta p_n}{L_p^2} = 0$$

The original boundary condition at $x = x_n$ still applies, given by Equation (8.11a) as

$$p_n(x_n) = p_{n0} \exp\left(\frac{eV_a}{kT}\right)$$

A second boundary condition needs to be determined. In many cases we assume that an ohmic contact exists at $x = (x_n + W_n)$, implying an infinite surface-recombination velocity and therefore an excess minority carrier concentration of zero. The second boundary condition is then written as

$$p_n(x = x_n + W_n) = p_{n0} \tag{8.30}$$

The general solution to Equation (8.9) is again given by Equation (8.12), which was

$$\delta p_n(x) = p_n(x) - p_{n0} = Ae^{x/L_p} + Be^{-x/L_p} \qquad (x \geq x_n)$$

In this case, because of the finite length of the n region, both terms of the general solution must be retained. Applying the boundary conditions of Equations (8.11b) and (8.30), the excess minority carrier concentration is given by

$$\delta p_n(x) = p_{n0}\left[\exp\left(\frac{eV_a}{kT}\right) - 1\right]\frac{\sinh\left[(x_n + W_n - x)/L_p\right]}{\sinh[W_n/L_p]} \tag{8.31}$$

Equation (8.31) is the general solution for the excess minority carrier hole concentration in the n region of a forward-biased pn junction. If $W_n \gg L_p$, the assumption for the long diode, Equation (8.31) reduces to the previous result given by Equation (8.14). If $W_n \ll L_p$, we can approximate the hyperbolic sine terms by

$$\sinh\left(\frac{x_n + W_n - x}{L_p}\right) \approx \left(\frac{x_n + W_n - x}{L_p}\right) \tag{8.32a}$$

and

$$\sinh\left(\frac{W_n}{L_p}\right) \approx \left(\frac{W_n}{L_p}\right) \tag{8.32b}$$

Then Equation (8.31) becomes

$$\delta p_n(x) = p_{n0}\left[\exp\left(\frac{eV_a}{kT}\right) - 1\right]\left(\frac{x_n + W_n - x}{W_n}\right) \tag{8.33}$$

The minority carrier concentration becomes a linear function of distance.

The minority carrier hole diffusion current density is given by

$$J_p = -eD_p \frac{d[\delta p_n(x)]}{dx}$$

so that in the short n region, we have

$$J_p(x) = \frac{eD_p p_{n0}}{W_n}\left[\exp\left(\frac{eV_a}{kT}\right) - 1\right] \tag{8.34}$$

The minority carrier hole diffusion current density now contains the length W_n in the denominator, rather than the diffusion length L_p. The diffusion current density is larger for a short diode than for a long diode since $W_n \ll L_p$. In addition, since the minority carrier concentration is approximately a linear function of distance through the n region, the minority carrier diffusion current density is a constant. This constant current implies that there is no recombination of minority carriers in the short region.

TEST YOUR UNDERSTANDING

TYU 8.1 The doping concentrations in a GaAs pn junction diode at $T = 300$ K are $N_d = 5 \times 10^{15}$ cm^{-3} and $N_a = 5 \times 10^{16}$ cm^{-3}. The minority carrier concentration at either space charge edge is to be no larger than 10 percent of the respective majority carrier concentration. Calculate the maximum forward-bias voltage that can be applied to this junction and still meet the required specifications.

[Ans. $V_a(\max) = 1.067$ V]

TYU 8.2 A silicon pn junction at $T = 300$ K has the following parameters: $N_a = 5 \times 10^{16}$ cm^{-3}, $N_d = 1 \times 10^{16}$ cm^{-3}, $D_n = 25$ cm^2/s, $D_p = 10$ cm^2/s, $\tau_{n0} = 5 \times 10^{-7}$ s, and $\tau_{p0} = 1 \times 10^{-7}$ s. The cross-sectional area is $A = 10^{-3}$ cm^2 and the forward-bias voltage is $V_a = 0.625$ V. Calculate the (*a*) minority electron diffusion current at the space charge edge, (*b*) minority hole diffusion current at the space charge edge, and (*c*) total current in the pn junction diode.

[Ans. (*a*) 0.154 mA; (*b*) 1.09 mA; (*c*) 1.24 mA]

TYU 8.3 Consider the silicon pn junction diode described in TYU 8.2. The p region is long and the n region is short with $W_n = 2$ μm. (*a*) Calculate the electron and hole currents in the depletion region. (*b*) Why has the hole current increased compared to that found in TYU 8.2?

[Ans. (*a*) $I_n = 0.154$ mA, $I_p = 5.44$ mA; (*b*) The hole density gradient has increased.]

8.2 | GENERATION–RECOMBINATION CURRENTS AND HIGH-INJECTION LEVELS

In the derivation of the ideal current–voltage relationship, we assumed low injection and neglected any effects occurring within the space charge region. High-level injection and other current components generated within the space charge region cause

the *I–V* relationship to deviate from the ideal expression. The additional currents are generated from the recombination processes discussed in Chapter 6.

8.2.1 Generation–Recombination Currents

The recombination rate of excess electrons and holes, given by the Shockley–Read–Hall recombination theory, was written as

$$R = \frac{C_n C_p N_t (np - n_i^2)}{C_n(n + n') + C_p(p + p')} \tag{8.35}$$

The parameters n and p are, as usual, the concentrations of electrons and holes, respectively.

Reverse-Biased Generation Current For a pn junction under reverse bias, we have argued that the mobile electrons and holes have essentially been swept out of the space charge region. Accordingly, within the space charge region, $n \approx p \approx 0$. The recombination rate from Equation (8.35) becomes

$$R = \frac{-C_n C_p N_t n_i^2}{C_n n' + C_p p'} \tag{8.36}$$

The negative sign implies a negative recombination rate; hence, we are really generating electron–hole pairs within the reverse-biased space charge region. The recombination of excess electrons and holes is the process whereby we are trying to reestablish thermal equilibrium. Since the concentration of electrons and holes is essentially zero within the reverse-biased space charge region, electrons and holes are being generated via the trap level to also try to reestablish thermal equilibrium. This generation process is schematically shown in Figure 8.12. As the electrons and holes

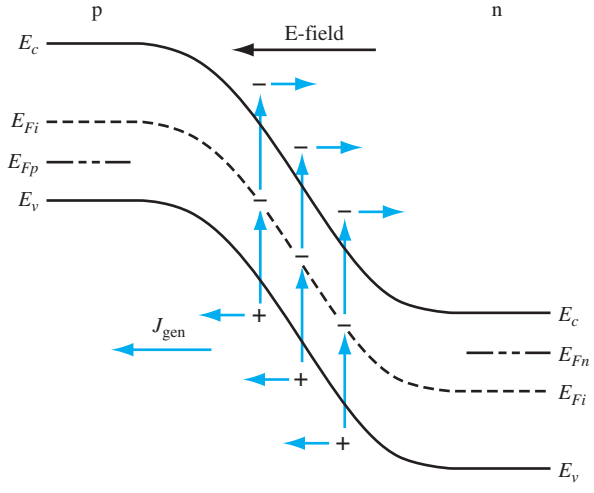


Figure 8.12 | Generation process in a reverse-biased pn junction.

are generated, they are swept out of the space charge region by the electric field. The flow of charge is in the direction of a reverse-biased current. This *reverse-biased generation current,* caused by the generation of electrons and holes in the space charge region, is in addition to the ideal reverse-biased saturation current.

We may calculate the density of the reverse-biased generation current by considering Equation (8.36). If we make a simplifying assumption and let the trap level be at the intrinsic Fermi level, then from Equations (6.92) and (6.97), we have that $n' = n_i$ and $p' = n_i$. Equation (8.36) now becomes

$$R = \frac{-n_i}{\frac{1}{N_t C_p} + \frac{1}{N_t C_n}} \tag{8.37}$$

Using the definitions of lifetimes from Equations (6.103) and (6.104), we may write Equation (8.37) as

$$R = \frac{-n_i}{\tau_{p0} + \tau_{n0}} \tag{8.38}$$

If we define a new lifetime as the average of τ_{p0} and τ_{n0}, or

$$\tau_0 = \frac{\tau_{p0} + \tau_{n0}}{2} \tag{8.39}$$

then the recombination rate can be written as

$$R = \frac{-n_i}{2\tau_0} \equiv -G \tag{8.40}$$

The negative recombination rate implies a generation rate, so G is the generation rate of electrons and holes in the space charge region.

The generation current density may be determined from

$$J_{\text{gen}} = \int_0^W e\, G dx \tag{8.41}$$

where the integral is over the space charge region. If we assume that the generation rate is constant throughout the space charge region, then we obtain

$$\boxed{J_{\text{gen}} = \frac{e n_i W}{2\tau_0}} \tag{8.42}$$

The total reverse-biased current density is the sum of the ideal reverse saturation current density and the generation current density, or

$$J_R = J_s + J_{\text{gen}} \tag{8.43}$$

The ideal reverse-saturation current density J_s is independent of the reverse-biased voltage. However, J_{gen} is a function of the depletion width W, which in turn is a function of the reverse-biased voltage. The actual reverse-biased current density, then, is no longer independent of the reverse-biased voltage.

EXAMPLE 8.6

Objective: Determine the relative magnitudes of the ideal reverse-saturation current density and the generation current density in a reverse-biased pn junction.

Consider a silicon pn junction at $T = 300$ K with parameters $D_n = 25\ \text{cm}^2/\text{s}$, $D_p = 10\ \text{cm}^2/\text{s}$, $N_a = N_d = 10^{16}\ \text{cm}^{-3}$, and $\tau_0 = \tau_{n0} = \tau_{p0} = 5 \times 10^{-7}$ s. Assume the diode is reverse biased at $V_R = 5$ V.

Solution

The ideal reverse-saturation current density was calculated in Example 8.2 and was found to be $J_s = 4.155 \times 10^{-11}\ \text{A/cm}^2$.

The built-in potential is found as

$$V_{bi} = V_t \ln\left(\frac{N_a N_d}{n_i^2}\right) = (0.0259)\ln\left[\frac{(10^{16})(10^{16})}{(1.5 \times 10^{10})^2}\right] = 0.695\ \text{V}$$

The depletion width is found to be

$$W = \left\{\frac{2\in_s(V_{bi} + V_R)}{e}\left(\frac{N_a + N_d}{N_a N_d}\right)\right\}^{1/2}$$

$$= \left\{\frac{2(11.7)(8.85 \times 10^{-14})(0.695 + 5)}{1.6 \times 10^{-19}}\left[\frac{10^{16} + 10^{16}}{(10^{16})(10^{16})}\right]\right\}^{1/2}$$

$$= 1.214 \times 10^{-4}\ \text{cm}$$

The generation current density is then found to be

$$J_{\text{gen}} = \frac{en_i W}{2\tau_0} = \frac{(1.6 \times 10^{-19})(1.5 \times 10^{10})(1.214 \times 10^{-4})}{2(5 \times 10^{-7})}$$

or

$$J_{\text{gen}} = 2.914 \times 10^{-7}\ \text{A/cm}^2$$

The ratio of the two currents is

$$\frac{J_{\text{gen}}}{J_s} = \frac{2.914 \times 10^{-7}}{4.155 \times 10^{-11}} \cong 7 \times 10^3$$

Comment

Comparing the solutions for the two current densities, it is obvious that, for the silicon pn junction diode at room temperature, the generation current density is approximately four orders of magnitude larger than the ideal saturation current density. The generation current is the dominant reverse-biased current in a silicon pn junction diode.

EXERCISE PROBLEM

Ex 8.6 Consider a GaAs pn junction diode at $T = 300$ K with parameters $N_d = 8 \times 10^{16}\ \text{cm}^{-3}$, $N_a = 2 \times 10^{15}\ \text{cm}^{-3}$, $D_n = 207\ \text{cm}^2/\text{s}$, $D_p = 9.80\ \text{cm}^2/\text{s}$, and $\tau_0 = \tau_{p0} = \tau_{n0} = 5 \times 10^{-8}$ s. (*a*) Calculate the ideal reverse-biased saturation current density. (*b*) Find the reverse-biased generation current density if the diode is reverse biased at $V_R = 5$ V. (*c*) Determine the ratio of J_{gen} to J_s.

[(*a*) $1.677 \times 10^{-17}\ \text{A/cm}^2$; (*b*) $9.166 \times 10^{-10}\ \text{A/cm}^2$; (*c*) 3.68×10^7]

Forward-Bias Recombination Current For the reverse-biased pn junction, electrons and holes are essentially completely swept out of the space charge region so that $n \approx p \approx 0$. Under forward bias, however, electrons and holes are injected across the space charge region, so we do, in fact, have some excess carriers in the space charge region. The possibility exists that some of these electrons and holes will recombine within the space charge region and not become part of the minority carrier distribution.

The recombination rate of electrons and holes is again given from Equation (8.35) as

$$R = \frac{C_n C_p N_t (np - n_i^2)}{C_n(n + n') + C_p(p + p')}$$

Dividing both numerator and denominator by $C_n C_p N_t$ and using the definitions of τ_{n0} and τ_{p0}, we may write the recombination rate as

$$R = \frac{np - n_i^2}{\tau_{p0}(n + n') + \tau_{n0}(p + p')} \tag{8.44}$$

Figure 8.13 shows the energy-band diagram of the forward-biased pn junction. Shown in the figure are the intrinsic Fermi level and the quasi-Fermi levels for electrons and holes. From the results of Chapter 6, we may write the electron concentration as

$$n = n_i \exp\left[\frac{E_{Fn} - E_{Fi}}{kT}\right] \tag{8.45}$$

and the hole concentration as

$$p = n_i \exp\left[\frac{E_{Fi} - E_{Fp}}{kT}\right] \tag{8.46}$$

where E_{Fn} and E_{Fp} are the quasi-Fermi levels for electrons and holes, respectively.

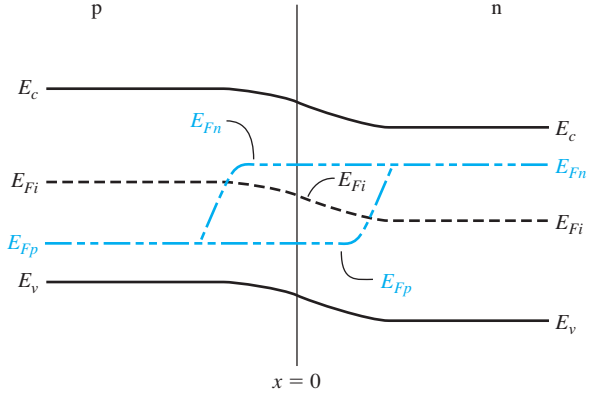


Figure 8.13 | Energy-band diagram of a forward-biased pn junction including quasi-Fermi levels.

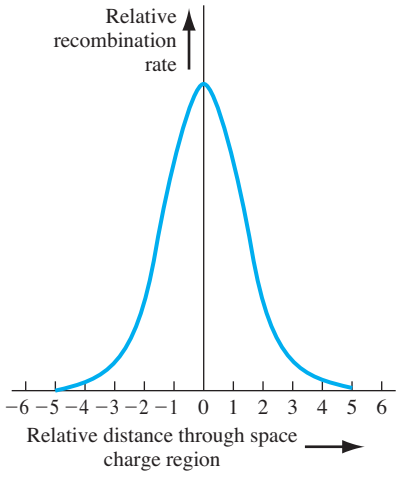


Figure 8.14 | Relative magnitude of the recombination rate through the space charge region of a forward-biased pn junction.

From Figure 8.13, we may note that

$$(E_{Fn} - E_{Fi}) + (E_{Fi} - E_{Fp}) = eV_a \tag{8.47}$$

where V_a is the applied forward-bias voltage. Again, if we assume that the trap level is at the intrinsic Fermi level, then $n' = p' = n_i$. Figure 8.14 shows a plot of the relative magnitude of the recombination rate as a function of distance through the space charge region. This plot was generated using Equations (8.44), (8.45), (8.46), and (8.47). A very sharp peak occurs at the metallurgical junction ($x = 0$).

At the center of the space charge region, we have

$$E_{Fn} - E_{Fi} = E_{Fi} - E_{Fp} = \frac{eV_a}{2} \tag{8.48}$$

Equations (8.45) and (8.46) then become

$$n = n_i \exp\left(\frac{eV_a}{2kT}\right) \tag{8.49}$$

and

$$p = n_i \exp\left(\frac{eV_a}{2kT}\right) \tag{8.50}$$

If we assume that $n' = p' = n_i$ and that $\tau_{n0} = \tau_{p0} = \tau_0$, then Equation (8.44) becomes

$$R_{\max} = \frac{n_i}{2\tau_0} \frac{[\exp(eV_a/kT) - 1]}{[\exp(eV_a/2kT) + 1]} \tag{8.51}$$

which is the maximum recombination rate for electrons and holes that occurs at the center of the forward-biased pn junction. If we assume that $V_a \gg kT/e$, we may

neglect the (-1) term in the numerator and the $(+1)$ term in the denominator. Equation (8.51) then becomes

$$R_{\max} = \frac{n_i}{2\tau_0} \exp\left(\frac{eV_a}{2kT}\right) \tag{8.52}$$

The recombination current density may be calculated from

$$J_{\text{rec}} = \int_0^W eR\, dx \tag{8.53}$$

where again the integral is over the entire space charge region. In this case, however, the recombination rate is not a constant through the space charge region. We have calculated the maximum recombination rate at the center of the space charge region, so we may write

$$J_{\text{rec}} = ex' \frac{n_i}{2\tau_0} \exp\left(\frac{eV_a}{2kT}\right) \tag{8.54}$$

where x' is a length over which the maximum recombination rate is effective. However, since τ_0 may not be a well-defined or known parameter, it is customary to write

$$\boxed{J_{\text{rec}} = \frac{eWn_i}{2\tau_0} \exp\left(\frac{eV_a}{2kT}\right) = J_{r0} \exp\left(\frac{eV_a}{2kT}\right)} \tag{8.55}$$

where W is the space charge width.

Total Forward-Bias Current The total forward-bias current density in the pn junction is the sum of the recombination and the ideal diffusion current densities. Figure 8.15 shows a plot of the minority carrier hole concentration in the neutral

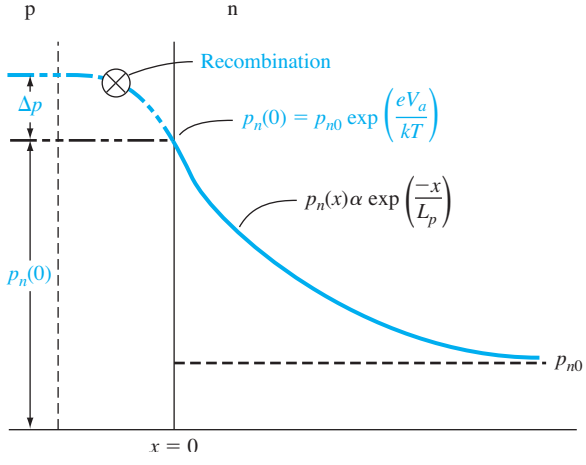


Figure 8.15 | Because of recombination, additional holes from the p region must be injected into the space charge region to establish the minority carrier hole concentration in the n region.

n region. This distribution yields the ideal hole diffusion current density and is a function of the minority carrier hole diffusion length and the applied junction voltage. The distribution is established as a result of holes being injected across the space charge region. If, now, some of the injected holes in the space charge region are lost due to recombination, then additional holes must be injected from the p region to make up for this loss. The flow of these additional injected carriers, per unit time, results in the recombination current. This added component is schematically shown in the figure.

The total forward-bias current density is the sum of the recombination and the ideal diffusion current densities, so we can write

$$J = J_{\text{rec}} + J_D \tag{8.56}$$

where J_{rec} is given by Equation (8.55) and J_D is given by

$$J_D = J_s \exp\left(\frac{eV_a}{kT}\right) \tag{8.57}$$

The (-1) term in Equation (8.27) has been neglected. The parameter J_s is the ideal reverse-saturation current density, and from previous discussion, the value of J_{r0} from the recombination current is larger than the value of J_s.

If we take the natural log of Equations (8.55) and (8.57), we obtain

$$\ln J_{\text{rec}} = \ln J_{r0} + \frac{eV_a}{2kT} = \ln J_{r0} + \frac{V_a}{2V_t} \tag{8.58a}$$

and

$$\ln J_D = \ln J_s + \frac{eV_a}{kT} = \ln J_s + \frac{V_a}{V_t} \tag{8.58b}$$

Figure 8.16 shows the recombination and diffusion current components plotted on a log current scale as a function of V_a/V_t. The slopes of the two curves are not the same. Also shown in the figure is the total current density—the sum of the two current components. We may notice that, at a low current density, the recombination current dominates, and at a higher current density, the ideal diffusion current dominates.

In general, the diode current–voltage relationship may be written as

$$\boxed{I = I_s\left[\exp\left(\frac{eV_a}{nkT}\right) - 1\right]} \tag{8.59}$$

where the parameter n is called the *ideality factor*. For a large forward-bias voltage, $n \approx 1$ when diffusion dominates, and for low forward-bias voltage, $n \approx 2$ when recombination dominates. There is a transition region where $1 < n < 2$.

8.2.2 High-Level Injection

In the derivation of the ideal diode *I–V* relationship, we assumed that low injection was valid. Low injection implies that the excess minority carrier concentrations are always much less than the majority carrier concentration.

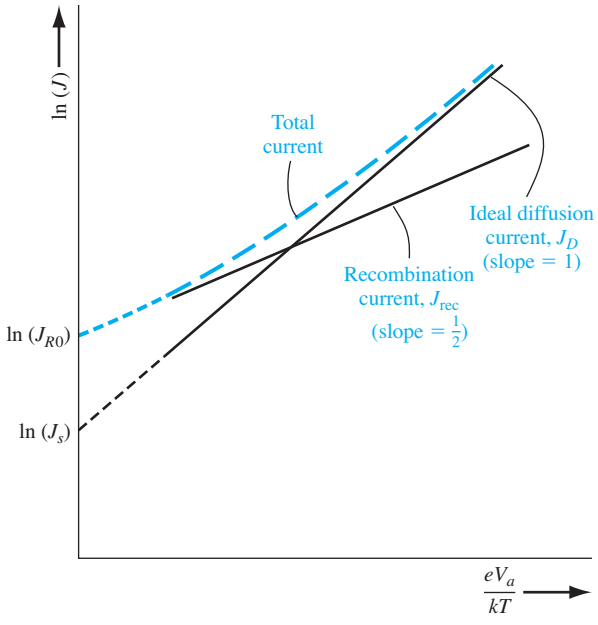


Figure 8.16 | Ideal diffusion, recombination, and total current in a forward-biased pn junction.

However, as the forward-bias voltage increases, the excess carrier concentrations increase and may become comparable or even greater than the majority carrier concentration. From Equation (8.18), we can write

$$np = n_i^2 \exp\left(\frac{V_a}{V_t}\right)$$

We have that $n = n_o + \delta n$ and $p = p_o + \delta p$, so that

$$(n_o + \delta n)(p_o + \delta p) = n_i^2 \exp\left(\frac{V_a}{V_t}\right) \tag{8.60}$$

Under high-level injection, we may have $\delta n > n_o$ and $\delta p > p_o$ so that Equation (8.60) becomes approximately

$$(\delta n)(\delta p) \cong n_i^2 \exp\left(\frac{V_a}{V_t}\right) \tag{861}$$

Since $\delta n = \delta p$, then

$$\delta n = \delta p \cong n_i \exp\left(\frac{V_a}{2V_t}\right) \tag{8.62}$$

The diode current is proportional to the excess carrier concentration so that, under high-level injection, we have

$$I \propto \exp\left(\frac{V_a}{2V_t}\right) \tag{8.63}$$

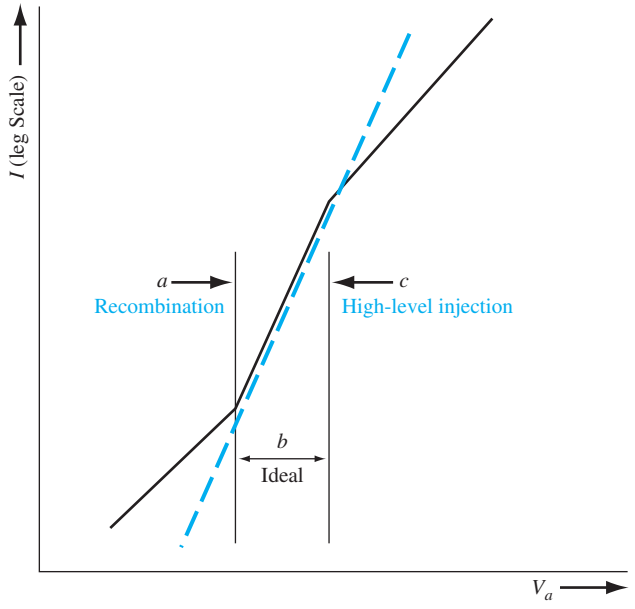


Figure 8.17 | Forward-bias current versus voltage from low forward bias to high forward bias.

In the high-level injection region, it takes a larger increase in diode voltage to produce a given increase in diode current.

The diode forward-bias current, from low-bias levels to high-bias levels, is plotted in Figure 8.17. This plot shows the effect of recombination at low-bias voltages and high-level injection at high-bias voltages.

TEST YOUR UNDERSTANDING

TYU 8.4 Consider a silicon pn junction diode at $T = 300$ K with parameters $N_a = 2 \times 10^{15}$ cm^{-3}, $N_d = 8 \times 10^{16}$ cm^{-3}, $D_p = 10$ cm^2/s, $D_n = 25$ cm^2/s, and $\tau_0 = \tau_{p0} = \tau_{n0} = 10^{-7}$ s. The diode is forward biased at $V_a = 0.35$ V. (*a*) Calculate the ideal diode current density. (*b*) Find the forward-biased recombination current density. (*c*) Determine the ratio of recombination current to the ideal diffusion current.

[Ans. (*a*) 2.137×10^{-4} A/cm^2; (*b*) 5.020×10^{-4} A/cm^2; (*c*) 2.35]

8.3 | SMALL-SIGNAL MODEL OF THE pn JUNCTION

We have been considering the dc characteristics of the pn junction diode. When semiconductor devices with pn junctions are used in linear amplifier circuits, for example, sinusoidal signals are superimposed on the dc currents and voltages, so that the small-signal characteristics of the pn junction become important.

8.3.1 Diffusion Resistance

The ideal current–voltage relationship of the pn junction diode was given by Equation (8.27), where J and J_s are current densities. If we multiply both sides of the equation by the junction cross-sectional area, we have

$$I_D = I_s\left[\exp\left(\frac{eV_a}{kT}\right) - 1\right] \tag{8.64}$$

where I_D is the diode current and I_s is the diode reverse-saturation current.

Assume that the diode is forward-biased with a dc voltage V_0 producing a dc diode current I_{DQ}. If we now superimpose a small, low-frequency sinusoidal voltage as shown in Figure 8.18, then a small sinusoidal current will be produced, superimposed on the dc current. The ratio of sinusoidal current to sinusoidal voltage is called the incremental conductance. In the limit of a very small sinusoidal current and voltage, the small-signal incremental conductance is just the slope of the dc current–voltage curve, or

$$g_d = \left.\frac{dI_D}{dV_a}\right|_{V_a=V_0} \tag{8.65}$$

The reciprocal of the incremental conductance is the incremental resistance, defined as

$$r_d = \left.\frac{dV_a}{dI_D}\right|_{I_D=I_{DQ}} \tag{8.66}$$

where I_{DQ} is the dc quiescent diode current.

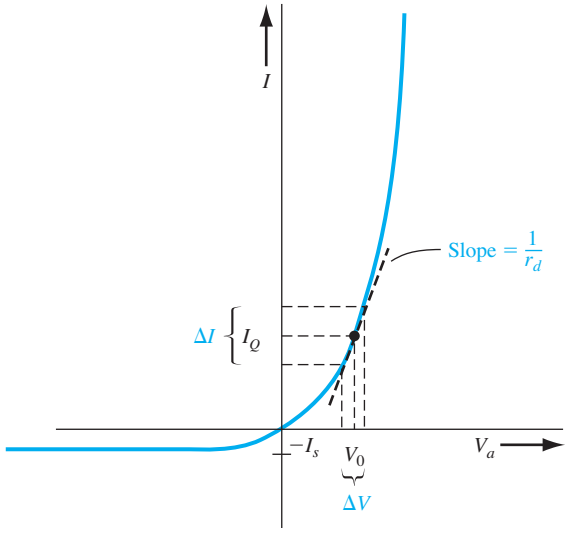


Figure 8.18 | Curve showing the concept of the small-signal diffusion resistance.

If we assume that the diode is biased sufficiently far in the forward-bias region, then the (-1) term can be neglected and the incremental conductance becomes

$$g_d = \left.\frac{dI_D}{dV_a}\right|_{V_a=V_0} = \left(\frac{e}{kT}\right) I_s \exp\left(\frac{eV_0}{kT}\right) \approx \frac{I_{DQ}}{V_t} \tag{8.67}$$

The small-signal incremental resistance is then the reciprocal function, or

$$\boxed{r_d = \frac{V_t}{I_{DQ}}} \tag{8.68}$$

The incremental resistance decreases as the bias current increases, and is inversely proportional to the slope of the *I–V* characteristic as shown in Figure 8.18. The incremental resistance is also known as the *diffusion resistance.*

8.3.2 Small-Signal Admittance

In the last chapter, we considered the pn junction capacitance as a function of the reverse-biased voltage. When the pn junction diode is forward-biased, another capacitance becomes a factor in the diode admittance. The small-signal admittance, or impedance, of the pn junction under forward bias is derived using the minority carrier diffusion current relations we have already considered.

Qualitative Analysis Before we delve into the mathematical analysis, we can qualitatively understand the physical processes that lead to a diffusion capacitance, which is one component of the junction admittance. Figure 8.19a schematically shows a pn junction forward biased with a dc voltage. A small ac voltage is also

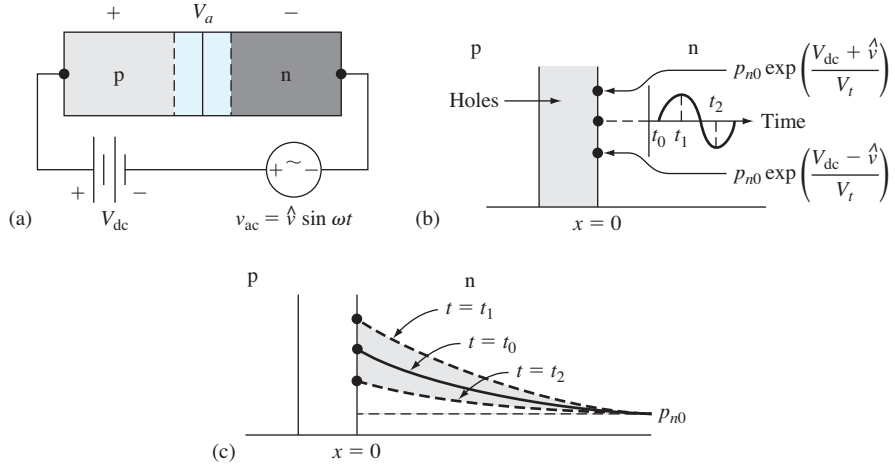


Figure 8.19 | (a) A pn junction with an ac voltage superimposed on a forward-biased dc value; (b) the hole concentration versus time at the space charge edge; (c) the hole concentration versus distance in the n region at three different times.

superimposed on the dc voltage so that the total forward-biased voltage can be written as $V_a = V_{dc} + \hat{v} \sin \omega t$.

As the voltage across the junction changes, the number of holes injected across the space charge region into the n region also changes. Figure 8.19b shows the hole concentration at the space charge edge as a function of time. At $t = t_0$, the ac voltage is zero so that the concentration of holes at $x = 0$ is just given by $p_n(0) = p_{n0} \exp(V_{dc}/V_t)$, which is what we have seen previously.

Now, as the ac voltage increases during its positive half cycle, the concentration of holes at $x = 0$ will increase and reach a peak value at $t = t_1$, which corresponds to the peak value of the ac voltage. When the ac voltage is on its negative half cycle, the total voltage across the junction decreases so that the concentration of holes at $x = 0$ decreases. The concentration reaches a minimum value at $t = t_2$, which corresponds to the time that the ac voltage reaches its maximum negative value. The minority carrier hole concentration at $x = 0$, then, has an ac component superimposed on the dc value as indicated in Figure 8.19b.

As previously discussed, the holes at the space charge edge ($x = 0$) diffuse into the n region where they recombine with the majority carrier electrons. We assume that the period of the ac voltage is large compared to the time it takes carriers to diffuse into the n region. The hole concentration as a function of distance into the n region can then be treated as a steady-state distribution. Figure 8.19c shows the steady-state hole concentrations at three different times. At $t = t_0$, the ac voltage is zero, so the $t = t_0$ curve corresponds to the hole distribution established by the dc voltage. The $t = t_1$ curve corresponds to the distribution established when the ac voltage has reached its peak positive value, and the $t = t_2$ curve corresponds to the distribution established when the ac voltage has reached its maximum negative value. The shaded areas represents the charge ΔQ that is alternately charged and discharged during the ac voltage cycle.

Exactly the same process is occurring in the p region with the electron concentration. The mechanism of charging and discharging of holes in the n region and electrons in the p region leads to a capacitance. This capacitance is called *diffusion capacitance*. The physical mechanism of this diffusion capacitance is different from that of the junction capacitance discussed in the last chapter. We show that the magnitude of the diffusion capacitance in a forward-biased pn junction is usually substantially larger than the junction capacitance.

Mathematical Analysis The minority carrier distribution in the pn junction will be derived for the case when a small sinusoidal voltage is superimposed on the dc junction voltage. We can then determine small signal, or ac, diffusion currents from these minority carrier functions. Figure 8.20 shows the minority carrier distribution in a pn junction when a forward-biased dc voltage is applied. The origin, $x = 0$, is set at the edge of the space charge region on the n side for convenience. The minority carrier hole concentration at $x = 0$ is given by Equation (8.7) as $p_n(0) = p_{n0} \exp(eV_a/kT)$, where V_a is the applied voltage across the junction.

Now let

$$V_a = V_0 + v_1(t) \tag{8.69}$$

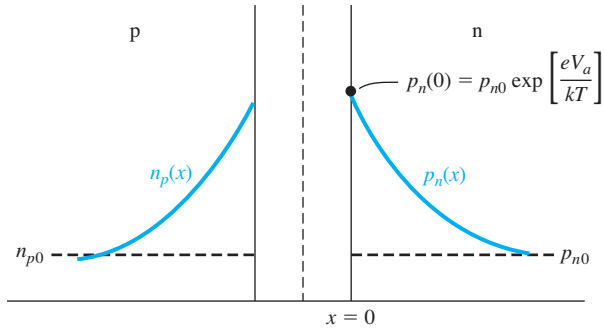


Figure 8.20 | The dc characteristics of a forward-biased pn junction used in the small-signal admittance calculations.

where V_0 is the dc quiescent bias voltage and $v_1(t)$ is the ac signal voltage that is superimposed on this dc level. We may now write

$$p_n(x = 0) = p_{n0} \exp\left\{\frac{e[V_0 + v_1(t)]}{kT}\right\} = p_n(0, t) \tag{8.70}$$

Equation (8.70) may be written as

$$p_n(0, t) = p_{dc} \exp\left[\frac{ev_1(t)}{kT}\right] \tag{8.71}$$

where

$$p_{dc} = p_{n0} \exp\left(\frac{eV_0}{kT}\right) \tag{8.72}$$

If we assume that $|v_1(t)| \ll (kT/e) = V_t$, then the exponential term in Equation (8.71) may be expanded into a Taylor series retaining only the linear terms, and the minority carrier hole concentration at $x = 0$ can be written as

$$p_n(0, t) \approx p_{dc}\left[1 + \frac{v_1(t)}{V_t}\right] \tag{8.73}$$

If we assume that the time-varying voltage $v_1(t)$ is a sinusoidal signal, we can write Equation (8.73) as

$$p_n(0, t) = p_{dc}\left(1 + \frac{\hat{V}_1}{V_t} e^{j\omega t}\right) \tag{8.74}$$

where $\hat{V}_1$ is the phasor of the applied sinusoidal voltage. Equation (8.74) will be used as the boundary condition in the solution of the time-dependent diffusion equation for the minority carrier holes in the n region.

In the neutral n region ($x > 0$), the electric field is assumed to be zero; thus, the behavior of the excess minority carrier holes is determined from the equation

$$D_p \frac{\partial^2(\delta p_n)}{\partial x^2} - \frac{\delta p_n}{\tau_{p0}} = \frac{\partial(\delta p_n)}{\partial t} \tag{8.75}$$

where δp_n is the excess hole concentration in the n region. We are assuming that the ac signal voltage $v_1(t)$ is sinusoidal. We then expect the steady-state solution for δp_n to be of the form of a sinusoidal solution superimposed on the dc solution, or

$$\delta p_n(x, t) = \delta p_0(x) + p_1(x)e^{j\omega t} \tag{8.76}$$

where $\delta p_0(x)$ is the dc excess carrier concentration and $p_1(x)$ is the magnitude of the ac component of the excess carrier concentration. The expression for $\delta p_0(x)$ is the same as that given in Equation (8.14).

Substituting Equation (8.76) into the differential Equation (8.75), we obtain

$$D_p\left\{\frac{\partial^2[\delta p_0(x)]}{\partial x^2} + \frac{\partial^2 p_1(x)}{\partial x^2}e^{j\omega t}\right\} - \frac{\delta p_0(x) + p_1(x)e^{j\omega t}}{\tau_{p0}} = j\omega p_1(x)e^{j\omega t} \tag{8.77}$$

We may rewrite this equation, combining the time-dependent and time-independent terms, as

$$\left\{\frac{D_p\,\partial^2[\delta p_0(x)]}{\partial x^2} - \frac{\delta p_0(x)}{\tau_{p0}}\right\} + \left[D_p\frac{\partial^2 p_1(x)}{\partial x^2} - \frac{p_1(x)}{\tau_{p0}} - j\omega p_1(x)\right]e^{j\omega t} = 0 \tag{8.78}$$

If the ac component, $p_1(x)$, is zero, then the first bracketed term is just the differential Equation (8.10), which is identically zero. Then we have, from the second bracketed term,

$$D_p\frac{d^2p_1(x)}{dx^2} - \frac{p_1(x)}{\tau_{p0}} - j\omega p_1(x) = 0 \tag{8.79}$$

Noting that $L_p^2 = D_p\tau_{p0}$, Equation (8.79) may be rewritten in the form

$$\frac{d^2p_1(x)}{dx^2} - \frac{(1 + j\omega\tau_{p0})}{L_p^2}p_1(x) = 0 \tag{8.80}$$

or

$$\frac{d^2p_1(x)}{dx^2} - C_p^2\,p_1(x) = 0 \tag{8.81}$$

where

$$C_p^2 = \frac{(1 + j\omega\tau_{p0})}{L_p^2} \tag{8.82}$$

The general solution to Equation (8.81) is

$$p_1(x) = K_1e^{-C_px} + K_2e^{+C_px} \tag{8.83}$$

One boundary condition is that $p_1(x \to +\infty) = 0$, which implies that the coefficient $K_2 = 0$. Then

$$p_1(x) = K_1e^{-C_px} \tag{8.84}$$

Applying the boundary condition at $x = 0$ from Equation (8.74), we obtain

$$p_1(0) = K_1 = p_{\text{dc}}\left(\frac{\hat{V}_1}{V_t}\right) \tag{8.85}$$

The hole diffusion current density can be calculated at $x = 0$. This is given by

$$J_p = -eD_p \left.\frac{\partial p_n}{\partial x}\right|_{x=0} \tag{8.86}$$

If we consider a homogeneous semiconductor, the derivative of the hole concentration will be just the derivative of the excess hole concentration. Then

$$J_p = -eD_p \left.\frac{\partial(\delta p_n)}{\partial x}\right|_{x=0} = -eD_p \left.\frac{\partial[\delta p_0(x)]}{\partial x}\right|_{x=0} - eD_p \left.\frac{\partial p_1(x)}{\partial x}\right|_{x=0} e^{j\omega t} \tag{8.87}$$

We can write this equation in the form

$$J_p = J_{p0} + j_p(t) \tag{8.88}$$

where

$$J_{p0} = -eD_p \left.\frac{\partial[\delta p_0(x)]}{\partial x}\right|_{x=0} = \frac{eD_p p_{n0}}{L_p}\left[\exp\left(\frac{eV_0}{kT}\right) - 1\right] \tag{8.89}$$

Equation (8.89) is the dc component of the hole diffusion current density and is exactly the same as in the ideal *I–V* relation derived previously.

The sinusoidal component of the diffusion current density is then found from

$$j_p(t) = \hat{J}_p e^{j\omega t} = -eD_p \left.\frac{\partial p_1(x)}{\partial x} e^{j\omega t}\right|_{x=0} \tag{8.90}$$

where $\hat{J}_p$ is the current density phasor. Combining Equations (8.90), (8.84), and (8.85), we have

$$\hat{J}_p = -eD_p(-C_p)\left[p_{\text{dc}}\left(\frac{\hat{V}_1}{V_t}\right)\right] e^{-C_p x}\Big|_{x=0} \tag{8.91}$$

We can write the total ac hole current phasor as

$$\hat{I}_p = A\hat{J}_p = eAD_p C_p p_{\text{dc}}\left(\frac{\hat{V}_1}{V_t}\right) \tag{8.92}$$

where A is the cross-sectional area of the pn junction. Substituting the expression for C_p, we obtain

$$\hat{I}_p = \frac{eAD_p p_{\text{dc}}}{L_p}\sqrt{1 + j\omega\tau_{p0}}\left(\frac{\hat{V}_1}{V_t}\right) \tag{8.93}$$

If we define

$$I_{p0} = \frac{eAD_p p_{dc}}{L_p} = \frac{eAD_p p_{n0}}{L_p}\exp\left(\frac{eV_0}{kT}\right) \tag{8.94}$$

then Equation (8.93) becomes

$$\hat{I}_p = I_{p0}\sqrt{1 + j\omega\tau_{p0}}\left(\frac{\hat{V}_1}{V_t}\right) \tag{8.95}$$

Going through the same type of analysis for the minority carrier electrons in the p region, we obtain

$$\hat{I}_n = I_{n0}\sqrt{1 + j\omega\tau_{n0}}\left(\frac{\hat{V}_1}{V_t}\right) \tag{8.96}$$

where

$$I_{n0} = \frac{eAD_n n_{p0}}{L_n} \exp\left(\frac{eV_0}{kT}\right) \tag{8.97}$$

The total ac current phasor is the sum of $\hat{I}_p$ and $\hat{I}_n$. The pn junction admittance is the total ac current phasor divided by the ac voltage phasor, or

$$Y = \frac{\hat{I}}{\hat{V}_1} = \frac{\hat{I}_p + \hat{I}_n}{\hat{V}_1} = \left(\frac{1}{V_t}\right)\left[I_{p0}\sqrt{1 + j\omega\tau_{p0}} + I_{n0}\sqrt{1 + j\omega\tau_{n0}}\right] \tag{8.98}$$

There is not a linear, lumped, finite, passive, bilateral network that can be synthesized to give this admittance function. However, we may make the following approximations. Assume that

$$\omega\tau_{p0} \ll 1 \tag{8.99a}$$

and

$$\omega\tau_{n0} \ll 1 \tag{8.99b}$$

These two assumptions imply that the frequency of the ac signal is not too large. Then we may write

$$\sqrt{1 + j\omega\tau_{p0}} \approx 1 + \frac{j\omega\tau_{p0}}{2} \tag{8.100a}$$

and

$$\sqrt{1 + j\omega\tau_{n0}} \approx 1 + \frac{j\omega\tau_{n0}}{2} \tag{8.100b}$$

Substituting Equations (8.100a) and (8.100b) into the admittance Equation (8.98), we obtain

$$Y = \left(\frac{1}{V_t}\right)\left[I_{p0}\left(1 + \frac{j\omega\tau_{p0}}{2}\right) + I_{n0}\left(1 + \frac{j\omega\tau_{n0}}{2}\right)\right] \tag{8.101}$$

If we combine the real and imaginary portions, we get

$$Y = \left(\frac{1}{V_t}\right)(I_{p0} + I_{n0}) + j\omega\left[\left(\frac{1}{2V_t}\right)(I_{p0}\tau_{p0} + I_{n0}\tau_{n0})\right] \tag{8.102}$$

Equation (8.102) may be written in the form

$$Y = g_d + j\omega C_d \tag{8.103}$$

The parameter g_d is called the *diffusion conductance* and is given by

$$g_d = \left(\frac{1}{V_t}\right)(I_{p0} + I_{n0}) = \frac{I_{DQ}}{V_t} \tag{8.104}$$

where I_{DQ} is the dc bias current. Equation (8.104) is exactly the same conductance as we obtained previously in Equation (8.67). The parameter C_d is called the *diffusion capacitance* and is given by

$$\boxed{C_d = \left(\frac{1}{2V_t}\right)(I_{p0}\tau_{p0} + I_{n0}\tau_{n0})} \tag{8.105}$$

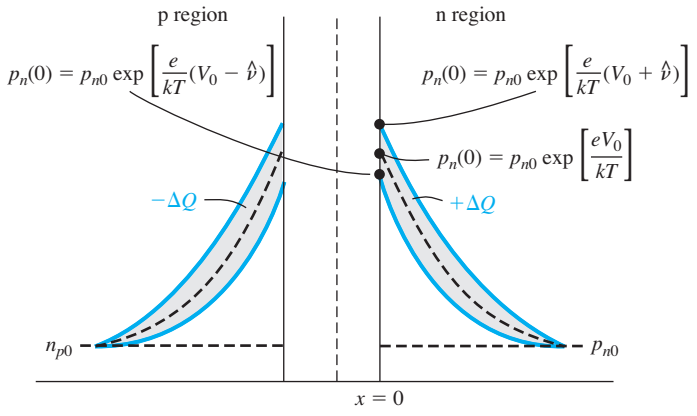


Figure 8.21 | Minority carrier concentration changes with changing forward-bias voltage.

The physics of the diffusion capacitance may be seen in Figure 8.21. The dc values of the minority carrier concentrations are shown along with the changes due to the ac component of voltage. The ΔQ charge is alternately being charged and discharged through the junction as the voltage across the junction changes. The change in the stored minority carrier charge as a function of the change in voltage is the diffusion capacitance. One consequence of the approximations $\omega\tau_{p0} \ll 1$ and $\omega\tau_{n0} \ll 1$ is that there are no "wiggles" in the minority carrier curves. The sinusoidal frequency is low enough so that the exponential curves are maintained at all times.

EXAMPLE 8.7

Objective: Calculate the small-signal admittance parameters of a pn junction diode.

This example is intended to give an indication of the magnitude of the diffusion capacitance as compared with the junction capacitance considered in the last chapter. The diffusion resistance will also be calculated. Assume that $N_a \gg N_d$ so that $p_{n0} \gg n_{p0}$. This assumption implies that $I_{p0} \gg I_{n0}$. Let $T = 300$ K, $\tau_{p0} = 10^{-7}$ s, and $I_{p0} = I_{DQ} = 1$ mA.

Solution

The diffusion capacitance, with these assumptions, is given by

$$C_d \approx \left(\frac{1}{2V_t}\right)(I_{p0}\tau_{p0}) = \frac{1}{(2)(0.0259)}(10^{-3})(10^{-7}) = 1.93 \times 10^{-9}\ \text{F}$$

The diffusion resistance is

$$r_d = \frac{V_t}{I_{DQ}} = \frac{0.0259\ \text{V}}{1\ \text{mA}} = 25.9\ \Omega$$

Comment

The value of 1.93 nF for the diffusion capacitance of a forward-biased pn junction is three to four orders of magnitude larger than the junction capacitance of the reverse-biased pn junction, which we calculated in Example 7.5.

EXERCISE PROBLEM

Ex 8.7 A silicon pn junction diode at $T = 300$ K has the following parameters: $N_d = 8 \times 10^{16}$ cm^{-3}, $N_a = 2 \times 10^{15}$ cm^{-3}, $D_n = 25$ cm^2/s, $D_p = 10$ cm^2/s, $\tau_{n0} = 5 \times 10^{-7}$s, and $\tau_{p0} = 10^{-7}$s. The cross-sectional area is $A = 10^{-3}$ cm^2. Determine the diffusion resistance and diffusion capacitance if the diode is forward biased at (*a*) $V_a = 0.550$ V and (*b*) $V_a = 0.610$ V.

[Ans. (*a*) $r_d = 118\ \Omega$, $C_d = 2.07$ nF; (*b*) $r_d = 11.6\ \Omega$, $C_d = 20.9$ nF]

The diffusion capacitance tends to dominate the capacitance terms in a forward-biased pn junction. The small-signal diffusion resistance can be fairly small if the diode current is a fairly large value. As the diode current decreases, the diffusion resistance increases. We will consider the impedance of forward-biased pn junctions again when we discuss bipolar transistors.

8.3.3 Equivalent Circuit

The small-signal equivalent circuit of the forward-biased pn junction is derived from Equation (8.103). This circuit is shown in Figure 8.22a. We need to add the junction capacitance, which will be in parallel with the diffusion resistance and diffusion capacitance. The last element we add, to complete the equivalent circuit, is a series resistance. The neutral n and p regions have finite resistances so the actual pn junction will include a series resistance. The complete equivalent circuit is given in Figure 8.22b.

The voltage across the actual junction is V_a and the total voltage applied to the pn diode is given by V_{app}. The junction voltage V_a is the voltage in the ideal current–voltage expression. We can write the expression

$$V_{app} = V_a + Ir_s \tag{8.106}$$

Figure 8.23 is a plot of the current–voltage characteristic from Equation (8.106) showing the effect of the series resistance. A larger applied voltage is required to

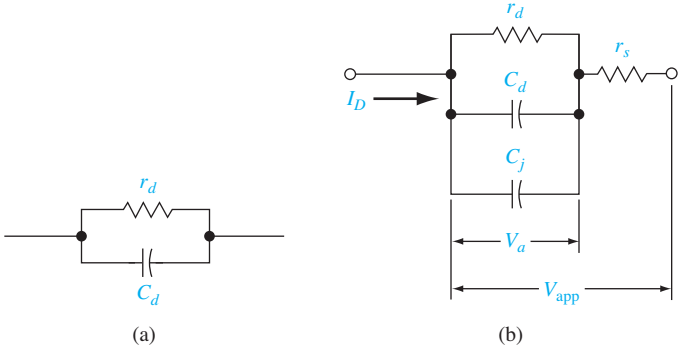


Figure 8.22 | (a) Small-signal equivalent circuit of ideal forward-biased pn junction diode; (b) complete small-signal equivalent circuit of pn junction.

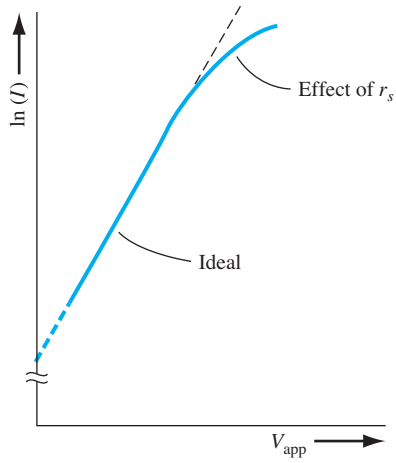


Figure 8.23 | Forward-biased *I–V* characteristics of a pn junction diode showing the effect of series resistance.

achieve the same current value when a series resistance is included. In most diodes, the series resistance will be negligible. In some semiconductor devices with pn junctions, however, the series resistance will be in a feedback loop; in these cases, the resistance is multiplied by a gain factor and becomes non-negligible.

TEST YOUR UNDERSTANDING

TYU 8.5 A GaAs pn junction diode at $T = 300$ K has the same parameters given in Ex 8.7 except that $D_n = 207$ cm^2/s and $D_p = 9.80$ cm^2/s. Determine the diffusion resistance and diffusion capacitance if the diode is forward biased at (*a*) $V_a = 0.970$ V and (*b*) $V_a = 1.045$ V.

[Ans. (*a*) $r_d = 263\ \Omega$, $C_d = 0.940$ nF; (*b*) $r_d = 14.6\ \Omega$, $C_d = 17.0$ nF]

TYU 8.6 A silicon pn junction diode at $T = 300$ K has the same parameters as those described in Ex 8.7. The neutral n-region and neutral p-region lengths are 0.01 cm. Estimate the series resistance of the diode (neglect ohmic contacts).

(Ans. $R = 66\ \Omega$)

*8.4 | CHARGE STORAGE AND DIODE TRANSIENTS

The pn junction is typically used as an electrical switch. In forward bias, referred to as the *on* state, a relatively large current can be produced by a small applied voltage; in reverse bias, referred to as the *off* state, only a very small current will exist. Of primary interest in circuit applications is the speed of the pn junction diode in switching states. We qualitatively discuss the transients that occur and the charge storage effects. We simply state the equations that describe the switching times without any mathematical derivations.

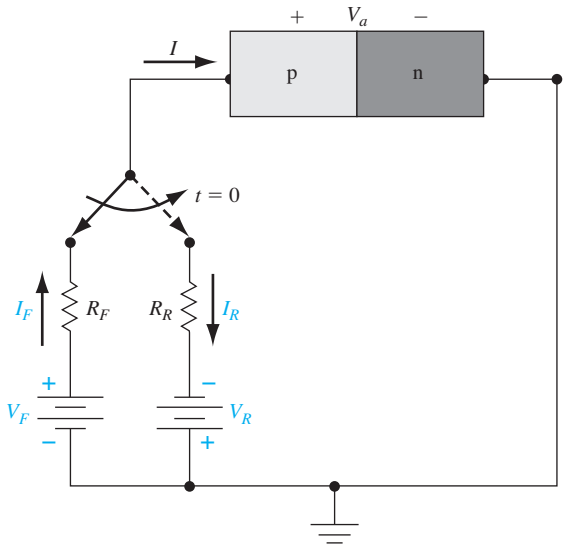


Figure 8.24 | Simple circuit for switching a diode from forward to reverse bias.

8.4.1 The Turn-off Transient

Suppose we want to switch a diode from the forward bias on state to the reverse-biased off state. Figure 8.24 shows a simple circuit that will switch the applied bias at $t = 0$. For $t < 0$, the forward-bias current is

$$I = I_F = \frac{V_F - V_a}{R_F} \tag{8.107}$$

The minority carrier concentrations in the device, for the applied forward voltage V_F, are shown in Figure 8.25a. There is excess minority carrier charge stored in both the p and n regions of the diode. The excess minority carrier concentrations at the space charge edges are supported by the forward-bias junction voltage V_a. When the voltage is switched from the forward- to the reverse-biased state, the excess minority carrier concentrations at the space charge edges can no longer be supported and they start to decrease, as shown in Figure 8.25b.

The collapse of the minority carrier concentrations at the edges of the space charge region leads to large concentration gradients and diffusion currents in the reverse-biased direction. If we assume, for the moment, that the voltage across the diode junction is small compared with V_R, then the reverse-biased current is limited to approximately

$$I = -I_R \approx \frac{-V_R}{R_R} \tag{8.108}$$

The junction capacitances do not allow the junction voltage to change instantaneously. If the current I_R were larger than this value, there would be a forward-bias voltage across the junction, which would violate our assumption of a reverse-biased current.

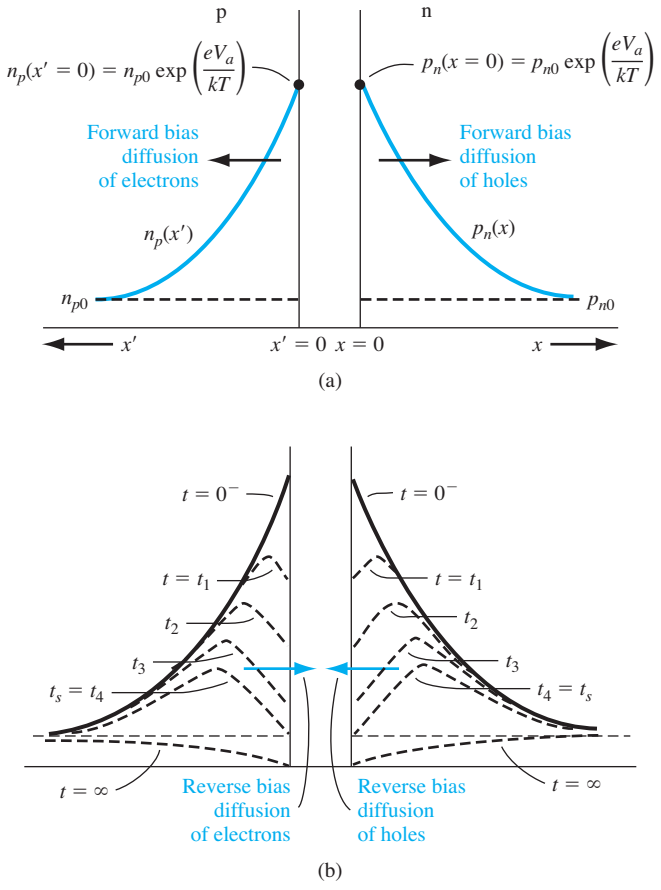


Figure 8.25 | (a) Steady-state forward-bias minority carrier concentrations; (b) minority carrier concentrations at various times during switching.

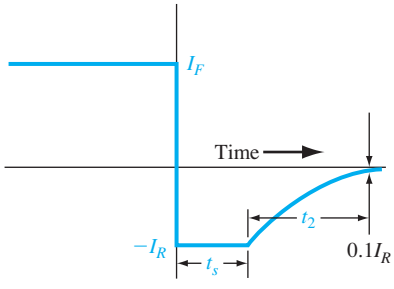


Figure 8.26 | Current characteristic versus time during diode switching.

If the current I_R were smaller than this value, there would be a reverse-biased voltage across the junction, which means that the junction voltage would have changed instantaneously. Since the reverse current is limited to the value given by Equation (8.108), the reverse-biased density gradient is constant; thus, the minority carrier concentrations at the space charge edge decrease with time as shown in Figure 8.25b.

This reverse current I_R will be approximately constant for $0^+ \leq t \leq t_s$, where t_s is called the *storage time.* The storage time is the length of time required for the minority carrier concentrations at the space charge edge to reach the thermal-equilibrium values. After this time, the voltage across the junction will begin to change. The current characteristic is shown in Figure 8.26. The reverse current is the flow of the stored minority carrier charge, which is the difference between the minority carrier concentrations at $t = 0^-$ and $t = \infty$, as shown in Figure 8.25b.

The storage time t_s can be determined by solving the time-dependent continuity equation. If we consider a one-sided p^+n junction, the storage time is determined from the equation

$$\text{erf}\sqrt{\frac{t_s}{\tau_{p0}}} = \frac{I_F}{I_F + I_R} \tag{8.109}$$

where erf (x) is known as the error function. An approximate solution for the storage time can be obtained as

$$t_s \approx \tau_{p0} \ln\left(1 + \frac{I_F}{I_R}\right) \tag{8.110}$$

The recovery phase for $t > t_s$ is the time required for the junction to reach its steady-state reverse-biased condition. The remainder of the excess charge is being removed and the space charge width is increasing to the reverse-biased value. The decay time t_2 is determined from

$$\text{erf}\sqrt{\frac{t_2}{\tau_{p0}}} + \frac{\exp(-t_2/\tau_{p0})}{\sqrt{\pi t_2/\tau_{p0}}} = 1 + 0.1\left(\frac{I_R}{I_F}\right) \tag{8.111}$$

The total turn-off time is the sum of t_s and t_2.

To switch the diode quickly, we need to be able to produce a large reverse current as well as have a small minority carrier lifetime. In the design of diode circuits, then, the designer must provide a path for the transient reverse-biased current pulse in order to be able to switch the diode quickly. These same effects will be considered when we discuss the switching of bipolar transistors.

8.4.2 The Turn-on Transient

The turn-on transient occurs when the diode is switched from its "off" state into the forward-bias "on" state. The turn-on can be accomplished by applying a forward-bias current pulse. The first stage of turn-on occurs very quickly and is the length of time required to narrow the space charge width from the reverse-biased value to its thermal-equilibrium value when $V_a = 0$. During this time, ionized donors and acceptors are neutralized as the space charge width narrows.

The second stage of the turn-on process is the time required to establish the minority carrier distributions. During this time the voltage across the junction is increasing toward its steady-state value. A small turn-on time is achieved if the minority carrier lifetime is small and if the forward-bias current is small.

TEST YOUR UNDERSTANDING

TYU 8.7 A one-sided p^+n silicon diode, which has a forward-bias current of $I_F = 1.75$ mA, is switched to reverse bias with an effective reverse-biased voltage of $V_R = 2$ V and an effective series resistance of $R_R = 4\ \text{k}\Omega$. The minority carrier hole lifetime is 10^{-7} s. (*a*) Determine the storage time t_s. (*b*) Calculate the decay time t_2. (*c*) What is the turn-off time of the diode?

[Ans. (*a*) 0.746×10^{-7} s; (*b*) 1.25×10^{-7} s; (*c*) $\approx 2 \times 10^{-7}$ s]

*8.5 | THE TUNNEL DIODE

The *tunnel diode* is a pn junction in which both the n and p regions are degenerately doped. As we discuss the operation of this device, we will find a region that exhibits a negative differential resistance. The tunnel diode was used in oscillator circuits in the past, but other types of solid-state devices are now used as high-frequency oscillators; thus, the tunnel diode is really only of academic interest. Nevertheless, this device does demonstrate the phenomenon of tunneling we discussed in Chapter 2.

Recall the degenerately doped semiconductors we discussed in Chapter 4: the Fermi level is in the conduction band of a degenerately doped n-type material and in the valence band of a degenerately doped p-type material. Then, even at $T = 0$ K, electrons will exist in the conduction band of the n-type material, and holes (empty states) will exist in the p-type material.

Figure 8.27 shows the energy-band diagram of a pn junction in thermal equilibrium for the case when both the n and p regions are degenerately doped. The depletion region width decreases as the doping increases and may be on the order of approximately 100 Å for the case shown in Figure 8.27. The potential barrier at the junction can be approximated by a triangular potential barrier, as shown in Figure 8.28. This potential barrier is similar to the potential barrier used in Chapter 2 to illustrate the tunneling phenomenon. The barrier width is small and the electric field in the space charge region is quite large; thus, a finite probability exists that an electron may tunnel through the forbidden band from one side of the junction to the other.

We may qualitatively determine the current–voltage characteristics of the tunnel diode by considering the simplified energy-band diagrams in Figure 8.29.

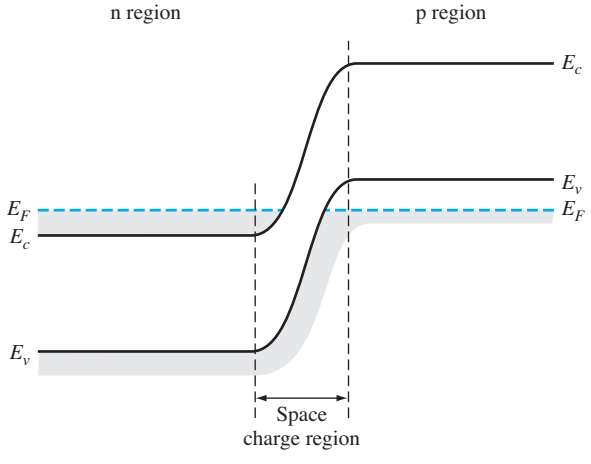


Figure 8.27 | Energy-band diagram of a pn junction in thermal equilibrium in which both the n and p regions are degenerately doped.

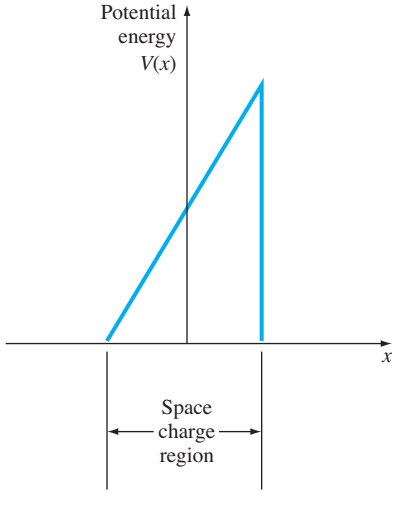


Figure 8.28 | Triangular potential barrier approximation of the potential barrier in the tunnel diode.

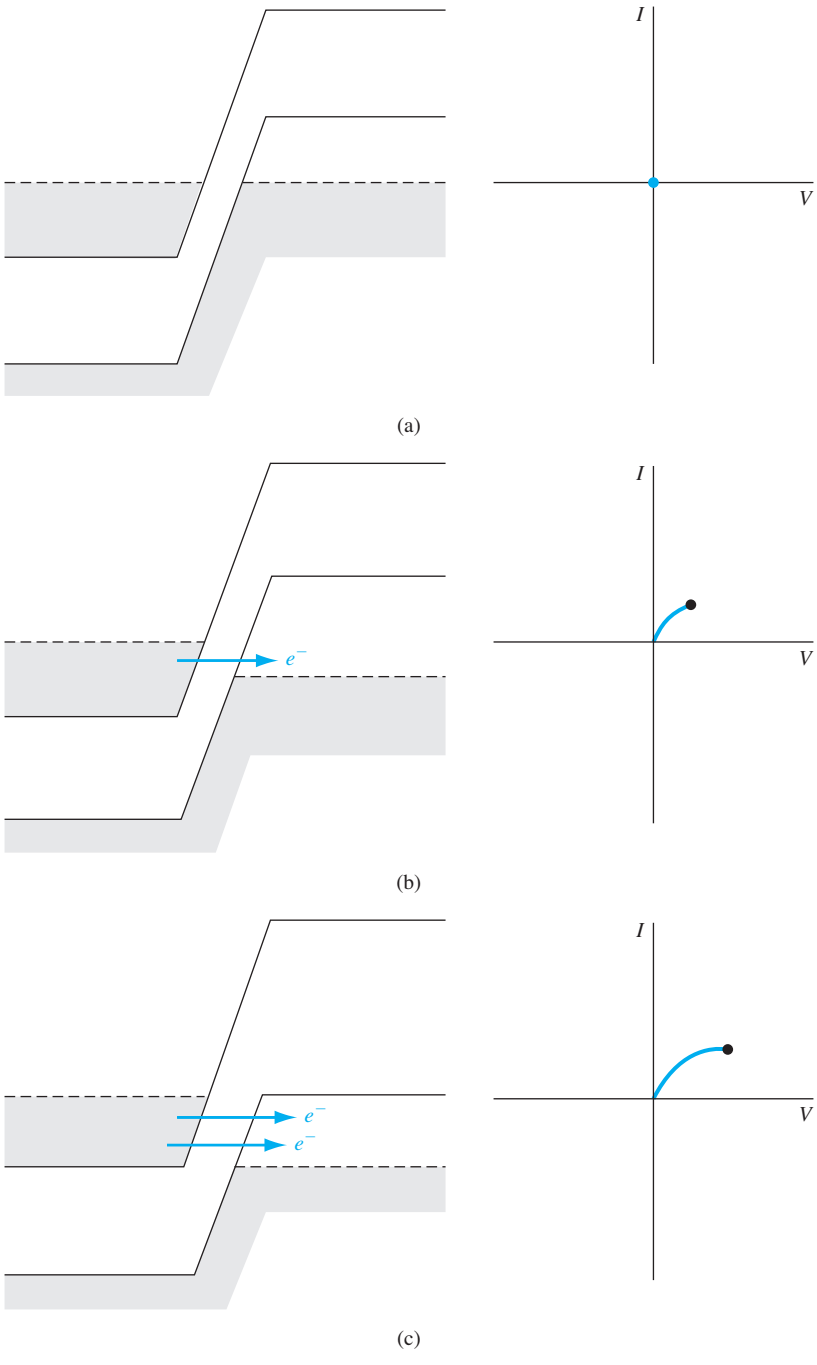


Figure 8.29 | Simplified energy-band diagrams and *I*–*V* characteristics of the tunnel diode at (a) zero bias; (b) a slight forward bias; (c) a forward bias producing maximum tunneling current. (*continued*)

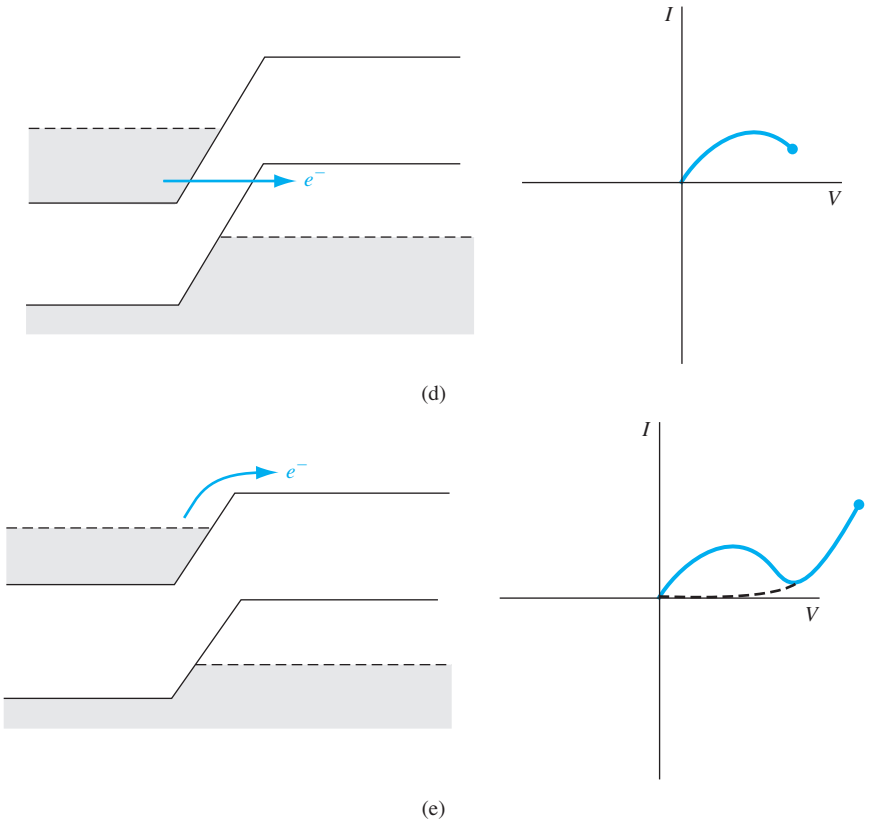


Figure 8.29 | *(concluded)* (d) A higher forward bias showing less tunneling current; (e) a forward bias for which the diffusion current dominates.

Figure 8.29a shows the energy-band diagram at zero bias, which produces zero current on the *I–V* diagram. If we assume, for simplicity, that we are near 0 K, then all energy states are filled below E_F on both sides of the junction.

Figure 8.29b shows the situation when a small forward-bias voltage is applied to the junction. Electrons in the conduction band of the n region are directly opposite to empty states in the valence band of the p region. There is a finite probability that some of these electrons will tunnel directly into the empty states, producing a forward-bias tunneling current as shown. With a slightly larger forward-bias voltage, as in Figure 8.29c, the maximum number of electrons in the n region will be opposite the maximum number of empty states in the p region; this will produce a maximum tunneling current.

As the forward-bias voltage continues to increase, the number of electrons on the n side directly opposite empty states on the p side decreases, as in Figure 8.29d, and the tunneling current will decrease. In Figure 8.29e, there are no electrons on the n side directly opposite to available empty states on the p side. For this forward-bias voltage, the tunneling current will be zero and the normal ideal diffusion current will exist in the device as shown in the *I–V* characteristics.

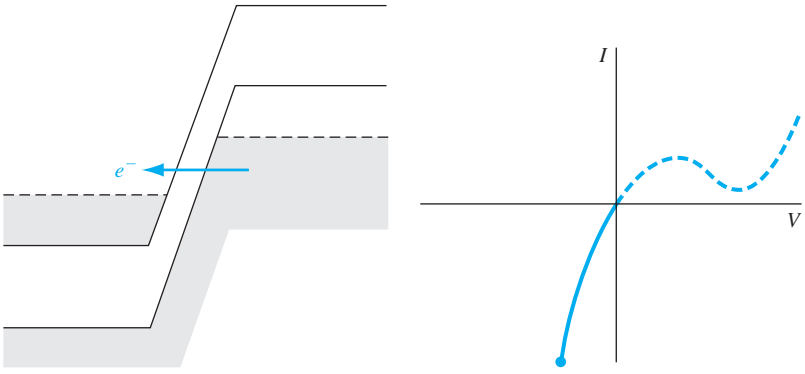


Figure 8.30 | (a) Simplified energy-band diagram of a tunnel diode with a reverse-biased voltage; (b) *I–V* characteristic of a tunnel diode with a reverse-biased voltage.

The portion of the curve showing a decrease in current with an increase in voltage is the region of differential negative resistance. The range of voltage and current for this region is quite small; thus, any power generated from an oscillator using this negative resistance property would also be fairly small.

A simplified energy-band diagram of the tunnel diode with an applied reverse-biased voltage is shown in Figure 8.30a. Electrons in the valence band on the p side are directly opposite empty states in the conduction band on the n side, so electrons can now tunnel directly from the p region into the n region, resulting in a large reverse-biased tunneling current. This tunneling current will exist for any reverse-biased voltage. The reverse-biased current will increase monotonically and rapidly with reverse-biased voltage as shown in Figure 8.30b.

8.6 | SUMMARY

- When a forward-bias voltage is applied across a pn junction (p region positive with respect to the n region), the potential barrier is lowered so that holes from the p region and electrons from the n region can flow across the junction.
- The boundary conditions relating the minority carrier hole concentration in the n region at the space charge edge and the minority carrier electron concentration in the p region at the space charge edge were derived.
- The holes that are injected into the n region and the electrons that are injected into the p region now become excess minority carriers. The behavior of the excess minority carrier is described by the ambipolar transport equation developed and described in Chapter 6. Solving the ambipolar transport equation and using the boundary conditions, the steady-state minority carrier hole and electron concentrations in the n region and p region, respectively, were derived.
- Gradients exist in the minority carrier hole and electron concentrations so that minority carrier diffusion currents exist in the pn junction. These diffusion currents yield the ideal current–voltage relationship of the pn junction diode.

- Excess carriers are generated in the space charge region of a reverse-biased pn junction. These carriers are swept out by the electric field and create the reverse-biased generation current that is another component of the reverse-biased diode current. Excess carriers recombine in the space charge region of a forward-biased pn junction. This recombination process creates the forward-bias recombination current that is another component of the forward-bias diode current.
- The small-signal equivalent circuit of the pn junction diode was developed. The two parameters of interest are the diffusion resistance and the diffusion capacitance.
- When a pn junction is switched from forward bias to reverse bias, the stored excess minority carrier charge must be removed from the junction. The time required to remove this charge is called the storage time and is a limiting factor in the switching speed of a diode.
- The *I–V* characteristics of a tunnel diode were developed showing a region of negative differential resistance.

GLOSSARY OF IMPORTANT TERMS

carrier injection The flow of carriers across the space charge region of a pn junction when a voltage is applied.

diffusion capacitance The capacitance of a forward-biased pn junction due to minority carrier storage effects.

diffusion conductance The ratio of a low-frequency, small-signal sinusoidal current to voltage in a forward-biased pn junction.

diffusion resistance The inverse of diffusion conductance.

forward bias The condition in which a positive voltage is applied to the p region with respect to the n region of a pn junction so that the potential barrier between the two regions is lowered below the thermal-equilibrium value.

generation current The reverse-biased pn junction current produced by the thermal generation of electron–hole pairs within the space charge region.

high-level injection The condition in which the excess carrier concentration becomes comparable to or greater than the majority carrier concentration.

"long" diode A pn junction diode in which both the neutral p and n regions are long compared with the respective minority carrier diffusion lengths.

recombination current The forward-bias pn junction current produced as a result of the flow of electrons and holes that recombine within the space charge region.

reverse saturation current The ideal reverse-biased current in a pn junction.

"short" diode A pn junction diode in which at least one of the neutral p or n regions is short compared to the respective minority carrier diffusion length.

storage time The time required for the excess minority carrier concentrations at the space charge edge to go from their steady-state values to zero when the diode is switched from forward to reverse bias.

CHECKPOINT

After studying this chapter, the reader should have the ability to:

- Describe the mechanism of charge flow across the space charge region of a pn junction when a forward-bias voltage is applied.

- State the boundary conditions for the minority carrier concentrations at the edge of the space charge region.
- Derive the expressions for the steady-state minority carrier concentrations in the pn junction.
- Derive the ideal current–voltage relationship for a pn junction diode.
- Describe the characteristics of a "short" diode.
- Describe generation and recombination currents in a pn junction.
- Define high-level injection and describe its effect on the diode *I–V* characteristics.
- Describe what is meant by diffusion resistance and diffusion capacitance.
- Describe the turn-off transient response in a pn junction.

REVIEW QUESTIONS

1. Sketch the energy bands in a zero-biased, reverse-biased, and forward-biased pn junction.
2. Write the boundary conditions for the excess minority carriers in a pn junction (a) under forward bias and (b) under reverse bias.
3. Sketch the steady-state minority carrier concentrations in a forward-biased pn junction.
4. Explain the procedure that is used in deriving the ideal current–voltage relationship in a pn junction diode.
5. Sketch the electron and hole currents through a forward-biased pn junction diode. Are currents near the junction primarily due to drift or diffusion? What about currents far from the junction?
6. What is the temperature dependence of the ideal reverse-saturation current?
7. What is meant by a "short" diode?
8. Explain the physical mechanism of the (a) generation current and (b) recombination current.
9. Sketch the forward-bias *I–V* characteristics of a pn junction diode showing the effects of recombination and high-level injection.
10. (a) Explain the physical mechanism of diffusion capacitance. (b) What is diffusion resistance?
11. If a forward-biased pn junction is switched off, explain what happens to the stored minority carriers. In which direction is the current immediately after the diode is switched off?

PROBLEMS

[**Note:** In the following problems, assume $T = 300$ K and the following parameters unless otherwise stated. For silicon pn junctions: $D_n = 25$ cm^2/s, $D_p = 10$ cm^2/s, $\tau_{n0} = 5 \times 10^{-7}$ s, $\tau_{p0} = 10^{-7}$ s. For GaAs pn junctions: $D_n = 205$ cm^2/s, $D_p = 9.8$ cm^2/s, $\tau_{n0} = 5 \times 10^{-8}$ s, $\tau_{p0} = 10^{-8}$ s.]

Section 8.1 pn Junction Current

8.1 (*a*) Consider an ideal pn junction diode at $T = 300$ K operating in the forward-bias region. Calculate the change in diode voltage that will cause a factor of 10 increase in current. (*b*) Repeat part (*a*) for a factor of 100 increase in current.

8.2 A silicon pn junction has impurity doping concentrations of $N_d = 2 \times 10^{15}$ cm^{-3} and $N_a = 8 \times 10^{15}$ cm^{-3}. Determine the minority carrier concentrations at the edges of the space charge region for (*a*) $V_a = 0.45$ V, (*b*) $V_a = 0.55$ V, and (*c*) $V_a = -0.55$ V.

8.3 The doping concentrations in a GaAs pn junction are $N_d = 10^{16}$ cm^{-3} and $N_a = 4 \times 10^{16}$ cm^{-3}. Find the minority carrier concentrations at the edges of the space charge region for (*a*) $V_a = 0.90$ V, (*b*) $V_a = 1.10$ V, and (*c*) $V_a = -0.95$ V.

8.4 (*a*) The doping concentrations in a silicon pn junction are $N_d = 5 \times 10^{15}$ cm^{-3} and $N_a = 5 \times 10^{16}$ cm^{-3}. The minority carrier concentration at either space charge edge is to be no larger than 10 percent of the respective majority carrier concentration. (*i*) Determine the maximum forward-bias voltage that can be applied to the junction and still meet the required specifications. (*ii*) Is the n-region or p-region concentration the factor that limits the forward-bias voltage? (*b*) Repeat part (*a*) if the doping concentrations are $N_d = 3 \times 10^{16}$ cm^{-3} and $N_a = 7 \times 10^{15}$ cm^{-3}.

8.5 Consider a GaAs pn junction with doping concentrations $N_a = 5 \times 10^{16}$ cm^{-3} and $N_d = 10^{16}$ cm^{-3}. The junction cross-sectional area is $A = 10^{-3}$ cm^2 and the applied forward-bias voltage is $V_a = 1.10$ V. Calculate the (*a*) minority electron diffusion current at the edge of the space charge region, (*b*) minority hole diffusion current at the edge of the space charge region, and (*c*) total current in the pn junction diode.

8.6 An n^+p silicon diode at $T = 300$ K has the following parameters: $N_d = 10^{18}$ cm^{-3}, $N_a = 10^{16}$ cm^{-3}, $D_n = 25$ cm^2/s, $D_p = 10$ cm^2/s, $\tau_{n0} = \tau_{p0} = 1$ μs, and $A = 10^{-4}$ cm^2. Determine the diode current for (*a*) a forward-bias voltage of 0.5 V and (*b*) a reverse-biased voltage of 0.5 V.

8.7 An ideal germanium pn junction diode has the following parameters: $N_a = 4 \times 10^{15}$ cm^{-3}, $N_d = 2 \times 10^{17}$ cm^{-3}, $D_p = 48$ cm^2/s, $D_n = 90$ cm^2/s, $\tau_{p0} = \tau_{n0} = 2 \times 10^{-6}$ s, and $A = 10^{-4}$ cm^2. Determine the diode current for (*a*) a forward-bias voltage of 0.25 V and (*b*) a reverse-biased voltage of 0.25 V.

8.8 A one-sided p^+n silicon diode has doping concentrations of $N_a = 5 \times 10^{17}$ cm^{-3} and $N_d = 8 \times 10^{15}$ cm^{-3}. The minority carrier lifetimes are $\tau_{n0} = 10^{-7}$ s and $\tau_{p0} = 8 \times 10^{-8}$ s. The cross-sectional area is $A = 2 \times 10^{-4}$ cm^2. Calculate the (*a*) reverse-biased saturation current, and (*b*) the forward-bias current at (*i*) $V_a = 0.45$ V, (*ii*) $V_a = 0.55$ V, and (*iii*) $V_a = 0.65$ V.

8.9 Calculate the applied reverse-biased voltage at which the ideal reverse current in a pn junction diode at $T = 300$ K reaches 90 percent of its reverse-saturation current value.

8.10 Fill in the missing data in the following table.

Case	V_a (V)	I (mA)	I_s (mA)	J_s (mA/cm^2)	A (cm^2)
1	0.65	0.50			2×10^{-4}
2	0.70		2×10^{-12}		1×10^{-3}
3		0.80		1×10^{-7}	1×10^{-4}
4	0.72	1.20		2×10^{-8}	

8.11 Consider an ideal silicon pn junction diode. (*a*) What must be the ratio of N_d/N_a so that 90 percent of the current in the depletion region is due to the flow of electrons? (*b*) Repeat part (*a*) if 80 percent of the current in the depletion region is due to the flow of holes.

8.12 A silicon pn junction diode is to be designed to operate at $T = 300$ K such that the diode current is $I = 10$ mA at a diode voltage of $V_D = 0.65$ V. The ratio of electron current to total current is to be 0.10 and the maximum current density is to be no more than 20 A/cm^2. Use the semiconductor parameters given in Example 8.2.

8.13 An ideal silicon pn junction at $T = 300$ K is under forward bias. The minority carrier lifetimes are $\tau_{n0} = 10^{-6}$ s and $\tau_{p0} = 10^{-7}$ s. The doping concentration in the n region is $N_d = 10^{16}$ cm^{-3}. Plot the ratio of hole current to the total current crossing the space charge region as the p-region doping concentration varies over the range $10^{15} \leq N_a \leq 10^{18}$ cm^{-3}. (Use a log scale for the doping concentrations.)

8.14 For a silicon pn junction at $T = 300$ K, assume $\tau_{p0} = 0.1\tau_{n0}$ and $\mu_n = 2.4\,\mu_p$. The ratio of electron current crossing the depletion region to the total current is defined as the electron injection efficiency. Determine the expression for the electron injection efficiency as a function of (*a*) N_d/N_a and (*b*) the ratio of n-type conductivity to p-type conductivity.

8.15 A silicon pn junction with a cross-sectional area of 10^{-4} cm^2 has the following properties at $T = 300$ K:

n region	p region
$N_d = 10^{17}$ cm^{-3}	$N_a = 5 \times 10^{15}$ cm^{-3}
$\tau_{p0} = 10^{-7}$ s	$\tau_{n0} = 10^{-6}$ s
$\mu_n = 850$ cm^2/V-s	$\mu_n = 1250$ cm^2/V-s
$\mu_p = 320$ cm^2/V-s	$\mu_p = 420$ cm^2/V-s

(*a*) Sketch the thermal equilibrium energy-band diagram of the pn junction, including the values of the Fermi level with respect to the intrinsic level on each side of the junction. (*b*) Calculate the reverse-saturation current I_s and determine the forward-bias current I at a forward-bias voltage of 0.5 V. (*c*) Determine the ratio of hole current to total current at the space charge edge x_n.

8.16 Consider an ideal silicon pn junction diode with the geometry shown in Figure P8.16. The doping concentrations are $N_a = 5 \times 10^{16}$ cm^{-3} and $N_d = 1.5 \times 10^{16}$ cm^{-3}, and the minority carrier lifetimes are $\tau_{n0} = 2 \times 10^{-7}$ s and $\tau_{p0} = 8 \times 10^{-8}$ s. The cross-sectional area is $A = 5 \times 10^{-4}$ cm^2. Calculate (*a*) the ideal reverse-saturation current due to holes, (*b*) the ideal reverse-saturation current due to electrons, (*c*) the hole concentration at $x = x_n$ for $V_a = 0.8V_{bi}$, (*d*) the electron current at $x = x_n$ for $V_a = 0.8\,V_{bi}$, and (*e*) the electron current at $x = x_n + (1/2)L_p$ for $V_a = 0.8\,V_{bi}$.

8.17 Consider the ideal long silicon pn junction shown in Figure P8.17. $T = 300$ K. The n region is doped with 10^{16} donor atoms per cm^3 and the p region is doped with 5×10^{16} acceptor atoms per cm^3. The minority carrier lifetimes are $\tau_{n0} = 0.05\ \mu$s and $\tau_{p0} = 0.01\ \mu$s. The minority carrier diffusion coefficients are $D_n = 23$ cm^2/s and $D_p = 8$ cm^2/s. The forward-bias voltage is $V_a = 0.610$ V. Calculate (*a*) the excess hole concentration as a function of x for $x \geq 0$, (*b*) the hole diffusion current density at $x = 3 \times 10^{-4}$ cm, and (*c*) the electron current density at $x = 3 \times 10^{-4}$ cm.

8.18 The limit of low injection is normally defined to be when the minority carrier concentration at the edge of the space charge region in the low-doped region becomes equal

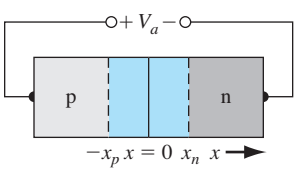


Figure P8.16 | Figure for Problem 8.16.

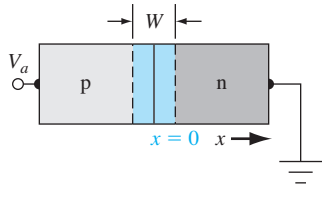


Figure P8.17 | Figure for Problem 8.17.

to one-tenth the majority carrier concentration in this region. Determine the value of the forward-bias voltage at which the limit of low injection is reached for the diode described in (*a*) Problem 8.7 and (*b*) Problem 8.8.

8.19 The cross-sectional area of a silicon pn junction is 10^{-3} cm^2. The temperature of the diode is $T = 300$ K, and the doping concentrations are $N_d = 10^{16}$ cm^{-3} and $N_a = 8 \times 10^{15}$ cm^{-3}. Assume minority carrier lifetimes of $\tau_{n0} = 10^{-6}$ s and $\tau_{p0} = 10^{-7}$ s. Calculate the total number of excess electrons in the p region and the total number of excess holes in the n region for (*a*) $V_a = 0.3$ V, (*b*) $V_a = 0.4$ V, and (*c*) $V_a = 0.5$ V.

8.20 Consider two ideal pn junctions at $T = 300$ K, having exactly the same electrical and physical parameters except for the bandgap energy of the semiconductor materials. The first pn junction has a bandgap energy of 0.525 eV and a forward-bias current of 10 mA with $V_a = 0.255$ V. For the second pn junction, "design" the bandgap energy so that a forward-bias voltage of $V_a = 0.32$ V will produce a current of 10 μA.

8.21 The reverse-biased saturation current is a function of temperature. (*a*) Assuming that I_s varies with temperature only from the intrinsic carrier concentration, show that we can write $I_s = CT^3 \exp(-E_g/kT)$ where C is a constant and a function only of the diode parameters. (*b*) Determine the increase in I_s as the temperature increases from $T = 300$ K to $T = 400$ K for a (*i*) germanium diode and (*ii*) silicon diode.

8.22 Assume that the mobilities, diffusion coefficients, and minority carrier lifetime parameters are independent of temperature (use the $T = 300$ K values). Assume that $\tau_{n0} = 10^{-6}$ s, $\tau_{p0} = 10^{-7}$ s, $N_d = 5 \times 10^{15}$ cm^{-3}, and $N_a = 5 \times 10^{16}$ cm^{-3}. Plot the ideal reverse saturation current density from $T = 200$ K to $T = 500$ K for (*a*) silicon, (*b*) germanium, and (*c*) gallium arsenide ideal pn junctions. (Use a log scale for the current density.)

8.23 An ideal silicon pn junction diode has a cross-sectional area of $A = 5 \times 10^{-4}$ cm^2. The doping concentrations are $N_a = 4 \times 10^{15}$ cm^{-3} and $N_d = 2 \times 10^{17}$ cm^{-3}. Assume that $E_g = 1.12$ eV as well as the diffusion coefficients and lifetimes are independent of temperature. The ratio of the magnitude of forward- to reverse-biased currents is to be no less than 2×10^4 with forward- and reverse-biased voltages of 0.50 V, and the maximum reverse-biased current is to be limited to 1.2 μA. Determine the maximum temperature at which the diode will meet these specifications and determine which specification is the limiting factor.

8.24 (*a*) A silicon pn junction diode has the geometry shown in Figure 8.11 in which the n region is "short" with a length $W_n = 0.7$ μm. The doping concentrations are $N_a = 2 \times 10^{17}$ cm^{-3} and $N_d = 2 \times 10^{15}$ cm^{-3}. The cross-sectional area is $A = 10^{-3}$ cm^2. Determine (*i*) the maximum forward-bias voltage such that low injection is still valid, and (*ii*) the resulting current at this forward-bias voltage. (*b*) Repeat part (*a*) if the doping concentrations are reversed such that $N_a = 2 \times 10^{15}$ cm^{-3} and $N_d = 2 \times 10^{17}$ cm^{-3}.

***8.25** A p$^+$n silicon diode is fabricated with a narrow n region as shown in Figure 8.11, in which $W_n < L_p$. Assume the boundary condition of $p_n = p_{n0}$ at $x = x_n + W_n$. (*a*) Derive the expression for the excess hole concentration $\delta p_n(x)$ as given by Equation (8.31). (*b*) Using the results of part (*a*), show that the current density in the diode is given by

$$J = \frac{eD_p p_{n0}}{L_p} \coth\left(\frac{W_n}{L_p}\right)\left[\exp\left(\frac{eV}{kT}\right) - 1\right]$$

*Asterisks next to problems indicate problems that are more difficult.

8.26 A silicon diode can be used to measure temperature by operating the diode at a fixed forward-bias current. The forward-bias voltage is then a function of temperature. At $T = 300$ K, the diode voltage is found to be 0.60 V. Determine the diode voltage at (*a*) $T = 310$ K and (*b*) $T = 320$ K.

8.27 A forward-biased silicon diode is to be used as a temperature sensor. The diode is forward biased with a constant current source and V_a is measured as a function of temperature. (*a*) Derive an expression for $V_a(T)$ assuming that D/L for electrons and holes, and E_g are independent of temperature. (*b*) If the diode is biased at $I_D = 0.1$ mA and if $I_s = 10^{-15}$ A at $T = 300$ K, plot V_a versus T for $20°C < T < 200°C$. (*c*) Repeat part (*b*) if $I_D = 1$ mA. (*d*) Determine any changes in the results of parts (*a*) through (*c*) if the change in bandgap energy with temperature is taken into account.

Section 8.2 Generation–Recombination Currents

8.28 Consider a silicon pn junction diode with an applied reverse-biased voltage of $V_R = 5$V. The doping concentrations are $N_a = N_d = 4 \times 10^{16}$ cm^{-3} and the cross-sectional area is $A = 10^{-4}$ cm^2. Assume minority carrier lifetimes of $\tau_0 = \tau_{n0} = \tau_{p0} = 10^{-7}$ s. Calculate the (*a*) ideal reverse-saturation current, (*b*) reverse-biased generation current, and (*c*) the ratio of the generation current to ideal saturation current.

8.29 Consider the diode described in Problem 8.28. Assume that all parameters except n_i are independent of temperature. (*a*) Determine the temperature at which I_s and I_{gen} will be equal. What are the values of I_s and I_{gen} at this temperature? (*b*) Calculate the forward-bias voltage at which the ideal diffusion current is equal to the recombination current at $T = 300$ K.

8.30 Consider a GaAs pn junction diode with a cross-sectional area of $A = 2 \times 10^{-4}$ cm^2 and doping concentrations of $N_a = N_d = 7 \times 10^{16}$ cm^{-3}. The electron and hole mobility values are $\mu_n = 5500$ cm^2/V-s and $\mu_p = 220$ cm^2/V-s, respectively, and the lifetime values are $\tau_0 = \tau_{n0} = \tau_{p0} = 2 \times 10^{-8}$ s. (*a*) Calculate the ideal diode current at a (*i*) reverse-biased voltage of $V_R = 3$ V, (*ii*) forward-bias voltage of $V_a = 0.6$ V, (*iii*) forward-bias voltage of $V_a = 0.8$ V, and (*iv*) forward-bias voltage of $V_a = 1.0$ V. (*b*) (*i*) Calculate the generation current at $V_R = 3$ V. Assuming the recombination current extrapolated to $V_a = 0$ is $I_{ro} = 6 \times 10^{-14}$ A, determine the generation current at (*ii*) $V_a = 0.6$ V, (*iii*) $V_a = 0.8$ V, and (*iv*) $V_a = 1.0$ V.

8.31 Consider the pn junction diode described in Problem 8.30. Plot the diode recombination current and the ideal diode current (on a log scale) versus forward-bias voltage over the range $0.1 \leq V_a \leq 1.0$ V.

8.32 A silicon pn junction diode at $T = 300$ K has the following parameters: $N_a = N_d = 10^{16}$ cm^{-3}, $\tau_{p0} = \tau_{n0} = \tau_0 = 5 \times 10^{-7}$ s, $D_p = 10$ cm^2/s, $D_n = 25$ cm^2/s, and a cross-sectional area of 10^{-4} cm^2. Plot the diode recombination current and the ideal diode current (on a log scale) versus forward-bias voltage over the range $0.1 \leq V_a \leq 0.6$ V.

8.33 Consider a GaAs pn diode at $T = 300$ K with $N_a = N_d = 10^{17}$ cm^{-3} and with a cross-sectional area of 5×10^{-3} cm^2. The minority carrier mobilities are $\mu_n = 3500$ cm^2/V-s and $\mu_p = 220$ cm^2/V-s. The electron–hole lifetimes are $\tau_{n0} = \tau_{p0} = \tau_0 = 10^{-8}$ s. Plot the diode forward-bias current including recombination current between diode voltages of $0.1 \leq V_D \leq 1.0$ V. Compare this plot to that for an ideal diode.

***8.34** Starting with Equation (8.44) and using the suitable approximations, show that the maximum recombination rate in a forward-biased pn junction is given by Equation (8.52).

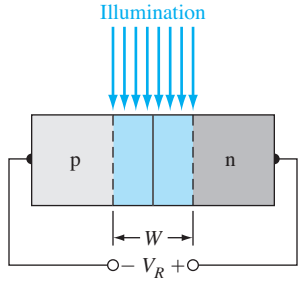


Figure P8.35 | Figure for Problem 8.35 and 8.36.

8.35 Consider, as shown in Figure P8.35, a uniformly doped silicon pn junction at $T = 300$ K with impurity doping concentrations of $N_a = N_d = 5 \times 10^{15}$ cm^{-3} and minority carrier lifetimes of $\tau_{n0} = \tau_{p0} = \tau_0 = 10^{-7}$ s. A reverse-biased voltage of $V_R = 10$ V is applied. A light source is incident only on the space charge region, producing an excess carrier generation rate of $g' = 4 \times 10^{19}$ cm^{-3} s^{-1}. Calculate the generation current density.

8.36 A long silicon pn junction diode has the following parameters: $N_d = 10^{18}$ cm^{-3}, $N_a = 3 \times 10^{16}$ cm^{-3}, $\tau_{n0} = \tau_{p0} = \tau_0 = 10^{-7}$ s, $D_n = 18$ cm^2/s, and $D_p = 6$ cm^2/s. A light source is incident on the space charge region such as shown in Figure P8.35, producing a generation current density of $J_G = 25$ mA/cm^2. The diode is open circuited. The generation current density forward biases the junction, inducing a forward-bias current in the opposite direction to the generation current. A steady-state condition is reached when the generation current density and forward-bias current density are equal in magnitude. What is the induced forward-bias voltage at this steady-state condition?

Section 8.3 Small-Signal Model of the pn Junction

8.37 (*a*) Calculate the small-signal diffusion capacitance and diffusion resistance of a silicon pn junction diode biased at $I_{DQ} = 1.2$ mA. Assume the minority carrier lifetimes are 0.5 μs in both the n and p regions. (*b*) Repeat part (*a*) for the case when the diode is biased at $I_{DQ} = 0.12$ mA.

8.38 Consider the diode described in Problem 8.37. A sinusoidal signal voltage with a peak value of 50 mV is superimposed on the dc forward-bias voltage. Determine the magnitude of charge that is being alternately charged and discharged in the n region.

8.39 Consider a p$^+$n silicon diode at $T = 300$ K. The diode is forward biased at a current of 1 mA. The hole lifetime in the n region is 10^{-7} s. Neglecting the depletion capacitance, calculate the diode impedance at frequencies of 10 kHz, 100 kHz, 1 MHz, and 10 MHz.

8.40 Consider a silicon pn junction with parameters as described in Problem 8.8. (*a*) Calculate and plot the depletion capacitance and diffusion capacitance over the voltage range $-10 \leq V_a \leq 0.75$ V. (*b*) Determine the voltage at which the two capacitances are equal.

8.41 Consider a p$^+$n silicon diode at $T = 300$ K. The slope of the diffusion capacitance versus forward-bias current is 2.5×10^{-6} F/A. Determine the hole lifetime and the diffusion capacitance at a forward-bias current of 1 mA.

8.42 A one-sided p^+n silicon diode has doping concentrations of $N_a = 4 \times 10^{17}$ cm^{-3} and $N_d = 8 \times 10^{15}$ cm^{-3}. The diode cross-sectional area is $A = 5 \times 10^{-4}$ cm^2. (*a*) The maximum diffusion capacitance is to be limited to 1 nF. Determine (*i*) the maximum current through the diode, (*ii*) the maximum forward-bias voltage, and (*iii*) the diffusion resistance. (*b*) Repeat part (*a*) if the maximum diffusion capacitance is limited to 0.25 nF.

8.43 A silicon pn junction diode at $T = 300$ K has a cross-sectional area of 10^{-2} cm^2. The length of the p region is 0.2 cm and the length of the n region is 0.1 cm. The doping concentrations are $N_d = 10^{15}$ cm^{-3} and $N_a = 10^{16}$ cm^{-3}. Determine (*a*) approximately the series resistance of the diode and (*b*) the current through the diode that will produce a 0.1 V drop across this series resistance.

8.44 We want to consider the effect of a series resistance on the forward-bias voltage required to achieve a particular diode current. (*a*) Assume the reverse-saturation current in a diode is $I_s = 10^{-10}$ A at $T = 300$ K. The resistivity of the n region is 0.2 Ω-cm and the resistivity of the p region is 0.1 Ω-cm. Assume the length of each neutral region is 10^{-2} cm and the cross-sectional area is 2×10^{-5} cm^2. Determine the required applied voltage to achieve a current of (*i*) 1 mA and (*ii*) 10 mA. (*b*) Repeat part (*a*) neglecting the series resistance.

8.45 (*a*) The reverse-saturation current in a diode is $I_s = 5 \times 10^{-12}$ A. The maximum small-signal diffusion resistance is to be $r_d = 32\ \Omega$. Determine the minimum forward-bias voltage that can be applied to meet this specification. (*b*) Repeat part (*a*) if the maximum small-signal diffusion resistance is to be $r_d = 60\ \Omega$.

8.46 (*a*) An ideal silicon pn junction diode at $T = 300$ K is forward biased at $V_a = +20$ mV. The reverse-saturation current is $I_s = 10^{-13}$ A. Calculate the small-signal diffusion resistance. (*b*) Repeat part (*a*) for an applied reverse-biased voltage of $V_a = -20$ mV.

Section 8.4 Charge Storage and Diode Transients

8.47 (*a*) In switching a pn junction from forward to reverse bias, assume that the ratio of reverse current, I_R, to forward current, I_F, is 0.2. Determine the ratio of storage time to minority carrier lifetime, t_s/τ_{p0}. (*b*) Repeat part (*a*) if the ratio of I_R to I_F is 1.0.

8.48 A pn junction is switched from forward to reverse bias. The storage time is specified to be $\tau_s = 0.3\,\tau_{p0}$. (*a*) Determine the required ratio of I_R to I_F to meet this specification. (*b*) Determine t_2/τ_{p0}.

8.49 Consider a diode with a junction capacitance of 18 pF at zero bias and 4.2 pF at a reverse-biased voltage of $V_R = 10$ V. The minority carrier lifetimes are 10^{-7} s. The diode is switched from a forward bias with a current of 2 mA to a reverse-biased voltage of 10 V applied through a 10 kΩ resistor. Estimate the turn-off time.

Section 8.5 The Tunnel Diode

8.50 Consider a silicon pn junction at $T = 300$ K with doping concentration of $N_d = N_a = 5 \times 10^{19}$ cm^{-3}. Assuming the abrupt junction approximation is valid, determine the space charge width at a forward-bias voltage of $V_a = 0.40$ V.

8.51 Sketch the energy-band diagram of an abrupt pn junction under zero bias in which the p region is degenerately doped and $E_C = E_F$ in the n region. Sketch the forward- and reverse-biased current–voltage characteristics. This diode is sometimes called a *backward diode*. Why?

Summary and Review

8.52 (*a*) Explain physically why the diffusion capacitance is not important in a reverse-biased pn junction. (*b*) Consider a silicon, a germanium, and gallium arsenide pn junction. If the total current density is the same in each diode under forward bias, discuss the expected relative values of electron and hole current densities.

***8.53** A silicon p^+n junction diode is to be designed to have a breakdown voltage of at least 60 V and to have a forward-bias current of $I_D = 50$ mA while still operating under low injection. The minority carrier lifetimes are $\tau_0 = \tau_{n0} = \tau_{p0} = 2 \times 10^{-7}$ s. Determine the doping concentrations and the minimum cross-sectional area.

***8.54** The donor and acceptor concentrations on either side of a silicon step junction are equal. (*a*) Derive an expression for the breakdown voltage in terms of the critical electric field and doping concentration. (*b*) If the breakdown voltage is to be $V_B = 50$ V, specify the range of allowed doping concentrations.

READING LIST

1. Dimitrijev, S. *Principles of Semiconductor Devices.* New York: Oxford University Press, 2006.
2. Kano, K. *Semiconductor Devices*. Upper Saddle River, NJ: Prentice Hall, 1998.
3. Muller, R. S., and T. I. Kamins. *Device Electronics for Integrated Circuits,* 2nd ed. New York: John Wiley and Sons, 1986.
4. Neudeck, G. W. *The PN Junction Diode.* Vol. 2 of the *Modular Series on Solid State Devices.* 2nd ed. Reading, MA: Addison-Wesley, 1989.

*5. Ng, K. K. *Complete Guide to Semiconductor Devices.* New York: McGraw-Hill, 1995.

6. Pierret, R. F. *Semiconductor Device Fundamentals.* Reading, MA: Addison-Wesley Publishing Co., 1996.
7. Roulston, D. J. *An Introduction to the Physics of Semiconductor Devices.* New York: Oxford University Press, 1999.
8. Shur, M. *Introduction to Electronic Devices*. New York: John Wiley and Sons, 1996.

*9. Shur, M. *Physics of Semiconductor Devices.* Englewood Cliffs, NJ: Prentice Hall, 1990.

10. Streetman, B. G., and S. K. Banerjee. *Solid State Electronic Devices,* 6th ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2006.
11. Sze, S. M. and K. K. Ng. Physics of Semiconductor Devices, 3rd ed. Hoboken, NJ: John Wiley and Sons, 2007.
12. Sze, S. M. *Semiconductor Devices: Physics and Technology,* 2nd ed. New York: John Wiley and Sons, 2001.

*13. Wang, S. *Fundamentals of Semiconductor Theory and Device Physics.* Englewood Cliffs, NJ: Prentice Hall, 1989.

14. Yang, E. S. *Microelectronic Devices.* New York: McGraw-Hill, 1988.

*Indicates references that are at an advanced level compared to this text.

CHAPTER 9

Metal–Semiconductor and Semiconductor Heterojunctions

In the preceding two chapters, we have considered the pn junction and assumed that the semiconductor material was the same throughout the structure. This type of junction is referred to as a *homojunction.* We developed the electrostatics of the junction and derived the current–voltage relationship. In this chapter, we consider the metal–semiconductor junction and the semiconductor heterojunction, in which the material on each side of the junction is not the same. These junctions can also produce diodes.

Semiconductor devices, or integrated circuits, must make contact with the outside world. This contact is made through nonrectifying metal–semiconductor junctions, or *ohmic contacts.* An ohmic contact is a low-resistance junction providing current conduction in both directions. We examine in this chapter the conditions that yield metal–semiconductor ohmic contacts. ■

9.0 | PREVIEW

In this chapter, we will:

- Determine the energy-band diagram of a metal–semiconductor junction.
- Investigate the electrostatics of the rectifying metal–semiconductor junction, which is known as the Schottky barrier diode.
- Derive the ideal current–voltage relation of the Schottky barrier diode.
- Discuss differences in the current transport mechanism between the Schottky barrier diode and pn junction diode, and discuss differences in turn-on voltage and switching times.
- Discuss ohmic contacts, which are low-resistance, nonrectifying metal–semiconductor junctions.
- Investigate the characteristics of a semiconductor heterojunction.

9.1 | THE SCHOTTKY BARRIER DIODE

One of the first practical semiconductor devices used in the early 1900s was the metal–semiconductor diode. This diode, also called a *point contact diode,* was made by touching a metallic whisker to an exposed semiconductor surface. These metal–semiconductor diodes were not easily reproduced or mechanically reliable and were replaced by the pn junction in the 1950s. However, semiconductor and vacuum technology is now used to fabricate reproducible and reliable metal–semiconductor contacts. In this section, we consider the metal–semiconductor rectifying contact, or Schottky barrier diode. In most cases, the rectifying contacts are made on n-type semiconductors; for this reason, we concentrate on this type of diode.

9.1.1 Qualitative Characteristics

The ideal energy-band diagram for a particular metal and n-type semiconductor before making contact is shown in Figure 9.1a. The vacuum level is used as a reference level. The parameter ϕ_m is the metal work function (measured in volts), ϕ_s is

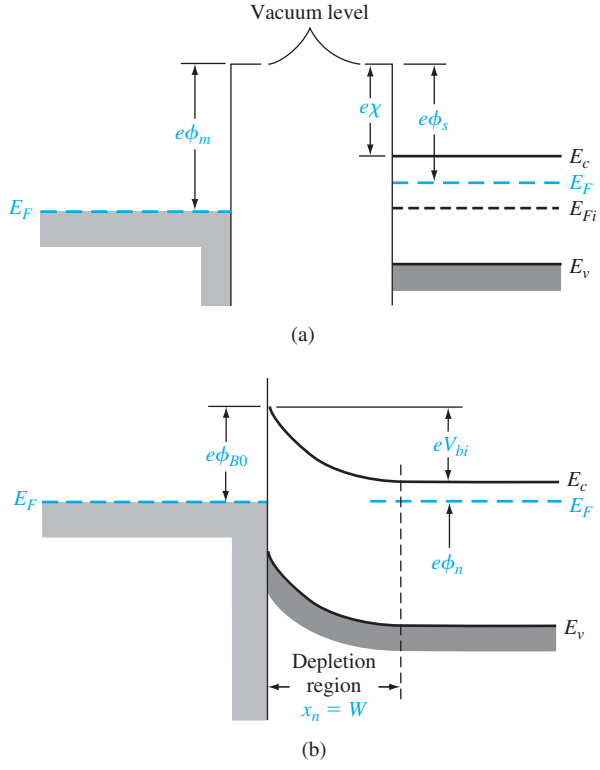


Figure 9.1 | (a) Energy-band diagram of a metal and semiconductor before contact; (b) ideal energy-band diagram of a metal–n-semiconductor junction for $\phi_m > \phi_s$.

Table 9.1 | Work functions of some elements

Element	Work function, ϕ_m
Ag, silver	4.26
Al, aluminum	4.28
Au, gold	5.1
Cr, chromium	4.5
Mo, molybdenum	4.6
Ni, nickel	5.15
Pd, palladium	5.12
Pt, platinum	5.65
Ti, titanium	4.33
W, tungsten	4.55

Table 9.2 | Electron affinity of some semiconductors

Element	Electron affinity, χ
Ge, germanium	4.13
Si, silicon	4.01
GaAs, gallium arsenide	4.07
AlAs, aluminum arsenide	3.5

the semiconductor work function, and χ is known as the *electron affinity*. The work functions of various metals are given in Table 9.1 and the electron affinities of several semiconductors are given in Table 9.2. In Figure 9.1a, we have assumed that $\phi_m > \phi_s$. The ideal thermal-equilibrium metal–semiconductor energy-band diagram, for this situation, is shown in Figure 9.1b. Before contact, the Fermi level in the semiconductor was above that in the metal. In order for the Fermi level to become a constant through the system in thermal equilibrium, electrons from the semiconductor flow into the lower energy states in the metal. Positively charged donor atoms remain in the semiconductor, creating a space charge region.

The parameter ϕ_{B0} is the ideal barrier height of the semiconductor contact, the potential barrier seen by electrons in the metal trying to move into the semiconductor. This barrier is known as the *Schottky barrier* and is given, ideally, by

$$\phi_{B0} = (\phi_m - \chi) \tag{9.1}$$

On the semiconductor side, V_{bi} is the built-in potential barrier. This barrier, similar to the case of the pn junction, is the barrier seen by electrons in the conduction band trying to move into the metal. The built-in potential barrier is given by

$$V_{bi} = \phi_{B0} - \phi_n \tag{9.2}$$

which makes V_{bi} a slight function of the semiconductor doping, as is the case in a pn junction.

If we apply a positive voltage to the semiconductor with respect to the metal, the semiconductor-to-metal barrier height increases, while ϕ_{B0} remains constant in this idealized case. This bias condition is the reverse bias. If a positive voltage is applied to the metal with respect to the semiconductor, the semiconductor-to-metal barrier V_{bi} is reduced while ϕ_{B0} again remains essentially constant. In this situation, electrons can more easily flow from the semiconductor into the metal since the barrier has been reduced. This bias condition is the forward bias. The energy-band diagrams for the reverse and forward bias are shown in Figures 9.2a,b where V_R is the magnitude of the reverse-biased voltage and V_a is the magnitude of the forward-bias voltage.

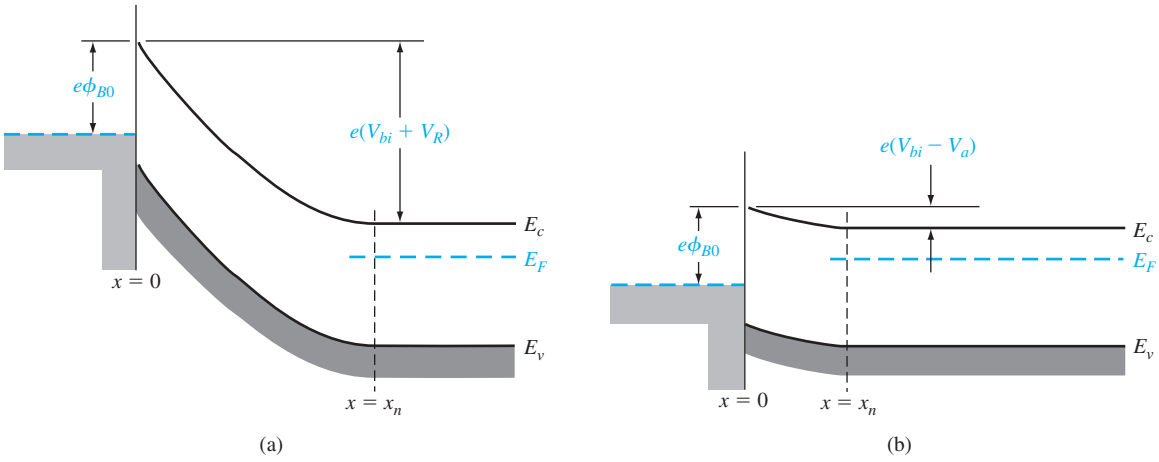


Figure 9.2 | Ideal energy-band diagram of a metal–semiconductor junction (a) under reverse bias and (b) under forward bias.

The energy-band diagrams versus voltage for the metal–semiconductor junction shown in Figure 9.2 are very similar to those of the pn junction given in the previous chapter. Because of this similarity, we expect the current–voltage characteristics of the Schottky barrier junction to be similar to the exponential behavior of the pn junction diode. The current mechanism here, however, is due to the flow of majority carrier electrons. In forward bias, the barrier seen by the electrons in the semiconductor is reduced, so majority carrier electrons flow more easily from the semiconductor into the metal. The forward-bias current is in the direction from metal to semiconductor: It is an exponential function of the forward-bias voltage V_a.

9.1.2 Ideal Junction Properties

We can determine the electrostatic properties of the junction in the same way as we do for the pn junction. The electric field in the space charge region is determined from Poisson's equation. We have that

$$\frac{d\mathrm{E}}{dx} = \frac{\rho(x)}{\epsilon_s} \tag{9.3}$$

where $\rho(x)$ is the space charge volume density and ϵ_s is the permittivity of the semiconductor. If we assume that the semiconductor doping is uniform, then by integrating Equation (9.3), we obtain

$$\mathrm{E} = \int \frac{eN_d}{\epsilon_s} dx = \frac{eN_d x}{\epsilon_s} + C_1 \tag{9.4}$$

where C_1 is a constant of integration. The electric field is zero at the space charge edge in the semiconductor, so the constant of integration can be found as

$$C_1 = -\frac{eN_d x_n}{\epsilon_s} \tag{9.5}$$

The electric field can then be written as

$$\mathrm{E} = -\frac{eN_d}{\epsilon_s}(x_n - x) \tag{9.6}$$

which is a linear function of distance, for the uniformly doped semiconductor, and reaches a peak value at the metal–semiconductor interface. Since the E-field is zero inside the metal, a negative surface charge must exist in the metal at the metal–semiconductor junction.

The space charge region width, W, may be calculated as we do for the pn junction. The result is identical to that of a one-sided p^+n junction. For the uniformly doped semiconductor, we have

$$W = x_n = \left[\frac{2\epsilon_s(V_{bi} + V_R)}{eN_d}\right]^{1/2} \tag{9.7}$$

where V_R is the magnitude of the applied reverse-biased voltage. We are again assuming an abrupt junction approximation.

EXAMPLE 9.1

Objective: Determine the theoretical barrier height, built-in potential barrier, and maximum electric field in a metal–semiconductor diode for zero applied bias.

Consider a contact between tungsten and n-type silicon doped to $N_d = 10^{16}$ cm^{-3} at $T = 300$ K.

■ Solution

The metal work function for tungsten (W) from Table 9.1 is $\phi_m = 4.55$ V and the electron affinity for silicon from Table 9.2 is $\chi = 4.01$ V. The barrier height is then

$$\phi_{B0} = \phi_m - \chi = 4.55 - 4.01 = 0.54 \text{ V}$$

where ϕ_{B0} is the ideal Schottky barrier height. We can calculate ϕ_n as

$$\phi_n = \frac{kT}{e}\ln\left(\frac{N_c}{N_d}\right) = 0.0259\ln\left(\frac{2.8 \times 10^{19}}{10^{16}}\right) = 0.206 \text{ V}$$

Then

$$V_{bi} = \phi_{B0} - \phi_n = 0.54 - 0.206 = 0.334 \text{ V}$$

The space charge width at zero bias is

$$x_n = \left[\frac{2\epsilon_s V_{bi}}{eN_d}\right]^{1/2} = \left[\frac{2(11.7)(8.85 \times 10^{-14})(0.334)}{(1.6 \times 10^{-19})(10^{16})}\right]^{1/2}$$

or

$$x_n = 0.208 \times 10^{-4} \text{ cm}$$

Then the maximum electric field is

$$|\mathrm{E}_{\max}| = \frac{eN_d x_n}{\epsilon_s} = \frac{(1.6 \times 10^{-19})(10^{16})(0.208 \times 10^{-4})}{(11.7)(8.85 \times 10^{-14})}$$

or finally

$$|\mathrm{E}_{\max}| = 3.21 \times 10^4 \text{ V/cm}$$

Comment

The values of space charge width and electric field are very similar to those obtained for a pn junction.

EXERCISE PROBLEM

Ex 9.1 Consider an ideal tungsten-to-n-type GaAs junction. Assume the GaAs is doped to a concentration of $N_d = 5 \times 10^{15}$ cm^{-3}. Determine the theoretical barrier height, the built-in potential barrier, and maximum electric field for the case of zero applied bias.

(Ans. $\phi_{B0} = 0.48$ V, $V_{bi} = 0.3623$ V, $|\mathcal{E}_{max}| = 2.24 \times 10^4$ V/cm)

A junction capacitance can also be determined in the same way as we do for the pn junction. We have that

$$C' = eN_d \frac{dx_n}{dV_R} = \left[\frac{e\epsilon_s N_d}{2\,(V_{bi} + V_R)}\right]^{1/2} \tag{9.8}$$

where C' is the capacitance per unit area. If we square the reciprocal of Equation (9.8), we obtain

$$\left(\frac{1}{C'}\right)^2 = \frac{2(V_{bi} + V_R)}{e\epsilon_s N_d} \tag{9.9}$$

We can use Equation (9.9) to obtain, to a first approximation, the built-in potential barrier V_{bi}, and the slope of the curve from Equation (9.9) to yield the semiconductor doping N_d. We can calculate the potential ϕ_n and then determine the Schottky barrier ϕ_{B0} from Equation (9.2).

EXAMPLE 9.2

Objective: To calculate the semiconductor doping concentration and Schottky barrier height from the silicon diode experimental capacitance data shown in Figure 9.3. Assume $T = 300$ K.

Solution

The intercept of the tungsten–silicon curve is approximately at $V_{bi} = 0.40$ V. From Equation (9.9), we can write

$$\frac{d(1/C')^2}{dV_R} \approx \frac{\Delta(1/C')^2}{\Delta V_R} = \frac{2}{e\epsilon_s N_d}$$

Then, from the figure, we have

$$\frac{\Delta(1/C')^2}{\Delta V_R} \approx 4.4 \times 10^{13}$$

so that

$$N_d = \frac{2}{(1.6 \times 10^{-19})(11.7)(8.85 \times 10^{-14})(4.4 \times 10^{13})} = 2.7 \times 10^{17}\ \text{cm}^{-3}$$

We can calculate

$$\phi_n = \frac{kT}{e} \ln\left(\frac{N_c}{N_d}\right) = (0.0259) \ln\left(\frac{2.8 \times 10^{19}}{2.7 \times 10^{17}}\right) = 0.12\ \text{V}$$

so that

$$\phi_{Bn} = V_{bi} + \phi_n = 0.40 + 0.12 = 0.52\ \text{V}$$

where ϕ_{Bn} is the actual Schottky barrier height.

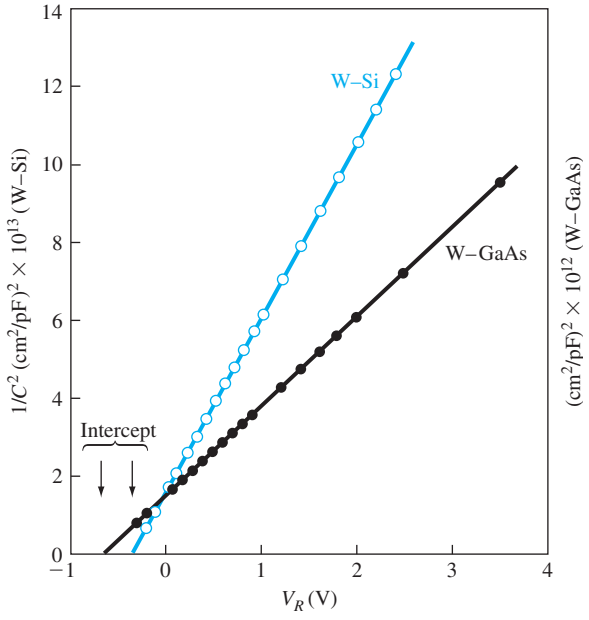


Figure 9.3 | $1/C^2$ versus V_R for W–Si and W–GaAs Schottky barrier diodes.
(From Sze and Ng [15].)

■ Comment

The experimental value of 0.52 V can be compared with the ideal barrier height of $\phi_{B0} = 0.54$ V found in Example 9.1. These results agree fairly well. For other metals, the discrepancy between experiment and theory is larger.

■ EXERCISE PROBLEM

Ex 9.2 Repeat Example 9.2 for the GaAs diode capacitance data shown in Figure 9.3.

(Ans. $V_{bi} \cong 0.64$ V, $N_d = 4.62 \times 10^{18}$ cm^{-3})

We can see that the built-in potential barrier of the gallium arsenide Schottky diode is larger than that of the silicon diode. This experimental result is normally observed for all types of metal contacts.

TEST YOUR UNDERSTANDING

TYU 9.1 Consider an ideal chromium-to-n-type silicon Schottky diode at $T = 300$ K. Assume the semiconductor is doped at a concentration of $N_d = 3 \times 10^{15}$ cm^{-3}. Determine the (*a*) ideal Schottky barrier height, (*b*) built-in potential barrier, (*c*) peak electric field with an applied reverse-biased voltage of $V_R = 5$ V, and (*d*) junction capacitance per unit area for $V_R = 5$ V.

[Ans. (*a*) $\phi_{B0} = 0.49$ V; (*b*) $V_{bi} = 0.253$ V; (*c*) $|E_{max}| = 6.98 \times 10^4$ V/cm; (*d*) $C' = 6.88 \times 10^{-9}$ F/cm^2]

TYU 9.2 Repeat TYU 9.1 for an ideal palladium-to-n-type GaAs Schottky diode with the same impurity concentration. [Ans. (a) $\phi_{B0} = 1.05$ V; (b) $V_{bi} = 0.919$ V; (c) $|E_{max}| = 7 \times 10^4$ V/cm; (d) $C' = 6.86 \times 10^{-9}$ F/cm^2]

9.1.3 Nonideal Effects on the Barrier Height

Schottky Barrier Lowering Several effects alter the actual Schottky barrier height from the theoretical value given by Equation (9.1). The first such effect that we consider is the Schottky effect, or image-force-induced lowering of the potential barrier.

An electron in a dielectric material at a distance x from the metal will create an electric field. The electric field can be determined by adding an image charge, $+e$, inside the metal located at the same distance, $|x|$, from the interface. This image effect is shown in Figure 9.4a. Note that the E–field lines are perpendicular to the metal surface as we expect. The force on the electron, due to the coulomb attraction with the image force, is

$$F = \frac{-e^2}{4\pi\epsilon_s(2x)^2} = -e\mathrm{E} \tag{9.10}$$

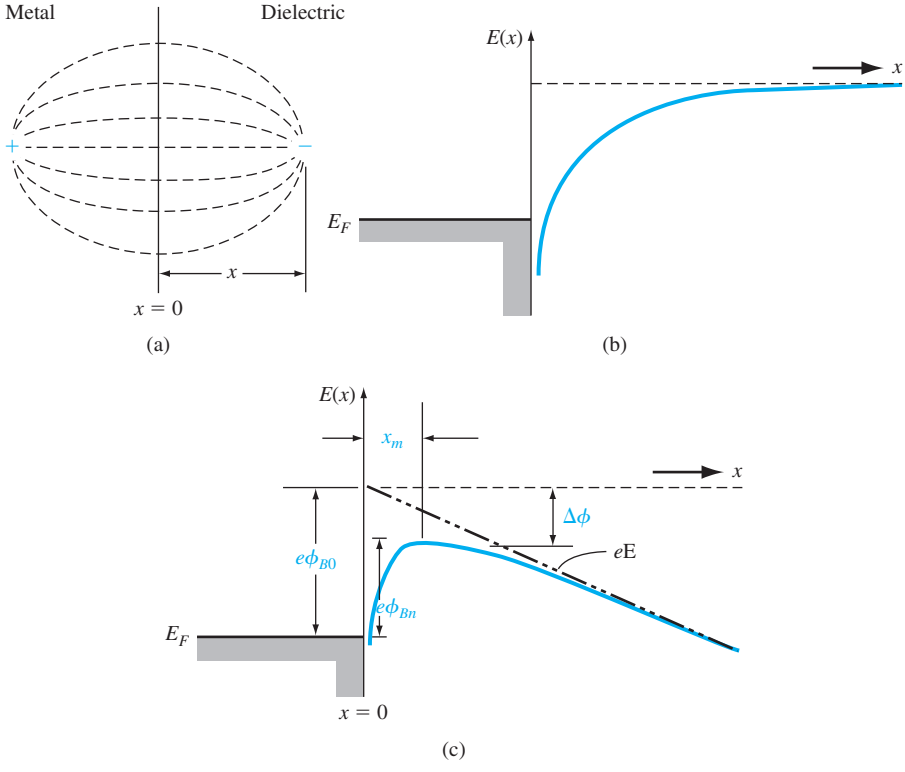


Figure 9.4 | (a) Image charge and electric field lines at a metal–dielectric interface. (b) Distortion of the potential barrier due to image forces with zero electric field and (c) with a constant electric field.

The potential can then be found as

$$-\phi(x) = +\int_x^\infty \mathrm{E}\,dx' = +\int_x^\infty \frac{e}{4\pi\epsilon_s \cdot 4(x')^2}\,dx' = \frac{-e}{16\pi\epsilon_s x} \tag{9.11}$$

where x' is the integration variable and where we have assumed that the potential is zero at $x = \infty$.

The potential energy of the electron is $-e\phi(x)$: Figure 9.4b is a plot of the potential energy assuming that no other electric fields exist. With an electric field present in the dielectric, the potential is modified and can be written as

$$-\phi(x) = \frac{-e}{16\pi\epsilon_s x} - \mathrm{E}x \tag{9.12}$$

The potential energy of the electron, including the effect of a constant electric field, is plotted in Figure 9.4c. The peak potential barrier is now lowered. This lowering of the potential barrier is the Schottky effect, or image force–induced lowering.

We can find the Schottky barrier lowering, $\Delta\phi$, and the position of the maximum barrier, x_m, from the condition that

$$\frac{d[e\phi(x)]}{dx} = 0 \tag{9.13}$$

We find that

$$x_m = \sqrt{\frac{e}{16\pi\epsilon_s \mathrm{E}}} \tag{9.14}$$

and

$$\boxed{\Delta\phi = \sqrt{\frac{e\mathrm{E}}{4\pi\epsilon_s}}} \tag{9.15}$$

EXAMPLE 9.3

Objective: Calculate the Schottky barrier lowering and the position of the maximum barrier height.

Consider a gallium arsenide metal–semiconductor contact in which the electric field in the semiconductor is assumed to be $\mathrm{E} = 6.8 \times 10^4$ V/cm.

■ **Solution**

The Schottky barrier lowering is given by Equation (9.15), which in this case yields

$$\Delta\phi = \sqrt{\frac{e\mathrm{E}}{4\pi\epsilon_s}} = \sqrt{\frac{(1.6 \times 10^{-19})(6.8 \times 10^4)}{4\pi\,(13.\,1)(8.85 \times 10^{-14})}} = 0.0273 \text{ V}$$

The position of the maximum barrier height is

$$x_m = \sqrt{\frac{e}{16\pi\epsilon_s \mathrm{E}}} = \sqrt{\frac{(1.6 \times 10^{-19})}{16\pi(13.1)(8.85 \times 10^{-14})(6.8 \times 10^4)}}$$

or

$$x_m = 2 \times 10^{-7} \text{ cm} = 20 \text{ Å}$$

Comment

Although the Schottky barrier lowering may seem like a small value, the barrier height and the barrier lowering will appear in exponential terms in the current–voltage relationship. A small change in the barrier height can thus have a significant effect on the current in a Schottky barrier diode.

EXERCISE PROBLEM

Ex 9.3 Consider the Schottky diode described in Example 9.1. Calculate the Schottky barrier lowering for a reverse-biased voltage of (*a*) $V_R = 1$ V and (*b*) $V_R = 5$ V.

[Ans. (*a*) $\Delta\phi = 0.0281$ V; (*b*) $\Delta\phi = 0.0397$ V]

Interface States Figure 9.5 shows the measured barrier heights in gallium arsenide and silicon Schottky diodes as a function of metal work functions. There is a monotonic relation between the measured barrier height and the metal work function, but the curves do not fit the simple relation given in Equation (9.1). The barrier height of the metal–semiconductor junction is determined by both the metal work function and the semiconductor surface or interface states.

A more detailed energy-band diagram of a metal to n-type semiconductor contact in thermal equilibrium is shown in Figure 9.6. We assume that a narrow interfacial layer of insulator exists between the metal and semiconductor. The interfacial layer can support a potential difference, but will be transparent to the flow of electrons between the metal and semiconductor. The semiconductor also shows a distribution of surface states at the metal–semiconductor interface. We assume that all states below

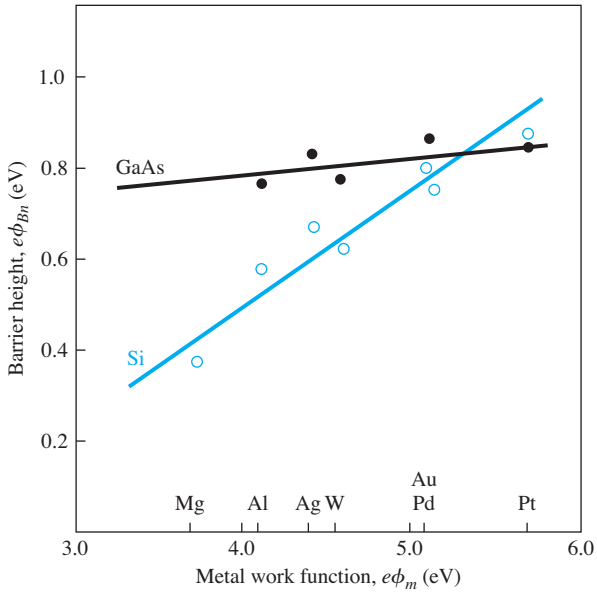


Figure 9.5 | Experimental barrier heights as a function of metal work functions for GaAs and Si.
(From Crowley and Sze [2].)

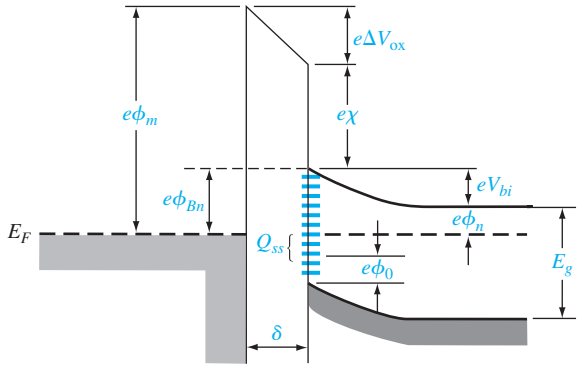


Figure 9.6 | Energy-band diagram of a metal–semiconductor junction with an interfacial layer and interface states.

the surface potential ϕ_0 are donor states, which will be neutral if the state contains an electron and positively charged if the state does not contain an electron. We also assume that all states above ϕ_0 are acceptor states, which will be neutral if the state does not contain an electron and negatively charged if the state contains an electron.

The diagram in Figure 9.6 shows some acceptor states above ϕ_0 and below E_F. These states tend to contain electrons and are negatively charged. We may assume that the surface state density is constant and equal to D_{it} states/cm^2-eV. The relation between the surface potential, surface state density, and other semiconductor parameters is found to be

$$(E_g - e\phi_0 - e\phi_{Bn}) = \frac{1}{eD_{it}}\sqrt{2e\epsilon_s N_d(\phi_{Bn} - \phi_n)} - \frac{\epsilon_i}{eD_{it}\delta}[\phi_m - (\chi + \phi_{Bn})] \tag{9.16}$$

We consider two limiting cases.

Case 1 Let $D_{it} \to \infty$. In this case, the right side of Equation (9.16) goes to zero. We then have

$$\phi_{Bn} = \frac{1}{e}(E_g - e\phi_0) \tag{9.17}$$

The barrier height is now fixed by the bandgap energy and the potential ϕ_0. The barrier height is totally independent of the metal work function and the semiconductor electron affinity. The Fermi level becomes "pinned" at the surface, at the surface potential ϕ_0.

Case 2 Let $D_{it}\delta \to 0$. Equation (9.16) reduces to

$$\phi_{Bn} = (\phi_m - \chi)$$

which is the original ideal expression.

The Schottky barrier height is a function of the electric field in the semiconductor through the barrier lowering effect. The barrier height is also a function of the surface states in the semiconductor. The barrier height, then, is modified from the ideal theoretical value. Since the surface state density is not predictable with any degree of certainty, the barrier height must be an experimentally determined parameter.

TEST YOUR UNDERSTANDING

TYU 9.3 Determine the Schottky barrier lowering and the position of the maximum barrier height for the junction described in TYU 9.1. Use the value of the electric field found in TYU 9.1. (Ans. $\Delta\phi = 0.0293$ eV, $x_m = 12$ Å)

9.1.4 Current–Voltage Relationship

The current transport in a metal–semiconductor junction is due mainly to majority carriers as opposed to minority carriers in a pn junction. The basic process in the rectifying contact with an n-type semiconductor is by transport of electrons over the potential barrier, which can be described by the thermionic emission theory.

The thermionic emission characteristics are derived by assuming that the barrier height is much larger than kT, so that the Maxwell–Boltzmann approximation applies and that thermal equilibrium is not affected by this process. Figure 9.7 shows the one-dimensional barrier with an applied forward-bias voltage V_a and two electron current density components. The current $J_{s\to m}$ is the electron current density due to the flow of electrons from the semiconductor into the metal, and the current $J_{m\to s}$ is the electron current density due to the flow of electrons from the metal into the semiconductor. The subscripts of the currents indicate the direction of electron flow. The conventional current direction is opposite to electron flow.

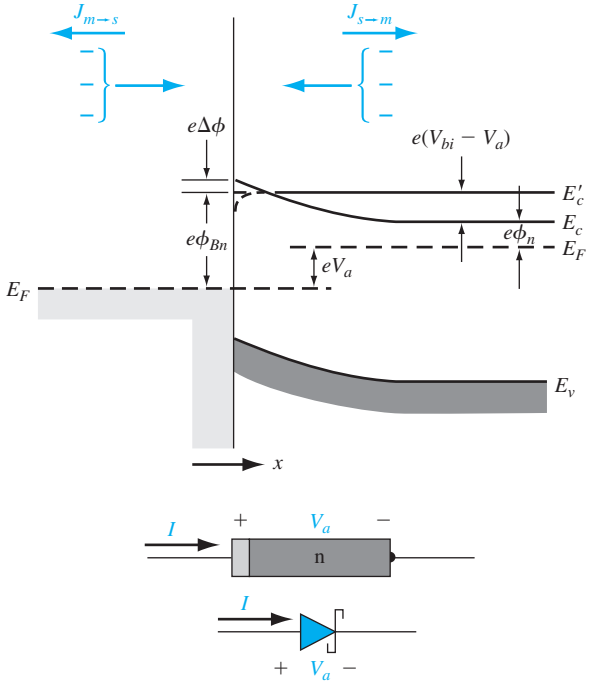


Figure 9.7 | Energy-band diagram of a forward-biased metal–semiconductor junction including the image lowering effect.

The current density $J_{s\to m}$ is a function of the concentration of electrons which have x-directed velocities sufficient to overcome the barrier. We may write

$$J_{s\to m} = e\int_{E_c'}^{\infty} v_x\, dn \tag{9.18}$$

where E_c' is the minimum energy required for thermionic emission into the metal, v_x is the carrier velocity in the direction of transport, and e is the magnitude of the electronic charge. The incremental electron concentration is given by

$$dn = g_c(E)\, f_F(E)\, dE \tag{9.19}$$

where $g_c(E)$ is the density of states in the conduction band and $f_F(E)$ is the Fermi–Dirac probability function. Assuming that the Maxwell–Boltzmann approximation applies, we may write

$$dn = \frac{4\pi(2m_n^*)^{3/2}}{h^3}\sqrt{E-E_c}\,\exp\left[\frac{-(E-E_F)}{kT}\right]dE \tag{9.20}$$

If all of the electron energy above E_c is assumed to be kinetic energy, then we have

$$\frac{1}{2}m_n^* v^2 = E - E_c \tag{9.21}$$

The net current density in the metal-to-semiconductor junction can be written as

$$J = J_{s\to m} - J_{m\to s} \tag{9.22}$$

which is defined to be positive in the direction from the metal to the semiconductor. We find that

$$\boxed{J = \left[A^*\,T^2\exp\left(\frac{-e\phi_{Bn}}{kT}\right)\right]\left[\exp\left(\frac{eV_a}{kT}\right)-1\right]} \tag{9.23}$$

where

$$\boxed{A^* \equiv \frac{4\pi e m_n^* k^2}{h^3}} \tag{9.24}$$

The parameter A^* is called the effective Richardson constant for thermionic emission.

Equation (9.23) can be written in the usual diode form as

$$\boxed{J = J_{sT}\left[\exp\left(\frac{eV_a}{kT}\right)-1\right]} \tag{9.25}$$

where J_{sT} is the reverse-saturation current density and is given by

$$\boxed{J_{sT} = A^*T^2\exp\left(\frac{-e\phi_{Bn}}{kT}\right)} \tag{9.26}$$

We may recall that the Schottky barrier height ϕ_{Bn} changes because of the image-force lowering. We have that $\phi_{Bn} = \phi_{B0} - \Delta\phi$. Then we can write Equation (9.26) as

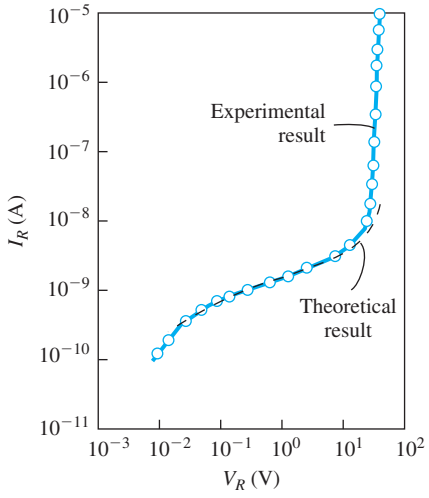


Figure 9.8 | Experimental and theoretical reverse-biased currents in a PtSi–Si diode. *(From Sze and Ng [15].)*

$$J_{sT} = A^*T^2 \exp\left(\frac{-e\phi_{B0}}{kT}\right) \exp\left(\frac{e\Delta\phi}{kT}\right) \tag{9.27}$$

The change in barrier height, $\Delta\phi$, will increase with an increase in the electric field, or with an increase in the applied reverse-biased voltage. Figure 9.8 shows a typical reverse-biased current–voltage characteristic of a Schottky barrier diode. The reverse-biased current increases with reverse-biased voltage because of the barrier lowering effect. This figure also shows the Schottky barrier diode going into breakdown.

EXAMPLE 9.4

Objective: Determine the effective Richardson constant from the current–voltage characteristics.

Consider the tungsten–silicon diode curve in Figure 9.9 and assume a barrier height of $\phi_{Bn} = 0.67$ V. From the figure, $J_{sT} \approx 6 \times 10^{-5}$ A/cm^2.

■ Solution

We have that

$$J_{sT} = A^* T^2 \exp\left(\frac{-e\phi_{Bn}}{kT}\right)$$

so that

$$A^* = \frac{J_{sT}}{T^2} \exp\left(\frac{+e\phi_{Bn}}{kT}\right)$$

Then

$$A^* = \frac{6 \times 10^{-5}}{(300)^2} \exp\left(\frac{0.67}{0.0259}\right) = 114 \text{ A/K}^2\text{-cm}^2$$

■ Comment

The experimentally determined value of A^* is a very strong function of ϕ_{Bn}, since ϕ_{Bn} is in the exponential term. A small change in ϕ_{Bn} will change the value of the Richardson constant substantially.

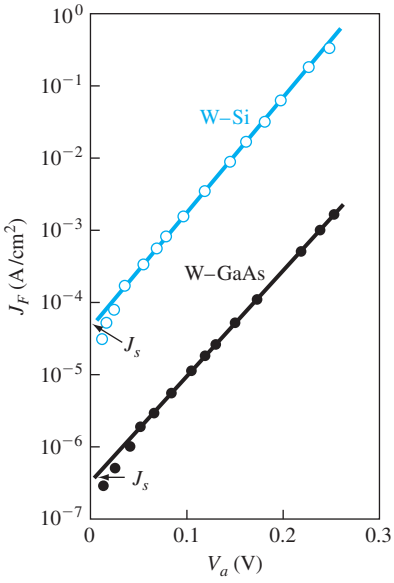


Figure 9.9 | Forward-bias current density J_F versus V_a for Wi–Si and W–GaAs diodes. *(From Sze and Ng [15].)*

■ EXERCISE PROBLEM

Ex 9.4 Calculate the ideal Richardson constant for a free electron.

($A^* = 120$ A/K^2-cm^2)

We may note that the reverse-saturation current densities of the tungsten–silicon and tungsten–gallium arsenide diodes in Figure 9.9 differ by approximately two orders of magnitude. This two order of magnitude difference will be reflected in the effective Richardson constant, assuming the barrier heights in the two diodes are essentially the same. The definition of the effective Richardson constant, given by Equation (9.24), contains the electron effective mass, which differs substantially between silicon and gallium arsenide. The fact that the effective mass is in the expression for the Richardson constant is a direct result of using the effective density of states function in the thermionic emission theory. The net result is that A^* and J_{sT} will vary widely between silicon and gallium arsenide.

9.1.5 Comparison of the Schottky Barrier Diode and the pn Junction Diode

Although the ideal current–voltage relationship of the Schottky barrier diode given by Equation (9.25) is of the same form as that of the pn junction diode, there are two important differences between a Schottky diode and a pn junction diode: The first is in the magnitudes of the reverse-saturation current densities and the second is in the switching characteristics.

The reverse-saturation current density of the Schottky barrier diode was given by Equation (9.26) and is

$$J_{sT} = A^* T^2 \exp\left(\frac{-e\phi_{Bn}}{kT}\right)$$

The ideal reverse-saturation current density of the pn junction diode can be written as

$$J_s = \frac{eD_n n_{po}}{L_n} + \frac{eD_p p_{no}}{L_p} \tag{9.28}$$

The form of the two equations is vastly different, and the current mechanism in the two devices is different. The current in a pn junction is determined by the diffusion of minority carriers while the current in a Schottky barrier diode is determined by thermionic emission of majority carriers over a potential barrier.

EXAMPLE 9.5

Objective: Calculate the ideal reverse-saturation current densities of a Schottky barrier diode and a pn junction diode.

Consider a tungsten barrier on silicon with a measured barrier height of $e\phi_{Bn} = 0.67$ eV. The effective Richardson constant is $A^* = 114$ A/K^2-cm^2. Let $T = 300$ K.

■ Solution

If we neglect the barrier lowering effect, we have for the Schottky barrier diode

$$J_{sT} = A^* T^2 \exp\left(\frac{-e\phi_{Bn}}{kT}\right) = (114)(300)^2 \exp\left(\frac{-0.67}{0.0259}\right) = 5.98 \times 10^{-5} \text{ A/cm}^2$$

Consider a silicon pn junction with the following parameters at $T = 300$ K.

$$N_a = 10^{18} \text{ cm}^{-3} \qquad N_d = 10^{16} \text{ cm}^{-3}$$
$$D_p = 10 \text{ cm}^2/\text{s} \qquad D_n = 25 \text{ cm}^2/\text{s}$$
$$\tau_{po} = 10^{-7} \text{ s} \qquad \tau_{no} = 10^{-7} \text{ s}$$

We can then calculate the following parameters:

$$L_p = 1.0 \times 10^{-3} \text{ cm} \qquad L_n = 1.58 \times 10^{-3} \text{ cm}$$
$$p_{no} = 2.25 \times 10^4 \text{ cm}^{-3} \qquad n_{po} = 2.25 \times 10^2 \text{ cm}^{-3}$$

The ideal reverse-saturation current density of the pn junction diode can be determined from Equation (9.28) as

$$J_s = \frac{(1.6 \times 10^{-19})(25)(2.25 \times 10^2)}{(1.58 \times 10^{-3})} + \frac{(1.6 \times 10^{-19})(10)(2.25 \times 10^4)}{(1.0 \times 10^{-3})}$$
$$= 5.7 \times 10^{-13} + 3.6 \times 10^{-11} = 3.66 \times 10^{-11} \text{ A/cm}^2$$

■ Comment

The ideal reverse-saturation current density of the Schottky barrier junction is orders of magnitude larger than that of the ideal pn junction diode.

■ EXERCISE PROBLEM

Ex 9.5 Using the results of Example 9.5, determine the forward-bias voltages required to produce a current of 10 μA in each diode. Assume each cross-sectional area is 10^{-4} cm^2.

(Ans. pn junction, $V_a = 0.5628$ V; Schottky junction, $V_a = 0.1922$ V)

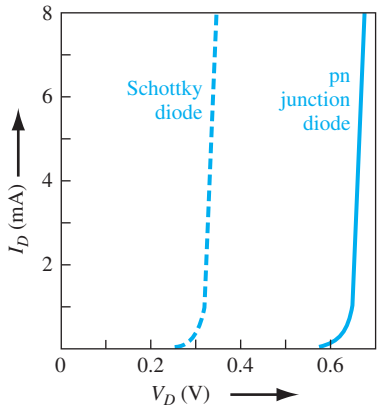


Figure 9.10 | Comparison of forward-bias *I–V* characteristics between a Schottky diode and a pn junction diode.

Recall that the reverse-biased current in a silicon pn junction diode is dominated by the generation current. A typical generation current density is approximately 10^{-7} A/cm^2, which is still two to three orders of magnitude less than the reverse-saturation current density of the Schottky barrier diode. A generation current also exists in the reverse-biased Schottky barrier diode; however, the generation current is negligible compared with the J_{sT} value.

Since $J_{sT} \gg J_s$, the forward-bias characteristics of the two types of diodes will also be different. Figure 9.10 shows typical *I–V* characteristics of a Schottky barrier diode and a pn junction diode. The effective turn-on voltage of the Schottky diode is less than that of the pn junction diode.

EXAMPLE 9.6

Objective: Calculate the forward-bias voltage required to induce a forward-bias current density of 10 A/cm^2 in a Schottky barrier diode and a pn junction diode.

Consider diodes with the parameters given in Example 9.5. We can assume that the pn junction diode will be sufficiently forward biased so that the ideal diffusion current will dominate. Let $T = 300$ K.

■ Solution

For the Schottky barrier diode, we have

$$J = J_{sT}\left[\exp\left(\frac{eV_a}{kT}\right) - 1\right]$$

Neglecting the (-1) term, we can solve for the forward-bias voltage. We find

$$V_a = \left(\frac{kT}{e}\right)\ln\left(\frac{J}{J_{sT}}\right) = V_t \ln\left(\frac{J}{J_{sT}}\right) = (0.0259)\ln\left(\frac{10}{5.98 \times 10^{-5}}\right) = 0.312 \text{ V}$$

For the pn junction diode, we have

$$V_a = V_t \ln\left(\frac{J}{J_s}\right) = (0.0259)\ln\left(\frac{10}{3.66 \times 10^{-11}}\right) = 0.682 \text{ V}$$

Comment

A comparison of the two forward-bias voltages shows that the Schottky barrier diode has a turn-on voltage that, in this case, is approximately 0.37 V smaller than the turn-on voltage of the pn junction diode.

EXERCISE PROBLEM

Ex 9.6 A pn junction diode and a Schottky diode have equal cross-sectional areas and have forward-bias currents of 0.5 mA. The reverse-saturation current of the Schottky diode is 5×10^{-7} A. The difference in forward-bias voltage between the two diodes is 0.30 V. Determine the reverse-saturation current of the pn junction diode.

(Ans. 4.66×10^{-12} A)

The actual difference between the turn-on voltages will be a function of the barrier height of the metal–semiconductor contact and the doping concentrations in the pn junction, but the relatively large difference will always be realized. We will consider one application that utilizes the difference in turn-on voltage in Chapter 12, in what is referred to as a *Schottky clamped transistor.*

The second major difference between a Schottky barrier diode and a pn junction diode is in the frequency response, or switching characteristics. In our discussion, we have considered the current in a Schottky diode as being due to the injection of majority carriers over a potential barrier. The energy-band diagram of Figure 9.1, for example, shows that there can be electrons in the metal directly adjacent to empty states in the semiconductor. If an electron from the valence band of the semiconductor were to flow into the metal, this effect would be equivalent to holes being injected into the semiconductor. This injection of holes would create excess minority carrier holes in the n region. However, calculations as well as measurements have shown that the ratio of the minority carrier hole current to the total current is extremely low in most cases.

The Schottky barrier diode, then, is a majority carrier device. This fact means that there is no diffusion capacitance associated with a forward-biased Schottky diode. The elimination of the diffusion capacitance makes the Schottky diode a higher-frequency device than the pn junction diode. Also, when switching a Schottky diode from forward to reverse bias, there is no minority carrier stored charge to remove, as is the case in the pn junction diode. Since there is no minority carrier storage time, the Schottky diodes can be used in fast-switching applications. A typical switching time for a Schottky diode is in the picosecond range, while for a pn junction it is normally in the nanosecond range.

TEST YOUR UNDERSTANDING

TYU 9.4 (*a*) The reverse-saturation currents of a pn junction and a Schottky diode are 10^{-14} A and 10^{-9} A, respectively. Determine the required forward-bias voltages in the pn junction diode and Schottky diode to produce a current of 100 μA in each diode. (*b*) Repeat part (*a*) for forward bias currents of 1 mA.

[Ans. (*a*) 0.596 V, 0.298 V; (*b*) 0.656 V, 0.358 V]

9.2 | METAL–SEMICONDUCTOR OHMIC CONTACTS

Contacts must be made between any semiconductor device, or integrated circuit, and the outside world. These contacts are made via *ohmic contacts.* Ohmic contacts are metal-to-semiconductor contacts, but in this case they are not rectifying contacts. An ohmic contact is a low-resistance junction providing conduction in both directions between the metal and the semiconductor. Ideally, the current through the ohmic contact is a linear function of applied voltage, and the applied voltage should be very small. Two general types of ohmic contacts are possible: The first type is the ideal nonrectifying barrier, and the second is the tunneling barrier. We define in this section a specific contact resistance that is used to characterize ohmic contacts.

9.2.1 Ideal Nonrectifying Barrier

We have considered an ideal metal-n-type semiconductor contact in Figure 9.1 for the case when $\phi_m > \phi_s$. Figure 9.11 shows the same ideal contact for the opposite case of $\phi_m < \phi_s$. In Figure 9.11a we see the energy levels before contact, and in Figure 9.11b, the barrier after contact for thermal equilibrium. To achieve thermal equilibrium in this junction, electrons flow from the metal into the lower energy states in the semiconductor, which makes the surface of the semiconductor more n type. The excess electron charge in the n-type semiconductor exists essentially as a surface charge density. If a positive voltage is applied to the metal, there is no barrier to electrons flowing from the semiconductor into the metal. If a positive voltage is applied to the semiconductor, the effective barrier height for electrons flowing from the metal into the semiconductor will be approximately $\phi_{Bn} = \phi_n$, which is fairly small for a moderately to heavily doped semiconductor. For this bias condition, electrons can easily flow from the metal into the semiconductor.

Figure 9.12a shows the energy-band diagram when a positive voltage is applied to the metal with respect to the semiconductor. Electrons can easily flow "downhill"

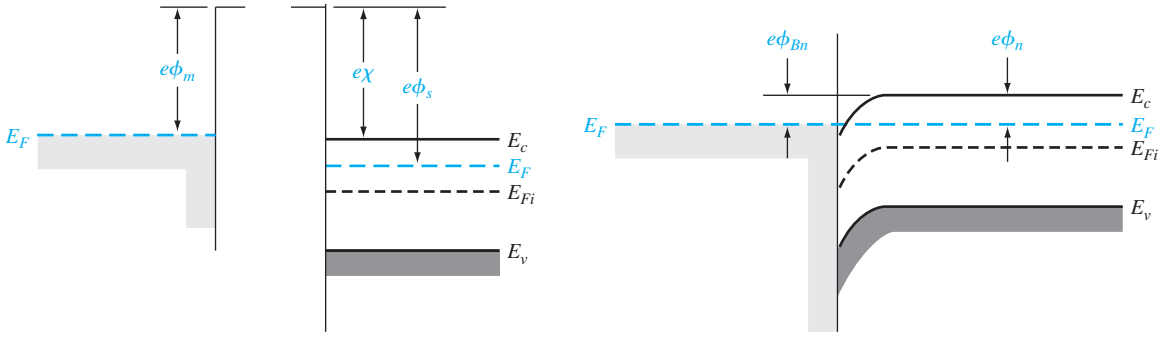


Figure 9.11 | Ideal energy-band diagram (a) before contact and (b) after contact for a metal-n-type semiconductor junction for $\phi_m < \phi_s$.

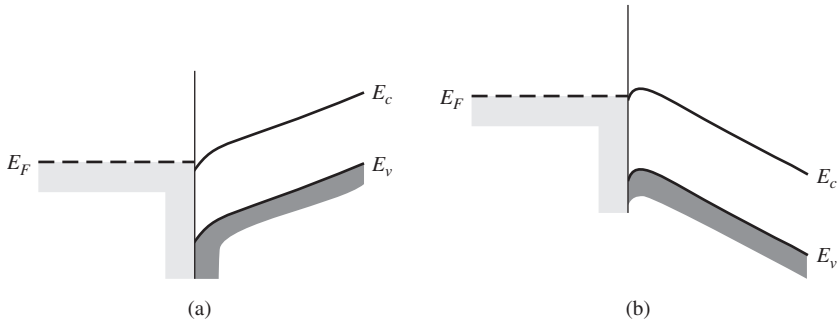


Figure 9.12 | Ideal energy-band diagram of a metal-n-type semiconductor ohmic contact (a) with a positive voltage applied to the metal and (b) with a positive voltage applied to the semiconductor.

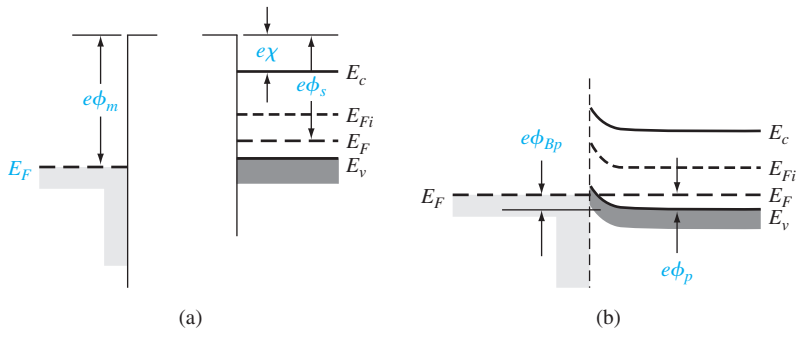


Figure 9.13 | Ideal energy-band diagram (a) before contact and (b) after contact for a metal–p-type semiconductor junction for $\phi_m < \phi_s$.

from the semiconductor into the metal. Figure 9.12b shows the case when a positive voltage is applied to the semiconductor with respect to the metal. Electrons can easily flow over the barrier from the metal into the semiconductor. This junction, then, is an ohmic contact.

Figure 9.13 shows an ideal nonrectifying contact between a metal and a p-type semiconductor. Figure 9.13a shows the energy levels before contact for the case when $\phi_m > \phi_s$. When contact is made, electrons from the semiconductor flow into the metal to achieve thermal equilibrium, leaving behind more empty states, or holes. The excess concentration of holes at the surface makes the surface of the semiconductor more p type. Electrons from the metal can readily move into the empty states in the semiconductor. This charge movement corresponds to holes flowing from the semiconductor into the metal. We can also visualize holes in the metal flowing into the semiconductor. This junction is also an ohmic contact.

The ideal energy bands shown in Figures 9.11 and 9.13 do not take into account the effect of surface states. If we assume that acceptor surface states exist in

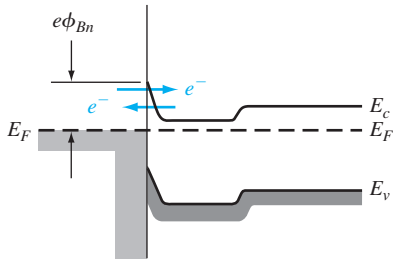


Figure 9.14 | Energy-band diagram of a heavily doped n-semiconductor-to-metal junction.

the upper half of the semiconductor bandgap, then, since all the acceptor states are below E_F for the case shown in Figure 9.11b, these surface states will be negatively charged and will alter the energy-band diagram. Similarly, if we assume that donor surface states exist in the lower half of the bandgap, then all of the donor states will be positively charged for the case shown in Figure 9.13b; the positively charged surface states will also alter this energy-band diagram. Therefore, if $\phi_m < \phi_s$ for the metal-n-type semiconductor contact, and if $\phi_m > \phi_s$ for the metal–p-type semiconductor contact, we may not necessarily form a good ohmic contact.

9.2.2 Tunneling Barrier

The space charge width in a rectifying metal–semiconductor contact is inversely proportional to the square root of the semiconductor doping. The width of the depletion region decreases as the doping concentration in the semiconductor increases; thus, as the doping concentration increases, the probability of tunneling through the barrier increases. Figure 9.14 shows a junction in which the metal is in contact with a heavily doped n-type epitaxial layer.

EXAMPLE 9.7

Objective: Calculate the space charge width for a Schottky barrier on a heavily doped semiconductor.

Consider silicon at $T = 300$ K doped at $N_d = 7 \times 10^{18}$ cm^{-3}. Assume a Shottky barrier with $\phi_{Bn} = 0.67$ V. For this case, we can assume that $V_{bi} \approx \phi_{B0}$. Neglect the barrier lowering effect.

■ Solution

From Equation (9.7), we have for zero applied bias

$$x_n = \left[\frac{2\epsilon_s V_{bi}}{eN_d}\right]^{1/2} = \left[\frac{2(11.7)(8.85 \times 10^{-14})(0.67)}{(1.6 \times 10^{-19})(7 \times 10^{18})}\right]^{1/2}$$

or

$$x_n = 1.1 \times 10^{-6} \text{ cm} = 110 \text{ Å}$$

Comment

In a heavily doped semiconductor, the depletion width is on the order of angstroms, so that tunneling is now a distinct possibility. For these types of barrier widths, tunneling may become the dominant current mechanism.

EXERCISE PROBLEM

Ex 9.7 Calculate the space charge width of a rectifying metal–GaAs–semiconductor junction. Assume the n-type doping concentration is $N_d = 7 \times 10^{18}$ cm^{-3} and the built-in potential barrier is $V_{bi} = 0.80$ V.

(Ans. $x_n = 128.7$ Å)

The tunneling current has the form

$$J_t \propto \exp\left(\frac{-e\phi_{Bn}}{E_{oo}}\right) \tag{9.29}$$

where

$$E_{oo} = \frac{e\hbar}{2}\sqrt{\frac{N_d}{\epsilon_s m_n^*}} \tag{9.30}$$

The tunneling current increases exponentially with doping concentration.

9.2.3 Specific Contact Resistance

A figure of merit of ohmic contacts is the specific contact resistance, R_c. This parameter is defined as the reciprocal of the derivative of current density with respect to voltage evaluated at zero bias. We may write

$$R_c = \left(\frac{\partial J}{\partial V}\right)^{-1}\Big|_{V=0} \qquad \Omega\text{-cm}^2 \tag{9.31}$$

We want R_c to be as small as possible for an ohmic contact.

For a rectifying contact with a low to moderate semiconductor doping concentration, the current–voltage relation is given by Equation (9.23) as

$$J_n = A^*T^2 \exp\left(\frac{-e\phi_{Bn}}{kT}\right)\left[\exp\left(\frac{eV}{kT}\right) - 1\right]$$

The thermionic emission current is dominant in this junction. The specific contact resistance for this case is then

$$R_c = \frac{\left(\frac{kT}{e}\right)\exp\left(\frac{+e\phi_{Bn}}{kT}\right)}{A^*T^2} \tag{9.32}$$

The specific contact resistance decreases rapidly as the barrier height decreases.

For a metal–semiconductor junction with a high impurity doping concentration, the tunneling process will dominate. From Equations (9.29) and (9.30), the specific contact resistance is found to be

$$R_c \propto \exp\left(\frac{+2\sqrt{\epsilon_s m_n^*}}{\hbar} \cdot \frac{\phi_{Bn}}{\sqrt{N_d}}\right) \tag{9.33}$$

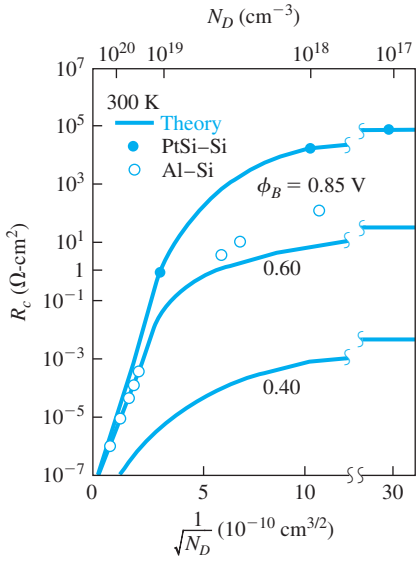


Figure 9.15 | Theoretical and experimental specific contact resistance as a function of doping.
(From Sze and Ng [15].)

which shows that the specific contact resistance is a very strong function of semiconductor doping.

Figure 9.15 shows a plot of the theoretical values of R_c as a function of semiconductor doping. For doping concentrations greater than approximately 10^{19} cm^{-3}, the tunneling process dominates and R_c shows the exponential dependence on N_d. For lower doping concentrations, the R_c values are dependent on the barrier heights and become almost independent of the doping. Also shown in the figure are experimental data for platinum silicide–silicon and aluminum–silicon junctions.

Equation (9.33) is the specific contact resistance of the tunneling junction, which corresponds to the metal–n$^+$ contact shown in Figure 9.14. However, the n$^+$n junction also has a specific contact resistance, since there is a barrier associated with this junction. For a fairly low doped n region, this contact resistance may actually dominate the total resistance of the junction.

The theory of forming ohmic contacts is straightforward. To form a good ohmic contact, we need to create a low barrier and use a highly doped semiconductor at the surface. However, the actual technology of fabricating good, reliable ohmic contacts is not as easy in practice as in theory. It is also more difficult to fabricate good ohmic contacts on wide-bandgap materials. In general, low barriers are not possible on these materials, so a heavily doped semiconductor at the surface must be used to form a tunneling contact. The formation of a tunneling junction requires diffusion, ion implantation, or perhaps epitaxial growth. The surface doping concentration in

the semiconductor may be limited to the impurity solubility, which is approximately 5×10^{19} cm^{-3} for n-type GaAs. Nonuniformities in the surface doping concentration may also prevent the theoretical limit of the specific contact resistance from being reached. In practice, a good deal of empirical processing is usually required before a good ohmic contact is obtained.

9.3 | HETEROJUNCTIONS

In the discussion of pn junctions in the previous chapters, we assumed that the semiconductor material is homogeneous throughout the structure. This type of junction is called a *homojunction.* When two different semiconductor materials are used to form a junction, the junction is called a *semiconductor heterojunction.*

As with many topics in this text, our goal is to provide the basic concepts concerning the heterojunction. The complete analysis of heterojunction structures involves quantum mechanics and detailed calculations that are beyond the scope of this text. The discussion of heterojunctions will, then, be limited to the introduction of some basic concepts.

9.3.1 Heterojunction Materials

Since the two materials used to form a heterojunction will have different energy bandgaps, the energy band will have a discontinuity at the junction interface. We may have an abrupt junction in which the semiconductor changes abruptly from a narrow-bandgap material to a wide-bandgap material. On the other hand, if we have a GaAs–$Al_xGa_{1-x}As$ system, for example, the value of x may continuously vary over a distance of several nanometers to form a graded heterojunction. Changing the value of x in the $Al_xGa_{1-x}As$ system allows us to engineer, or design, the bandgap energy.

In order to have a useful heterojunction, the lattice constants of the two materials must be well matched. The lattice match is important because any lattice mismatch can introduce dislocations resulting in interface states. For example, germanium and gallium arsenide have lattice constants matched to within approximately 0.13 percent. Germanium–gallium arsenide heterojunctions have been studied quite extensively. More recently, gallium arsenide–aluminum gallium arsenide (GaAs–AlGaAs) junctions have been investigated quite thoroughly, since the lattice constants of GaAs and the AlGaAs system vary by no more than 0.14 percent.

9.3.2 Energy-Band Diagrams

In the formation of a heterojunction with a narrow-bandgap material and a wide-bandgap material, the alignment of the bandgap energies is important in determining the characteristics of the junction. Figure 9.16 shows three possible situations. In Figure 9.16a, we see the case when the forbidden bandgap of the wide-gap material completely overlaps the bandgap of the narrow-gap material. This case, called *straddling,* applies to most heterojunctions. We consider only this case here. The other possibilities are called *staggered* and *broken gap* and are shown in Figure 9.16b,c.

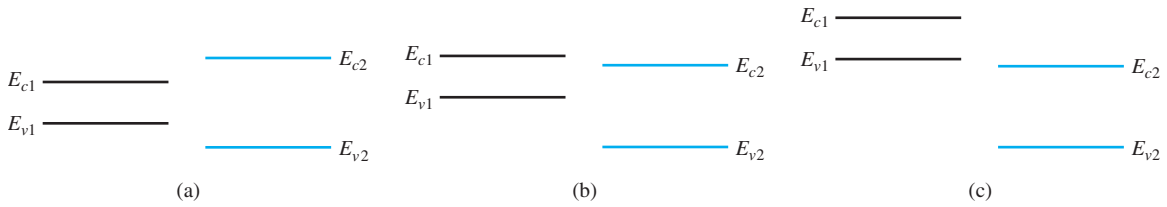


Figure 9.16 | Relation between narrow-bandgap and wide-bandgap energies: (a) straddling, (b) staggered, and (c) broken gap.

There are four basic types of heterojunction. Those in which the dopant type changes at the junction are called *anisotype*. We can form nP or Np junctions, where the capital letter indicates the larger-bandgap material. Heterojunctions with the same dopant type on either side of the junction are called *isotype*. We can form nN and pP isotype heterojunctions.

Figure 9.17 shows the energy-band diagrams of isolated n-type and P-type materials, with the vacuum level used as a reference. The electron affinity of the wide-bandgap material is less than that of the narrow-bandgap material. The difference between the two conduction band energies is denoted by ΔE_c, and the difference between the two valence band energies is denoted by ΔE_v. From Figure 9.17, we can see that

$$\Delta E_c = e(\chi_n - \chi_P) \tag{9.34a}$$

and

$$\Delta E_c + \Delta E_v = E_{gP} - E_{gn} = \Delta E_g \tag{9.34b}$$

In the ideal abrupt heterojunction using nondegenerately doped semiconductors, the vacuum level is parallel to both conduction bands and valence bands. If the vacuum level is continuous, then the same ΔE_c and ΔE_v discontinuities will exist at the

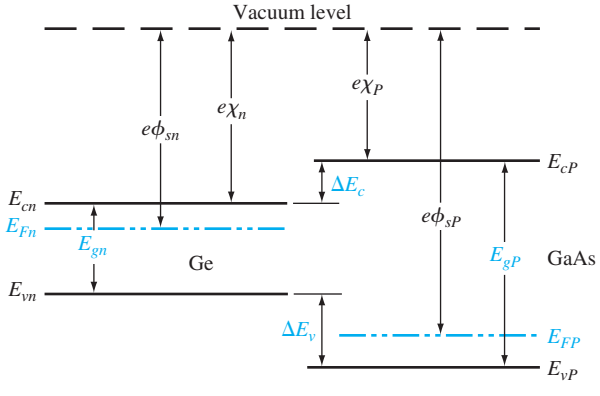


Figure 9.17 | Energy-band diagrams of a narrow-bandgap and a wide-bandgap material before contact.

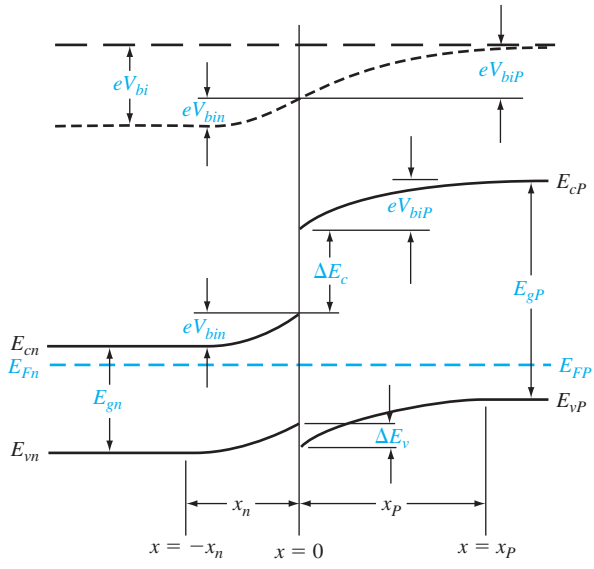


Figure 9.18 | Ideal energy-band diagram of an nP heterojunction in thermal equilibrium.

heterojunction interface. This ideal situation is known as the *electron affinity rule*. There is still some uncertainty about the applicability of this rule, but it provides a good starting point for the discussion of heterojunctions.

Figure 9.18 shows a general ideal nP heterojunction in thermal equilibrium. In order for the Fermi levels in the two materials to become aligned, electrons from the narrow-gap n region and holes from the wide-gap P region must flow across the junction. As in the case of a homojunction, this flow of charge creates a space charge region in the vicinity of the metallurgical junction. The space charge width into the n-type region is denoted by x_n and the space charge width into the P-type region is denoted by x_P. The discontinuities in the conduction and valence bands and the change in the vacuum level are shown in the figure.

9.3.3 Two-Dimensional Electron Gas

Before we consider the electrostatics of the heterojunction, we will discuss a unique characteristic of an isotype junction. Figure 9.19 shows the energy-band diagram of an nN GaAs–AlGaAs heterojunction in thermal equilibrium. The AlGaAs can be moderately to heavily doped n type, while the GaAs can be more lightly doped or even intrinsic. As mentioned previously, to achieve thermal equilibrium, electrons from the wide-bandgap AlGaAs flow into the GaAs, forming an accumulation layer of electrons in the potential well adjacent to the interface. One basic quantum-mechanical result that we have found previously is that the energy of an electron contained in a potential well is quantized. The phrase *two-dimensional electron gas*

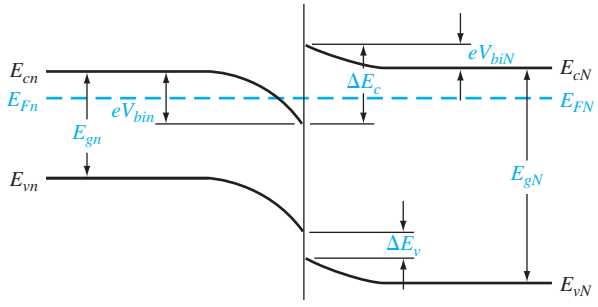


Figure 9.19 | Ideal energy-band diagram of an nN heterojunction in thermal equilibrium.

refers to the condition in which the electrons have quantized energy levels in one spatial direction (perpendicular to the interface), but are free to move in the other two spatial directions.

The potential function near the interface can be approximated by a triangular potential well. Figure 9.20a shows the conduction band edges near the abrupt junction interface and Figure 9.20b shows the approximation of the triangular potential well. We can write

$$V(x) = eEz \qquad z > 0 \tag{9.35a}$$

$$V(z) = \infty \qquad z < 0 \tag{9.35b}$$

Schrodinger's wave equation can be solved using this potential function. The quantized energy levels are shown in Figure 9.20b. Higher energy levels are usually not considered.

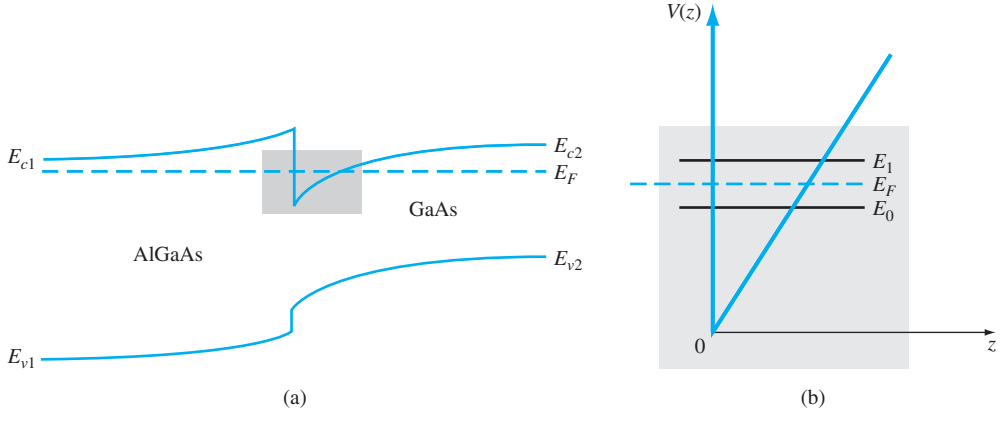


Figure 9.20 | (a) Conduction-band edge at N-AlGaAs, n-GaAs heterojunction; (b) triangular well approximation with discrete electron energies.

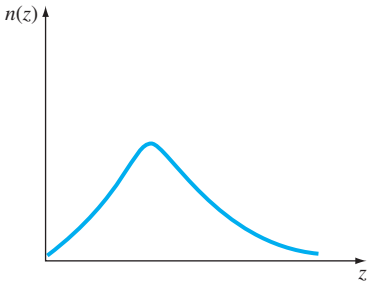


Figure 9.21 | Electron density in triangular potential well.

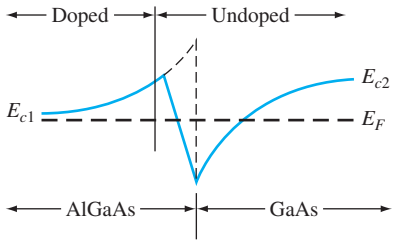


Figure 9.22 | Conduction-band edge at a graded heterojunction.

The qualitative distribution of electrons in the potential well is shown in Figure 9.21. A current parallel to the interface will be a function of this electron concentration and of the electron mobility. Since the GaAs can be lightly doped or intrinsic, the two-dimensional electron gas is in a region of low impurity doping so that impurity scattering effects are minimized. The electron mobility will be much larger than if the electrons were in the same region as the ionized donors.

The movement of the electrons parallel to the interface will still be influenced by the coulomb attraction of the ionized impurities in the AlGaAs. The effect of these forces can be further reduced by using a graded AlGaAs–GaAs heterojunction. The graded layer is $Al_xGa_{1-x}As$ in which the mole fraction x varies with distance. In this case, an intrinsic layer of graded AlGaAs can be sandwiched between the N-type AlGaAs and the intrinsic GaAs. Figure 9.22 shows the conduction-band edges across a graded AlGaAs–GaAs heterojunction in thermal equilibrium. The electrons in the potential well are further separated from the ionized impurities so that the electron mobility is increased above that in an abrupt heterojunction.

*9.3.4 Equilibrium Electrostatics

We now consider the electrostatics of the nP heterojunction that is shown in Figure 9.18. As in the case of the homojunction, potential differences exist across the space charge regions in both the n region and the P region. These potential differences correspond to the built-in potential barriers on either side of the junction. The built-in potential barrier for this ideal case is defined as shown in Figure 9.18 to be the potential difference across the vacuum level. The built-in potential barrier is the sum of the potential differences across each of the space charge regions. The heterojunction built-in potential barrier, however, is not equal to the difference between the conduction bands across the junction or the difference between the valence bands across the junction, as we defined for the homojunction.

Ideally, the total built-in potential barrier V_{bi} can be found as the difference between the work functions, or

$$V_{bi} = \phi_{sP} - \phi_{sn} \tag{9.36}$$

Equation (9.36), from Figure 9.17, can be written as

$$eV_{bi} = [e\chi_P + E_{gP} - (E_{FP} - E_{vP})] - [e\chi_n + E_{gn} - (E_{Fn} - E_{vn})] \quad (9.37a)$$

or

$$eV_{bi} = e(\chi_P - \chi_n) + (E_{gP} - E_{gn}) + (E_{Fn} - E_{vn}) - (E_{FP} - E_{vP}) \quad (9.37b)$$

which can be expressed as

$$eV_{bi} = -\Delta E_c + \Delta E_g + kT \ln\left(\frac{N_{vn}}{p_{no}}\right) - kT \ln\left(\frac{N_{vP}}{p_{po}}\right) \quad (9.38)$$

Finally, we can write Equation (9.38) as

$$eV_{bi} = \Delta E_v + kT \ln\left(\frac{p_{po}}{p_{no}} \cdot \frac{N_{vn}}{N_{vP}}\right) \quad (9.39)$$

where p_{po} and p_{no} are the hole concentrations in the P and n materials, respectively, and N_{vn} and N_{vP} are the effective density of states functions in the n and P materials, respectively. We can also obtain an expression for the built-in potential barrier in terms of the conduction band shift as

$$eV_{bi} = -\Delta E_c + kT \ln\left(\frac{n_{no}}{n_{po}} \cdot \frac{N_{cP}}{N_{cn}}\right) \quad (9.40)$$

EXAMPLE 9.8

Objective: Determine ΔE_c, ΔE_v, and V_{bi} for an n–Ge to P–GaAs heterojunction using the electron affinity rule.

Consider n-type Ge doped with $N_d = 10^{16}$ cm^{-3} and P-type GaAs doped with $N_a = 10^{16}$ cm^{-3}. Let $T = 300$ K so that $n_i = 2.4 \times 10^{13}$ cm^{-3} for Ge.

■ Solution

From Equation (9.34a), we have

$$\Delta E_c = e(\chi_n - \chi_P) = e(4.13 - 4.07) = 0.06 \text{ eV}$$

and from Equation (9.34b), we have

$$\Delta E_v = \Delta E_g - \Delta E_c = (1.43 - 0.67) - 0.06 = 0.70 \text{ eV}$$

To determine V_{bi} using Equation (9.39), we need to determine p_{no} in Ge, or

$$p_{no} = \frac{n_i^2}{N_d} = \frac{(2.4 \times 10^{13})^2}{10^{16}} = 5.76 \times 10^{10} \text{ cm}^{-3}$$

Then

$$eV_{bi} = 0.70 + (0.0259) \ln\left[\frac{(10^{16})(6 \times 10^{18})}{(5.76 \times 10^{10})(7 \times 10^{18})}\right]$$

or, finally,

$$V_{bi} \approx 1.0 \text{ V}$$

■ Comment

There is a nonsymmetry in the ΔE_c and ΔE_v values that will tend to make the potential barriers seen by electrons and holes different. This nonsymmetry does not occur in homojunctions.

EXERCISE PROBLEM

Ex 9.8 Repeat Example 9.8 for an n-Ge-to-P-GaAs heterojunction. The Ge is doped with $N_d = 10^{15}$ cm^{-3} donors and the GaAs doped with $N_a = 10^{15}$ cm^{-3} acceptors. Let $T = 300$ K. (Λ 689.0 = $^{ig}\Lambda$ ·suA)

We can determine the electric field and potential in the junction from Poisson's equation in exactly the same way as we do for the homojunction. For homogeneous doping on each side of the junction, we have in the n region

$$\mathrm{E}_n = \frac{eN_{dn}}{\epsilon_n}(x_n + x) \qquad (-x_n \le x < 0) \tag{9.41a}$$

and in the P region

$$\mathrm{E}_P = \frac{eN_{aP}}{\epsilon_P}(x_P - x) \qquad (0 < x \le x_P) \tag{9.41b}$$

where ϵ_n and ϵ_P are the permittivities of the n and P materials, respectively. We may note that $\mathrm{E}_n = 0$ at $x = -x_n$ and $\mathrm{E}_P = 0$ at $x = x_P$. The electric flux density D is continuous across the junction, so

$$\epsilon_n \mathrm{E}_n(x = 0) = \epsilon_P \mathrm{E}_P(x = 0) \tag{9.42a}$$

which gives

$$N_{dn}x_n = N_{aP}x_P \tag{9.42b}$$

Equation (9.42b) simply states that the net negative charge in the P region is equal to the net positive charge in the n region—the same condition we had in a pn homojunction. We are neglecting any interface states that may exist at the heterojunction.

The electric potential can be found by integrating the electric field through the space charge region so that the potential difference across each region can be determined. We find that

$$V_{bin} = \frac{eN_{dn}x_n^2}{2\epsilon_n} \tag{9.43a}$$

and

$$V_{biP} = \frac{eN_{aP}x_P^2}{2\epsilon_P} \tag{9.43b}$$

Equation (9.42b) can be rewritten as

$$\frac{x_n}{x_P} = \frac{N_{aP}}{N_{dn}} \tag{9.44}$$

The ratio of the built-in potential barriers can then be determined as

$$\frac{V_{bin}}{V_{biP}} = \frac{\epsilon_P}{\epsilon_n} \cdot \frac{N_{dn}}{N_{aP}} \cdot \frac{x_n^2}{x_P^2} = \frac{\epsilon_P N_{aP}}{\epsilon_n N_{dn}} \tag{9.45}$$

Assuming that ϵ_n and ϵ_P are of the same order of magnitude, the larger potential difference is across the lower-doped region.

The total built-in potential barrier is

$$V_{bi} = V_{bin} + V_{biP} = \frac{eN_{dn}x_n^2}{2\epsilon_n} + \frac{eN_{aP}x_P^2}{2\epsilon_P} \tag{9.46}$$

If we solve for x_P, for example, from Equation (9.42b) and substitute into Equation (9.46), we can solve for x_n as

$$x_n = \left[\frac{2\epsilon_n\epsilon_P N_{aP}V_{bi}}{eN_{dn}(\epsilon_n N_{dn} + \epsilon_P N_{aP})}\right]^{1/2} \tag{9.47a}$$

We can also find

$$x_P = \left[\frac{2\epsilon_n\epsilon_P N_{dn}V_{bi}}{eN_{aP}(\epsilon_n N_{dn} + \epsilon_P N_{aP})}\right]^{1/2} \tag{9.47b}$$

The total depletion width is found to be

$$W = x_n + x_P = \left[\frac{2\epsilon_n\epsilon_P(N_{dn} + N_{aP})^2 V_{bi}}{eN_{dn}N_{aP}(\epsilon_n N_{dn} + \epsilon_P N_{aP})}\right]^{1/2} \tag{9.48}$$

If a reverse-biased voltage is applied across the heterojunction, the same equations apply if V_{bi} is replaced by $V_{bi} + V_R$. Similarly, if a forward bias is applied, the same equations also apply if V_{bi} is replaced by $V_{bi} - V_a$. As explained earlier, V_R is the magnitude of the reverse-biased voltage and V_a is the magnitude of the forward-bias voltage.

As in the case of a homojunction, a change in depletion width with a change in junction voltage yields a junction capacitance. We can find for the nP junction

$$C_j' = \left[\frac{eN_{dn}N_{aP}\epsilon_n\epsilon_P}{2(\epsilon_n N_{dn} + \epsilon_P N_{aP})(V_{bi} + V_R)}\right]^{1/2} \qquad (\text{F/cm}^2) \tag{9.49}$$

A plot of $(1/C_j')^2$ versus V_R again yields a straight line. The extrapolation of this plot of $(1/C_j')^2 = 0$ is used to find the built-in potential barrier, V_{bi}.

Figure 9.18 shows the ideal energy-band diagram for the nP abrupt heterojunction. The experimentally determined values of ΔE_c and ΔE_v may differ from the ideal values determined using the electron affinity rule. One possible explanation for this difference is that most heterojunctions have interface states. If we assume that the electrostatic potential is continuous through the junction, then the electric flux density will be discontinuous at the heterojunction due to the surface charge trapped in the interface states. The interface states will then change the energy-band diagram of the semiconductor heterojunction just as they changed the energy-band diagram of the metal–semiconductor junction. Another possible explanation for the deviation from the ideal is that as the two materials are brought together to form the heterojunction, the electron orbitals of each material begin to interact with each other, resulting in a transition region of a few angstroms at the interface. The energy bandgap is then continuous through this transition region and not a characteristic of either material. However, we still have the relation

$$\Delta E_c + \Delta E_v = \Delta E_g \tag{9.50}$$

for the straddling type of heterojunction, although the ΔE_c and ΔE_v values may differ from those determined from the electron affinity rule.

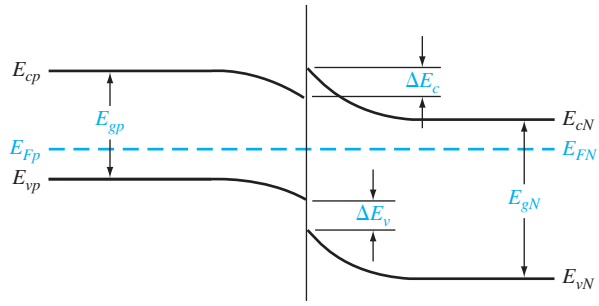


Figure 9.23 | Ideal energy-band diagram of an Np heterojunction in thermal equilibrium.

We may consider the general characteristics of the energy-band diagrams of the other types of heterojunction. Figure 9.23 shows the energy-band diagram of an Np heterojunction. The same ΔE_c and ΔE_v discontinuities exist, although the general shape of the conduction band, for example, is different in the nP and the Np junctions. This difference in energy bands will influence the *I–V* characteristics of the two junctions.

The other two types of heterojunctions are the nN and the pP isotype junctions. The energy-band diagram of the nN junction is shown in Figure 9.19. To achieve thermal equilibrium, electrons from the wide-bandgap material will flow into the narrow-bandgap material. A positive space charge region exists in the wide-bandgap material and an accumulation layer of electrons now exists at the interface in the narrow-bandgap material. Since there are a large number of allowed energy states in the conduction band, we expect the space charge width x_n and the built-in potential barrier V_{bin} to be small in the narrow-gap material. The energy-band diagram of the pP heterojunction in thermal equilibrium is shown in Figure 9.24. To achieve thermal equilibrium, holes from the wide-bandgap material will flow into the narrow-bandgap material, creating an accumulation layer of holes in the narrow-bandgap material at the interface. These types of isotype heterojunctions are obviously not possible in a homojunction.

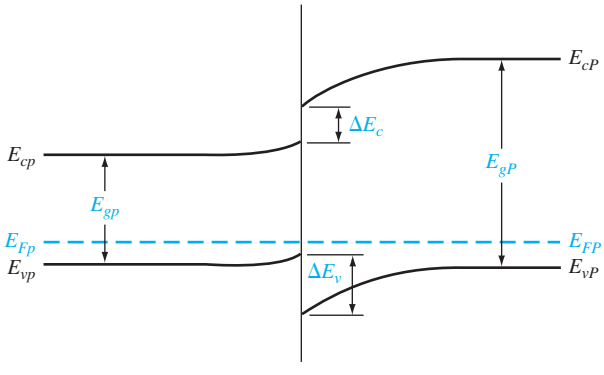


Figure 9.24 | Ideal energy-band diagram of a pP heterojunction in thermal equilibrium.

*9.3.5 Current–Voltage Characteristics

The ideal current–voltage characteristics of a pn homojunction have been developed in Chapter 8. Since the energy-band diagram of a heterojunction is more complicated than that of a homojunction, we would expect the *I–V* characteristics of the two junctions to differ.

One immediate difference between a homojunction and a heterojunction is in the barrier heights seen by the electrons and holes. Since the built-in potential barrier for electrons and holes in a homojunction is the same, the relative magnitude of the electron and hole currents is determined by the relative doping levels. In a heterojunction, the barrier heights seen by electrons and holes are not the same. The energy-band diagrams in Figures 9.18 and 9.23 demonstrate that the barrier heights for electrons and holes in a heterojunction can be significantly different. The barrier height for electrons in Figure 9.18 is larger than that for holes, so we would expect the current due to electrons to be insignificant compared to the hole current. If the barrier height for electrons is 0.2 eV larger than that for holes, the electron current will be approximately a factor of 10^4 smaller than the hole current, assuming all other parameters are equal. The opposite situation exists for the band diagram shown in Figure 9.23.

The conduction-band edge in Figure 9.23 and the valence-band edge in Figure 9.18 are somewhat similar to that of a rectifying metal–semiconductor contact. We derive the current–voltage characteristics of a heterojunction, in general, on the basis of thermionic emission of carriers over the barrier, as we do in the case of metal–semiconductor junction. We can then write

$$J = A^* T^2 \exp\left(\frac{-E_w}{kT}\right) \tag{9.51}$$

where E_w is an effective barrier height. The barrier height can be increased or reduced by an applied potential across the junction as in the case of a pn homojunction or a Schottky barrier junction. The heterojunction *I–V* characteristics, however, may need to be modified to include diffusion effects and tunneling effects. Another complicating factor is that the effective mass of a carrier changes from one side of the junction to the other. Although the actual derivation of the *I–V* relationship of the heterojunction is complex, the general form of the *I–V* equation is still similar to that of a Schottky barrier diode and is generally dominated by one type of carrier.

9.4 | SUMMARY

- A metal on a lightly doped semiconductor can produce a rectifying contact that is known as a Schottky barrier diode. The ideal barrier height between the metal and semiconductor is the difference between the metal work function and the semiconductor electron affinity.
- When a positive voltage is applied to an n-type semiconductor with respect to the metal (reverse bias), the barrier between the semiconductor and metal increases so that there is essentially no flow of charged carriers. When a positive voltage is applied to the metal with respect to an n-type semiconductor (forward bias), the barrier between the semiconductor and metal is lowered so that electrons can easily flow from the semiconductor into the metal by a process called thermionic emission.

- The ideal current–voltage relationship of the Schottky barrier diode is the same as that of the pn junction diode. However, since the current mechanism is different from that of the pn junction diode, the switching speed of the Schottky diode is faster. In addition, the reverse saturation current of the Schottky diode is larger than that of the pn junction diode, so a Schottky diode requires less forward bias voltage to achieve a given current compared to a pn junction diode.
- Metal–semiconductor junctions can also form ohmic contacts, which are low-resistance junctions providing conduction in both directions with very little voltage drop across the junction.
- Semiconductor heterojunctions are formed between two semiconductor materials with different bandgap energies. One useful property of a heterojunction is the creation of a potential well at the interface. Electrons are confined to the potential well in the direction perpendicular to the interface, but are free to move in the other two directions.

GLOSSARY OF IMPORTANT TERMS

anisotype junction A heterojunction in which the type of dopant changes at the metallurgical junction.

electron affinity rule The rule stating that, in an ideal heterojunction, the discontinuity at the conduction band is the difference between the electron affinities in the two semiconductors.

heterojunction The junction formed by the contact between two different semiconductor materials.

image force–induced lowering The lowering of the peak potential barrier at the metal–semiconductor junction due to an electric field.

isotype junction A heterojunction in which the type of dopant is the same on both sides of the junction.

ohmic contact A low-resistance metal–semiconductor contact providing conduction in both directions between the metal and semiconductor.

Richardson constant The parameter A^* in the current–voltage relation of a Schottky diode.

Schottky barrier height The potential barrier ϕ_{Bn} from the metal to semiconductor in a metal–semiconductor junction.

Schottky effect Another term for image force–induced lowering.

specific contact resistance The inverse of the slope of the J versus V curve of a metal–semiconductor contact evaluated at $V = 0$.

thermionic emission The process by which charge flows over a potential barrier as a result of carriers with sufficient thermal energy.

tunneling barrier A thin potential barrier in which the current is dominated by the tunneling of carriers through the barrier.

two-dimensional electron gas (2-DEG) The accumulation layer of electrons contained in a potential well at a heterojunction interface. The electrons are free to move in the "other" two spatial directions.

CHECKPOINT

After studying this chapter, the reader should have the ability to:

- Sketch the energy-band diagram of zero-biased, reverse-biased, and forward-biased Schottky barrier diodes.
- Describe the charge flow in a forward-biased Schottky barrier diode.

- Explain the Schottky barrier lowering and its effect on the reverse saturation current in a Schottky barrier diode.
- Explain the effect of interface states on the characteristics of a Schottky barrier diode.
- Describe one effect of a larger reverse saturation current in a Schottky barrier diode compared to that of a pn junction diode.
- Describe what is meant by an ohmic contact.
- Draw the energy-band diagram of an nN heterojunction.
- Explain what is meant by a two-dimensional electron gas.

REVIEW QUESTIONS

1. What is the ideal Schottky barrier height? Indicate the Schottky barrier height on an energy-band diagram.
2. Using an energy-band diagram, indicate the effect of the Schottky barrier lowering.
3. What is the mechanism of charge flow in a forward-biased Schottky barrier diode?
4. Compare the forward-biased current–voltage characteristic of a Schottky barrier diode to that of pn junction diode.
5. Explain the difference in switching characteristics between a Schottky diode and a pn junction diode. Discuss charge storage effects.
6. Sketch the ideal energy-band diagram of a metal–semiconductor junction in which $\phi_m < \phi_s$. Explain why this is an ohmic contact.
7. Sketch the energy-band diagram of a tunneling junction. Why is this an ohmic contact?
8. What is a heterojunction?
9. What is a 2-D electron gas?

PROBLEMS

(In the following problems, assume $A^* = 120$ A/K^2-cm^2 for silicon and $A^* = 1.12$ A/K^2-cm^2 for gallium arsenide Schottky diodes unless otherwise stated.)

Section 9. 1 The Schottky Barrier Diode

9.1 Consider a contact between Al and n Si doped at $N_d = 10^{16}$ cm^{-3}. $T = 300$ K. (*a*) Draw the energy-band diagrams of the two materials before the junction is formed. (*b*) Draw the ideal energy band at zero bias after the junction is formed. (*c*) Calculate ϕ_{B0}, x_d, and E_{max} for part (*b*). (*d*) Repeat parts (*b*) and (*c*) using the data in Figure 9.5.

9.2 (*a*) A Schottky barrier diode formed on n-type silicon has a doping concentration of $N_d = 5 \times 10^{15}$ cm^{-3} and a barrier height of $\phi_{B0} = 0.65$ V. Determine the built-in potential barrier V_{bi}. (*b*) If the doping concentration changes to $N_d = 10^{16}$ cm^{-3}, determine the values of ϕ_{B0} and V_{bi}. Do these values increase, decrease, or remain the same? (*c*) Repeat part (*b*) if the doping concentration is $N_d = 10^{15}$ cm^{-3}.

9.3 Gold is deposited on n-type silicon forming an ideal rectifying junction. The doping concentration is $N_d = 10^{16}$ cm^{-3}. Assume $T = 300$ K. Determine the theoretical values of (*a*) ϕ_{B0}, (*b*) V_{bi}, and (*c*) x_n and $|E_{max}|$ at (*i*) $V_R = 1$ V and (*ii*) $V_R = 5$ V.

9.4 A Schottky diode is formed by depositing gold on n-type GaAs that is doped at a concentration of $N_d = 5 \times 10^{15}$ cm^{-3}. For $T = 300$ K, determine the theoretical values of (*a*) ϕ_{B0}, (*b*) ϕ_n, (*c*) V_{bi}, and (*d*) x_n and $|E_{max}|$ at (*i*) $V_R = 1$ V and (*ii*) $V_R = 5$ V.

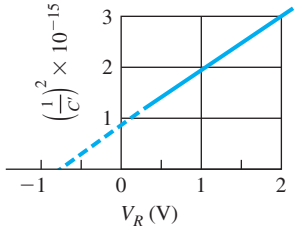


Figure P9.7 | Figure for Problem 9.7.

9.5 Repeat Problem 9.4, parts (*b*) through (*d*), if the experimentally determined barrier height is found to be $\phi_{Bn} = 0.88$ V.

9.6 (*a*) A Pt–n-type silicon junction with $N_d = 10^{15}$ cm^{-3} has a cross-sectional area of $A = 10^{-4}$ cm^2. Let $T = 300$ K. Using the data shown in Figure 9.5, determine the junction capacitance at (*i*) $V_R = 1$ V and (*ii*) $V_R = 5$ V. (*b*) Repeat part (*a*) for a doping concentration of $N_d = 10^{16}$ cm^{-3}.

9.7 A Schottky diode with n-type GaAs at $T = 300$ K yields the $1/C'^2$ versus V_R plot shown in Figure P9.7, where C' is the capacitance per cm^2. Determine (*a*) V_{bi}, (*b*) N_d, (*c*) ϕ_n, and (*d*) ϕ_{B0}.

9.8 Consider a W–n-type silicon Schottky barrier at $T = 300$ K with $N_d = 5 \times 10^{15}$ cm^{-3}. Use the data in Figure 9.5 to determine the barrier height. (*a*) Determine V_{bi}, x_n, and $|E_{max}|$ for (*i*) $V_R = 1$ V and (*ii*) $V_R = 5$ V. (*b*) Using the values of $|E_{max}|$ from part (*a*), determine the Schottky barrier lowering parameters $\Delta\phi$ and x_m.

9.9 Starting with Equation (9.12), derive Equations (9.14) and (9.15).

9.10 A Au-n-GaAs Schottky diode at $T = 300$ K is doped at a concentration of $N_d = 10^{16}$ cm^{-3}. Use the data in Figure 9.5 to determine the barrier height. Then determine (*a*) V_{bi}, x_n, and $|E_{max}|$ for zero bias. (*b*) Determine the reverse-biased voltage at which the Schottky barrier lowering, $\Delta\phi$, will be 5 percent of the barrier height.

9.11 Consider n-type silicon doped at $N_d = 10^{16}$ cm^{-3} with a gold contact to form a Schottky diode. Investigate the effect of Schottky barrier lowering. (*a*) Plot the Schottky barrier lowering $\Delta\phi$ versus reverse-biased voltage over the range $0 \le V_R \le 50$ V. (*b*) Plot the ratio $J_{sT}(V_R)/J_{sT}(V_R = 0)$ over the same range of reverse-biased voltage.

***9.12** The energy-band diagram of a Schottky diode is shown in Figure 9.6. Assume the following parameters:

$\phi_m = 5.2$ V	$\phi_n = 0.10$ V	$\phi_0 = 0.60$ V
$E_g = 1.43$ eV	$\delta = 25$ Å	$\epsilon_i = \epsilon_0$
$\epsilon_s = (13.1)\epsilon_0$	$\chi = 4.07$ V	$N_d = 10^{16}$ cm^{-3}
		$D_{it} = 10^{13}$ eV^{-1} cm^{-2}

(*a*) Determine the theoretical barrier height ϕ_{B0} without interface states. (*b*) Determine the barrier height with interface states. (*c*) Repeat parts (*a*) and (*b*) if $\phi_m = 4.5$ V.

*Asterisks next to problems indicate problems that are more difficult.

***9.13** A Schottky barrier diode contains interface states and an interfacial layer. Assume the following parameters:

$$\phi_m = 4.75 \text{ V} \qquad \phi_n = 0.164 \text{ V} \qquad \phi_0 = 0.230 \text{ V}$$
$$E_g = 1.\ 12 \text{ eV} \qquad \delta = 20 \text{ Å} \qquad \epsilon_i = \epsilon_0$$
$$\epsilon_s = (11.7)\epsilon_0 \qquad \chi = 4.01 \text{ V} \qquad N_d = 5 \times 10^{16} \text{ cm}^{-3}$$
$$\phi_{B0} = 0.60 \text{ V}$$

Determine the interface state density, D_{it}, in units of $\text{eV}^{-1}\ \text{cm}^{-2}$.

9.14 A Schottky diode at $T = 300$ K is formed with Pt on n-type silicon with a doping concentration of $N_d = 5 \times 10^{15}\ \text{cm}^{-3}$. The barrier height is found to be $\phi_{Bn} = 0.89$ V. Determine (*a*) ϕ_n, (*b*) V_{bi}, (*c*) J_{sT}, and (*d*) V_a such that $J_n = 5\ \text{A/cm}^2$. (Neglect the barrier lowering effect.)

9.15 (*a*) Consider a Schottky diode at $T = 300$ K that is formed with tungsten on n-type silicon. Use Figure 9.5 to determine the barrier height. Assume a doping concentration of $N_d = 10^{16}\ \text{cm}^{-3}$ and assume a cross-sectional area $A = 10^{-4}\ \text{cm}^2$. Determine the forward-bias voltage required to induce a current of (*i*) 10 μA, (*ii*) 100 μA, and (*iii*) 1 mA. (*b*) Repeat part (*a*) for a temperature of $T = 350$ K. (Neglect the barrier lowering effect.)

9.16 An Au-n-GaAs Schottky diode at $T = 300$ K has a doping concentration of $N_d = 10^{16}\ \text{cm}^{-3}$. (*a*) Using Figure 9.5, determine the barrier height. (*b*) Calculate the reverse-biased saturation current J_{sT}. (*c*) Determine the forward-bias voltage required to induce a current density of $J_n = 10\ \text{A/cm}^2$. (*d*) What is the change in forward-bias voltage necessary to double the current density? (Neglect the Schottky barrier lowering.)

9.17 (*a*) Consider an Au–n-type GaAs Schottky diode with a cross-sectional area of $10^{-4}\ \text{cm}^2$. Plot the forward-bias current–voltage characteristics over a voltage range of $0 \le V_D \le 0.5$ V. Plot the current on a log scale. (*b*) Repeat part (*a*) for an Au–n-type silicon Schottky diode. (*c*) What conclusions can be drawn from these results?

9.18 A Schottky diode at $T = 300$ K is formed between tungsten and n-type silicon doped at $N_d = 10^{16}\ \text{cm}^{-3}$. The cross-sectional area is $A = 10^{-4}\ \text{cm}^2$. Determine the reverse-biased saturation current at (*a*) $V_R = 2$ V and (*b*) $V_R = 4$ V. (Take into account the Schottky barrier lowering.)

***9.19** Starting with the basic current equation given by Equation (9.18), derive the relation given by Equation (9.23).

9.20 The reverse-saturation current densities in a pn junction diode and a Schottky diode are $10^{-11}\ \text{A/cm}^2$ and $6 \times 10^{-8}\ \text{A/cm}^2$, respectively, at $T = 300$ K. The cross-sectional area of the Schottky diode is $A = 10^{-4}\ \text{cm}^2$. The current in each diode is 0.80 mA. The difference in forward-bias voltages between the two diodes is 0.285 V. Determine (*a*) the voltage applied to each diode and (*b*) the cross-sectional area of the pn junction diode.

9.21 A pn junction diode and a Schottky diode each have cross-sectional areas of $A = 8 \times 10^{-4}\ \text{cm}^2$. The reverse saturation current densities at $T = 300$ K for the pn junction diode and Schottky diode are $8 \times 10^{-13}\ \text{A/cm}^2$ and $6 \times 10^{-9}\ \text{A/cm}^2$, respectively. Determine the required forward-bias voltage in each diode to yields currents of (*a*)150 μA, (*b*) 700 μA, and (*c*) 1.2 mA.

9.22 (*a*) The two diodes described in Problem 9.21 are connected in series and are driven by a constant current source of 0.80 mA. Determine (*i*) the current in each diode and

(*ii*) the voltage across each diode. (*b*) Repeat part (*a*) for the case when the diodes are connected in parallel.

9.23 A Schottky diode and a pn junction diode have cross-sectional areas of $A = 7 \times 10^{-4}$ cm^2. The reverse-saturation current densities at $T = 300$ K of the Schottky diode and pn junction are 4×10^{-8} A/cm^2 and 3×10^{-12} A/cm^2, respectively. A forward-bias current of 0.8 mA is required in each diode. (*a*) Determine the forward-bias voltage required across each diode. (*b*) If the voltage from part (*a*) is maintained across each diode, determine the current in each diode if the temperature is increased to 400 K. (Take into account the temperature dependence of the reverse-saturation currents. Assume $E_g = 1.12$ eV for the pn junction diode and $\phi_{B0} = 0.82$ V for the Schottky diode.)

9.24 Compare the current–voltage characteristics of a Schottky barrier diode and a pn junction diode. Use the results of Example 9.5 and assume diode areas of 5×10^{-4} cm^2. Plot the current–voltage characteristics on a linear scale over a current range of $0 \leq I_D \leq 10$ mA.

Section 9.2 Metal–Semiconductor Ohmic Contacts

9.25 The contact resistance of an ohmic contact is $R_c = 10^{-4}$ Ω-cm^2. Determine the junction resistance if the cross-sectional area is (*a*) 10^{-3} cm^2, (*b*) 10^{-4} cm^2, and (*c*) 10^{-5} cm^2.

9.26 (*a*) The contact resistance of an ohmic contact is $R_c = 5 \times 10^{-5}$ Ω-cm^2. The cross-sectional area of the junction is 10^{-5} cm^2. Determine the voltage across the junction if the current is (*i*) $I = 1$ mA and (*ii*) $I = 100$ μA. (*b*) Repeat part (*a*) if the cross-sectional area is 10^{-6} cm^2.

9.27 An ohmic contact between a metal and silicon may be formed that has a very low barrier height. (*a*) Determine the value of ϕ_{Bn} that will produce a contact resistance of $R_c = 5 \times 10^{-5}$ Ω-cm^2 at $T = 300$ K. (*b*) Repeat part (*a*) for a contact resistance of $R_c = 5 \times 10^{-6}$ Ω-cm^2.

9.28 A metal, with a work function $\phi_m = 4.2$ V, is deposited on an n-type silicon semiconductor with $\chi_s = 4.0$ V and $E_g = 1.12$ eV. Assume no interface states exist at the junction. Let $T = 300$ K. (*a*) Sketch the energy-band diagram for zero bias for the case when no space charge region exists at the junction. (*b*) Determine N_d so that the condition in part (*a*) is satisfied. (*c*) What is the potential barrier height seen by electrons in the metal moving into the semiconductor?

9.29 Consider the energy-band diagram of a silicon Schottky junction under zero bias shown in Figure P9.29. Let $\phi_{B0} = 0.7$ V and $T = 300$ K. Determine the doping required so that $x_d = 50$ Å at the point where the potential is $\phi_{B0}/2$ below the peak value. (Neglect the barrier lowering effect.)

9.30 A metal–semiconductor junction is formed between a metal with a work function of 4.3 eV and p-type silicon with an electron affinity of 4.0 eV. The acceptor doping concentration in the silicon is $N_a = 5 \times 10^{16}$ cm^{-3}. Assume $T = 300$ K. (*a*) Sketch the thermal equilibrium energy-band diagram. (*b*) Determine the height of the Schottky barrier. (*c*) Sketch the energy-band diagram with an applied reverse-biased voltage of $V_R = 3$ V. (*d*) Sketch the energy-band diagram with an applied forward-bias voltage of $V_a = 0.25$ V.

9.31 (*a*) Consider a metal–semiconductor junction formed between a metal with a work function of 4.65 eV and Ge with an electron affinity of 4.13 eV. The doping

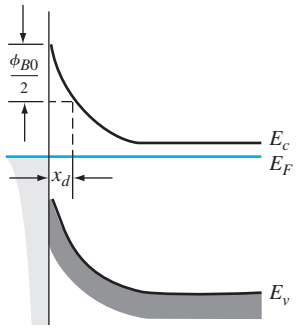


Figure P9.29 | Figure for Problem 9.29.

concentration in the Ge material is $N_d = 6 \times 10^{13}$ cm^{-3} and $N_a = 3 \times 10^{13}$ cm^{-3}. Assume $T = 300$ K. Sketch the zero bias energy-band diagram and determine the Schottky barrier height. (*b*) Repeat part (*a*) if the metal work function is 4.35 eV.

Section 9.3 Heterojunctions

9.32 Sketch the energy-band diagrams of an abrupt $Al_{0.3}Ga_{0.7}As$–GaAs heterojunction for: (*a*) N^+–AlGaAs, intrinsic GaAs, (*b*) N^+–AlGaAs, p–GaAs, and (*c*) P^+–AlGaAs, n^+–GaAs. Assume $E_g = 1.85$ eV for $Al_{0.3}Ga_{0.7}As$ and assume $\Delta E_c = \frac{2}{3}\Delta E_g$.

9.33 Repeat Problem 9.32 assuming the ideal electron affinity rule. Determine ΔE_c and ΔE_v.

***9.34** Starting with Poisson's equation, derive Equation (9.48) for an abrupt heterojunction.

Summary and Review

***9.35** (*a*) Derive an expression for dV_a/dT as a function of current density in a Schottky diode. Assume the minority carrier current is negligible. (*b*) Compare dV_a/dT for a GaAs Schottky diode to that for a Si Schottky diode. (*c*) Compare dV_a/dT for a Si Schottky diode to that for a Si pn junction diode.

9.36 The $(1/C_j)^2$ versus V_R data are measured for two Schottky diodes with equal areas. One diode is fabricated with 1 Ω-cm silicon and the other diode with 5 Ω-cm silicon. The plots intersect the voltage axis as $V_R = -0.5$ V for diode A and at $V_R = -1.0$ V for diode B. The slope of the plot for diode A is 1.5×10^{18} (F^2-V)$^{-1}$ and that for diode B is 1.5×10^{17} (F^2-V)$^{-1}$. Determine which diode has the higher metal work function and which diode has the lower resistivity silicon.

***9.37** Both Schottky barrier diodes and ohmic contacts are to be fabricated by depositing a particular metal on a silicon integrated circuit. The work function of the metal is 4.5 V. Considering the ideal metal–semiconductor contact, determine the allowable range of doping concentrations for each type of contact. Consider both p- and n-type silicon regions.

9.38 Consider an n-GaAs–p-AlGaAs heterojunction in which the bandgap offsets are $\Delta E_c = 0.3$ eV and $\Delta E_v = 0.15$ eV. Discuss the difference in the expected electron and hole currents when the junction is forward biased.

READING LIST

1. Anderson, R. L. "Experiments on Ge–GaAs Heterojunctions." *Solid-State Electronics* 5, no. 5 (September–October 1962), pp. 341–351.
2. Crowley, A. M., and S. M. Sze. "Surface States and Barrier Height of Metal–Semiconductor Systems." *Journal of Applied Physics* 36 (1965), p. 3212.
3. Hu, C. C. *Modern Semiconductor Devices for Integrated Circuits.* Upper Saddle River, NJ: Pearson Prentice Hall, 2010.
4. MacMillan, H. F., H. C. Hamaker, G. F. Virshup, and J. G. Werthen. "Multijunction III-V Solar Cells: Recent and Projected Results." *Twentieth IEEE Photovoltaic Specialists Conference* (1988), pp. 48–54.
5. Michaelson, H. B. "Relation between an Atomic Electronegativity Scale and the Work Function." *IBM Journal of Research and Development* 22, no. 1 (January 1978), pp. 72–80.
6. Pierret, R. F. *Semiconductor Device Fundamentals.* Reading, MA: Addison-Wesley, 1996.
7. Rideout, V. L. "A Review of the Theory, Technology and Applications of Metal–Semiconductor Rectifiers." *Thin Solid Films* 48, no. 3 (February 1, 1978), pp. 261–291.
8. Roulston, D. J. *Bipolar Semiconductor Devices.* New York: McGraw-Hill, 1990.

*9. Shur, M. *GaAs Devices and Circuits.* New York: Plenum Press, 1987.

10. ———. *Introduction to Electronic Devices.* New York: John Wiley and Sons, 1996.

*11. ———. *Physics of Semiconductor Devices.* Englewood Cliffs, NJ: Prentice Hall, 1990.

*12. Singh, J. *Physics of Semiconductors and Their Heterostructures.* New York: McGraw-Hill, 1993.

13. ———. *Semiconductor Devices: Basic Principles.* New York: John Wiley and Sons, 2001.
14. Streetman, B. G., and S. K. Banerjee. *Solid State Electronic Devices.* 6th ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2006.
15. Sze, S. M., and K. K. Ng. *Physics of Semiconductor Devices,* 3rd ed. Hoboken, NJ: John Wiley and Sons, 2007.

*16. Wang, S. *Fundamentals of Semiconductor Theory and Device Physics.* Englewood Cliffs, NJ: Prentice Hall, 1989.

*17. Wolfe, C. M., N. Holonyak, Jr., and G. E. Stillman. *Physical Properties of Semiconductors.* Englewood Cliffs, NJ: Prentice Hall, 1989.

18. Yang, E. S. *Microelectronic Devices.* New York: McGraw-Hill, 1988.

*19. Yuan, J. S. *SiGe, GaAs, and InP Heterojunction Bipolar Transistors.* New York: John Wiley and Sons, 1999.

*Indicates references that are at an advanced level compared to this text.

CHAPTER 10

Fundamentals of the Metal–Oxide–Semiconductor Field-Effect Transistor

The single-junction semiconductor devices that we have considered, including the pn homojunction diode, can be used to produce rectifying current–voltage characteristics and to form electronic switching circuits. The transistor is a multijunction semiconductor device that, in conjunction with other circuit elements, is capable of current gain, voltage gain, and signal power gain. The basic transistor action is the control of current at one terminal by the voltage applied across the other two terminals of the device.

The **M**etal–**O**xide–**S**emiconductor **F**ield-**E**ffect **T**ransistor (MOSFET) is one of two major types of transistors. The fundamental physics of the MOSFET is developed in this chapter. The MOSFET is used extensively in digital circuit applications where, because of its small size, millions of devices can be fabricated in a single integrated circuit.

Two complementary configurations of MOS transistors, the n-channel MOSFET and the p-channel MOSFET, can be fabricated. Electronic circuit design becomes very versatile when the two types of devices are used in the same circuit. These circuits are referred to as complementary MOS (CMOS) circuits. ■

10.0 | PREVIEW

In this chapter, we will:

- Study the characteristics of energy bands as a function of applied voltage in the metal–oxide–semiconductor structure known as the MOS capacitor. The MOS capacitor is the heart of the MOSFET.
- Discuss the concept of surface inversion in the semiconductor of the MOS capacitor.

- Define and derive the expression for the threshold voltage, which is a basic parameter of the MOSFET.
- Discuss various physical structures of MOSFETs, including enhancement and depletion mode devices.
- Derive the ideal current–voltage relationship of the MOSFET.
- Develop the small-signal equivalent circuit of the MOSFET. This circuit is used to relate small-signal currents and voltages in analog circuits.
- Derive the frequency limiting factors of the MOSFET.

10.1 | THE TWO-TERMINAL MOS STRUCTURE

The heart of the MOSFET is the MOS capacitor shown in Figure 10.1. The metal may be aluminum or some other type of metal, although in many cases, it is actually a high-conductivity polycrystalline silicon that has been deposited on the oxide; however, the term metal is usually still used. The parameter t_{ox} in the figure is the thickness of the oxide and ϵ_{ox} is the permittivity of the oxide.

10.1.1 Energy-Band Diagrams

The physics of the MOS structure can be more easily explained with the aid of the simple parallel-plate capacitor. Figure 10.2a shows a parallel-plate capacitor with the top plate at a negative voltage with respect to the bottom plate. An insulator material separates the two plates. With this bias, a negative charge exists on the top plate, a positive charge exists on the bottom plate, and an electric field is induced between the two plates as shown. The capacitance per unit area for this geometry is

$$C' = \frac{\epsilon}{d} \tag{10.1}$$

where ϵ is the permittivity of the insulator and d is the distance between the two plates. The magnitude of the charge per unit area on either plate is

$$Q' = C'V \tag{10.2}$$

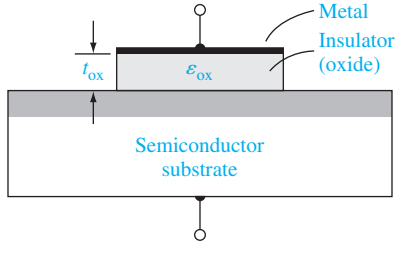


Figure 10.1 | The basic MOS capacitor structure.

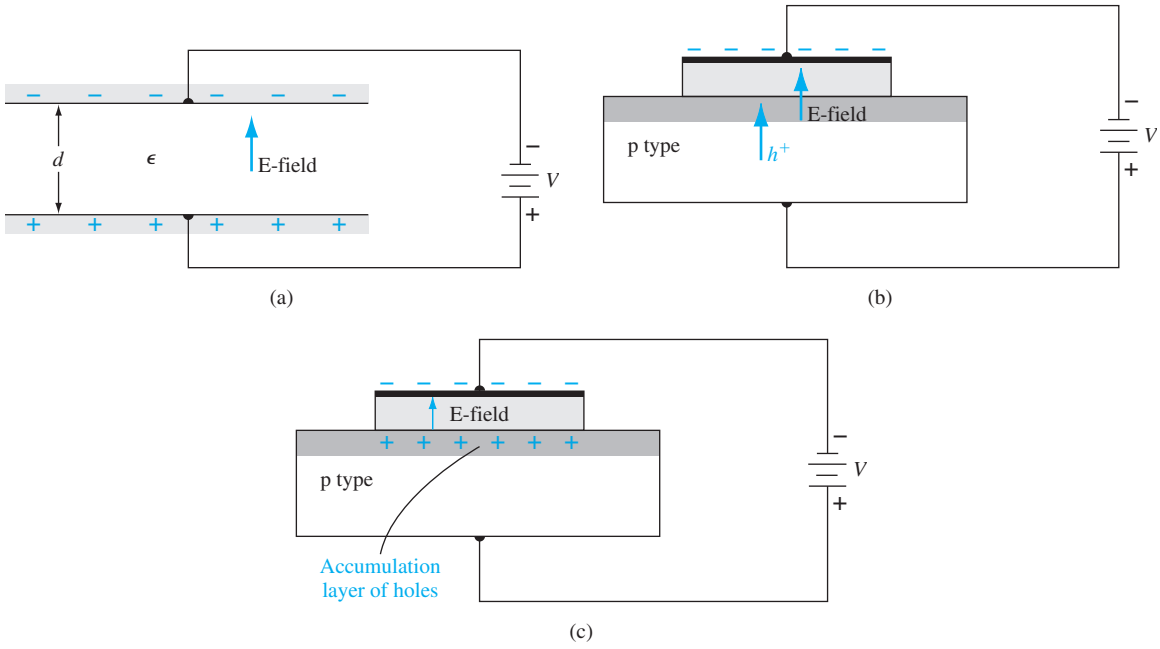


Figure 10.2 | (a) A parallel-plate capacitor showing the electric field and conductor charges. (b) A corresponding MOS capacitor with a negative gate bias showing the electric field and charge flow. (c) The MOS capacitor with an accumulation layer of holes.

where the prime indicates charge or capacitance per unit area. The magnitude of the electric field is

$$\mathrm{E} = \frac{V}{d} \tag{10.3}$$

Figure 10.2b shows a MOS capacitor with a p-type semiconductor substrate. The top metal gate is at a negative voltage with respect to the semiconductor substrate. From the example of the parallel-plate capacitor, we can see that a negative charge will exist on the top metal plate and an electric field will be induced with the direction shown in the figure. If the electric field were to penetrate into the semiconductor, the majority carrier holes would experience a force toward the oxide–semiconductor interface. Figure 10.2c shows the equilibrium distribution of charge in the MOS capacitor with this particular applied voltage. An *accumulation layer* of holes at the oxide–semiconductor junction corresponds to the positive charge on the bottom "plate" of the MOS capacitor.

Figure 10.3a shows the same MOS capacitor in which the polarity of the applied voltage is reversed. A positive charge now exists on the top metal plate and the induced electric field is in the opposite direction as shown. If the electric field penetrates the semiconductor in this case, majority carrier holes will experience a force away from the oxide–semiconductor interface. As the holes are pushed away from

the interface, a negative space charge region is created because of the fixed ionized acceptor atoms. The negative charge in the induced depletion region corresponds to the negative charge on the bottom "plate" of the MOS capacitor. Figure 10.3b shows the equilibrium distribution of charge in the MOS capacitor with this applied voltage.

The energy-band diagrams of the MOS capacitor with a p-type substrate for various gate biases are shown in Figure 10.4. Figure 10.4a shows the ideal case when zero bias is applied across the MOS device. The energy bands in the semiconductor are flat indicating no net charge exists in the semiconductor. This condition is known as *flat band* and is discussed in more detail later in the chapter.

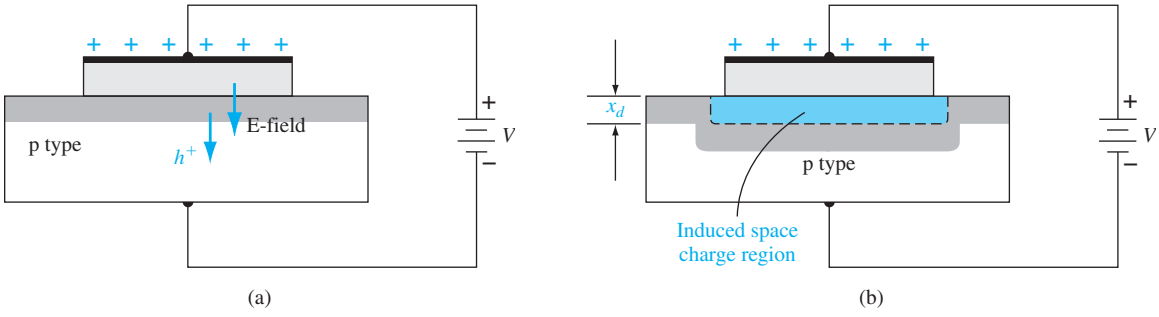


Figure 10.3 | The MOS capacitor with a moderate positive gate bias, showing (a) the electric field and charge flow and (b) the induced space charge region.

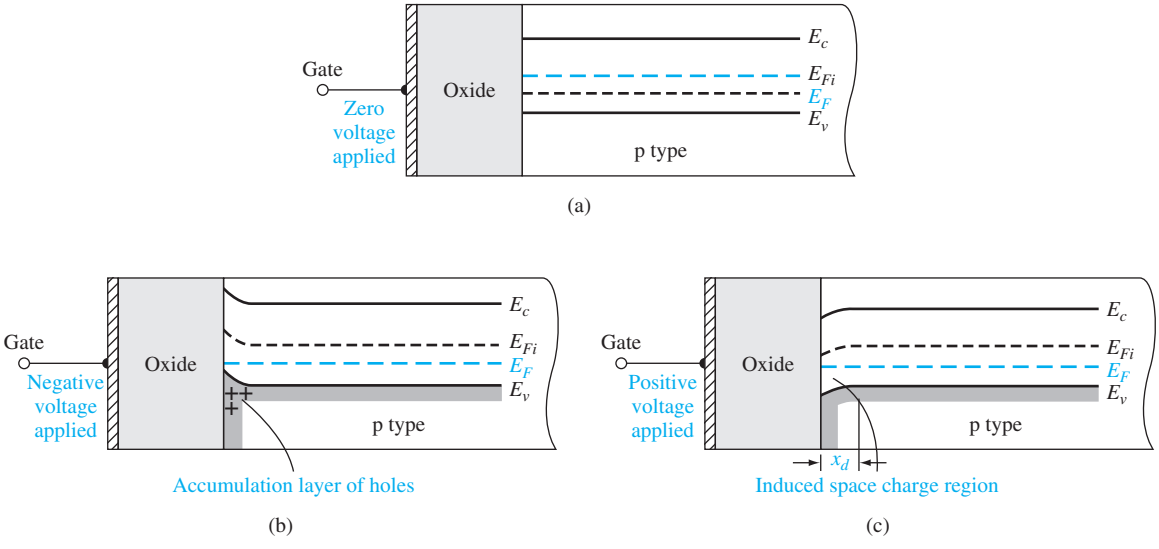


Figure 10.4 | The energy-band diagram of a MOS capacitor with a p-type substrate for (a) a zero applied gate bias showing the *ideal* case, (b) a negative gate bias, and (c) a moderate positive gate bias.

Figure 10.4b shows the energy-band diagram for the case when a negative bias is applied to the gate. (Remember that positive electron energy is plotted "upward" and positive voltage is plotted "downward.") The valence-band edge is closer to the Fermi level at the oxide–semiconductor interface than in the bulk material, which implies that there is an accumulation of holes. The semiconductor surface appears to be more p-type than the bulk material. The Fermi level is a constant in the semiconductor since the MOS system is in thermal equilibrium and there is no current through the oxide.

Figure 10.4c shows the energy-band diagram of the MOS system when a positive voltage is applied to the gate. The conduction- and valence-band edges bend as shown in the figure, indicating a space charge region similar to that in a pn junction. The conduction band and intrinsic Fermi levels move closer to the Fermi level. The induced space charge width is x_d.

Now consider the case when a still larger positive voltage is applied to the top metal gate of the MOS capacitor. We expect the induced electric field to increase in magnitude and the corresponding positive and negative charges on the MOS capacitor to increase. A larger negative charge in the MOS capacitor implies a larger induced space charge region and more band bending. Figure 10.5 shows such a condition. The intrinsic Fermi level at the surface is now below the Fermi level. The conduction band at the surface is now close to the Fermi level, whereas the valence band is close to the Fermi level in the bulk semiconductor. This result implies that the surface in the semiconductor adjacent to the oxide–semiconductor interface is n type. By applying a sufficiently large positive gate voltage, we have inverted the surface of the semiconductor from a p-type to an n-type semiconductor. We have created an *inversion layer* of electrons at the oxide–semiconductor interface.

In the MOS capacitor structure that we have just considered, we assumed a p-type semiconductor substrate. The same type of energy-band diagrams can be constructed for a MOS capacitor with an n-type semiconductor substrate. Figure 10.6a shows the MOS capacitor structure with a positive voltage applied to the top gate terminal. A positive charge exists on the top gate and an electric field is induced with the direction shown in the figure. An accumulation layer of electrons will be induced in the n-type substrate. The case when a negative voltage is applied to the top gate

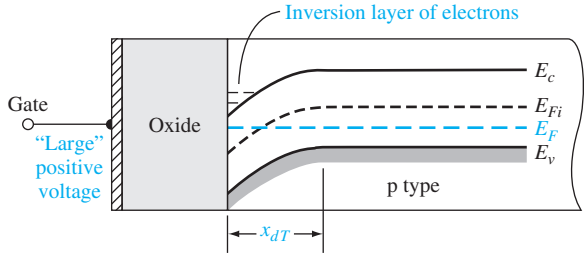


Figure 10.5 | The energy-band diagram of the MOS capacitor with a p-type substrate for a "large" positive gate bias.

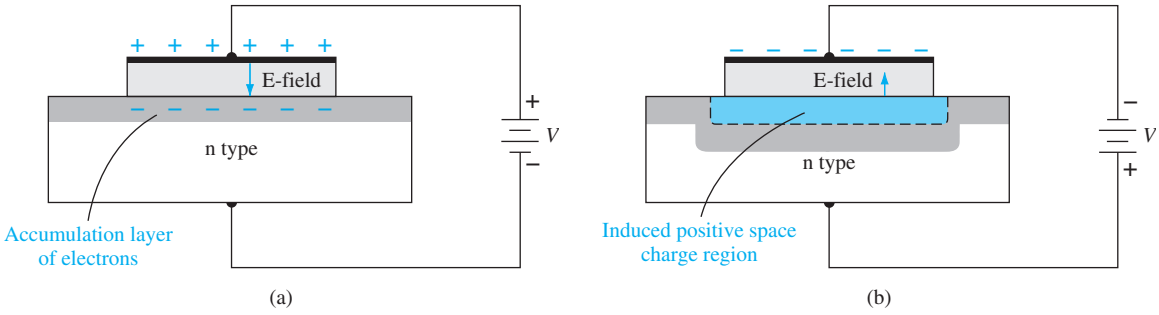


Figure 10.6 | The MOS capacitor with an n-type substrate for (a) a positive gate bias and (b) a moderate negative gate bias.

is shown in Figure 10.6b. A positive space charge region is induced in the n-type semiconductor in this situation.

The energy-band diagrams for this MOS capacitor with the n-type substrate are shown in Figure 10.7. Figure 10.7a shows the case when a positive voltage is applied to the gate and an accumulation layer of electrons is formed. Figure 10.7b shows the energy bands when a negative voltage is applied to the gate. The conduction and valence bands now bend upward indicating that a space charge region has been induced in the n-type substrate. Figure 10.7c shows the energy bands when a larger negative voltage is applied to the gate. The conduction and valence bands are bent even more and the intrinsic Fermi level has moved above the Fermi level. The valence band at the surface is now close to the Fermi level, whereas the conduction band is close to the Fermi level in the bulk semiconductor. This result implies that the semiconductor surface adjacent to the oxide–semiconductor interface is p type. By applying a sufficiently large negative voltage to the gate of the MOS capacitor, the semiconductor surface has been inverted from n type to p type. An inversion layer of holes has been induced at the oxide–semiconductor interface.

10.1.2 Depletion Layer Thickness

We may calculate the width of the induced space charge region adjacent to the oxide–semiconductor interface. Figure 10.8 shows the space charge region in a p-type semiconductor substrate. The potential ϕ_{fp} is the difference (in V) between E_{Fi} and E_F and is given by

$$\phi_{fp} = V_t \ln\left(\frac{N_a}{n_i}\right) \tag{10.4}$$

where N_a is the acceptor doping concentration and n_i is the intrinsic carrier concentration.

The potential ϕ_s is called the surface potential; it is the difference (in V) between E_{Fi} measured in the bulk semiconductor and E_{Fi} measured at the surface. The surface

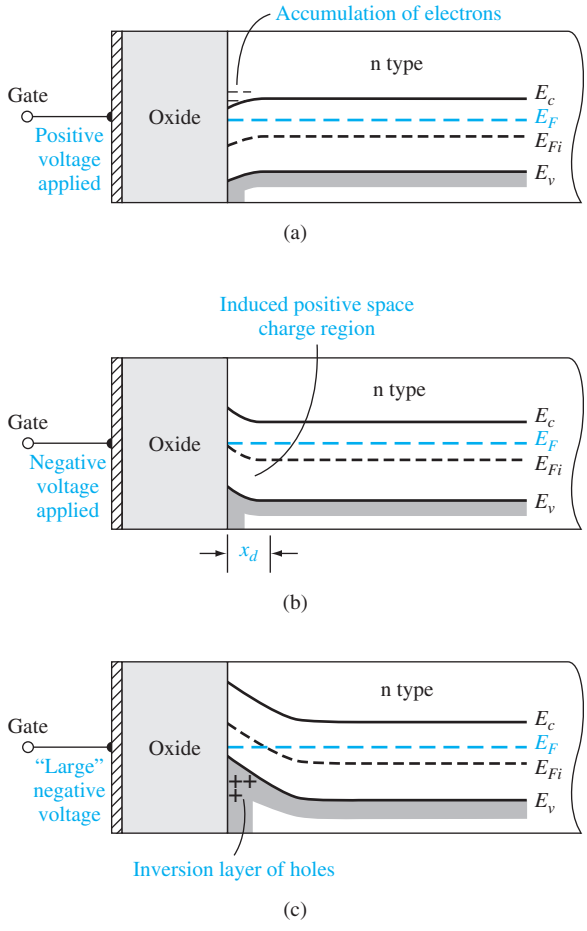


Figure 10.7 | The energy-band diagram of the MOS capacitor with an n-type substrate for (a) a positive gate bias, (b) a moderate negative bias, and (c) a "large" negative gate bias.

potential is the potential difference across the space charge layer. The space charge width can now be written in a form similar to that of a one-sided pn junction. We can write that

$$x_d = \left(\frac{2\epsilon_s \phi_s}{eN_a}\right)^{1/2} \tag{10.5}$$

where ϵ_s is the permittivity of the semiconductor. Equation (10.5) assumes that the abrupt depletion approximation is valid.

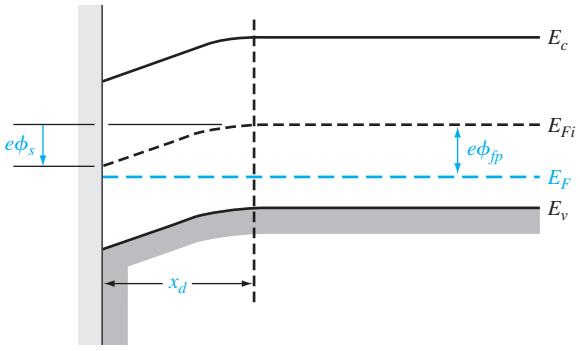


Figure 10.8 | The energy-band diagram in the p-type semiconductor, indicating surface potential.

Figure 10.9 shows the energy bands for the case in which $\phi_s = 2\phi_{fp}$. The Fermi level at the surface is as far above the intrinsic level as the Fermi level is below the intrinsic level in the bulk semiconductor. The electron concentration at the surface is the same as the hole concentration in the bulk material. This condition is known as the *threshold inversion point.* The applied gate voltage creating this condition is known as the *threshold voltage.* If the gate voltage increases above this threshold value, the conduction band will bend slightly closer to the Fermi level, but the change in the conduction band at the surface is now only a slight function of gate voltage. The electron concentration at the surface, however, is an exponential function of the surface potential. The surface potential may increase by a few (kT/e) volts, which will change the electron concentration by orders of magnitude, but the space charge width changes only slightly. In this case, then, the space charge region has essentially reached a maximum width.

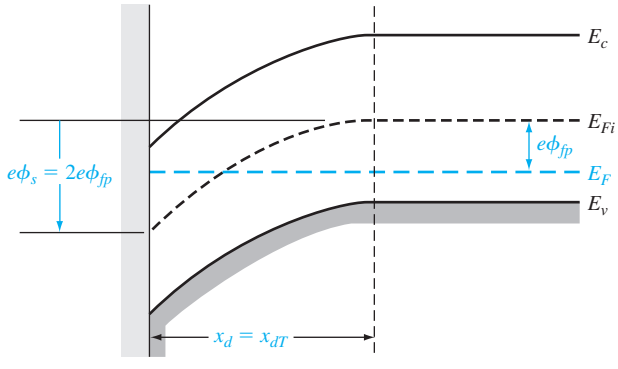


Figure 10.9 | The energy-band diagram in the p-type semiconductor at the threshold inversion point.

The maximum space charge width, x_{dT}, at this inversion transition point can be calculated from Equation (10.5) by setting $\phi_s = 2\phi_{fp}$. Then

$$x_{dT} = \left(\frac{4\epsilon_s \phi_{fp}}{eN_a}\right)^{1/2} \tag{10.6}$$

EXAMPLE 10.1

Objective: Calculate the maximum space charge width for a given semiconductor doping concentration.

Consider silicon at $T = 300$ K doped to $N_a = 10^{16}$ cm^{-3}. The intrinsic carrier concentration is $n_i = 1.5 \times 10^{10}$ cm^{-3}.

■ Solution

From Equation (10.4), we have

$$\phi_{fp} = V_t \ln\left(\frac{N_a}{n_i}\right) = (0.0259)\ln\left(\frac{10^{16}}{1.5 \times 10^{10}}\right) = 0.3473 \text{ V}$$

Then the maximum space charge width is

$$x_{dT} = \left[\frac{4\epsilon_s \phi_{fp}}{eN_a}\right]^{1/2} = \left[\frac{4(11.7)(8.85 \times 10^{-14})(0.3473)}{(1.6 \times 10^{-19})(10^{16})}\right]^{1/2}$$

or

$$x_{dT} \cong 0.30 \times 10^{-4} \text{ cm} = 0.30\ \mu\text{m}$$

■ Comment

The maximum induced space charge width is on the same order of magnitude as pn junction space charge widths.

■ EXERCISE PROBLEM

Ex 10.1 Consider an oxide-to-p-type silicon junction at $T = 300$ K. The impurity doping concentration in the silicon is $N_a = 2 \times 10^{15}$ cm^{-3}. Calculate the maximum space charge width. Does the space charge width increase or decrease as the p-type doping concentration decreases?

(Ans. $x_{dT} = 0.629\ \mu$m, increase)

We have been considering a p-type semiconductor substrate. The same maximum induced space charge region width occurs in an n-type substrate. Figure 10.10 is the energy-band diagram at the threshold voltage with an n-type substrate. We can write

$$\phi_{fn} = V_t \ln\left(\frac{N_d}{n_i}\right) \tag{10.7}$$

and

$$x_{dT} = \left(\frac{4\epsilon_s \phi_{fn}}{eN_d}\right)^{1/2} \tag{10.8}$$

Note that we are always assuming the parameters ϕ_{fp} and ϕ_{fn} to be positive quantities.

Figure 10.11 is a plot of x_{dT} at $T = 300$ K as a function of doping concentration in silicon. The semiconductor doping can be either n type or p type.

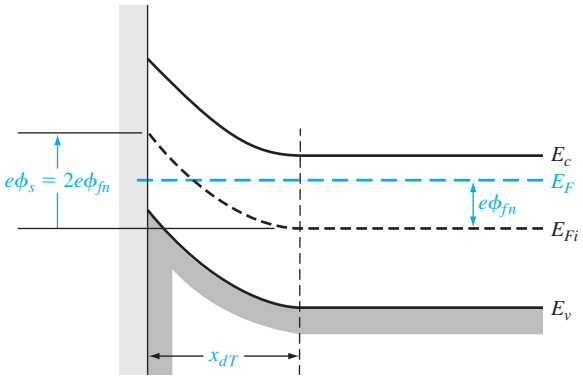


Figure 10.10 | The energy-band diagram in the n-type semiconductor at the threshold inversion point.

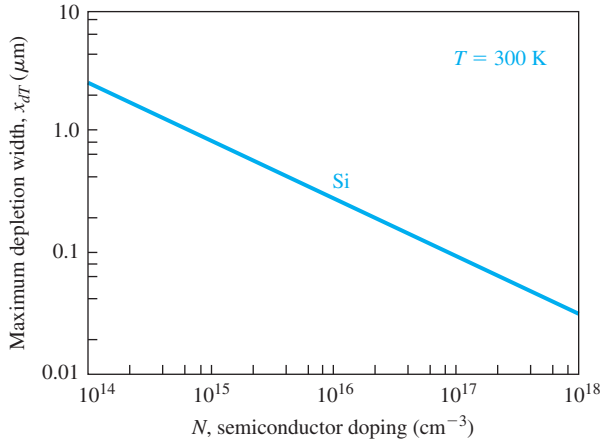


Figure 10.11 | Maximum induced space charge region width versus semiconductor doping.

10.1.3 Surface Charge Density

From the results of Chapter 4, the electron concentration in the conduction band can be written in the form

$$n = n_i \exp\left[\frac{E_F - E_{Fi}}{kT}\right] \tag{10.9}$$

For a p-type semiconductor substrate, the electron inversion charge density can then be written as (see Figure 10.9)

$$n_s = n_i \exp\left[\frac{e(\phi_{fp} + \Delta\phi_s)}{kT}\right] = n_i \exp\left[\frac{\phi_{fp} + \Delta\phi_s}{V_t}\right] \tag{10.10a}$$

or

$$n_s = n_i \exp\left(\frac{\phi_{fp}}{V_t}\right) \cdot \exp\left(\frac{\Delta\phi_s}{V_t}\right) \tag{10.10b}$$

where $\Delta\phi_s$, is the surface potential greater than $2\phi_{fp}$.

We may note that

$$n_{st} = n_i \exp\left(\frac{\phi_{fp}}{V_t}\right) \tag{10.11}$$

where n_{st} is the surface charge density at the threshold inversion point. The electron inversion charge density can then be written as

$$n_s = n_{st} \exp\left(\frac{\Delta\phi_s}{V_t}\right) \tag{10.12}$$

Figure 10.12 shows the electron inversion charge density as a function of surface potential for the case when the threshold inversion charge density is $n_{st} = 10^{16}\ \text{cm}^{-3}$. We may note that the inversion charge density increases by a factor of 10 with a 60-mV increase in surface potential. As discussed previously, the electron inversion charge density increases rapidly with small increases in surface potential, which means that the space charge width essentially reaches a maximum value.

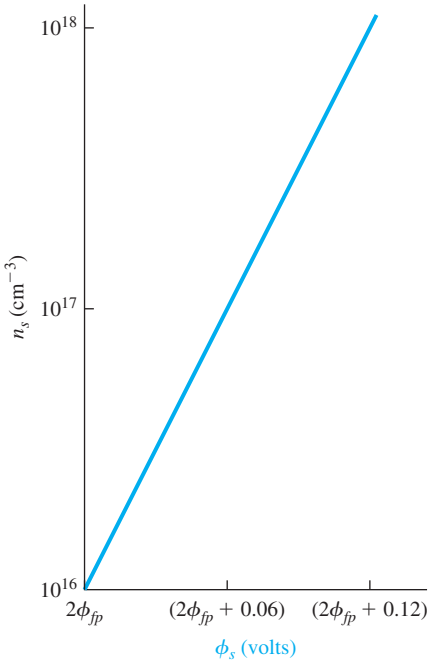


Figure 10.12 | Electron inversion charge density as a function of surface potential.

10.1.4 Work Function Differences

We have been concerned, so far, with the energy-band diagrams of the semiconductor material. Figure 10.13a shows the energy levels in the metal, silicon dioxide (SiO_2), and silicon relative to the vacuum level. The metal work function is ϕ_m and the electron affinity is χ. The parameter χ_i is the oxide electron affinity and, for SiO_2, $\chi_i = 0.9$ V.

Figure 10.13b shows the energy-band diagram of the entire MOS structure with zero gate voltage applied. The Fermi level is a constant through the entire system at thermal equilibrium. We may define ϕ'_m as a modified metal work function—the potential required to inject an electron from the metal into the conduction band of the oxide. Similarly, χ' is defined as a modified electron affinity. The voltage V_{ox0} is the potential drop across the oxide for zero applied gate voltage and is not necessarily zero because of the difference between ϕ_m and χ. The potential ϕ_{s0} is the surface potential for this case.

If we sum the energies from the Fermi level on the metal side to the Fermi level on the semiconductor side, we have

$$e\phi'_m + eV_{ox0} = e\chi' + \frac{E_g}{2} - e\phi_{s0} + e\phi_{fp} \tag{10.13}$$

Equation (10.13) can be rewritten as

$$V_{ox0} + \phi_{s0} = -\left[\phi'_m - \left(\chi' + \frac{E_g}{2e} + \phi_{fp}\right)\right] \tag{10.14}$$

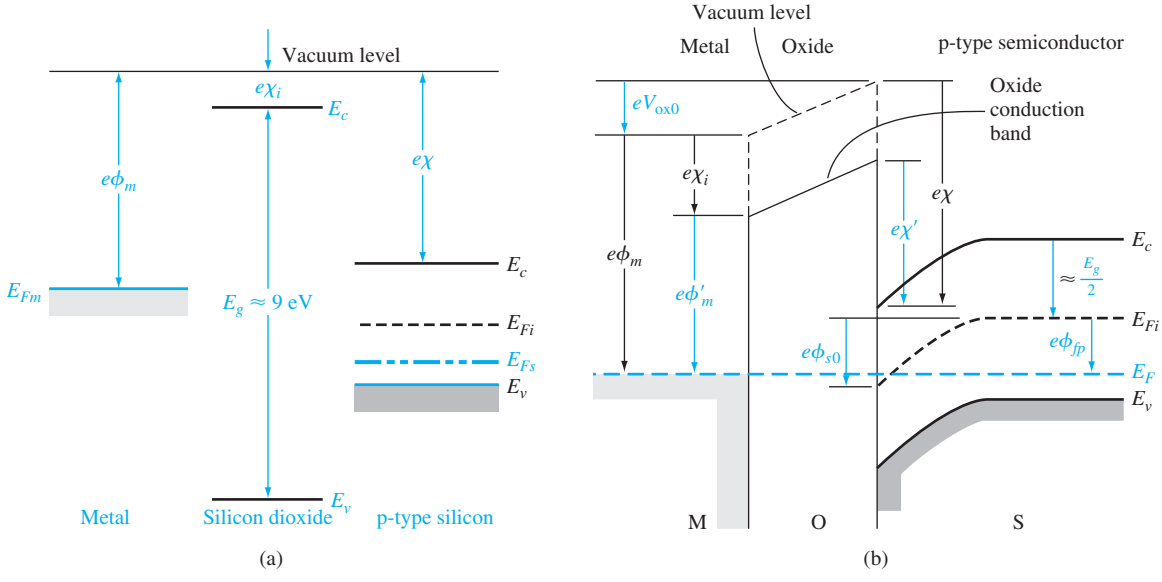


Figure 10.13 | (a) Energy levels in a MOS system prior to contact and (b) energy-band diagram through the MOS structure in thermal equilibrium after contact.

We can define a potential ϕ_{ms} as

$$\phi_{ms} \equiv \left[\phi'_m - \left(\chi' + \frac{E_g}{2e} + \phi_{fp}\right)\right] \tag{10.15}$$

which is known as the metal–semiconductor work function difference.

EXAMPLE 10.2

Objective: Determine the metal–semiconductor work function difference, ϕ_{ms}, for a given MOS system and semiconductor doping.

For an aluminum–silicon dioxide junction, $\phi'_m = 3.20$ V and, for a silicon–silicon dioxide junction, $\chi' = 3.25$ V. We may assume that $E_g = 1.12$ V. Let the p-type doping be $N_a = 10^{15}$ cm^{-3}.

■ Solution

For silicon at $T = 300$ K, we may calculate ϕ_{fp} as

$$\phi_{fp} = V_t \ln\left(\frac{N_a}{n_i}\right) = (0.0259)\ln\left(\frac{10^{15}}{1.5 \times 10^{10}}\right) = 0.288 \text{ V}$$

Then the metal–semiconductor work function difference is

$$\phi_{ms} = \phi'_m - \left(\chi' + \frac{E_g}{2e} + \phi_{fp}\right) = 3.20 - (3.25 + 0.560 + 0.288)$$

or

$$\phi_{ms} = -0.898 \text{ V}$$

■ Comment

The value of ϕ_{ms} will become more negative as the doping of the p-type substrate increases.

■ EXERCISE PROBLEM

Ex 10.2 Repeat Example 10.2 for a semiconductor doping of $N_a = 10^{16}$ cm^{-3}.

(Ans. $\phi_{ms} = -0.957$ V)

Degenerately doped polysilicon deposited on the oxide is also often used as the metal gate. Figure 10.14a shows the energy-band diagram of a MOS capacitor with an n^+ polysilicon gate and a p-type substrate. Figure 10.14b shows the energy-band

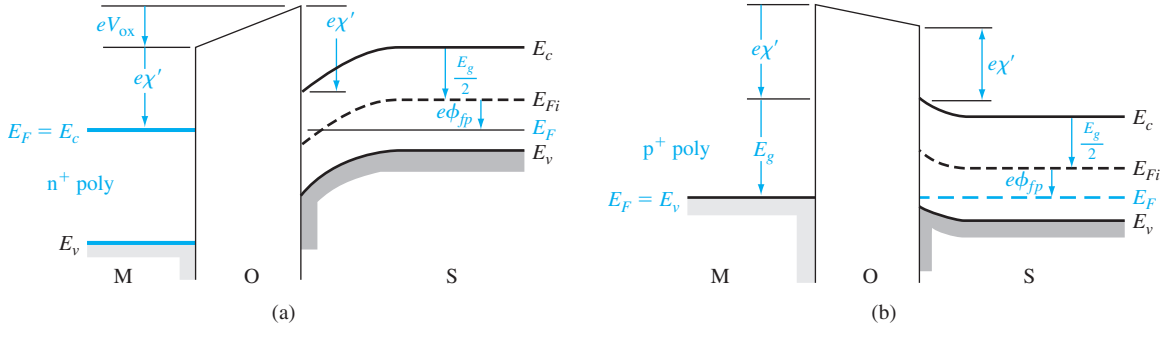


Figure 10.14 | Energy-band diagram through the MOS structure with a p-type substrate at zero gate bias for (a) an n^+ polysilicon gate and (b) a p^+ polysilicon gate.

diagram for the case of a p^+ polysilicon gate and the p-type silicon substrate. In the degenerately doped polysilicon, we will initially assume that $E_F = E_c$ for the n^+ case and $E_F = E_v$ for the p^+ case.

For the n^+ polysilicon gate, the metal–semiconductor work function difference can be written as

$$\phi_{ms} = \left[\chi' - \left(\chi' + \frac{E_g}{2e} + \phi_{fp}\right)\right] = -\left(\frac{E_g}{2e} + \phi_{fp}\right) \tag{10.16}$$

and for the p^+ polysilicon gate, we have

$$\phi_{ms} = \left[\left(\chi' + \frac{E_g}{e}\right) - \left(\chi' + \frac{E_g}{2e} + \phi_{fp}\right)\right] = \left(\frac{E_g}{2e} - \phi_{fp}\right) \tag{10.17}$$

However, for degenerately doped n^+ polysilicon and p^+ polysilicon, the Fermi level can be above E_c and below E_v, respectively, by 0.1 to 0.2 V. The experimental ϕ_{ms} values will then be slightly different from the values calculated by using Equations (10.16) and (10.17).

We have been considering a p-type semiconductor substrate. We may also have an n-type semiconductor substrate in a MOS capacitor. Figure 10.15 shows the energy-band diagram of the MOS capacitor with a metal gate and the n-type semiconductor substrate, for the case when a negative voltage is applied to the gate. The metal–semiconductor work function difference for this case is defined as

$$\phi_{ms} = \phi'_m - \left(\chi' + \frac{E_g}{2e} - \phi_{fn}\right) \tag{10.18}$$

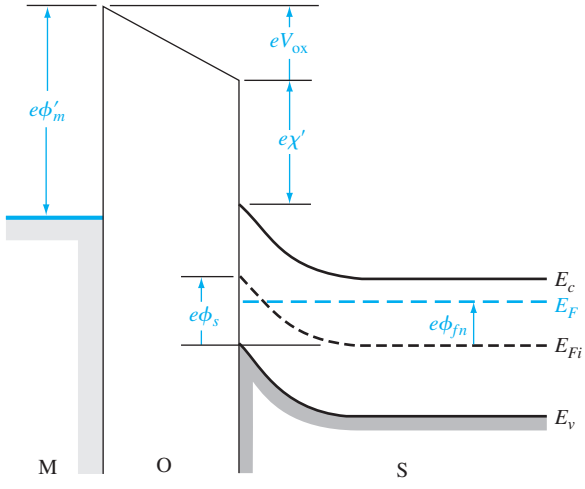


Figure 10.15 | Energy-band diagram through the MOS structure with an n-type substrate for a negative applied gate bias.

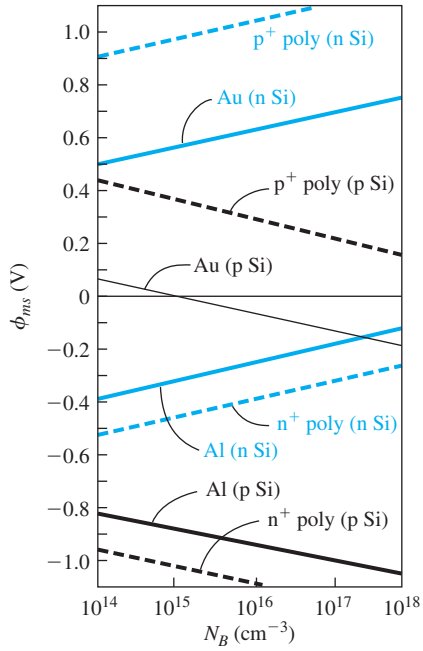


Figure 10.16 | Metal–semiconductor work function difference versus doping for aluminum, gold, and n^- and p^- polysilicon gates.
(*From Sze [17] and Werner [20].*)

where ϕ_{fn} is assumed to be a positive value. We will have similar expressions for n^+ and p^+ polysilicon gates.

Figure 10.16 shows the work function differences as a function of semiconductor doping for the various types of gates. We may note that the magnitudes of ϕ_{ms} for the polysilicon gates are somewhat larger than Equations (10.16) and (10.17) predict. This difference again is because the Fermi level is not equal to the conduction-band energy for the n^+ gate and is not equal to the valence-band energy for the p^+ gate. The metal–semiconductor work function difference becomes important in the flat-band and threshold voltage parameters discussed next.

10.1.5 Flat-Band Voltage

The *flat-band voltage* is defined as the applied gate voltage such that there is no band bending in the semiconductor and, as a result, zero net space charge in this region. Figure 10.17 shows this flat-band condition. Because of the work function difference and possible trapped charge in the oxide, the voltage across the oxide for this case is not necessarily zero.

We have implicitly been assuming that there is zero net charge density in the oxide material. This assumption may not be valid—a net fixed charge density,

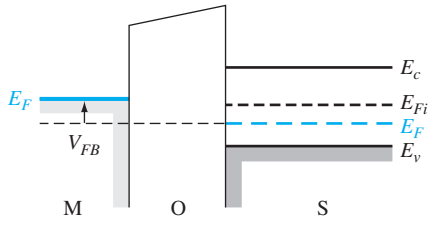


Figure 10.17 | Energy-band diagram of a MOS capacitor at flat band.

usually positive, may exist in the insulator. The positive charge has been identified with broken or dangling covalent bonds near the oxide–semiconductor interface. During the thermal formation of SiO_2, oxygen diffuses through the oxide and reacts near the Si–SiO_2 interface to form the SiO_2. Silicon atoms may also break away from the silicon material just prior to reacting to form SiO_2. When the oxidation process is terminated, excess silicon may exist in the oxide near the interface, resulting in the dangling bonds. The magnitude of this oxide charge seems, in general, to be a strong function of the oxidizing conditions such as oxidizing ambient and temperature. The charge density can be altered to some degree by annealing the oxide in an argon or nitrogen atmosphere. However, the charge is rarely zero.

The net fixed charge in the oxide appears to be located fairly close to the oxide–semiconductor interface. We will assume in the analysis of the MOS structure that an equivalent trapped charge per unit area, Q'_{ss}, is located in the oxide directly adjacent to the oxide–semiconductor interface. For the moment, we will ignore any other oxide-type charges that may exist in the device. The parameter Q'_{ss} is usually given in terms of number of electronic charges per unit area.

Equation (10.14), for zero applied gate voltage, can be written as

$$V_{ox0} + \phi_{s0} = -\phi_{ms} \tag{10.19}$$

If a gate voltage is applied, the potential drop across the oxide and the surface potential will change. We can then write

$$V_G = \Delta V_{ox} + \Delta\phi_s = (V_{ox} - V_{ox0}) + (\phi_s - \phi_{s0}) \tag{10.20}$$

Using Equation (10.19), we have

$$V_G = V_{ox} + \phi_s + \phi_{ms} \tag{10.21}$$

Figure 10.18 shows the charge distribution in the MOS structure for the flat-band condition. There is zero net charge in the semiconductor, and we can assume that an equivalent fixed surface charge density exists in the oxide. The charge density on the metal is Q'_m, and from charge neutrality we have

$$Q'_m + Q'_{ss} = 0 \tag{10.22}$$

We can relate Q'_m to the voltage across the oxide by

$$V_{ox} = \frac{Q'_m}{C_{ox}} \tag{10.23}$$

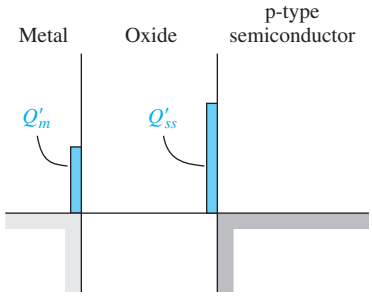


Figure 10.18 | Charge distribution in a MOS capacitor at flat band.

where C_{ox} is the oxide capacitance per unit area.[1] Substituting Equation (10.22) into Equation (10.23), we have

$$V_{ox} = \frac{-Q'_{ss}}{C_{ox}} \tag{10.24}$$

In the flat-band condition, the surface potential is zero, or $\phi_s = 0$. Then from Equation (10.21), we have

$$V_G = \boxed{V_{FB} = \phi_{ms} - \frac{Q'_{ss}}{C_{ox}}} \tag{10.25}$$

Equation (10.25) is the flat-band voltage for this MOS device.

EXAMPLE 10.3

Objective: Calculate the flat-band voltage for a MOS capacitor with a p-type semiconductor substrate.

Consider a MOS capacitor with a p-type silicon substrate doped to $N_a = 10^{16}$ cm^{-3}, a silicon dioxide insulator with a thickness of $t_{ox} = 20$ nm $= 200$ Å, and an n$^+$ polysilicon gate. Assume that $Q'_{ss} = 5 \times 10^{10}$ electronic charges per cm^2.

■ Solution

The work function difference, from Figure 10.16, is $\phi_{ms} \cong -1.1$V. The oxide capacitance is found to be

$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} = \frac{(3.9)(8.85 \times 10^{-14})}{200 \times 10^{-8}} = 1.726 \times 10^{-7} \text{ F/cm}^2$$

The equivalent oxide charge density is

$$Q'_{ss} = (5 \times 10^{10})(1.6 \times 10^{-19}) = 8 \times 10^{-9} \text{ C/cm}^2$$

The flat-band voltage is then determined to be

$$V_{FB} = \phi_{ms} - \frac{Q'_{ss}}{C_{ox}} = -1.1 \frac{-8 \times 10^{-9}}{1.726 \times 10^{-7}} = -1.15 \text{ V}$$

[1]Although we will, in general, use the primed notation for capacitance per unit area or charge per unit area, we will omit, for convenience, the prime on the oxide capacitance per unit area parameter.

Comment

The applied gate voltage required to achieve the flat-band condition for this p-type substrate is negative. If the amount of fixed oxide charge increases, the flat-band voltage becomes even more negative.

EXERCISE PROBLEM

Ex 10.3 Repeat Example 10.3 for a doping concentration of $N_a = 2 \times 10^{15}$ cm^{-3}, an oxide thickness of $t_{ox} = 4$ nm $= 40$ Å, and $Q'_{ss} = 2 \times 10^{10}$ electronic charges per cm^2. What is the value of the metal–semiconductor work function difference?

(Ans. $\phi_{ms} \cong -1.034$ V, $V_{FB} = -1.03$ V)

10.1.6 Threshold Voltage

The threshold voltage is defined as the applied gate voltage required to achieve the threshold inversion point. The threshold inversion point, in turn, is defined as the condition when the surface potential is $\phi_s = 2\phi_{fp}$ for the p-type semiconductor and $\phi_s = 2\phi_{fn}$ for the n-type semiconductor. These conditions are shown in Figures 10.9 and 10.10. The threshold voltage will be derived in terms of the electrical and geometrical properties of the MOS capacitor.

Figure 10.19 shows the charge distribution through the MOS device at the threshold inversion point for a p-type semiconductor substrate. The space charge width has reached its maximum value. We will assume that there is an equivalent oxide charge Q'_{ss} and the positive charge on the metal gate at threshold is Q'_{mT}. The prime on the charge terms indicates charge per unit area. Even though we are assuming that the surface has been inverted, we will neglect the inversion layer charge at this threshold inversion point. From conservation of charge, we can write

$$Q'_{mT} + Q'_{ss} = |Q'_{SD}(\text{max})| \tag{10.26}$$

where

$$\boxed{|Q'_{SD}(\text{max})| = eN_a x_{dT}} \tag{10.27}$$

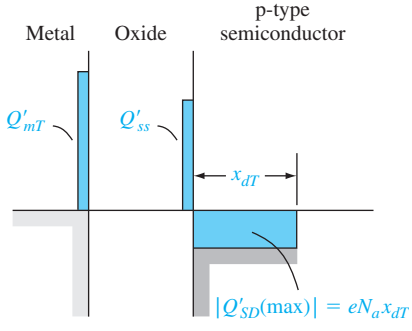


Figure 10.19 | Charge distribution in a MOS capacitor with a p-type substrate at the threshold inversion point.

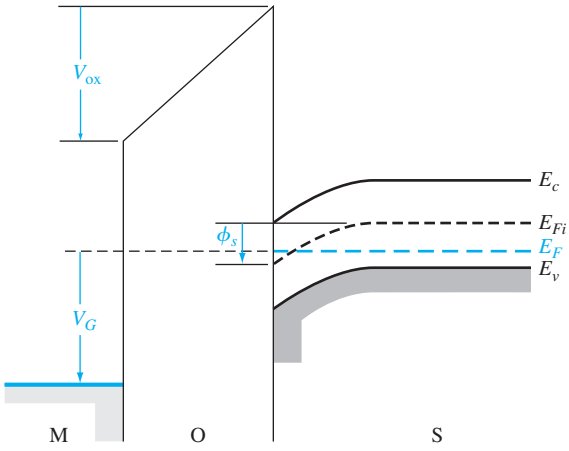


Figure 10.20 | Energy-band diagram through the MOS structure with a positive applied gate bias.

and is the magnitude of the maximum space charge density per unit area of the depletion region.

The energy-band diagram of the MOS system with an applied positive gate voltage is shown in Figure 10.20. As we mentioned, an applied gate voltage will change the voltage across the oxide and will change the surface potential. We had from Equation (10.20) that

$$V_G = \Delta V_{ox} + \Delta\phi_s = V_{ox} + \phi_s + \phi_{ms}$$

At threshold, we can define $V_G = V_{TN}$, where V_{TN} is the threshold voltage that creates the electron inversion layer charge. The surface potential is $\phi_s = 2\phi_{fp}$ at threshold, so Equation (10.20) can be written as

$$V_{TN} = V_{oxT} + 2\phi_{fp} + \phi_{ms} \tag{10.28}$$

where V_{oxT} is the voltage across the oxide at this threshold inversion point.

The voltage V_{oxT} can be related to the charge on the metal and to the oxide capacitance by

$$V_{oxT} = \frac{Q'_{mT}}{C_{ox}} \tag{10.29}$$

where again C_{ox} is the oxide capacitance per unit area. Using Equation (10.26), we can write

$$V_{oxT} = \frac{Q'_{mT}}{C_{ox}} = \frac{1}{C_{ox}}\left(|Q'_{SD}\,(\max)| - Q'_{ss}\right) \tag{10.30}$$

Finally, the threshold voltage can be written as

$$\boxed{V_{TN} = \frac{|Q'_{SD}\,(\max)|}{C_{ox}} - \frac{Q'_{ss}}{C_{ox}} + \phi_{ms} + 2\phi_{fp}} \tag{10.31a}$$

or

$$V_{TN} = \left(|Q'_{SD}(\max)| - Q'_{ss}\right)\left(\frac{t_{ox}}{\epsilon_{ox}}\right) + \phi_{ms} + 2\phi_{fp} \tag{10.31b}$$

Using the definition of flat-band voltage from Equation (10.25), we can also express the threshold voltage as

$$V_{TN} = \frac{|Q'_{SD}(\text{ max})|}{C_{\text{ox}}} + V_{FB} + 2\phi_{fp} \tag{10.31c}$$

For a given semiconductor material, oxide material, and gate metal, the threshold voltage is a function of semiconductor doping, oxide charge Q'_{ss}, and oxide thickness.

EXAMPLE 10.4

Objective: Calculate the threshold voltage of a MOS system using an aluminum gate.

Consider a p-type silicon substrate at $T = 300$ K doped to $N_a = 10^{15}$ cm^{-3}. Let $Q'_{ss} = 10^{10}$ cm^{-2}, $t_{ox} = 12$ nm $= 120$ Å, and assume the oxide is silicon dioxide.

■ Solution

From Figure 10.16, we find $\phi_{ms} \cong -0.88$ V. The other parameters are

$$\phi_{fp} = V_t \ln\left(\frac{N_a}{n_i}\right) = (0.0259)\ln\left(\frac{10^{15}}{1.5 \times 10^{10}}\right) = 0.2877 \text{ V}$$

and

$$x_{dT} = \left\{\frac{4\epsilon_s \phi_{fp}}{eN_a}\right\}^{1/2} = \left\{\frac{4(11.7)(8.85 \times 10^{-14})(0.2877)}{(1.6 \times 10^{-19})(10^{15})}\right\}^{1/2} = 8.63 \times 10^{-5} \text{ cm}$$

Then

$$|Q'_{SD}\,(\text{max})| = eN_a\, x_{dT} = (1.6 \times 10^{-19})(10^{15})(8.63 \times 10^{-5}) = 1.381 \times 10^{-8} \text{ C/cm}^2$$

The threshold voltage is now found to be

$$V_{TN} = (|Q'_{SD}\,(\text{max})| - Q'_{ss})\left(\frac{t_{ox}}{\epsilon_{ox}}\right) + \phi_{ms} + 2\phi_{fp}$$

$$= [(1.381 \times 10^{-8}) - (10^{10})(1.6 \times 10^{-19})] \cdot \left[\frac{120 \times 10^{-8}}{(3.9)(8.85 \times 10^{-14})}\right]$$

$$+ (-0.88) + 2(0.2877)$$

or

$$V_{TN} = -0.262 \text{ V}$$

■ Comment

In this example, the semiconductor is fairly lightly doped, which, in conjunction with the positive charge in the oxide and the work function difference, is sufficient to induce an electron inversion layer charge even with zero applied gate voltage. This condition makes the threshold voltage negative.

■ EXERCISE PROBLEM

Ex 10.4 Determine the metal–semiconductor work function difference and the threshold voltage for a silicon MOS device at $T = 300$ K for the following parameters: p$^+$ polysilicon gate, $N_a = 2 \times 10^{16}$ cm^{-3}, $t_{ox} = 8$ nm $= 80$ Å, and $Q'_{ss} = 2 \times 10^{10}$ cm^{-2}.

(Ans. $\phi_{ms} \cong -0.28$ V, $V_{TN} = +1.16$ V)

A negative threshold voltage for a p-type substrate implies a depletion mode device. A negative voltage must be applied to the gate in order to make the inversion

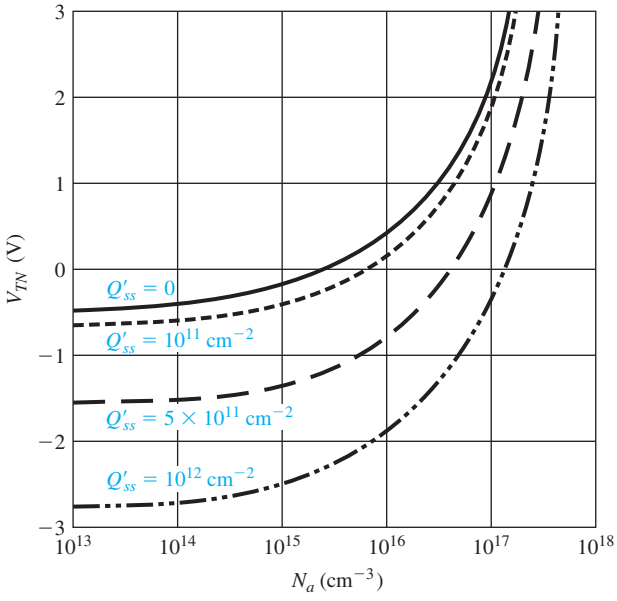


Figure 10.21 | Threshold voltage of an n-channel MOSFET versus the p-type substrate doping concentration for various values of oxide trapped charge ($t_{ox} = 500$ Å, aluminum gate).

layer charge equal to zero, whereas a positive gate voltage will induce a larger inversion layer charge.

Figure 10.21 is a plot of the threshold voltage V_{TN} as a function of the acceptor doping concentration for various positive oxide charge values. We may note that the p-type semiconductor must be somewhat heavily doped in order to obtain an enhancement mode device.

The previous derivation of the threshold voltage assumed a p-type semiconductor substrate. The same type of derivation can be done with an n-type semiconductor substrate, where a negative gate voltage can induce an inversion layer of holes at the oxide–semiconductor interface.

Figure 10.15 shows the energy-band diagram of the MOS structure with an n-type substrate and with an applied negative gate voltage. The threshold voltage for this case can be derived and is given by

$$V_{TP} = \left(-|Q'_{SD}\,(\text{max})| - Q'_{ss}\right)\left(\frac{t_{ox}}{\epsilon_{ox}}\right) + \phi_{ms} - 2\phi_{fn} \tag{10.32}$$

where

$$\phi_{ms} = \phi'_m - \left(\chi' + \frac{E_g}{2e} - \phi_{fn}\right) \tag{10.33a}$$

$$|Q'_{SD}\,(\text{max})| = eN_d x_{dT} \tag{10.33b}$$

$$x_{dT} = \left\{\frac{4\epsilon_s \phi_{fn}}{eN_d}\right\}^{1/2} \tag{10.33c}$$

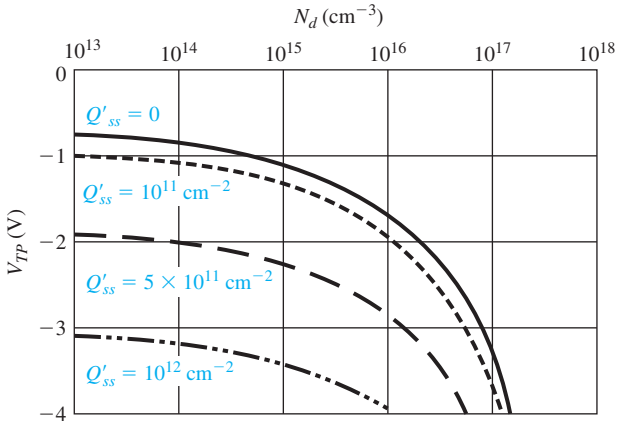


Figure 10.22 | Threshold voltage of a p-channel MOSFET versus the n-type substrate doping concentration for various values of oxide trapped charge ($t_{ox} = 500$ Å, aluminum gate).

and

$$\phi_{fn} = V_t \ln\left(\frac{N_d}{n_i}\right) \tag{10.33d}$$

We may note that x_{dT} and ϕ_{fn} are defined as positive quantities. We may also note that the notation of V_{TP} is the threshold voltage that will induce an inversion layer of holes. We will later drop the N and P subscript notation on the threshold voltage, but, for the moment, the notation may be useful for clarity.

Figure 10.22 is a plot of V_{TP} versus doping concentration for several values of Q'_{ss}. We may note that, for all values of positive oxide charge, this MOS capacitor is always an enhancement mode device. As the Q'_{ss} charge increases, the threshold voltage becomes more negative, which means that it takes a larger applied gate voltage to create the inversion layer of holes at the oxide–semiconductor interface.

DESIGN EXAMPLE 10.5

Objective: Determine the gate material and design the semiconductor doping concentration to yield a specified threshold voltage.

Consider a MOS device with silicon dioxide and an n-type silicon substrate. The oxide thickness is $t_{ox} = 12$ nm $= 120$ Å and the oxide charge is $Q'_{ss} = 2 \times 10^{10}$ cm^{-2}. The threshold voltage is to be approximately $V_{TP} = -0.3$ V.

■ Solution

The solution to this design problem is not straight forward, since the doping concentration parameter appears in the terms ϕ_{fn}, x_{dT}, Q'_{sD} (max), and ϕ_{ms}. The threshold voltage, then, is a nonlinear function of N_d. We resort to trial and error to obtain a solution.

Figure 10.22 shows the threshold voltage for an aluminum gate system. Since the required threshold voltage in this problem is less negative than the values shown in Figure 10.22,

we require a metal–semiconductor work function difference value that is more positive than for the aluminum gate. We therefore choose a p^+ polysilicon gate.

Consider a doping concentration of $N_d = 10^{17}$ cm^{-3}. From Figure 10.16, the metal–semiconductor work function difference is $\phi_{ms} \cong +1.1$ V. The remaining parameters are found to be

$$\phi_{fn} = V_t \ln\left(\frac{N_d}{n_i}\right) = (0.0259)\ln\left(\frac{10^{17}}{1.5 \times 10^{10}}\right) = 0.407 \text{ V}$$

$$x_{dT} = \left(\frac{4\epsilon_s \phi_{fn}}{eN_d}\right)^{1/2} = \left\{\frac{4(11.7)(8.85 \times 10^{-14})(0.407)}{(1.6 \times 10^{-19})(10^{17})}\right\}^{1/2}$$

$$= 1.026 \times 10^{-5} \text{ cm}$$

$$|Q'_{SD}(\max)| = eN_d x_{dT} = (1.6 \times 10^{-19})(10^{17})(1.026 \times 10^{-5})$$

$$= 1.642 \times 10^{-7} \text{ C/cm}^2$$

The threshold voltage is determined to be

$$V_{TP} = [-|Q'_{SD}(\max)| - Q'_{ss}] \cdot \left(\frac{t_{ox}}{\epsilon_{ox}}\right) + \phi_{ms} - 2\phi_{fn}$$

or

$$V_{TP} = \frac{[-(1.642 \times 10^{-7}) - (2 \times 10^{10})(1.6 \times 10^{-19})] \cdot (120 \times 10^{-8})}{(3.9)(8.85 \times 10^{-14})} + 1.1 - 2(0.407)$$

which yields

$$V_{TP} = -0.296 \text{ V} \cong -0.3 \text{ V}$$

Comment

The negative threshold voltage, with the n-type substrate, implies that this MOS capacitor is an enhancement mode device. The inversion layer charge is zero with zero applied gate voltage, and a negative gate voltage must be applied to induce the hole inversion charge.

EXERCISE PROBLEM

Ex 10.5 Consider a MOS capacitor with silicon dioxide and an n-type silicon substrate at $T = 300$ K with the following parameters: p^+ polysilicon gate, $N_d = 2 \times 10^{16}$ cm^{-3}, $t_{ox} = 20$ nm $= 200$ Å, and $Q'_{ss} = 5 \times 10^{10}$ cm^{-2}. Determine the threshold voltage. Is the capacitor an enhancement mode or depletion mode device?

(Ans. $V_{TP} = -0.12$ V, enhancement mode)

TEST YOUR UNDERSTANDING

TYU 10.1 (*a*) Consider an oxide-to-n-type silicon junction at $T = 300$ K. The impurity doping concentration in the silicon is $N_d = 8 \times 10^{15}$ cm^{-3}. Calculate the maximum space charge width in the silicon. (*b*) Repeat part (*a*) for a doping concentration of $N_d = 4 \times 10^{16}$ cm^{-3}.

[Ans. (*a*) 0.332 μm; (*b*) 0.158 μm]

TYU 10.2 Consider an n^+ polysilicon gate in a MOS structure with a p-type silicon substrate. The doping concentration of the silicon is $N_a = 3 \times 10^{16}$ cm^{-3}. Using Equation (10.16), find the value of ϕ_{ms}.

(Ans. $\phi_{ms} = -0.936$ V)

TYU 10.3 Repeat TYU 10.2 for a p^+ polysilicon gate using Equation (10.17).

(Ans. $\phi_{ms} = +0.184$ V)

TYU 10.4 Consider the MOS capacitor described in Exercise TYU 10.3. The oxide thickness is $t_{ox} = 16$ nm $= 160$ Å and the oxide charge density is $Q'_{ss} = 8 \times 10^{10}$ cm^{-2}. Determine the flat-band voltage.

(Ans. $+0.125$ V)

TYU 10.5 Consider an n^+ polysilicon gate on silicon dioxide with a p-type silicon substrate doped to $N_a = 3 \times 10^{16}$ cm^{-3}. Assume $Q'_{ss} = 5 \times 10^{10}$ cm^{-2}. Determine the required oxide thickness such that the threshold voltage is $V_{TN} = +0.65$ V.

(Ans. $t_{ox} = 4.52$ nm $= 452$ Å)

10.2 | CAPACITANCE–VOLTAGE CHARACTERISTICS

As mentioned previously, the MOS capacitor structure is the heart of the MOSFET. A great deal of information about the MOS device and the oxide–semiconductor interface can be obtained from the capacitance versus voltage or *C–V* characteristics of the device. The capacitance of a device is defined as

$$C = \frac{dQ}{dV} \tag{10.34}$$

where dQ is the magnitude of the differential change in charge on one plate as a function of the differential change in voltage dV across the capacitor. The capacitance is a small-signal or ac parameter and is measured by superimposing a small ac voltage on an applied dc gate voltage. The capacitance, then, is measured as a function of the applied dc gate voltage.

10.2.1 Ideal *C–V* Characteristics

First we will consider the ideal *C–V* characteristics of the MOS capacitor and then discuss some of the deviations that occur from these idealized results. We will initially assume that there is zero charge trapped in the oxide and also that there is no charge trapped at the oxide–semiconductor interface.

There are three operating conditions of interest in the MOS capacitor: accumulation, depletion, and inversion. Figure 10.23a shows the energy-band diagram of a MOS capacitor with a p-type substrate for the case when a negative voltage is applied to the gate, inducing an accumulation layer of holes in the semiconductor at the oxide–semiconductor interface. A small differential change in voltage across the MOS structure will cause a differential change in charge on the metal gate and also in the hole accumulation charge, as shown in Figure 10.23b. The differential changes in charge density occur at the edges of the oxide, as in a parallel-plate capacitor. The

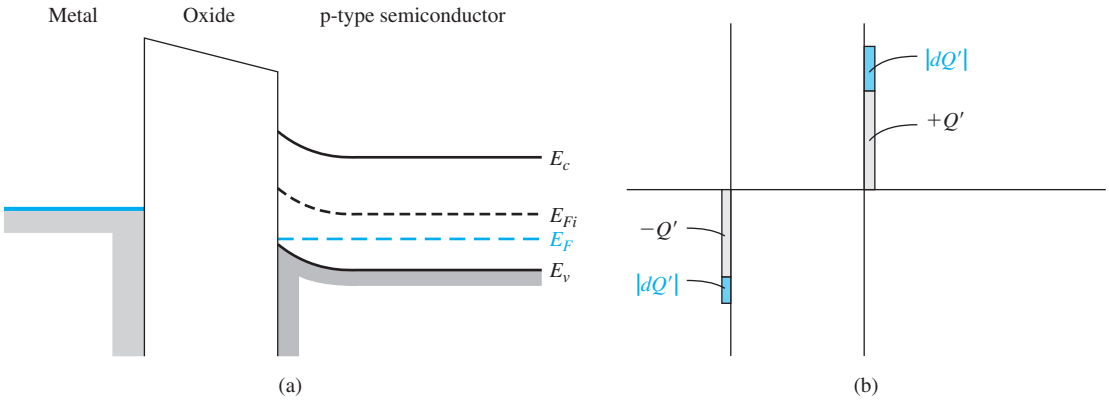


Figure 10.23 | (a) Energy-band diagram through a MOS capacitor for the accumulation mode. (b) Differential charge distribution at accumulation for a differential change in gate voltage.

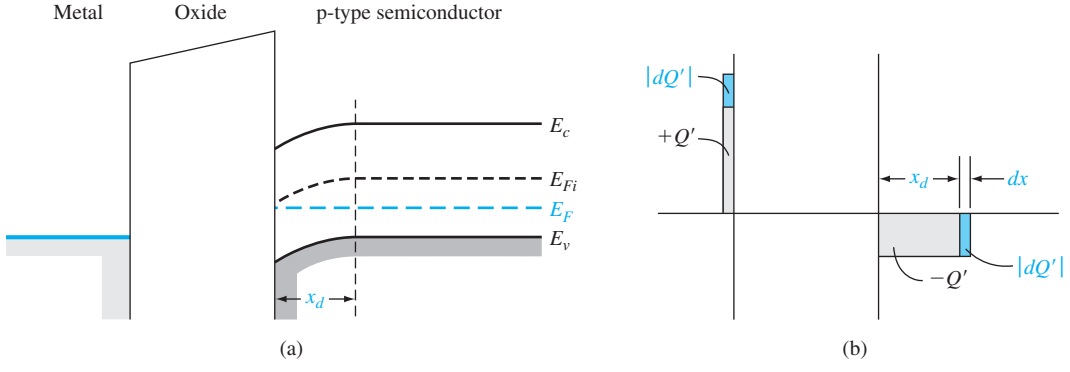


Figure 10.24 | (a) Energy-band diagram through a MOS capacitor for the depletion mode. (b) Differential charge distribution at depletion for a differential change in gate voltage.

capacitance C' per unit area of the MOS capacitor for this accumulation mode is just the oxide capacitance, or

$$C'(\text{acc}) = C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} \tag{10.35}$$

Figure 10.24a shows the energy-band diagram of the MOS device when a small positive voltage is applied to the gate, inducing a space charge region in the semiconductor; Figure 10.24b shows the charge distribution through the device for this condition. The oxide capacitance and the capacitance of the depletion region are in series. A small differential change in voltage across the capacitor will cause a differential change in the space charge width. The corresponding differential changes

in charge densities are shown in the figure. The total capacitance of the series combination is

$$\frac{1}{C'(\text{depl})} = \frac{1}{C_{\text{ox}}} + \frac{1}{C'_{SD}} \tag{10.36a}$$

or

$$C'(\text{depl}) = \frac{C_{\text{ox}} C'_{SD}}{C_{\text{ox}} + C'_{SD}} \tag{10.36b}$$

Since $C_{\text{ox}} = \epsilon_{\text{ox}}/t_{\text{ox}}$ and $C'_{SD} = \epsilon_s/x_d$, Equation (10.36b) can be written as

$$C'(\text{depl}) = \frac{C_{\text{ox}}}{1 + \dfrac{C_{\text{ox}}}{C'_{SD}}} = \frac{\epsilon_{\text{ox}}}{t_{\text{ox}} + \left(\dfrac{\epsilon_{\text{ox}}}{\epsilon_s}\right) x_d} \tag{10.37}$$

As the space charge width increases, the total capacitance $C'(\text{depl})$ decreases.

We had defined the threshold inversion point to be the condition when the maximum depletion width is reached, but there is essentially zero inversion charge density. This condition will yield a minimum capacitance C'_{min}, which is given by

$$C'_{\text{min}} = \frac{\epsilon_{\text{ox}}}{t_{\text{ox}} + \left(\dfrac{\epsilon_{\text{ox}}}{\epsilon_s}\right) x_{dT}} \tag{10.38}$$

Figure 10.25a shows the energy-band diagram of this MOS device for the inversion condition. In the ideal case, a small incremental change in the voltage across the MOS capacitor will cause a differential change in the inversion layer charge density. The space charge width does not change. If the inversion charge can respond to the change in capacitor voltage as indicated in Figure 10.25b, then the capacitance is again just the oxide capacitance, or

$$C'(\text{inv}) = C_{\text{ox}} = \frac{\epsilon_{\text{ox}}}{t_{\text{ox}}} \tag{10.39}$$

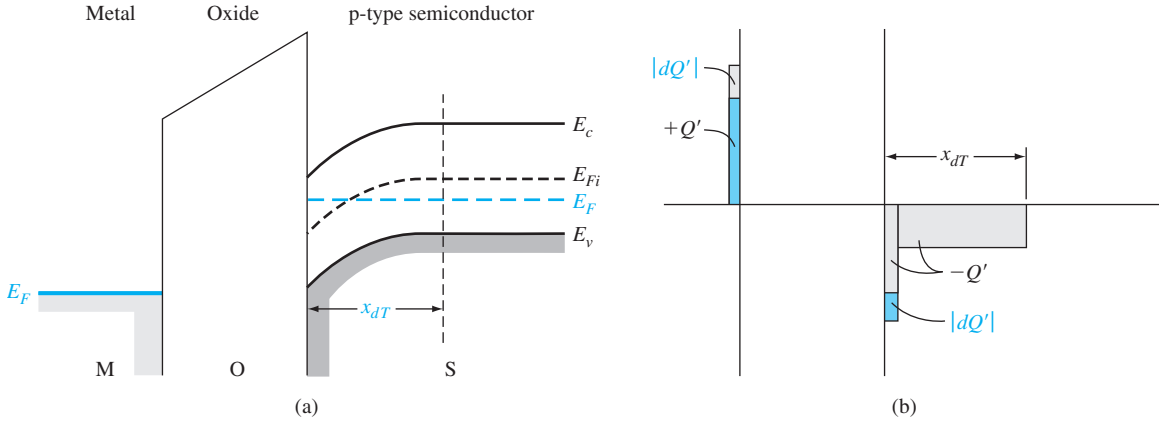


Figure 10.25 | (a) Energy-band diagram through a MOS capacitor for the inversion mode. (b) Differential charge distribution at inversion for a low-frequency differential change in gate voltage.

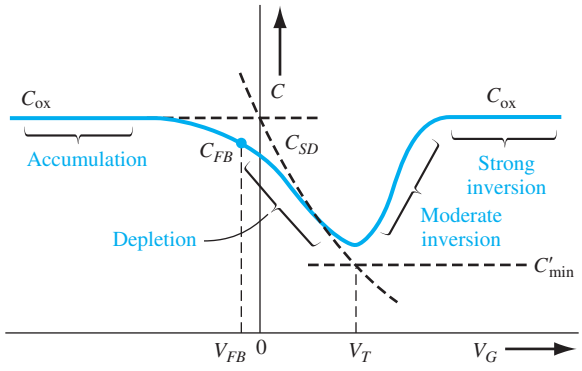


Figure 10.26 | Ideal low-frequency capacitance versus gate voltage of a MOS capacitor with a p-type substrate. Individual capacitance components are also shown.

Figure 10.26 shows the ideal capacitance versus gate voltage, or *C–V*, characteristics of the MOS capacitor with a p-type substrate. The three dashed segments correspond to the three components C_{ox}, C'_{SD}, and C'_{min}. The solid curve is the ideal net capacitance of the MOS capacitor. Moderate inversion, which is indicated in the figure, is the transition region between the point when only the space charge density changes with gate voltage and when only the inversion charge density changes with gate voltage.

The point on the curve that corresponds to the flat-band condition is of interest. The flat-band condition occurs between the accumulation and depletion conditions. The capacitance at flat band is given by

$$C'_{FB} = \frac{\epsilon_{ox}}{t_{ox} + \left(\frac{\epsilon_{ox}}{\epsilon_s}\right)\sqrt{\left(\frac{kT}{e}\right)\left(\frac{\epsilon_s}{eN_a}\right)}} \tag{10.40}$$

We may note that the flat-band capacitance is a function of oxide thickness as well as semiconductor doping. The general location of this point on the *C–V* plot is shown in Figure 10.26.

EXAMPLE 10.6

Objective: Calculate C_{ox}, C'_{min}, and C'_{FB} for a MOS capacitor.

Consider a p-type silicon substrate at $T = 300$ K doped to $N_a = 10^{16}$ cm^{-3}.

The oxide is silicon dioxide with a thickness of $t_{ox} = 18$ nm $= 180$ Å, and the gate is aluminum.

■ Solution

The oxide capacitance is

$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} = \frac{(3.9)(8.85 \times 10^{-14})}{180 \times 10^{-8}} = 1.9175 \times 10^{-7} \text{ F/cm}^2$$

To find the minimum capacitance, we need to calculate

$$\phi_{fp} = V_t \ln\left(\frac{N_a}{n_i}\right) = (0.0259)\ln\left(\frac{10^{16}}{1.5 \times 10^{10}}\right) = 0.3473 \text{ V}$$

and

$$x_{dT} = \left\{\frac{4\epsilon_s \phi_{fp}}{eN_a}\right\}^{1/2} = \left\{\frac{4(11.7)(8.85 \times 10^{-14})(0.3473)}{(1.6 \times 10^{-19})(10^{16})}\right\}^{1/2}$$

$$\cong 0.30 \times 10^{-4} \text{ cm}$$

Then

$$C'_{\min} = \frac{\epsilon_{ox}}{t_{ox} + \left(\frac{\epsilon_{ox}}{\epsilon_s}\right)x_{dT}} = \frac{(3.9)(8.85 \times 10^{-14})}{180 \times 10^{-8} + \left(\frac{3.9}{11.7}\right)(0.30 \times 10^{-4})}$$

$$= 2.925 \times 10^{-8} \text{ F/cm}^2$$

We may note that

$$\frac{C'_{\min}}{C_{ox}} = \frac{2.925 \times 10^{-8}}{1.9175 \times 10^{-7}} = 0.1525$$

The flat-band capacitance is

$$C'_{FB} = \frac{\epsilon_{ox}}{t_{ox} + \left(\frac{\epsilon_{ox}}{\epsilon_s}\right)\sqrt{\frac{V_t \epsilon_s}{eN_a}}}$$

$$= \frac{(3.9)(8.85 \times 10^{-14})}{180 \times 10^{-8} + \left(\frac{3.9}{11.7}\right)\sqrt{\frac{(0.0259)(11.7)(8.85 \times 10^{-14})}{(1.6 \times 10^{-19})(10^{16})}}}$$

$$= 1.091 \times 10^{-7} \text{ F/cm}^2$$

We also note that

$$\frac{C'_{FB}}{C_{ox}} = \frac{1.091 \times 10^{-7}}{1.9175 \times 10^{-7}} = 0.569$$

Comment

The ratios of $C'_{\min}/C_{ox}$ and C'_{FB}/C_{ox} are typical values obtained in C–V plots.

EXERCISE PROBLEM

Ex 10.6 Consider a MOS capacitor with the following parameters: n^+ polysilicon gate, $N_a = 3 \times 10^{16}$ cm^{-3}, $t_{ox} = 8$ nm $= 80$ Å, and $Q'_{ss} = 2 \times 10^{10}$ cm^{-2}. Determine the ratios $C'_{\min}/C_{ox}$ and C'_{FB}/C_{ox}.

(Ans. $C'_{\min}/C_{ox} = 0.1178$, $C'_{FB}/C_{ox} = 0.5094$)

If we assume the oxide capacitance per unit area is $C_{ox} = 1.9175 \times 10^{-7}$ F/cm^2 and the channel length and width are $L = 2$ μm and $W = 20$ μm, respectively, then the total gate oxide capacitance is

$$C_{oxT} = C_{ox} LW = (1.9175 \times 10^{-7})(2 \times 10^{-4})(20 \times 10^{-4})$$

$$= 7.67 \times 10^{-14} \text{ F} = 0.0767 \text{ pF} = 76.7 \text{ fF}$$

The total oxide capacitance in a typical MOS device is quite small.

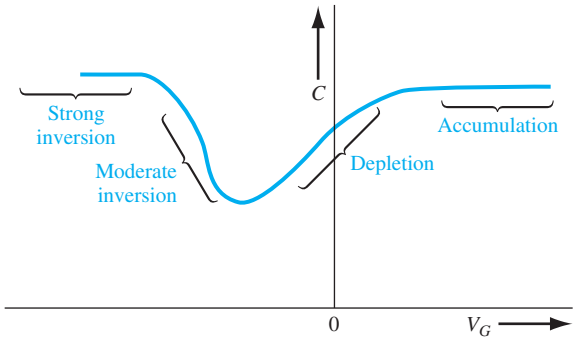


Figure 10.27 | Ideal low-frequency capacitance versus gate voltage of a MOS capacitor with an n-type substrate.

The same type of ideal *C–V* characteristics is obtained for a MOS capacitor with an n-type substrate by changing the sign of the voltage axis. The accumulation condition is obtained for a positive gate bias and the inversion condition is obtained for a negative gate bias. This ideal curve is shown in Figure 10.27.

10.2.2 Frequency Effects

Figure 10.25a shows the MOS capacitor with a p-type substrate and biased in the inversion condition. We have argued that a differential change in the capacitor voltage in the ideal case causes a differential change in the inversion layer charge density. However, we must consider the source of electrons that produces a change in the inversion charge density.

There are two sources of electrons that can change the charge density of the inversion layer. The first source is by diffusion of minority carrier electrons from the p-type substrate across the space charge region. This diffusion process is the same as that in a reverse-biased pn junction that generates the ideal reverse saturation current. The second source of electrons is by thermal generation of electron–hole pairs within the space charge region. This process is again the same as that in a reverse-biased pn junction generating the reverse-biased generation current. Both of these processes generate electrons at a particular rate. The electron concentration in the inversion layer, then, cannot change instantaneously. If the ac voltage across the MOS capacitor changes rapidly, the change in the inversion layer charge will not be able to respond. The *C–V* characteristics will then be a function of the frequency of the ac signal used to measure the capacitance.

In the limit of a very high frequency, the inversion layer charge will not respond to a differential change in capacitor voltage. Figure 10.28 shows the charge distribution in the MOS capacitor with a p-type substrate. At a high-signal frequency, the differential change in charge occurs at the metal and in the space charge width in the semiconductor. The capacitance of the MOS capacitor is then C'_{min}, which we discussed earlier.

The high-frequency and low-frequency limits of the *C–V* characteristics are shown in Figure 10.29. In general, high frequency corresponds to a value on the order of 1 MHz and low frequency corresponds to values in the range of 5 to 100 Hz. Typically, the high-frequency characteristics of the MOS capacitor are measured.

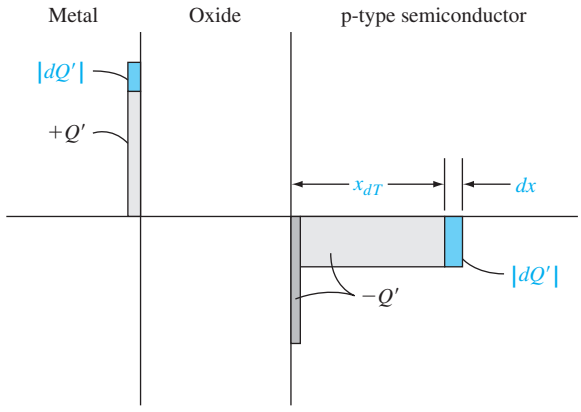


Figure 10.28 | Differential charge distribution at inversion for a high-frequency differential change in gate voltage.

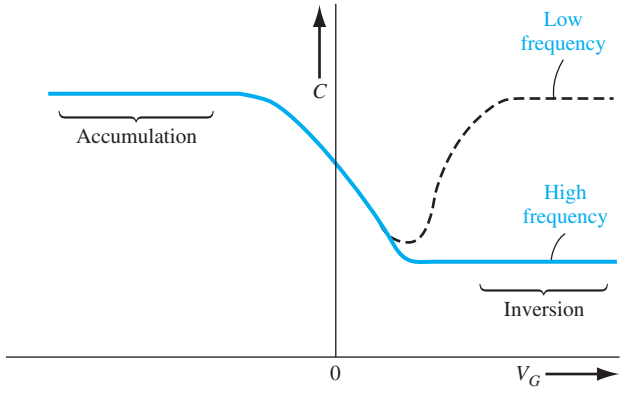


Figure 10.29 | Low-frequency and high-frequency capacitance versus gate voltage of a MOS capacitor with a p-type substrate.

10.2.3 Fixed Oxide and Interface Charge Effects

In all of the discussion concerning *C–V* characteristics so far, we have assumed an ideal oxide in which there are no fixed oxide or oxide–semiconductor interface charges. These two types of charges will change the *C–V* characteristics.

We previously discussed how the fixed oxide charge affects the threshold voltage. This charge will also affect the flat-band voltage. The flat-band voltage from Equation (10.25) is given by

$$V_{FB} = \phi_{ms} - \frac{Q'_{ss}}{C_{ox}}$$

where Q'_{ss} is the equivalent fixed oxide charge and ϕ_{ms} is the metal–semiconductor work function difference. The flat-band voltage shifts to more negative voltages for a positive fixed oxide charge. Since the oxide charge is not a function of gate voltage, the curves

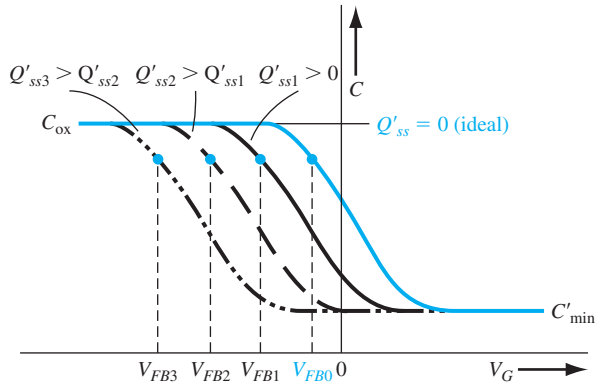


Figure 10.30 | High-frequency capacitance versus gate voltage of a MOS capacitor with a p-type substrate for several values of effective trapped oxide charge.

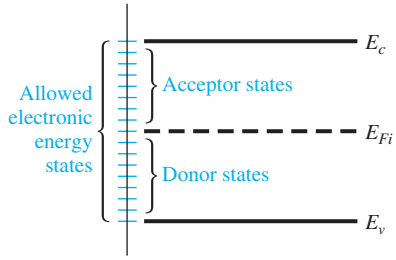


Figure 10.31 | Schematic diagram showing interface states at the oxide–semiconductor interface.

show a parallel shift with oxide charge, and the shape of the *C–V* curves remains the same as the ideal characteristics. Figure 10.30 shows the high-frequency characteristics of a MOS capacitor with a p-type substrate for several values of fixed positive oxide charge.

The *C–V* characteristics can be used to determine the equivalent fixed oxide charge. For a given MOS structure, ϕ_{ms} and C_{ox} are known, so the ideal flat-band voltage and flat-band capacitance can be calculated. The experimental value of flat-band voltage can be measured from the *C–V* curve, and the value of fixed oxide charge can then be determined. The *C–V* measurements are a valuable diagnostic tool to characterize a MOS device. This characterization is especially useful in the study of radiation effects on MOS devices, for example.

We first encountered oxide–semiconductor interface states in Chapter 9 in the discussion of Schottky barrier diodes. Figure 10.31 shows the energy-band diagram of a semiconductor at the oxide–semiconductor interface. The periodic nature of the semiconductor is abruptly terminated at the interface so that allowed electronic energy levels will exist within the forbidden bandgap. These allowed energy states are referred to as interface states. Charge can flow between the semiconductor and interface states, in contrast to the fixed oxide charge. The net charge in these interface states is a function of the position of the Fermi level in the bandgap.

In general, acceptor states exist in the upper half of the bandgap and donor states exist in the lower half of the bandgap. An acceptor state is neutral if the Fermi level is below the state and becomes negatively charged if the Fermi level is above the state. A donor state is neutral if the Fermi level is above the state and becomes positively charged if the Fermi level is below the state. The charge of the interface states is then a function of the gate voltage applied across the MOS capacitor.

Figure 10.32a shows the energy-band diagram in a p-type semiconductor of a MOS capacitor biased in the accumulation condition. In this case, there is a net positive charge trapped in the donor states. Now let the gate voltage change to produce the energy-band diagram shown in Figure 10.32b. The Fermi level corresponds to

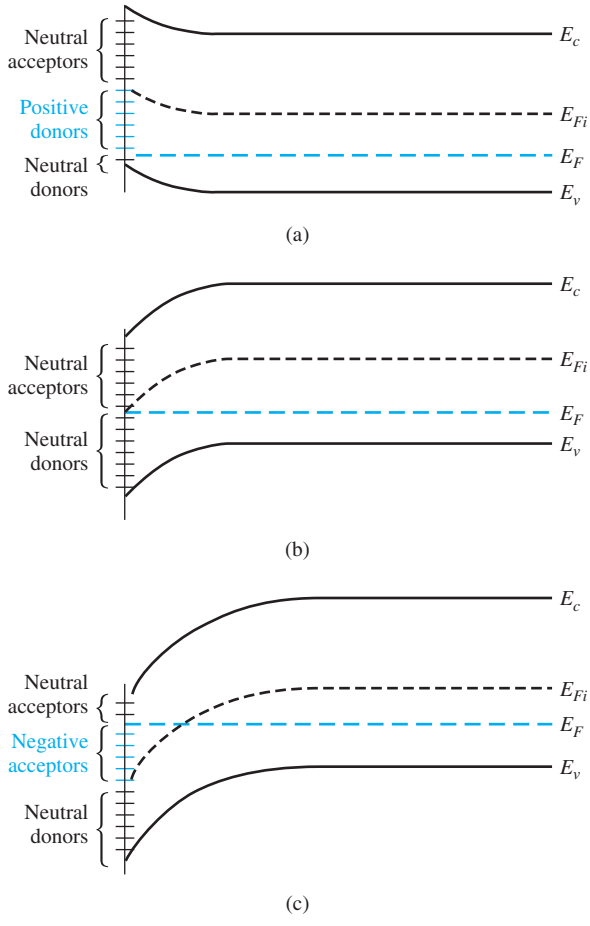


Figure 10.32 | Energy-band diagram in a p-type semiconductor showing the charge trapped in the interface states when the MOS capacitor is biased (a) in accumulation, (b) at midgap, and (c) at inversion.

the intrinsic Fermi level at the surface; thus, all interface states are neutral. This particular bias condition is known as *midgap*. Figure 10.32c shows the condition at inversion in which there is now a net negative charge in the acceptor states.

The net charge in the interface states changes from positive to negative as the gate voltage sweeps from the accumulation, depletion, to the inversion condition. We noted that the *C–V* curves shifted in the negative gate voltage direction due to positive fixed oxide charge. When interface states are present, the amount and direction of the shift change as we sweep through the gate voltage, since the amount and sign of the interface trapped charge change. The *C–V* curves now become "smeared out" as shown in Figure 10.33.

Again, the *C–V* measurements can be used as a diagnostic tool in semiconductor device process control. For a given MOS device, the ideal *C–V* curve can be determined.

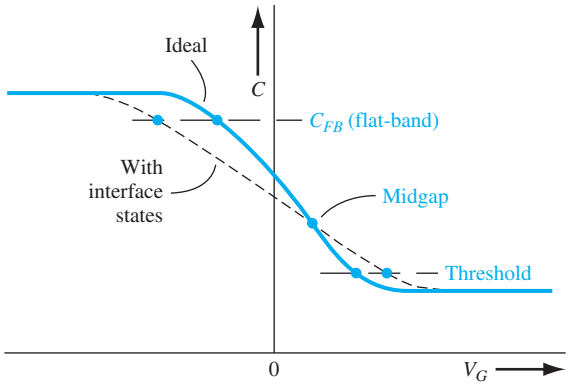


Figure 10.33 | High-frequency C–V characteristics of a MOS capacitor showing effects of interface states.

Any "smearing out" in the experimental curve indicates the presence of interface states and any parallel shift indicates the presence of fixed oxide charge. The amount of smearing out can be used to determine the density of interface states. These types of measurement are extremely useful in the study of radiation effects on MOS devices.

10.3 | THE BASIC MOSFET OPERATION

The current in a MOSFET is due to the flow of charge in the inversion layer or channel region adjacent to the oxide–semiconductor interface. We have discussed the creation of the inversion layer charge in enhancement-type MOS capacitors. We may also have depletion-type devices in which a channel already exists at zero gate voltage.

10.3.1 MOSFET Structures

There are four basic MOSFET device types. Figure 10.34 shows an n-channel enhancement mode MOSFET. Implicit in the enhancement mode notation is the idea that the semiconductor substrate is not inverted directly under the oxide with zero gate voltage. A positive gate voltage induces the electron inversion layer, which then "connects" the n-type source and the n-type drain regions. The source terminal is the source of carriers that flow through the channel to the drain terminal. For this n-channel device, electrons flow from the source to the drain so the conventional current will enter the drain and leave the source. The conventional circuit symbol for this n-channel enhancement mode device is also shown in the figure.

Figure 10.35 shows an n-channel depletion mode MOSFET. An n-channel region exists under the oxide with 0 V applied to the gate. However, we have shown that the threshold voltage of a MOS device with a p-type substrate may be negative; this means that an electron inversion layer already exists with zero gate voltage applied. Such a device is also considered to be a depletion mode device. The n-channel shown in the figure can be an electron inversion layer or an intentionally doped n region. The conventional circuit symbol for the n-channel depletion mode MOSFET is also shown in the figure.

Figures 10.36a, b show a p-channel enhancement mode MOSFET and a p-channel depletion mode MOSFET. In the p-channel enhancement mode device, a negative

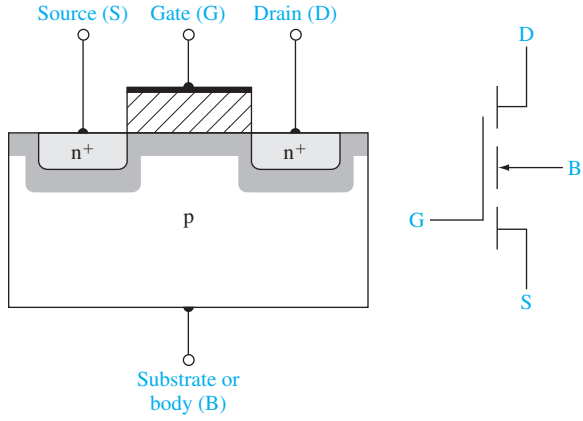


Figure 10.34 | Cross section and circuit symbol for an n-channel enhancement mode MOSFET.

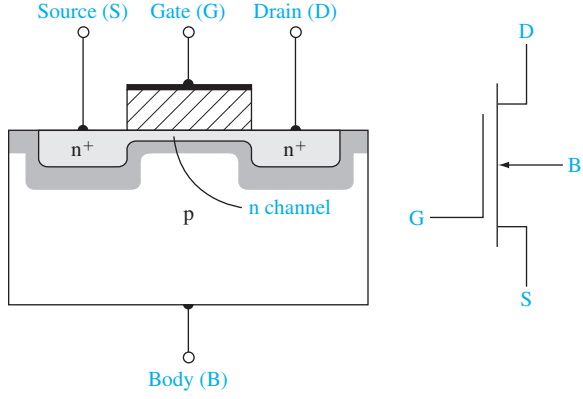


Figure 10.35 | Cross section and circuit symbol for an n-channel depletion mode MOSFET.

gate voltage must be applied to create an inversion layer of holes that will "connect" the p-type source and drain regions. Holes flow from the source to the drain, so the conventional current will enter the source and leave the drain. A p-channel region exists in the depletion mode device even with zero gate voltage. The conventional circuit symbols are shown in the figure.

10.3.2 Current–Voltage Relationship—Concepts

Figure 10.37a shows an n-channel enhancement mode MOSFET with a gate-to-source voltage that is less than the threshold voltage and with only a very small drain-to-source voltage. The source and substrate, or body, terminals are held at ground potential. With this bias configuration, there is no electron inversion layer, the drain-to-substrate pn junction is reverse biased, and the drain current is zero (disregarding pn junction leakage currents).

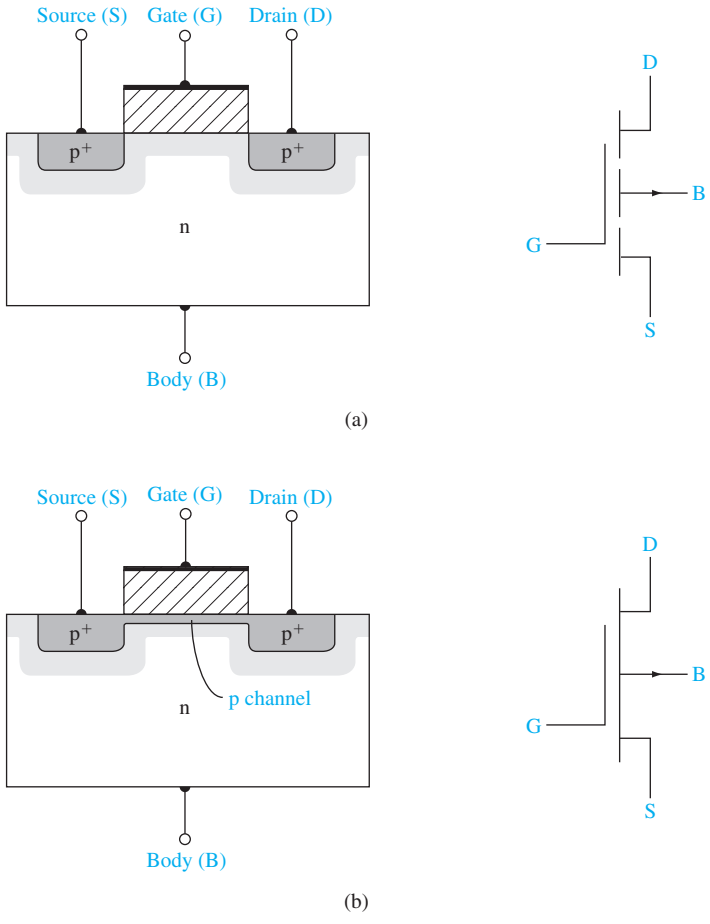


Figure 10.36 | Cross section and circuit symbol for (a) a p-channel enhancement mode MOSFET and (b) a p-channel depletion mode MOSFET.

Figure 10.37b shows the same MOSFET with an applied gate voltage such that $V_{GS} > V_T$. An electron inversion layer has been created so that when a small drain voltage is applied, the electrons in the inversion layer will flow from the source to the positive drain terminal. The conventional current enters the drain terminal and leaves the source terminal. In this ideal case, there is no current through the oxide to the gate terminal.

For small V_{DS} values, the channel region has the characteristics of a resistor, so we can write

$$I_D = g_d V_{DS} \tag{10.41}$$

where g_d is defined as the channel conductance in the limit as $V_{DS} \to 0$. The channel conductance is given by

$$g_d = \frac{W}{L} \cdot \mu_n |Q_n'| \tag{10.42}$$

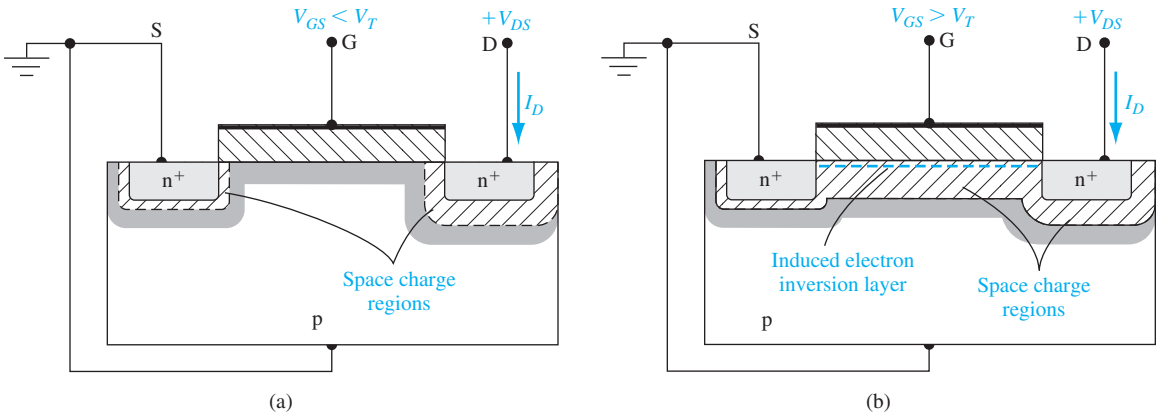


Figure 10.37 | The n-channel enhancement mode MOSFET (a) with an applied gate voltage $V_{GS} < V_T$ and (b) with an applied gate voltage $V_{GS} > V_T$.

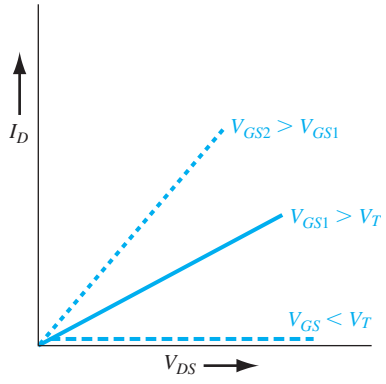


Figure 10.38 | I_D versus V_{DS} characteristics for small values of V_{DS} at three V_{GS} voltages.

where μ_n is the mobility of the electrons in the inversion layer and $|Q'_n|$ is the magnitude of the inversion layer charge per unit area. The inversion layer charge is a function of the gate voltage; thus, the basic MOS transistor action is the modulation of the channel conductance by the gate voltage. The channel conductance, in turn, determines the drain current. We will initially assume that the mobility is a constant; we will discuss mobility effects and variations in the next chapter.

The I_D versus V_{DS} characteristics, for small values of V_{DS}, are shown in Figure 10.38. When $V_{GS} < V_T$, the drain current is zero. As V_{GS} becomes larger than V_T, channel inversion charge density increases, which increases the channel conductance. A larger value of g_d produces a larger initial slope of the I_D versus V_{DS} characteristic as shown in the figure.

Figure 10.39a shows the basic MOS structure for the case when $V_{GS} > V_T$ and the applied V_{DS} voltage is small. The thickness of the inversion channel layer in the

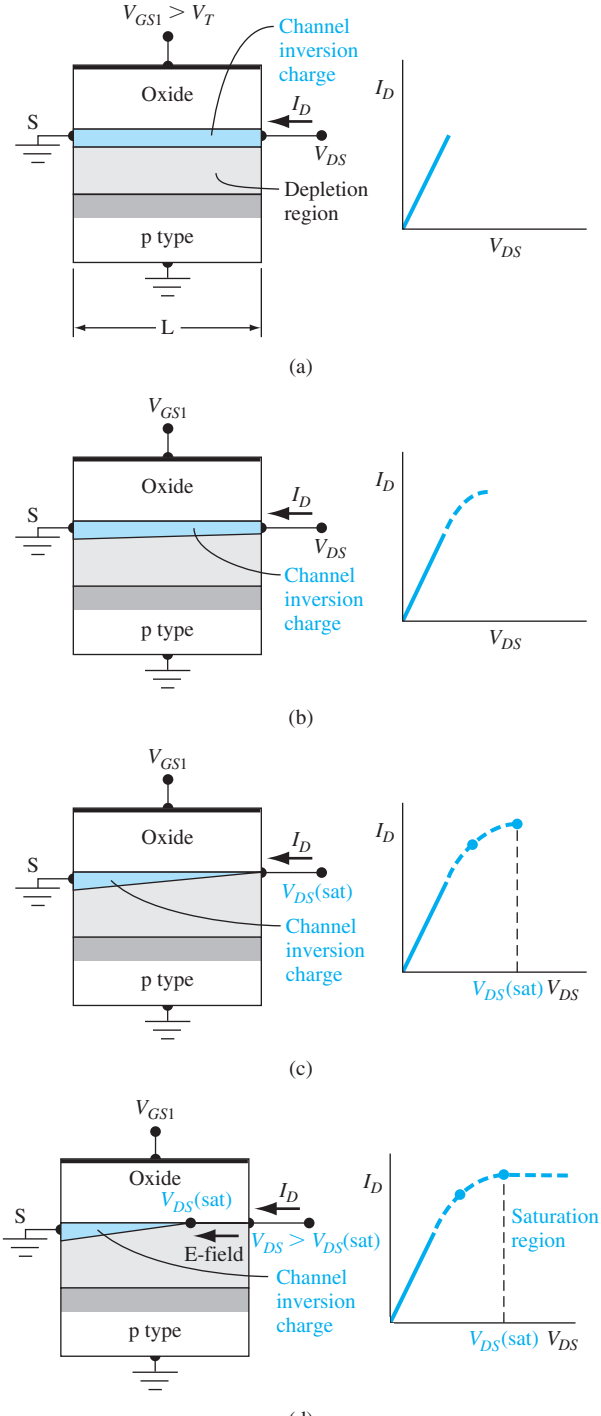


Figure 10.39 | Cross section and I_D versus V_{DS} curve when $V_{GS} < V_T$ for (a) a small V_{DS} value, (b) a larger V_{DS} value, (c) a value of $V_{DS} = V_{DS}$(sat), and (d) a value of $V_{DS} > V_{DS}$(sat).

figure qualitatively indicates the relative charge density, which is essentially constant along the entire channel length for this case. The corresponding I_D versus V_{DS} curve is shown in the figure.

Figure 10.39b shows the situation when the V_{DS} value increases. As the drain voltage increases, the voltage drop across the oxide near the drain terminal decreases, which means that the induced inversion charge density near the drain also decreases. The incremental conductance of the channel at the drain decreases, which then means that the slope of the I_D versus V_{DS} curve will decrease. This effect is shown in the I_D versus V_{DS} curve in the figure.

When V_{DS} increases to the point where the potential drop across the oxide at the drain terminal is equal to V_T, the induced inversion charge density is zero at the drain terminal. This effect is schematically shown in Figure 10.39c. At this point, the incremental conductance at the drain is zero, which means that the slope of the I_D versus V_{DS} curve is zero. We can write

$$V_{GS} - V_{DS}(\text{sat}) = V_T \tag{10.43a}$$

or

$$\boxed{V_{DS}(\text{sat}) = V_{GS} - V_T} \tag{10.43b}$$

where $V_{DS}(\text{sat})$ is the drain-to-source voltage producing zero inversion charge density at the drain terminal.

When V_{DS} becomes larger than the $V_{DS}(\text{sat})$ value, the point in the channel at which the inversion charge is just zero moves toward the source terminal. In this case, electrons enter the channel at the source, travel through the channel toward the drain, and then, at the point where the charge goes to zero, the electrons are injected into the space charge region where they are swept by the E-field to the drain contact. If we assume that the change in channel length ΔL is small compared to the original length L, then the drain current will be a constant for $V_{DS} > V_{DS}(\text{sat})$. The region of the I_D versus V_{DS} characteristic is referred to as the *saturation region.* Figure 10.39d shows this region of operation.

When V_{GS} changes, the I_D versus V_{DS} curve will change. We saw that, if V_{GS} increases, the initial slope of I_D versus V_{DS} increases. We can also note from Equation (10.43b) that the value of $V_{DS}(\text{sat})$ is a function of V_{GS}. We can generate the family of curves for this n-channel enhancement mode MOSFET as shown in Figure 10.40.

Figure 10.41 shows an n-channel depletion mode MOSFET. If the n-channel region is actually an induced electron inversion layer created by the metal– semiconductor work function difference and fixed charge in the oxide, the current–voltage characteristics are exactly the same as we have discussed, except that V_T is a negative quantity. We may also consider the case when the n-channel region is actually an n-type semiconductor region. In this type of device, a negative gate voltage will induce a space charge region under the oxide, reducing the thickness of the n-channel region. The reduced thickness decreases the channel conductance, which reduces the drain current. A positive gate voltage will create an electron accumulation layer, which increases the drain current. One basic requirement for this device is that the channel

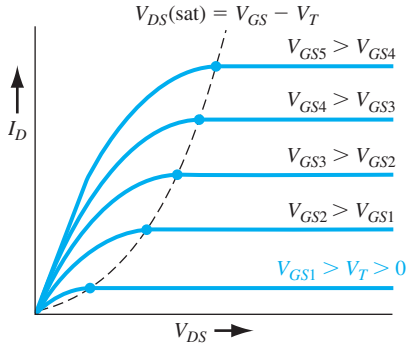


Figure 10.40 | Family of I_D versus V_{DS} curves for an n-channel enhancement mode MOSFET.

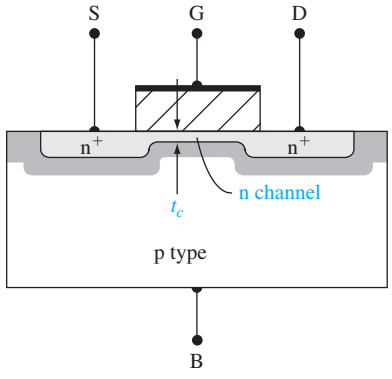


Figure 10.41 | Cross section of an n-channel depletion mode MOSFET.

thickness t_c must be less than the maximum induced space charge width in order to be able to turn the device off. The general I_D versus V_{DS} family of curves for an n-channel depletion mode MOSFET is shown in Figure 10.42.

In the next section, we derive the ideal current–voltage relation for the n-channel MOSFET. In the nonsaturation region, we obtain

$$I_D = \frac{W\mu_n C_{ox}}{2L}\left[2(V_{GS} - V_T)V_{DS} - V_{DS}^2\right] \tag{10.44a}$$

which can be written as

$$I_D = \frac{k_n'}{2} \cdot \frac{W}{L} \cdot \left[2(V_{GS} - V_T)V_{DS} - V_{DS}^2\right] \tag{10.44b}$$

or

$$I_D = K_n\left[2(V_{GS} - V_T)V_{DS} - V_{DS}^2\right] \tag{10.44c}$$

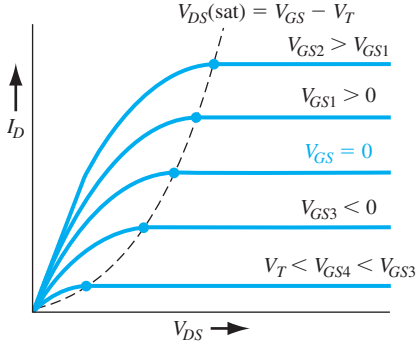


Figure 10.42 | Family of I_D versus V_{DS} curves for an n-channel depletion mode MOSFET.

The parameter $k_n' = \mu_n C_{ox}$ is called the *process conduction parameter* for the n-channel MOSFET and has units of A/V^2. The parameter $K_n = (W\mu_n C_{ox})/2L = (k_n'/2) \cdot (W/L)$ is called the *conduction parameter* for the n-channel MOSFET and also has units of A/V^2.

When the transistor is biased in the saturation region, the ideal currernt–voltage relation is given by

$$I_D = \frac{W\mu_n C_{ox}}{2L}(V_{GS} - V_T)^2 \tag{10.45a}$$

which can be written as

$$I_D = \frac{k_n'}{2} \cdot \frac{W}{L} \cdot (V_{GS} - V_T)^2 \tag{10.45b}$$

or

$$I_D = K_n(V_{GS} - V_T)^2 \tag{10.45c}$$

In general, for a given technology, the process conduction parameter, k_n', is a constant. From Equations (10.44b) and (10.45b), then, we see that the design of a MOSFET, in terms of current capability, is determined by the width-to-length parameter.

The operation of a p-channel device is the same as that of the n-channel device, except the charge carrier is the hole and the conventional current direction and voltage polarities are reversed.

*10.3.3 Current–Voltage Relationship—Mathematical Derivation

In the previous section, we qualitatively discussed the current–voltage characteristics. In this section, we derive the mathematical relation between the drain current, the gate-to-source voltage, and the drain-to-source voltage. Figure 10.43 shows the geometry of the device that we use in this derivation.

In this analysis, we make the following assumptions:

1. The current in the channel is due to drift rather than diffusion.
2. There is no current through the gate oxide.
3. A gradual channel approximation is used in which $\partial E_y/\partial y \gg \partial E_x/\partial x$. This approximation means that E_x is essentially a constant.
4. Any fixed oxide charge is an equivalent charge density at the oxide–semiconductor interface.
5. The carrier mobility in the channel is constant.

We start the analysis with Ohm's law, which can be written as

$$J_x = \sigma E_x \tag{10.46}$$

where σ is the channel conductivity and E_x is the electric field along the channel created by the drain-to-source voltage. The channel conductivity is given by $\sigma = e\mu_n n(y)$, where μ_n is the electron mobility and $n(y)$ is the electron concentration in the inversion layer.

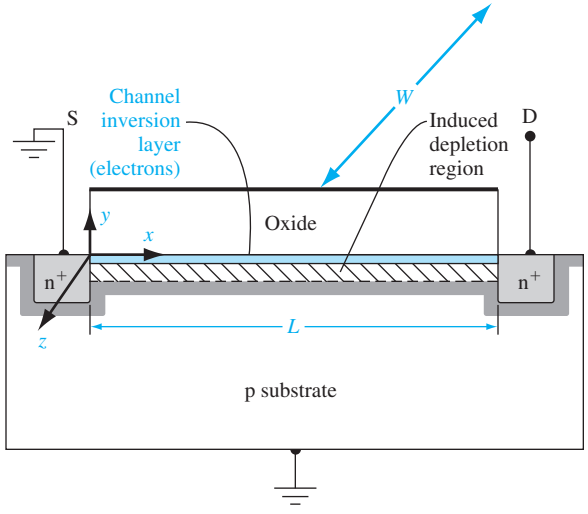


Figure 10.43 | Geometry of a MOSFET for I_D versus V_{DS} derivation.

The total channel current is found by integrating J_x over the cross-sectional area in the y and z directions. Then

$$I_x = \int_y \int_z J_x \, dy \, dz \tag{10.47}$$

We may write that

$$Q'_n = -\int \text{en}(y) \, dy \tag{10.48}$$

where Q'_n is the inversion layer charge per unit area and is a negative quantity for this case.

Equation (10.47) then becomes

$$I_x = -W\mu_n Q'_n \mathrm{E}_x \tag{10.49}$$

where W is the channel width, the result of integrating over z.

Two concepts we use in the currernt–voltage derivation are charge neutrality and Gauss's law. Figure 10.44 shows the charge densities through the device for $V_{GS} > V_T$. The charges are all given in terms of charge per unit area. Using the concept of charge neutrality, we can write

$$Q'_m + Q'_{ss} + Q'_n + Q'_{SD}(\text{max}) = 0 \tag{10.50}$$

The inversion layer charge and induced space charge are negative for this n-channel device.

Gauss's law can be written as

$$\oint_s \epsilon \mathrm{E}_n \, dS = Q_T \tag{10.51}$$

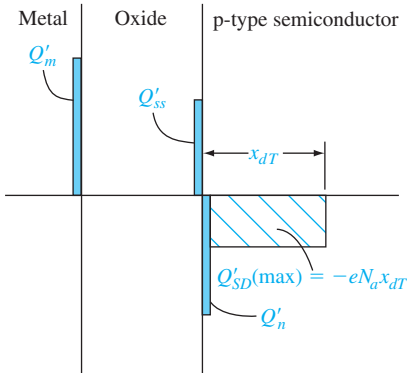


Figure 10.44 | Charge distribution in the n-channel enhancement mode MOSFET for $V_{GS} > V_T$.

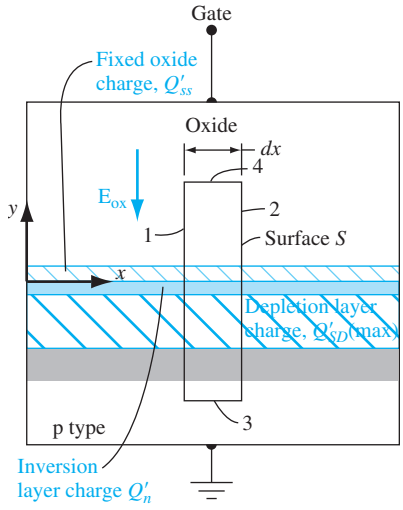


Figure 10.45 | Geometry for applying Gauss's law.

where the integral is over a closed surface. Q_T is the total charge enclosed by the surface, and E_n is the outward directed normal component of the electric field crossing the surface S. Gauss's law is applied to the surface defined in Figure 10.45. Since the surface must be enclosed, we must take into account the two end surfaces in the x-y plane. However, there is no z-component of the electric field so these two end surfaces do not contribute to the integral of Equation (10.51).

Now consider the surfaces labeled 1 and 2 in Figure 10.45. From the gradual channel approximation, we assume that E_x is essentially a constant along the channel length. This assumption means that E_x into surface 2 is the same as E_x out of surface 1. Since the integral in Equation (10.51) involves the outward component of the E-field, the contributions of surfaces 1 and 2 cancel each other. Surface 3 is in the neutral p region, so the electric field is zero at this surface.

Surface 4 is the only surface that contributes to Equation (10.51). Taking into account the direction of the electric field in the oxide, Equation (10.51) becomes

$$\oint_s \epsilon E_n \, dS = -\epsilon_{ox} E_{ox} W \, dx = Q_T \tag{10.52}$$

where ϵ_{ox} is the permittivity of the oxide. The total charge enclosed is

$$Q_T = [Q'_{ss} + Q'_n + Q'_{SD}(\text{max})] W \, dx \tag{10.53}$$

Combining Equations (10.52) and (10.53), we have

$$-\epsilon_{ox} E_{ox} = Q'_{ss} + Q'_n + Q'_{SD}(\text{max}) \tag{10.54}$$

We now need an expression for E_{ox}. Figure 10.46a shows the oxide and channel. We assume that the source is at ground potential. The voltage V_x is the potential in the channel at a point x along the channel length. The potential difference across the oxide at x is a function of V_{GS}, V_x, and the metal–semiconductor work function difference.

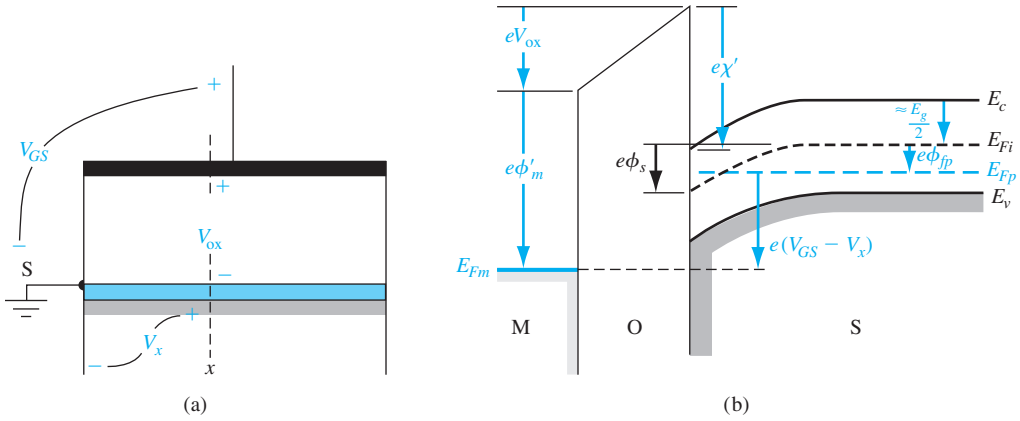


Figure 10.46 | (a) Potentials at a point x along the channel. (b) Energy-band diagram through the MOS structure at the point x.

The energy-band diagram through the MOS structure at point x is shown in Figure 10.46b. The Fermi level in the p-type semiconductor is E_{Fp} and the Fermi level in the metal is E_{Fm}. We have

$$E_{Fp} - E_{Fm} = e(V_{GS} - V_x) \tag{10.55}$$

Considering the potential barriers, we can write

$$V_{GS} - V_x = (\phi'_m + V_{ox}) - \left(\chi' + \frac{E_g}{2e} - \phi_s + \phi_{fp}\right) \tag{10.56}$$

which can also be written as

$$V_{GS} - V_x = V_{ox} + 2\phi_{fp} + \phi_{ms} \tag{10.57}$$

where ϕ_{ms} is the metal–semiconductor work function difference, and $\phi_s = 2\phi_{fp}$ for the inversion condition.

The electric field in the oxide is

$$\mathrm{E}_{ox} = \frac{V_{ox}}{t_{ox}} \tag{10.58}$$

Combining Equations (10.54), (10.57), and (10.58), we find that

$$\begin{aligned} -\epsilon_{ox}\,\mathrm{E}_{ox} &= \frac{-\epsilon_{ox}}{t_{ox}}[(V_{GS} - V_x) - (\phi_{ms} + 2\phi_{fp})] \\ &= Q'_{ss} + Q'_n + Q'_{SD}(\max) \end{aligned} \tag{10.59}$$

The inversion charge density, Q'_n, from Equation (10.59) can be substituted into Equation (10.49) and we obtain

$$I_x = -W\mu_n C_{ox}\frac{dV_x}{dx}[(V_{GS} - V_x) - V_T] \tag{10.60}$$

where $\mathrm{E}_x = -dV_x/dx$ and V_T is the threshold voltage defined by Equation (10.31b).

We can now integrate Equation (10.60) over the length of the channel. We have

$$\int_0^L I_x\,dx = -W\mu_n C_{ox}\int_{V_x(0)}^{V_x(L)}[(V_{GS} - V_T) - V_x]\,dV_x \tag{10.61}$$

We are assuming a constant mobility μ_n. For the n-channel device, the drain current enters the drain terminal and is a constant along the entire channel length. Letting $I_D = -I_x$, Equation (10.61) becomes

$$I_D = \frac{W\mu_n C_{ox}}{2L}[2(V_{GS} - V_T)V_{DS} - V_{DS}^2] \tag{10.62}$$

Equation (10.62) is valid for $V_{GS} \geq V_T$ and for $0 \leq V_{DS} \leq V_{DS}(\text{sat})$.

Equation (10.62) can also be written as

$$I_D = \frac{k_n'}{2} \cdot \frac{W}{L} \cdot [2(V_{GS} - V_T)V_{DS} - V_{DS}^2] = K_n[2(V_{GS} - V_T)\, V_{DS} - V_{DS}^2] \tag{10.63}$$

where k_n' is the process conduction parameter and K_n *i*s the conduction parameter. These parameters are described and defined in Equations (10.44b) and (10.44c).

Figure 10.47 shows plots of Equation (10.62) as a function of V_{DS} for several values of V_{GS}. We can find the value of V_{DS} at the peak current value from $\partial I_D/\partial V_{DS} = 0$. Then, using Equation (10.62), the peak current occurs when

$$V_{DS} = V_{GS} - V_T \tag{10.64}$$

This value of V_{DS} is just $V_{DS}(\text{sat})$, the point at which saturation occurs. For $V_{DS} > V_{DS}(\text{sat})$, the ideal drain current is a constant and is equal to

$$I_D(\text{sat}) = \frac{W\mu_n C_{ox}}{2L}[2(V_{GS} - V_T)V_{DS}(\text{sat}) - V_{DS}^2(\text{sat})] \tag{10.65}$$

Using Equation (10.64) for $V_{DS}(\text{sat})$, Equation (10.65) becomes

$$I_D(\text{sat}) = \frac{W\mu_n C_{ox}}{2L}(V_{GS} - V_T)^2 \tag{10.66}$$

Equation (10.66) can also be written as

$$I_D = \frac{k_n'}{2} \cdot \frac{W}{L} \cdot (V_{GS} - V_T)^2 = K_n (V_{GS} - V_T)^2 \tag{10.67}$$

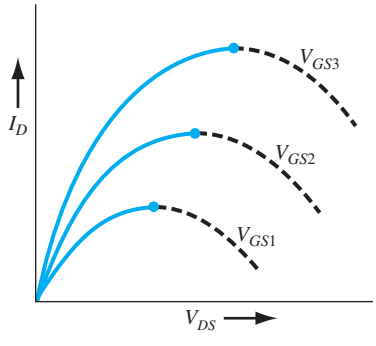


Figure 10.47 | Plots of I_D versus V_{DS} from Equation (10.62).

Equation (10.62) is the ideal current–voltage relationship of the n-channel MOSFET in the nonsaturation region for $0 \leq V_{DS} \leq V_{DS}(\text{sat})$, and Equation (10.66) is the ideal current–voltage relationship of the n-channel MOSFET in the saturation region for $V_{DS} \geq V_{DS}(\text{sat})$. These *I–V* expressions were explicitly derived for an n-channel enhancement mode device. However, these same equations apply to an n-channel depletion mode MOSFET in which the threshold voltage V_T is a negative quantity.

EXAMPLE 10.7

Objective: Design the width of a MOSFET such that a specified current is induced for a given applied bias.

Consider an ideal n-channel MOSFET with parameters $L = 1.25\ \mu\text{m}$, $\mu_n = 650\ \text{cm}^2/\text{V-s}$, $C_{ox} = 6.9 \times 10^{-8}\ \text{F/cm}^2$, and $V_T = 0.65\ \text{V}$. Design the channel width W such that $I_D(\text{sat}) = 4\ \text{mA}$ for $V_{GS} = 5\ \text{V}$.

■ Solution

For the transition biased in the saturation region, we have, from Equation (10.66),

$$I_D(\text{sat}) = \frac{W\mu_n C_{ox}}{2L}(V_{GS} - V_T)^2$$

or

$$4 \times 10^{-3} = \frac{W(650)(6.9 \times 10^{-8})}{2(1.25 \times 10^{-4})} \cdot (5 - 0.65)^2 = 3.39\,W$$

Then

$$W = 11.8\ \mu\text{m}$$

■ Comment

The current capability of a MOSFET is directly proportional to the channel width W. The current handling capability can be increased by increasing W.

■ EXERCISE PROBLEM

Ex 10.7 The parameters of an n-channel silicon MOSFET are $\mu_n = 650\ \text{cm}^2/\text{V-s}$, $t_{ox} = 8\ \text{nm} = 80\ \text{Å}$, $W/L = 12$, and $V_T = 0.40\ \text{V}$. If the transistor is biased in the saturation region, find the drain current for (*a*) $V_{GS} = 0.8\ \text{V}$, (*b*) $V_{GS} = 1.2\ \text{V}$, and (*c*) $V_{GS} = 1.6\ \text{V}$.

[(*a*) 0.269 mA; (*b*) 1.077 mA; (*c*) 2.423 mA]

We can use the *I–V* relations to experimentally determine the mobility and threshold voltage parameters. From Equation (10.62), we can write, for very small values of V_{DS},

$$I_D = \frac{W\mu_n C_{ox}}{L}(V_{GS} - V_T)V_{DS} \tag{10.68}$$

Figure 10.48a shows a plot of Equation (10.68) as a function of V_{GS} for constant V_{DS}. A straight line is fitted through the points. The deviation from the straight line at low values of V_{GS} is due to subthreshold conduction and the deviation at higher values of V_{GS}

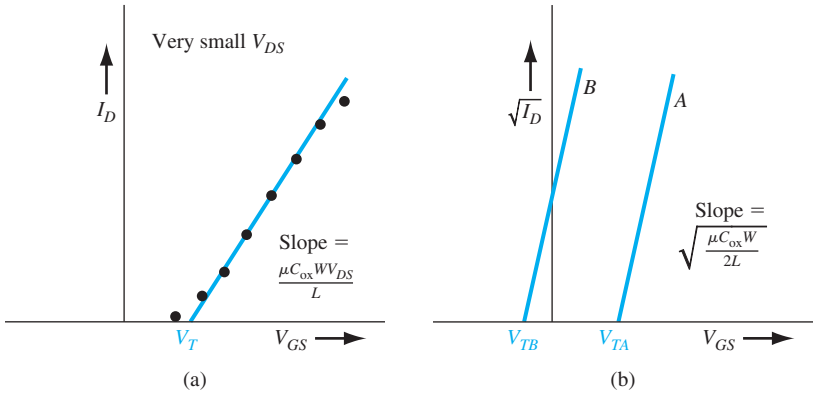


Figure 10.48 | (a) I_D versus V_{GS} (for small V_{DS}) for enhancement mode MOSFET. (b) Ideal $\sqrt{I_D}$ versus V_{GS} in saturation region for enhancement mode (curve A) and depletion mode (curve B) n-channel MOSFETs.

is due to mobility being a function of gate voltage. Both of these effects will be considered in the next chapter. The extrapolation of the straight line to zero current gives the threshold voltage, and the slope is proportional to the inversion carrier mobility.

Now consider the case when the transistor is biased in the saturation region. If we take the square root of Equation (10.66), we obtain

$$\sqrt{I_{D(\text{sat})}} = \sqrt{\frac{W\mu_n C_{\text{ox}}}{2L}}(V_{GS} - V_T) \tag{10.69}$$

Figure 10.48b is a plot of Equation (10.69). In the ideal case, we can obtain the same information from both curves. However, as we will see in the next chapter, the threshold voltage may be a function of V_{DS} in short-channel devices. Since Equation (10.69) applies to devices biased in the saturation region, the V_T parameter in this equation may differ from the extrapolated value determined in Figure 10.48a. In general, the nonsaturation current–voltage characteristics will produce the more reliable data.

EXAMPLE 10.8

Objective: Determine the inversion carrier mobility from experimental results.

Consider an n-channel MOSFET with $W = 15\ \mu\text{m}$, $L = 2\ \mu\text{m}$, and $C_{\text{ox}} = 6.9 \times 10^{-8}\ \text{F/cm}^2$. Assume that the drain current in the nonsaturation region for $V_{DS} = 0.10$ V is $I_D = 35\ \mu\text{A}$ at $V_{GS} = 1.5$ V and $I_D = 75\ \mu\text{A}$ at $V_{GS} = 2.5$ V.

■ Solution

From Equation (10.68), we can write

$$I_{D2} - I_{D1} = \frac{W\mu_n C_{\text{ox}}}{L}(V_{GS2} - V_{GS1})V_{DS}$$

so that

$$75 \times 10^{-6} - 35 \times 10^{-6} = \left(\frac{15}{2}\right)\mu_n(6.9 \times 10^{-8})(2.5 - 1.5)(0.10)$$

which yields

$$\mu_n = 773 \text{ cm}^2/\text{V-s}$$

We can then determine

$$V_T = 0.625 \text{ V}$$

Comment

The mobility of carriers in the inversion layer is less than that in the bulk semiconductor due to the surface scattering effect. We will discuss this effect in the next chapter.

EXERCISE PROBLEM

Ex 10.8 An n-channel silicon MOSFET has the following parameters: $W = 6\ \mu$m, $L = 1.5\ \mu$m, and $t_{ox} = 8$ nm $= 80$ Å. When the transistor is biased in the saturation region, the drain current is $I_D(sat) = 0.132$ mA at $V_{GS} = 1.0$ V and $I_D(sat) = 0.295$ mA at $V_{GS} = 1.25$ V. Determine the electron mobility and the threshold voltage.

(Ans. $\mu_n = 600 \text{ cm}^2/\text{V-s}$, $V_T = 0.496$ V)

The current–voltage relationship of a p-channel device can be obtained by the same type of analysis. Figure 10.49 shows a p-channel enhancement mode MOSFET. The voltage polarities and current direction are the reverse of those in the n-channel device. We may note the change in the subscript notation for this device. For the current direction shown in the figure, the *I–V* relation for the p-channel MOSFET biased in the nonsaturation region is

$$I_D = \frac{W\mu_p C_{ox}}{2L}\left[2(V_{SG} + V_T)V_{SD} - V_{SD}^2\right] \tag{10.70}$$

Equation (10.70) is valid for $0 \le V_{SD} \le V_{SD}(sat)$.

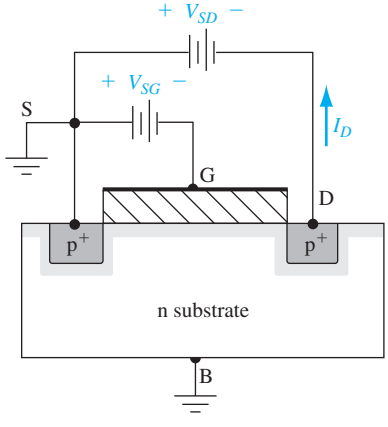


Figure 10.49 | Cross section and bias configuration for a p-channel enhancement mode MOSFET.

Equation (10.70) can also be written as

$$I_D = \frac{k_p'}{2} \cdot \frac{W}{L} \cdot \left[2(V_{SG} + V_T)V_{SD} - V_{SD}^2\right] = K_p\left[2(V_{SG} + V_T)V_{SD} - V_{SD}^2\right] \tag{10.71}$$

where $k_p' = \mu_p C_{ox}$ is the *process conduction parameter* for the p-channel MOSFET and $K_p = (W\mu_p C_{ox})/(2L) = (k_p'/2) \cdot (W/L)$ is the *conduction parameter.*

When the transistor is biased in the saturation region, the *I–V* relation is given by

$$\boxed{I_D(\text{sat}) = \frac{W\mu_p C_{\text{ox}}}{2L}(V_{SG} + V_T)^2} \tag{10.72}$$

Equation (10.72) is valid for $V_{SD} \geq V_{SD}\,(sat)$.

Equation (10.72) can also be written as

$$I_D = \frac{k_p'}{2} \cdot \frac{W}{L} \cdot (V_{SG} + V_T)^2 = K_p(V_{SG} + V_T)^2 \tag{10.73}$$

The source-to-drain saturation voltage is given by

$$V_{SD}(sat) = V_{SG} + V_T \tag{10.74}$$

Note the change in the sign in front of V_T and note that the mobility is now the mobility of the holes in the hole inversion layer charge. Keep in mind that V_T is negative for a p-channel enhancement mode MOSFET and positive for a depletion mode p-channel device.

One assumption we made in the derivation of the currernt–voltage relationship was that the charge neutrality condition given by Equation (10.50) was valid over the entire length of the channel. We implicitly assumed that $Q_{SD}'(\text{max})$ was constant along the length of the channel. The space charge width, however, varies between source and drain due to the drain-to-source voltage; it is widest at the drain when $V_{DS} > 0$. A change in the space charge density along the channel length must be balanced by a corresponding change in the inversion layer charge. An increase in the space charge width means that the inversion layer charge is reduced, implying that the drain current and drain-to-source saturation voltage are less than the ideal values. The actual saturation drain current may be as much as 20 percent less than the predicted value due to this bulk charge effect.

10.3.4 Transconductance

The MOSFET transconductance is defined as the change in drain current with respect to the corresponding change in gate voltage, or

$$g_m = \frac{\partial I_D}{\partial V_{GS}} \tag{10.75}$$

The transconductance is sometimes referred to as the transistor gain.

If we consider an n-channel MOSFET operating in the nonsaturation region, then using Equation (10.62), we have

$$g_{mL} = \frac{\partial I_D}{\partial V_{GS}} = \frac{W\mu_n C_{ox}}{L} \cdot V_{DS} \tag{10.76}$$

The transconductance increases linearly with V_{DS} but is independent of V_{GS} in the nonsaturation region.

The *I–V* characteristics of an n-channel MOSFET in the saturation region are given by Equation (10.66). The transconductance in this region of operation is given by

$$g_{ms} = \frac{\partial I_D(\text{sat})}{\partial V_{GS}} = \frac{W\mu_n C_{ox}}{L}(V_{GS} - V_T) \tag{10.77}$$

In the saturation region, the transconductance is a linear function of V_{GS} and is independent of V_{DS}.

The transconductance is a function of the geometry of the device as well as of carrier mobility and threshold voltage. The transconductance increases as the width of the device increases, and it also increases as the channel length and oxide thickness decrease. In the design of MOSFET circuits, the size of the transistor, in particular the channel width W, is an important engineering design parameter.

10.3.5 Substrate Bias Effects

In all of our analyses so far, the substrate, or body, has been connected to the source and held at ground potential. In MOSFET circuits, the source and body may not be at the same potential. Figure 10.50a shows an n-channel MOSFET and the associated double-subscripted voltage variables. The source-to-substrate pn junction must always be zero or reverse biased, so V_{SB} must always be greater than or equal to zero.

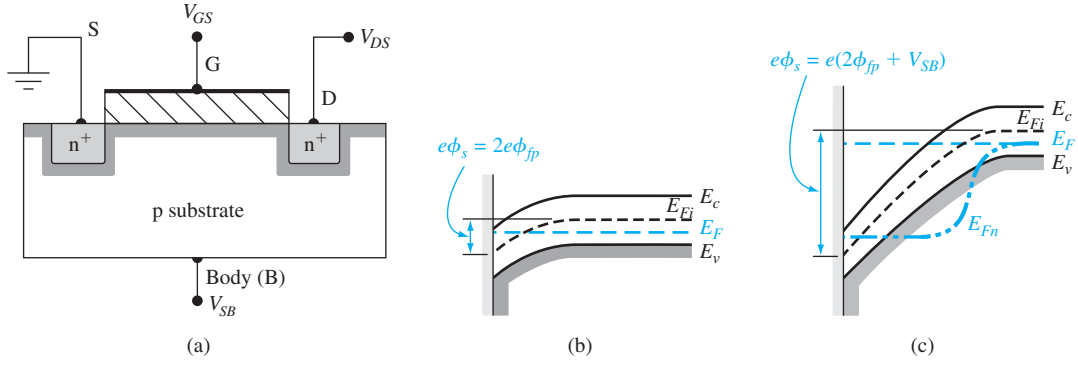


Figure 10.50 | (a) Applied voltages on an n-channel MOSFET. (b) Energy-band diagram at inversion point when $V_{SB} = 0$. (c) Energy-band diagram at inversion point when $V_{SB} > 0$ is applied.

If $V_{SB} = 0$, threshold is defined as the condition when $\phi_s = 2\phi_{fp}$ as we discussed previously and as shown in Figure 10.50b. When $V_{SB} > 0$ the surface will still try to invert when $\phi_s = 2\phi_{fp}$. However, these electrons are at a higher potential energy than are the electrons in the source. The newly created electrons will move laterally and flow out of the source terminal. When $\phi_s = 2\phi_{fp} + V_{SB}$, the surface reaches an equilibrium inversion condition. The energy-band diagram for this condition is shown in Figure 10.50c. The curve represented as E_{Fn} is the Fermi level from the p substrate through the reverse-biased source–substrate junction to the source contact.

The space charge region width under the oxide increases from the original x_{dT} value when a reverse-biased source–substrate junction voltage is applied. With an applied $V_{SB} > 0$, there is more charge associated with this region. Considering the charge neutrality condition through the MOS structure, the positive charge on the top metal gate must increase to compensate for the increased negative space charge in order to reach the threshold inversion point. So when $V_{SB} > 0$, the threshold voltage of the n-channel MOSFET increases.

When $V_{SB} = 0$, we had

$$Q'_{SD}\,(\text{max}) = -eN_a x_{dT} = -\sqrt{2e\epsilon_s N_a(2\phi_{fp})} \tag{10.78}$$

When $V_{SB} > 0$, the space charge width increases and we now have

$$Q'_{SD} = -eN_a x_d = -\sqrt{2e\epsilon_s N_a(2\phi_{fp} + V_{SB})} \tag{10.79}$$

The change in the space charge density is then

$$\Delta Q'_{SD} = -\sqrt{2e\epsilon_s N_a}\left[\sqrt{2\phi_{fp} + V_{SB}} - \sqrt{2\phi_{fp}}\right] \tag{10.80}$$

To reach the threshold condition, the applied gate voltage must be increased. The change in threshold voltage can be written as

$$\Delta V_T = -\frac{\Delta Q'_{SD}}{C_{\text{ox}}} = \frac{\sqrt{2e\epsilon_s N_a}}{C_{\text{ox}}}\left[\sqrt{2\phi_{fp} + V_{SB}} - \sqrt{2\phi_{fp}}\right] \tag{10.81}$$

where $\Delta V_T = V_T(V_{SB} > 0) - V_T(V_{SB} = 0)$. We may note that V_{SB} must always be positive so that, for the n-channel device, ΔV_T is always positive. The threshold voltage of the n-channel MOSFET will increase as a function of the source–substrate junction voltage.

From Equation (10.81), we may define

$$\gamma = \frac{\sqrt{2e\epsilon_s N_a}}{C_{ox}} \tag{10.82}$$

where γ is defined as the *body-effect coefficient*. Equation (10.81) may then be written as

$$\Delta V_T = \gamma\left[\sqrt{2\phi_{fp} + V_{SB}} - \sqrt{2\phi_{fp}}\right] \tag{10.83}$$

EXAMPLE 10.9

Objective: Calculate the body-effect coefficient and the change in the threshold voltage due to an applied source-to-body voltage.

Consider an n-channel silicon MOSFET at $T = 300$ K. Assume the substrate is doped to $N_a = 3 \times 10^{16}$ cm^{-3} and assume the oxide is silicon dioxide with a thickness of $t_{ox} = 20$ nm $= 200$ Å. Let $V_{SB} = 1$ V.

■ Solution

We can calculate that

$$\phi_{fp} = V_t \ln\left(\frac{N_a}{n_i}\right) = (0.0259)\ln\left(\frac{3 \times 10^{16}}{1.5 \times 10^{10}}\right) = 0.3758 \text{ V}$$

and

$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} = \frac{(3.9)(8.85 \times 10^{-14})}{200 \times 10^{-8}} = 1.726 \times 10^{-7} \text{ F/cm}^2$$

From Equation (10.82), we find the body-effect coefficient to be

$$\gamma = \frac{\sqrt{2e\epsilon_s N_a}}{C_{ox}} = \frac{[2(1.6 \times 10^{-19})(11.7)(8.85 \times 10^{-14})(3 \times 10^{16})]^{1/2}}{1.726 \times 10^{-7}}$$

or

$$\gamma = 0.5776 \text{ V}^{1/2}$$

The change in threshold voltage for $V_{SB} = 1$ V is found to be

$$\begin{aligned}\Delta V_T &= \gamma\left[\sqrt{2\phi_{fp} + V_{SB}} - \sqrt{2\phi_{fp}}\right] \\ &= (0.5776)\left[\sqrt{2(0.3758) + 1} - \sqrt{2(0.3758)}\right] \\ &= (0.5776)[1.3235 - 0.8669] = 0.264 \text{ V}\end{aligned}$$

■ Comment

Figure 10.51 shows plots of $\sqrt{I_D(sat)}$ versus V_{GS} for various applied values of V_{SB}. The original threshold voltage is assumed to be $V_{TO} = 0.64$ V.

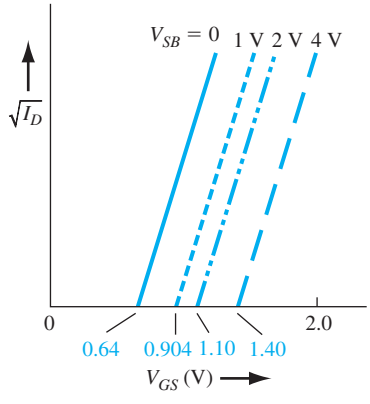


Figure 10.51 | Plots of $\sqrt{I_D}$ versus V_{GS} at several values of V_{SB} for an n-channel MOSFET.

EXERCISE PROBLEM

Ex 10.9 A silicon MOSFET has the following parameters: $N_a = 10^{16}$ cm^{-3} and $t_{ox} = 12$ nm $= 120$ Å. Calculate (*a*) the body-effect coefficient and (*b*) the change in threshold voltage for (*i*) $V_{SB} = 1$ and (*ii*) $V_{SB} = 2$ V.

[Ans. (*a*) $\gamma = 0.200\ \text{V}^{1/2}$; (*b*) (*i*) $\Delta V_T = 0.093\text{7 V}$, (*ii*) $\Delta V_T = 0.162$ V]

If a body or substrate bias is applied to a p-channel device, the threshold voltage is shifted to more negative values. Because the threshold voltage of a p-channel enhancement mode MOSFET is negative, a body voltage will increase the applied negative gate voltage required to create inversion. The same general observation was made for the n-channel MOSFET.

TEST YOUR UNDERSTANDING

TYU 10.6 The silicon n-channel MOSFET described in Exercise Problem Ex 10.7 is to be redesigned by changing the W/L ratio such that $I_D = 100$ μA when the transistor is biased in the saturation regin with $V_{GS} = 1.0$ V.

(Ans. $W/L = 1.98$)

TYU 10.7 The parameters of a p-channel MOSFET are $\mu_p = 310$ cm^2/V-s, $t_{ox} = 220$ Å, $W/L = 60$, and $V_T = -0.40$ V. If the transistor is biased in the saturation region, find the drain current for $V_{SG} = 1$, 1.5, and 2 V.

(Ans. $I_D = 0.526$, 1.77 and 3.74 mA)

TYU 10.8 The p-channel MOSFET in TYU 10.7 is to be redesigned by changing the (W/L) ratio such that $I_D = 200$ μA when the transistor is biased in the saturation region with $V_{SG} = 1.25$ V.

(Ans. $W/L = 11.4$)

TYU 10.9 Repeat Exercise Problem Ex 10.9 for a substrate impurity doping concentration of $N_a = 10^{15}$ cm^{-3}.

[Ans. (*a*) $\gamma = 0.0633\ \text{V}^{1/2}$; (*b*) (*i*) $\Delta V_T = 0.0314$ V, (*ii*) $\Delta V_T = 0.0536$ V]

10.4 | FREQUENCY LIMITATIONS

In many applications, the MOSFET is used in a linear amplifier circuit. A small-signal equivalent circuit for the MOSFET is needed in order to mathematically analyze the electronic circuit. The equivalent circuit contains capacitances and resistances that introduce frequency effects. We initially develop a small-signal equivalent circuit and then discuss the physical factors that limit the frequency response of the MOSFET. A transistor cutoff frequency, which is a figure of merit, is then defined and an expression derived for this factor.

10.4.1 Small-Signal Equivalent Circuit

The small-signal equivalent circuit of the MOSFET is constructed from the basic MOSFET geometry. A model based on the inherent capacitances and resistances

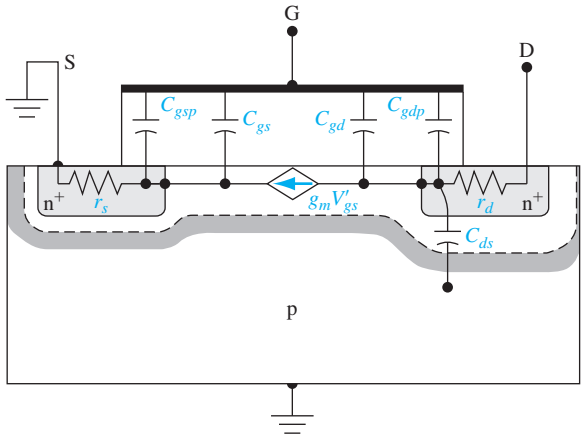


Figure 10.52 | Inherent resistances and capacitances in the n-channel MOSFET structure.

within the transistor structure, along with elements that represent the basic device equations, is shown in Figure 10.52. One simplifying assumption we will make in the equivalent circuit is that the source and substrate are both tied to ground potential.

Two of the capacitances connected to the gate are inherent in the device. These capacitances are C_{gs} and C_{gd}, which represent the interaction between the gate and the channel charge near the source and drain terminals, respectively. The remaining two gate capacitances, C_{gsp} and C_{gdp}, are parasitic or overlap capacitances. In real devices, the gate oxide will overlap the source and drain contacts because of tolerance or fabrication factors. As we will see, the drain overlap capacitance—C_{gdp}, in particular—will lower the frequency response of the device. The parameter C_{ds} is the drain-to-substrate pn junction capacitance, and r_s and r_d are the series resistances associated with the source and drain terminals. The small-signal channel current is controlled by the internal gate-to-source voltage through the transconductance.

The small-signal equivalent circuit for the n-channel common-source MOSFET is shown in Figure 10.53. The voltage V'_{gs} is the internal gate-to-source voltage that

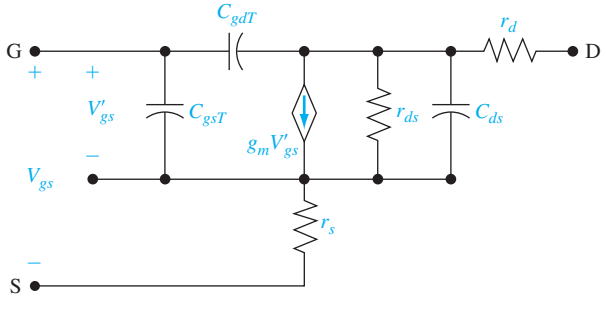


Figure 10.53 | Small-signal equivalent circuit of a common-source n-channel MOSFET.

controls the channel current. The parameters C_{gsT} and C_{gdT} are the total gate-to-source and total gate-to-drain capacitances. One parameter, r_{ds}, shown in Figure 10.53, is not shown in Figure 10.52. This resistance is associated with the slope I_D versus V_{DS}. In the ideal MOSFET biased in the saturation region, I_D is independent of V_{DS} so that r_{ds} would be infinite. In short-channel-length devices, in particular, r_{ds} is finite because of channel length modulation, which we will consider in the next chapter.

A simplified small-signal equivalent circuit valid at low frequency is shown in Figure 10.54. The series resistances, r_s and r_d, have been neglected, so the drain current is essentially only a function of the gate-to-source voltage through the transconductance. The input gate impedance is infinite in this simplified model.

The source resistance r_s can have a significant effect on the transistor characteristics. Figure 10.55 shows a simplified, low-frequency equivalent circuit including r_s but neglecting r_{ds}. The drain current is given by

$$I_d = g_m V'_{gs} \tag{10.84}$$

and the relation between V_{gs} and V'_{gs} can be found from

$$V_{gs} = V'_{gs} + (g_m V'_{gs}) r_s = (1 + g_m r_s) V'_{gs} \tag{10.85}$$

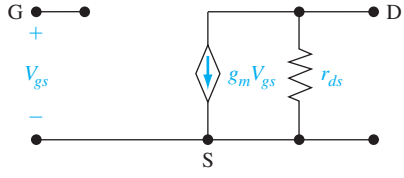


Figure 10.54 | Simplified, low-frequency small-signal equivalent circuit of a common-source n-channel MOSFET.

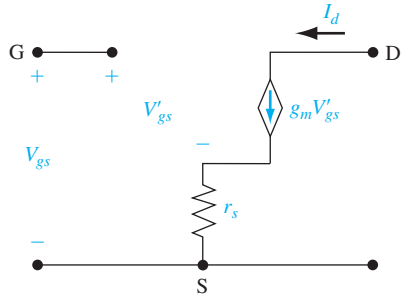


Figure 10.55 | Simplified, low-frequency small-signal equivalent circuit of common-source n-channel MOSFET including source resistance r_s.

The drain current from Equation (10.84) can now be written as

$$I_d = \left(\frac{g_m}{1 + g_m r_s}\right) V_{gs} = g'_m V_{gs} \tag{10.86}$$

The source resistance reduces the effective transconductance or transistor gain.

The equivalent circuit of the p-channel MOSFET is exactly the same as that of the n-channel except that all voltage polarities and current directions are reversed. The same capacitances and resistances that are in the n-channel model apply to the p-channel model.

10.4.2 Frequency Limitation Factors and Cutoff Frequency

There are two basic frequency limitation factors in the MOSFET. The first factor is the channel transit time. If we assume that carriers are traveling at their saturation drift velocity v_{sat}, then the transit time is $\tau_t = L/v_{\text{sat}}$ where L is the channel length. If $v_{\text{sat}} = 10^7$ cm/s and $L = 1$ μm, then $\tau_t = 10$ ps, which translates into a maximum frequency of 100 GHz. This frequency is much larger than the typical maximum frequency response of a MOSFET. The transit time of carriers through the channel is usually not the limiting factor in the frequency responses of MOSFETs.

The second limiting factor is the gate or capacitance charging time. If we neglect r_s, r_d, r_{ds}, and C_{ds}, the resulting equivalent small-signal circuit is shown in Figure 10.56 where R_L is a load resistance.

The input gate impedance in this equivalent circuit is no longer infinite. Summing currents at the input gate node, we have

$$I_i = j\omega C_{gsT} V_{gs} + j\omega C_{gdT}(V_{gs} - V_d) \tag{10.87}$$

where I_i is the input current. Likewise, summing currents at the output drain node, we have

$$\frac{V_d}{R_L} + g_m V_{gs} + j\omega C_{gdT}(V_d - V_{gs}) = 0 \tag{10.88}$$

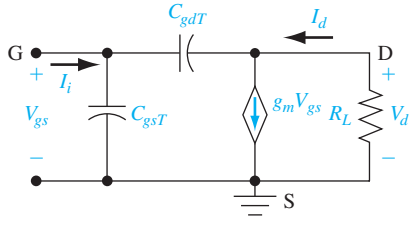


Figure 10.56 | High-frequency small-signal equivalent circuit of common-source n-channel MOSFET.

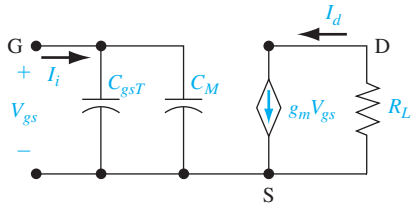


Figure 10.57 | Small-signal equivalent circuit including Miller capacitance.

Combining Equations (10.87) and (10.88) to eliminate the voltage variable V_d, we can determine the input current as

$$I_i = j\omega\left[C_{gsT} + C_{gdT}\left(\frac{1 + g_m R_L}{1 + j\omega\, R_L C_{gdT}}\right)\right]V_{gs} \tag{10.89}$$

Normally, $\omega R_L C_{gdT}$ is much less than unity; therefore, we may neglect the $(j\omega R_L C_{gdT})$ term in the denominator. Equation (10.89) then simplifies to

$$I_i = j\omega[C_{gsT} + C_{gdT}(1 + g_m R_L)]V_{gs} \tag{10.90}$$

Figure 10.57 shows the equivalent circuit with the equivalent input impedance described by Equation (10.90). The parameter C_M is the Miller capacitance and is given by

$$C_M = C_{gdT}(1 + g_m R_L) \tag{10.91}$$

The serious effect of the drain overlap capacitance now becomes apparent. When the transistor is operating in the saturation region, C_{gd} essentially becomes zero, but C_{gdp} is a constant. This parasitic capacitance is multiplied by the gain of the transistor and can become a significant factor in the input impedance.

The cutoff frequency f_T is defined to be the frequency at which the magnitude of the current gain of the device is unity, or when the magnitude of the input current I_i is equal to the ideal load current I_d. From Figure 10.57, we can see that

$$I_i = j\omega(C_{gsT} + C_M)V_{gs} \tag{10.92}$$

and the ideal load current is

$$I_d = g_m V_{gs} \tag{10.93}$$

The magnitude of the current gain is then

$$\left|\frac{I_d}{I_i}\right| = \frac{g_m}{2\pi f(C_{gsT} + C_M)} \tag{10.94}$$

Setting the magnitude of the current gain equal to unity at the cutoff frequency, we find

$$f_T = \frac{g_m}{2\pi\,(C_{gsT} + C_M)} = \frac{g_m}{2\pi\, C_G} \tag{10.95}$$

where C_G is the equivalent input gate capacitance.

In the ideal MOSFET, the overlap or parasitic capacitances, C_{gsp} and C_{gdp}, are zero. Also, when the transistor is biased in the saturation region, C_{gd} approaches zero and C_{gs} is approximately $C_{ox}WL$. The transconductance of the ideal MOSFET biased in the saturation region and assuming a constant mobility is given by Equation (10.77) as

$$g_{ms} = \frac{W\mu_n C_{ox}}{L}(V_{GS} - V_T)$$

Then, for this ideal case, the cutoff frequency is

$$f_T = \frac{g_m}{2\pi C_G} = \frac{\dfrac{W\mu_n C_{ox}}{L}(V_{GS} - V_T)}{2\pi(C_{ox}WL)} = \frac{\mu_n(V_{GS} - V_T)}{2\pi L^2} \tag{10.96}$$

EXAMPLE 10.10

Objective: Calculate the cutoff frequency of an ideal MOSFET with a constant mobility.

Assume that the electron mobility in an n-channel device is $\mu_n = 400$ cm^2/V-s and that the channel length is $L = 4$ μm. Also assume that $V_T = 1$ V and $V_{GS} = 3$ V.

■ Solution

From Equation (10.96), the cutoff frequency is

$$f_T = \frac{\mu_n(V_{GS} - V_T)}{2\pi L^2} = \frac{400(3-1)}{2\pi\,(4 \times 10^{-4})^2} = 796 \text{ MHz}$$

■ Comment

In an actual MOSFET, the effect of the parasitic capacitance will substantially reduce the cutoff frequency from that calculated in this example.

■ EXERCISE PROBLEM

Ex 10.10 An n-channel silicon MOSFET has the following parameters: $\mu_n = 420$ cm^2/V-s, $t_{ox} = 18$ nm $= 180$ Å, $L = 1.2$ μm, $W = 24$ μm, and $V_T = 0.4$ V. The transistor is biased in the saturation region at $V_{GS} = 1.5$ V. Determine the cutoff frequency.

(Ans. $f_T = 5.11$ GHz)

TEST YOUR UNDERSTANDING

TYU 10.10 Consider the n-channel MOSFET described in Exercise Problem Ex 10.10. The transistor is connected to an effective load resistance of $R_L = 100$ kΩ. Calculate the ratio of Miller capacitance C_M to gate-to-drain capacitance C_{gdT}.

(Ans. 178)

*10.5 | THE CMOS TECHNOLOGY

The primary objective of this book is to present the basic physics of semiconductor materials and devices without considering in detail the various fabrication processes; this important subject is left to other books. However, there is one MOS technology

that is used extensively, for which the basic fabrication techniques must be considered in order to understand essential characteristics of these devices and circuits. The MOS technology we consider briefly is the complementary MOS, or CMOS, process.

We have considered the physics of both n-channel and p-channel enhancement mode MOSFETs. Both devices are used in a CMOS inverter, which is the basis of CMOS digital logic circuits. The dc power dissipation in a digital circuit can be reduced to very low levels by using a complementary p-channel and n-channel pair.

It is necessary to form electrically isolated p- and n-substrate regions in an integrated circuit to accommodate the n- and p-channel transistors. The p-well process has been a commonly used technique for CMOS circuits. The process starts with a fairly low doped n-type silicon substrate in which the p-channel MOSFET will be fabricated. A diffused p region, called a p well, is formed in which the n-channel MOSFET will be fabricated. In most cases, the p-type substrate doping level must be larger than the n-type substrate doping level to obtain the desired threshold voltages. The larger p doping can easily compensate the initial n doping to form the p well. A simplified cross section of the p-well CMOS structure is shown in Figure 10.58a. The notation FOX stands for field oxide, which is a relatively thick oxide separating the devices. The FOX prevents either the n or p substrate from becoming inverted and helps maintain isolation between the two devices. In practice, additional processing steps must be included; for example, providing connections so that the p well and n substrate can be electrically connected to the appropriate voltages. The n substrate

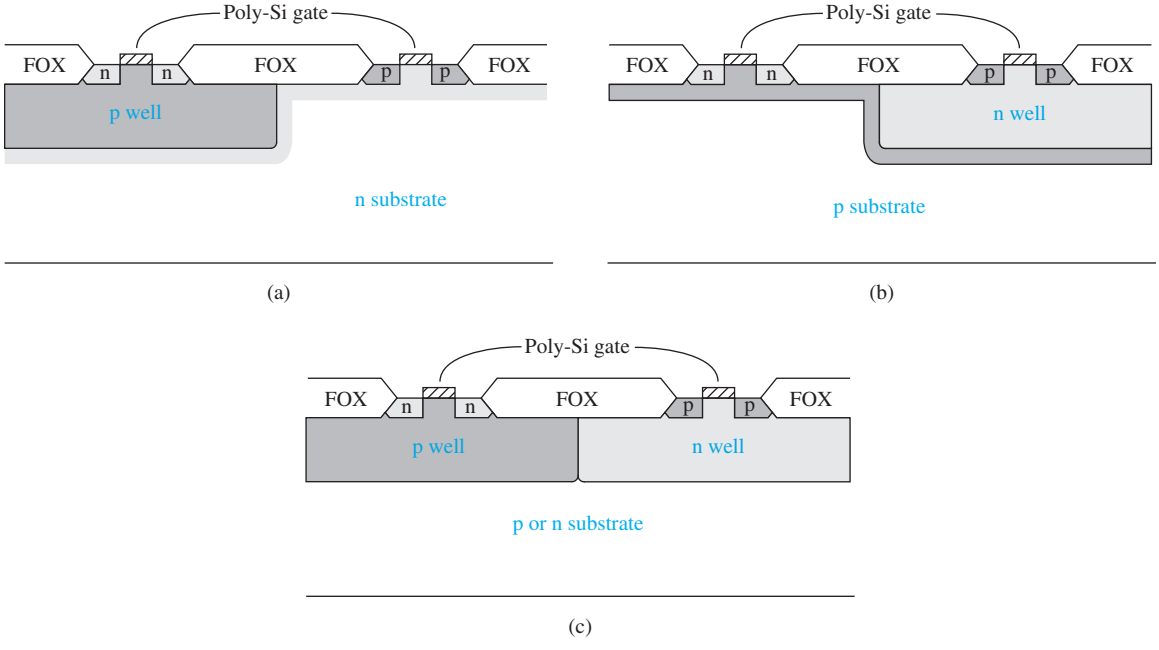


Figure 10.58 | CMOS structures: (a) p well, (b) n well, and (c) twin well.
(From Yang [22].)

must always be at a higher potential than the p well; therefore, this pn junction will always be reverse biased.

With ion implantation now being extensively used for threshold voltage control, both the n-well CMOS process and twin-well CMOS process can be used. The n-well CMOS process, shown in Figure 10.58b, starts with an optimized p-type substrate that is used to form the n-channel MOSFETs. (The n-channel MOSFETs, in general, have superior characteristics, so this starting point should yield excellent n-channel devices.) The n well is then added, in which the p-channel devices are fabricated. The n-well doping can be controlled by ion implantation.

The twin-well CMOS process, shown in Figure 10.58c, allows both the p-well and n-well regions to be optimally doped to control the threshold voltage and transconductance of each transistor. The twin-well process allows a higher packing density because of self-aligned channel stops.

One major problem in CMOS circuits has been latch-up. *Latch-up* refers to a high-current, low-voltage condition that may occur in a four-layer pnpn structure. Figure 10.59a shows the circuit of a CMOS inverter and Figure 10.59b shows a simplified integrated circuit layout of the inverter circuit. In the CMOS layout, p^+ source to n substrate to p well to n^+ source forms such a four-layer structure.

The equivalent circuit of this four-layer structure is shown in Figure 10.60. The silicon-controlled rectifier action involves the interaction of the parasitic pnp and npn bipolar transistors. Bipolar transistors are discussed in Chapter 12. The npn transistor corresponds to the vertical n^+-source to p-well to n-substrate structure and the pnp transistor corresponds to the lateral p-well to n-substrate to p^+-source structure. Under normal CMOS operation, both parasitic bipolar transistors are cut off. However, under certain conditions, avalanche breakdown may occur in the p-well to n-substrate junction, driving both bipolar transistors into saturation. This high-current, low-voltage condition—latch-up—can sustain itself by positive feedback. The condition can prevent the CMOS circuit from operating and can also cause permanent damage and burnout of the circuit.

Latch-up can be prevented if the product $\beta_n \beta_p$ is less than unity at all times, where β_n and β_p are the common-emitter current gains of the npn and pnp parasitic

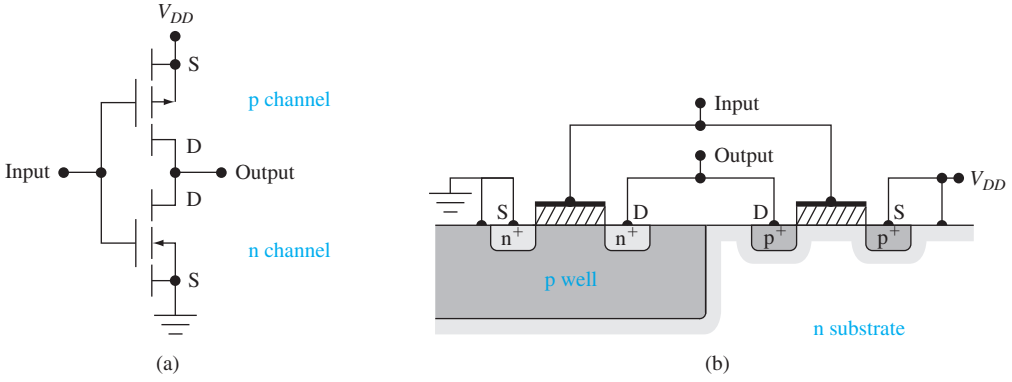


Figure 10.59 | (a) CMOS inverter circuit. (b) Simplified integrated circuit cross section of CMOS inverter.

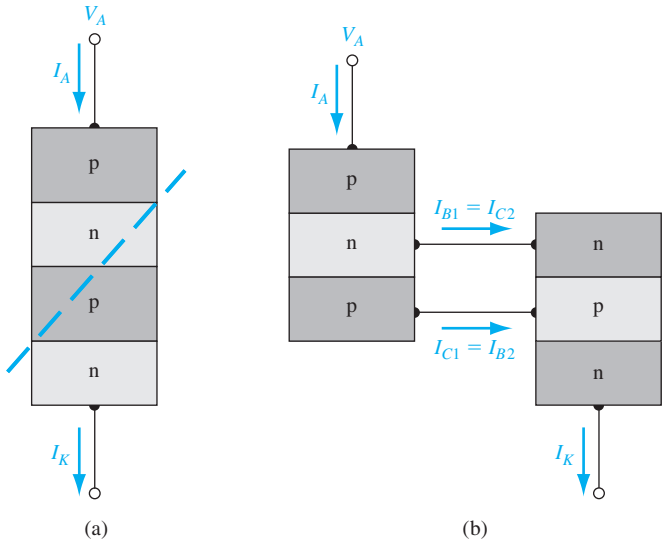


Figure 10.60 | (a) The splitting of the basic pnpn structure. (b) The two-transistor equivalent circuit of the four-layered pnpn device.

bipolar transistors, respectively. One method of preventing latch-up is to "kill" the minority carrier lifetime. Minority carrier lifetime degradation can be accomplished by gold doping or neutron irradiation, either of which introduces deep traps within the semiconductor. The deep traps increase the excess minority carrier recombination rate and reduce current gain. A second method of preventing latch-up is by using proper circuit layout techniques. If the two bipolar transistors can be effectively decoupled, then latch-up can be minimized or prevented. The two parasitic bipolar transistors can also be decoupled by using a different fabrication technology. The silicon-on-insulator technology, for example, allows the n-channel and the p-channel MOSFETs to be isolated from each other by an insulator. This isolation decouples the parasitic bipolar transistors.

10.6 | SUMMARY

- The fundamental physics and characteristics of the metal-oxide–semiconductor field-effect transistor (MOSFET) have been considered in this chapter.
- The heart of the MOSFET is the MOS capacitor. The energy bands in the semiconductor adjacent to the oxide–semiconductor interface bend, depending upon the voltage applied to the gate.
- An inversion layer of electrons can be created at the oxide–semiconductor surface in a p-type semiconductor by applying a sufficiently positive gate voltage, and an inversion layer of holes can be created at the oxide–semiconductor surface in an n-type semiconductor by applying a sufficiently negative gate voltage.
- The threshold voltage is the applied gate voltage required to reach the threshold inversion point. The flat-band voltage was defined and discussed.

- The n-channel MOSFET, both enhancement mode and depletion mode, and the p-channel MOSFET, both enhancement mode and depletion mode, were described.
- The basic transistor action is the modulation of the current at the drain terminal by the gate-to-source voltage.
- The ideal MOSFET current–voltage relations were derived.
- The body-effect coefficient was defined and discussed. The expression for the shift in threshold voltage due to the body effect was derived.
- A small-signal equivalent circuit of the MOSFET was developed.
- Various physical factors in the MOSFET that affect the frequency limitations were discussed. An expression for the cutoff frequency was developed.
- The CMOS technology was briefly considered.

GLOSSARY OF IMPORTANT TERMS

accumulation layer charge The induced charge directly under an oxide that is in excess of the thermal-equilibrium majority carrier concentration.

channel conductance The ratio of drain current to drain-to-source voltage in the limit as $V_{DS} \rightarrow 0$.

channel conductance modulation The process whereby the channel conductance varies with gate-to-source voltage.

CMOS Complementary MOS; the technology that uses both p- and n-channel devices in an electronic circuit fabricated in a single semiconductor chip.

conduction parameter The multiplying coefficient of the voltage terms to obtain the MOSFET drain current.

cutoff frequency The signal frequency at which the input ac gate current is equal to the output ac drain current.

depletion mode MOSFET The type of MOSFET in which a gate voltage must be applied to turn the device off.

enhancement mode MOSFET The type of MOSFET in which a gate voltage must be applied to turn the device on.

equivalent fixed oxide charge The effective fixed charge in the oxide, Q'_{ss}, directly adjacent to the oxide–semiconductor interface.

field-effect The phenomenon by which an electric field perpendicular to the surface of a semiconductor can modulate the conductance.

flat-band voltage The gate voltage that must be applied to create the flat-band condition in which there is no space charge region in the semiconductor under the oxide.

interface states The allowed electronic energy states within the bandgap energy at the oxide–semiconductor interface.

inversion layer charge The induced charge directly under the oxide, which is the opposite type compared with the semiconductor doping.

inversion layer mobility The mobility of carriers in the inversion layer.

metal–semiconductor work function difference The parameter ϕ_{ms}, a function of the difference between the metal work function and semiconductor electron affinity.

oxide capacitance The ratio of oxide permittivity to oxide thickness, which is the capacitance per unit area, C_{ox}.

process conduction parameter The product of carrier mobility and oxide capacitance.

saturation The condition in which the inversion charge density is zero at the drain and the drain current is no longer a function of the drain-to-source voltage.

strong inversion The condition in which the inversion charge density is larger than the magnitude of the semiconductor doping concentration.

threshold inversion point The condition in which the inversion charge density is equal in magnitude to the semiconductor doping concentration.

threshold voltage The gate voltage that must be applied to achieve the threshold inversion point.

transconductance The ratio of an incremental change in drain current to the corresponding incremental change in gate voltage.

weak inversion The condition in which the inversion charge density is less than the magnitude of the semiconductor doping concentration.

CHECKPOINT

After studying this chapter, the reader should have the ability to:

- Sketch the energy-band diagrams in the semiconductor of the MOS capacitor under various bias conditions.
- Describe the process by which an inversion layer of charge is created in a MOS capacitor.
- Discuss the reason the space charge width reaches a maximum value once the inversion layer is formed.
- Discuss what is meant by the metal–semiconductor work function difference and why this value is different between aluminum, n^+ polysilicon, and p^+ polysilicon gates.
- Describe what is meant by flat-band voltage.
- Define threshold voltage.
- Sketch the *C–V* characteristics of a MOS capacitor with p-type and n-type semiconductor substrates under high-frequency and low-frequency conditions.
- Discuss the effects of fixed trapped oxide charge and interface states on the *C–V* characteristics.
- Sketch the cross sections of n-channel and p-channel MOSFET structures.
- Explain the basic operation of the MOSFET.
- Discuss the *I–V* characteristics of the MOSFET when biased in the nonsaturation and saturation regions.
- Describe the substrate bias effects on the threshold voltage.
- Sketch the small-signal equivalent circuit, including capacitances, of the MOSFET, and explain the physical origin of each capacitance.
- Discuss the condition that defines the cutoff frequency of a MOSFET.
- Sketch the cross section of a CMOS structure.
- Discuss what is meant by latch-up in a CMOS structure.

REVIEW QUESTIONS

1. Sketch the energy-band diagrams in a MOS capacitor with an n-type substrate in accumulation, depletion, and inversion modes.
2. Describe what is meant by an inversion layer of charge. Describe how an inversion layer of charge can be formed in a MOS capacitor with a p-type substrate.

3. Why does the space charge region in the semiconductor of a MOS capacitor reach a maximum width once the inversion layer is formed?
4. Define surface potential. Does the surface potential change significantly with gate voltage once threshold is reached?
5. Sketch the energy-band diagram through a MOS structure with a p-type substrate and an n^+ polysilicon gate under zero bias.
6. Define the flat-band voltage. Sketch the energy-band diagram in a MOS capacitor at flat band.
7. Define the threshold voltage. What is the surface potential at the threshold voltage?
8. Sketch the *C–V* characteristics of a MOS capacitor with an n-type substrate under the low-frequency condition. How do the characteristics change for the high-frequency condition?
9. Indicate the approximate capacitance at flat band on the *C–V* characteristic of a MOS capacitor with a p-type substrate under the high-frequency condition.
10. What is the effect on the *C–V* characteristics of a MOS capacitor with a p-type substrate if the amount of positive trapped oxide charge increases?
11. Qualitatively sketch the inversion charge density in the channel region when the transistor is biased in the nonsaturation region. Repeat for the case when the transistor is biased in the saturation region.
12. Define V_{DS}(sat).
13. Define enhancement mode and depletion mode for both n-channel and p-channel devices.
14. Sketch the charge distribution through a MOS capacitor with a p-type substrate when biased in the inversion mode. Write the charge neutrality equation.
15. Discuss why the threshold voltage changes when a reverse-biased source-to-substrate voltage is applied to a MOSFET.

PROBLEMS

(*Note:* In the following problems, assume the semiconductor and oxide in the MOS system are silicon and silicon dioxide, respectively, and assume the temperature is $T = 300$ K unless otherwise stated. Use Figure 10.16 to determine the metal–semiconductor work function difference.)

Section 10.1 The Two-Terminal MOS Structure

10.1 The dc charge distributions of four ideal MOS capacitors are shown in Figure P10.1. For each case: (*a*) Is the semiconductor n or p type? (*b*) Is the device biased in the accumulation, depletion, or inversion mode? (*c*) Draw the energy-band diagram in the semiconductor region.

10.2 (*a*) Calculate the maximum space charge width x_{dT} and the maximum space charge density $|Q'_{SD}(\max)|$ in a MOS capacitor with a p-type silicon substrate at $T = 300$ K for doping concentrations of (*i*) $N_a = 7 \times 10^{15}$ cm^{-3} and (*ii*) $N_a = 3 \times 10^{16}$ cm^{-3}. (*b*) Repeat part (*a*) for $T = 350$ K.

10.3 (*a*) Consider a MOS capacitor at $T = 300$ K with an n-type silicon substrate. Determine the silicon doping concentration such that $|Q'_{SD}(\max)| = 1.25 \times 10^{-8}$ C/cm^{-2}. (*b*) What is the surface potential that results in the maximum space charge width?

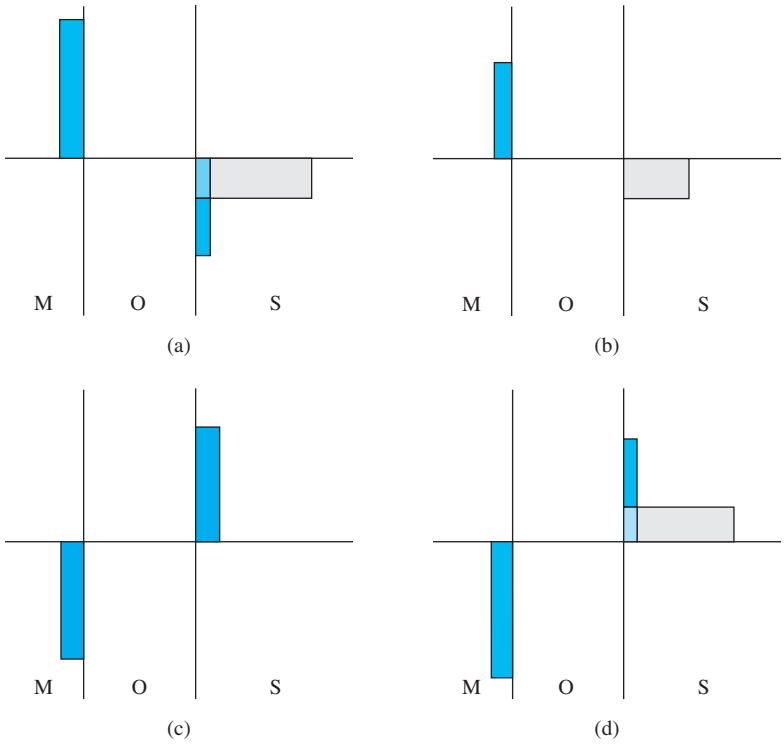


Figure P10.1 | Figure for Problem 10.1.

10.4 Determine the metal–semiconductor work function difference ϕ_{ms} in a MOS structure with p-type silicon for the case when the gate is (*a*) aluminum, (*b*) n^- polysilicon, and (*c*) p^+ polysilicon. Let $N_a = 6 \times 10^{15}$ cm^{-3}.

10.5 The silicon impurity doping concentration in an aluminum–silicon dioxide-silicon MOS device is $N_a = 4 \times 10^{16}$ cm^{-3}. Using the parameters in Example 10.2, determine the metal–semiconductor work function difference ϕ_{ms}.

10.6 Consider a MOS capacitor with an n-type silicon substrate. A metal–semiconductor work function difference of $\phi_{ms} = -0.30$ V is required. Determine the silicon doping concentration required to meet this specification when the gate is (*a*) n^+ polysilicon, (*b*) p^+ polysilicon, and (*c*) aluminum. If a particular gate cannot meet this specification, explain why.

10.7 (*a*) Consider the MOS capacitor described in Problem 10.5. For an oxide thickness of $t_{ox} = 20$ nm $= 200$ Å and an oxide charge of $Q'_{ss} = 5 \times 10^{10}$ cm^{-2}, calculate the flat-band voltage. (*b*) Repeat part (*a*) for an oxide thickness of $t_{ox} = 8$ nm $= 80$ Å.

10.8 (*a*) Consider an n^+ polysilicon–silicon dioxide–n-type silicon MOS structure. Let $N_d = 4 \times 10^{15}$ cm^{-3}. Calculate the ideal flat-band voltage for $t_{ox} = 20$ nm $= 200$ Å. (*b*) Considering the results of part (*a*), determine the shift in flat-band voltage for (*i*) $Q'_{ss} = 4 \times 10^{10}$ cm^{-2} and (*ii*) $Q'_{ss} = 10^{11}$ cm^{-2}. (*c*) Repeat parts (*a*) and (*b*) for an oxide thickness of $t_{ox} = 12$ nm $= 120$ Å.

10.9 Consider an aluminum gate–silicon dioxide–p-type silicon MOS structure with $t_{ox} = 450$ Å. The silicon doping is $N_a = 2 \times 10^{16}$ cm^{-3} and the flat-band voltage is $V_{FB} = -1.0$ V. Determine the fixed oxide charge Q'_{ss}.

10.10 Consider a MOS device with a p-type silicon substrate with $N_a = 2 \times 10^{16}$ cm^{-3}. The oxide thickness is $t_{ox} = 15$ nm $= 150$ Å and the equivalent oxide charge is $Q'_{ss} = 7 \times 10^{10}$ cm^{-2}. Calculate the threshold voltage for (*a*) an n$^+$ polysilicon gate, (*b*) a p$^+$ polysilicon gate, and (*c*) an aluminum gate.

10.11 Repeat Problem 10.10 for an n-type silicon substrate with a doping of $N_d = 3 \times 10^{15}$ cm^{-3}.

10.12 A 400-Å oxide is grown on p-type silicon with $N_a = 5 \times 10^{15}$ cm^{-3}. The flat-band voltage is -0.9 V. Calculate the surface potential at the threshold inversion point as well as the threshold voltage assuming negligible oxide charge. Also find the maximum space charge width for this device.

10.13 A MOS device with an aluminum gate is fabricated on a p-type silicon substrate. The oxide thickness is $t_{ox} = 22$ nm $= 220$ Å and the trapped oxide charge is $Q'_{ss} = 4 \times 10^{10}$ cm^{-2}. The measured threshold voltage is $V_T = +0.45$ V. Determine the p-type doping concentration.

10.14 Consider a MOS device with the following parameters: p$^+$ polysilicon gate, n-type silicon substrate, $t_{ox} = 18$ nm $= 180$ Å, and $Q'_{ss} = 4 \times 10^{10}$ cm^{-2}. Determine the silicon doping concentration such that the threshold voltage is in the range $-0.35 \leq V_{TP} \leq -0.25$ V.

10.15 Repeat Problem 10.13 for an n-type silicon substrate if the measured threshold voltage is $V_T = -0.975$ V. Determine the n-type doping concentration.

10.16 An n$^+$ polysilicon gate–silicon dioxide–silicon MOS capacitor has an oxide thickness of $t_{ox} = 18$ nm $= 180$ Å and a doping of $N_a = 10^{15}$ cm^{-3}. The oxide charge density is $Q'_{ss} = 6 \times 10^{10}$ cm^{-2}. Calculate the (*a*) flat-band voltage and (*b*) threshold voltage.

10.17 An n-channel depletion mode MOSFET with an n$^+$ polysilicon gate is shown in Figure 10.41. The n-channel doping is $N_d = 10^{15}$ cm^{-3} and the oxide thickness is $t_{ox} = 500$ Å. The equivalent fixed oxide charge is $Q'_{ss} = 10^{10}$ cm^{-2}. The n-channel thickness t_c is equal to the maximum induced space charge width. (Disregard the space charge region at the n-channel–p-substrate junction.) (*a*) Determine the channel thickness t_c and (*b*) calculate the threshold voltage.

10.18 Consider a MOS capacitor with an n$^+$ polysilicon gate and n-type silicon substrate. Assume $N_a = 10^{16}$ cm^{-3} and let $E_F - E_c = 0.2$ eV in the n$^+$ polysilicon. Assume the oxide has a thickness of $t_{ox} = 300$ Å. Also assume that χ' (polysilicon) = χ' (single-crystal silicon). (*a*) Sketch the energy-band diagrams (*i*) for $V_G = 0$ and (*ii*) at flat band. (*b*) Calculate the metal–semiconductor work function difference. (*c*) Calculate the threshold voltage for the ideal case of zero fixed oxide charge and zero interface states.

***10.19** The threshold voltage of an n-channel MOSFET is given by Equation (10.31a). Plot V_T versus temperature over the range $200 \leq T \leq 450$ K. Consider both an aluminum gate and an n$^+$ polysilicon gate. Assume the work functions are independent of temperature and use device parameters similar to those in Example 10.4.

***10.20** Plot the threshold voltage of an n-channel MOSFET versus p-type substrate doping concentration similar to Figure 10.21. Consider both n$^+$ and p$^+$ polysilicon gates. Use reasonable device parameters.

---

*Asterisks next to problems indicate problems that are more difficult.

***10.21** Plot the threshold voltage of a p-channel MOSFET versus n-type substrate doping concentration similar to Figure 10.22. Consider both n^+ and p^+ polysilicon gates. Use reasonable device parameters.

10.22 Consider an NMOS device with the parameters given in Problem 10.12. Plot V_T versus t_{ox} over the range $20 \leq t_{ox} \leq 500$ Å.

Section 10.2 Capacitance–Voltage Characteristics

10.23 An ideal MOS capacitor with an n^+ polysilicon gate has a silicon dioxide thickness of $t_{ox} = 12$ nm $= 120$ Å on a p-type silicon substrate doped at $N_a = 10^{16}$ cm^{-3}. Determine the capacitance C_{ox}, C'_{FB}, C'_{min}, and C'(inv) at (*a*) $f = 1$ Hz and (*b*) $f = 1$ MHz. (*c*) Determine V_{FB} and V_T. (*d*) Sketch C'/C_{ox} versus V_G for parts (*a*) and (*b*).

10.24 Repeat Problem 10.23 for an ideal MOS capacitor with a p^+ polysilicon gate and an n-type silicon substrate doped at $N_d = 5 \times 10^{14}$ cm^{-3}.

***10.25** Using superposition, show that the shift in the flat-band voltage due to a fixed charge distribution $\rho(x)$ in the oxide is given by

$$\Delta V_{FB} = -\frac{1}{C_{ox}} \int_0^{t_{ox}} \frac{x\rho(x)}{t_{ox}} dx$$

10.26 Using the results of Problem 10.25, calculate the shift in flat-band voltage for $t_{ox} = 20$ nm $= 200$ Å for the following oxide charge distributions:
(*a*) $Q'_{ss} = 8 \times 10^{10}$ cm^{-2} is entirely located at the oxide–semiconductor interface,
(*b*) $Q'_{ss} = 8 \times 10^{10}$ cm^{-2} is uniformly distributed throughout the oxide, and
(*c*) $Q'_{ss} = 8 \times 10^{10}$ cm^{-2} forms a triangular distribution with the peak at the oxide–semiconductor interface and is zero at the metal–oxide interface.

10.27 An ideal MOS capacitor is fabricated by using intrinsic silicon and an n^+ polysilicon gate. (*a*) Sketch the energy-band diagram through the MOS structure under flat-band conditions. (*b*) Sketch the low-frequency *C–V* characteristics from negative to positive gate voltage.

10.28 Consider a MOS capacitor with a p-type substrate. Assume that donor-type interface traps exist only at midgap (i.e., at E_{Fi}). Sketch the high-frequency *C–V* curve from accumulation to inversion. Compare this sketch to the ideal *C–V* plot.

10.29 Consider an SOS capacitor as shown in Figure P10.29. Assume the SiO_2 is ideal (no trapped charge) and has a thickness of $t_{ox} = 500$ Å. The doping concentrations are $N_d = 10^{16}$ cm^{-3} and $N_a = 10^{16}$ cm^{-3}. (*a*) Sketch the energy-band diagram through the device for (*i*) flat band, (*ii*) $V_G = +3$ V, and (*iii*) $V_G = -3$ V. (*b*) Calculate the flat-band voltage. (*c*) Estimate the voltage across the oxide for (*i*) $V_G = +3$ V and (*ii*) $V_G = -3$ V. (*d*) Sketch the high-frequency *C–V* characteristic curve.

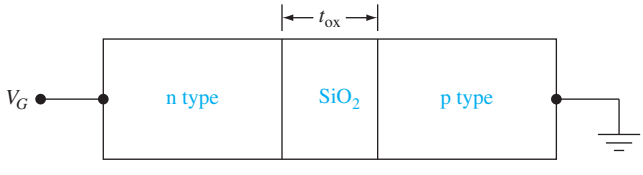


Figure P10.29 | Figure for Problem 10.29.

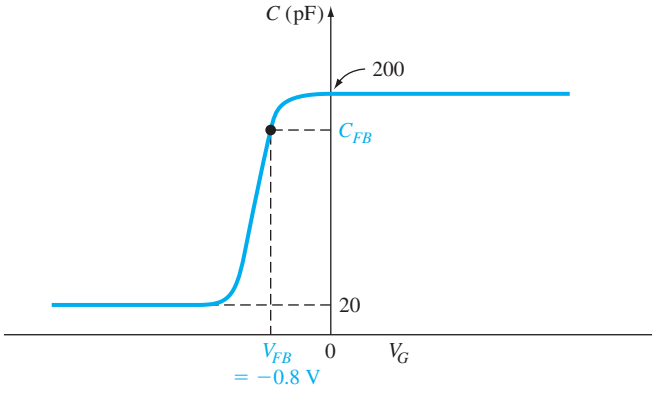


Figure P10.30 | Figure for Problem 10.30.

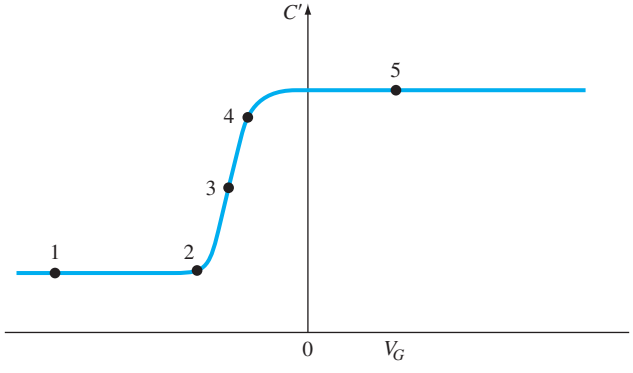


Figure P10.31 | Figure for Problem 10.31.

10.30 The high-frequency *C–V* characteristic curve of a MOS capacitor is shown in Figure P10.30. The area of the device is 2×10^{-3} cm^2. The metal–semiconductor work function difference is $\phi_{ms} = -0.50$ V, the oxide is SiO_2, the semiconductor is silicon, and the semiconductor doping concentration is 2×10^{16} cm^{-3}. (*a*) Is the semiconductor n or p type? (*b*) What is the oxide thickness? (*c*) What is the equivalent trapped oxide charge density? (*d*) Determine the flat-band capacitance.

10.31 Consider the high-frequency *C–V* plot shown in Figure P10.31. (*a*) Indicate which points correspond to flat-band, inversion, accumulation, threshold, and depletion modes. (*b*) Sketch the energy-band diagram in the semiconductor for each condition.

Section 10.3 The Basic MOSFET Operation

10.32 An expression that includes the inversion charge density is given by Equation (10.59). Consider the definition of threshold voltage and show that the inversion charge density goes to zero at the drain terminal at saturation. (Hint: Let $V_x = V_{DS} = V_{DS}(\text{sat})$.)

10.33 Consider an n-channel MOSFET with the following parameters: $k_n' = 0.18$ mA/V^2, $W/L = 8$, and $V_T = 0.4$ V. Determine the drain current I_D for (*a*) $V_{GS} = 0.8$ V, $V_{DS} = 0.2$ V; (*b*) $V_{GS} = 0.8$ V, $V_{DS} = 1.2$ V; (*c*) $V_{GS} = 0.8$ V, $V_{DS} = 2.5$ V; and (*d*) $V_{GS} = 1.2$ V, $V_{DS} = 2.5$ V.

10.34 A p-channel MOSFET has the following parameters: $k_p' = 0.10$ mA/V^2, $W/L = 15$, and $V_T = -0.4$ V. Calculate the drain current I_D for (*a*) $V_{SG} = 0.8$ V, $V_{SD} = 0.25$ V; (*b*) $V_{SG} = 0.8$ V, $V_{SD} = 1.0$ V; (*c*) $V_{SG} = 1.2$ V, $V_{SD} = 1.0$ V; and (*d*) $V_{SG} = 1.2$ V, $V_{SD} = 2.0$ V.

10.35 The parameters of an n-channel MOSFET are $k_n' = 0.6$ mA/V^2 and $V_T = 0.8$ V. The drain current is 1 mA with applied voltages of $V_{GS} = 1.4$ V, $V_{SB} = 0$, and $V_{DS} = 4$ V. (*a*) What is the W/L value? (*b*) What is the value of I_D for $V_{GS} = 1.85$ V, $V_{SB} = 0$, and $V_{DS} = 6$ V? (*c*) Determine the value of I_D for $V_{GS} = 1.2$ V, $V_{SB} = 0$, and $V_{DS} = 0.15$ V.

10.36 Consider a p-channel MOSFET with the following parameters: $k_p' = 0.12$ mA/V^2 and $W/L = 20$. The drain current is 100 μA with applied voltages of $V_{SG} = 0$, $V_{BS} = 0$, and $V_{SD} = 1.0$ V. (*a*) Determine the V_T value. (*b*) Determine the drain current I_D for $V_{SG} = 0.4$ V, $V_{SB} = 0$, and $V_{SD} = 1.5$ V. (*c*) What is the value of I_D for $V_{SG} = 0.6$ V, $V_{SB} = 0$, and $V_{SD} = 0.15$ V?

10.37 An ideal n-channel MOSFET has the following parameters: $V_T = 0.45$ V, $\mu_n = 425$ cm^2/V-s, $t_{ox} = 11$ nm $= 110$ Å, $W = 20$ μm, and $L = 1.2$ μm. (*a*) Plot I_D versus V_{DS} for $0 \le V_{DS} \le 3$ V and for $V_{GS} = 0$, 0.6, 1.2, 1.8, and 2.4 V. Indicate on each curve the $V_{DS}(sat)$ point. (*b*) Plot $\sqrt{I_D(sat)}$ versus V_{GS} for $0 \le V_{GS} \le 2.4$ V. (*c*) Plot I_D versus V_{GS} for $0 \le V_{GS} \le 2.4$ V and for $V_{DS} = 0.1$ V.

10.38 Consider an ideal p-channel MOSFET with the following parameters: $V_T = -0.35$ V, $\mu_p = 210$ cm^2/V-s, $t_{ox} = 11$ nm $= 110$ Å, $W = 35$ μm, and $L = 1.2$ μm. (*a*) Plot I_D versus V_{SD} for $0 \le V_{SD} \le 3$ V and for $V_{SG} = 0$, 0.6, 1.2, 1.8, and 2.4 V. Indicate on each curve the $V_{SD}(sat)$ point. (*b*) Plot $\sqrt{I_D(sat)}$ versus V_{SG} for $0 \le V_{SG} \le 2.4$ V. (*c*) Plot I_D versus V_{SG} for $0 \le V_{SG} \le 2.4$ V and for $V_{SD} = 0.1$ V.

10.39 Consider an n-channel MOSFET with the same parameters as described in Problem 10.37 except that $V_T = -0.8$ V. (*a*) Plot I_D versus V_{DS} for $0 \le V_{DS} \le 3$ V and for $V_{GS} = -0.8$, 0, +0.8, and +1.6 V. (*b*) Plot $\sqrt{I_D(sat)}$ versus V_{GS} for $-0.8 \le V_{GS} \le 1.6$ V.

10.40 Consider an n-channel enhancement mode MOSFET biased as shown in Figure P10.40. Sketch the current–voltage characteristics, I_D versus V_{DS}, for (*a*) $V_{GD} = 0$, (*b*) $V_{GD} = V_T/2$, and (*c*) $V_{GD} = 2V_T$.

10.41 Figure P10.41 shows the cross section of an NMOS device that includes source and drain resistances. These resistances take into account the bulk n^+ semiconductor resistance and the ohmic contact resistance. The current–voltage relations can be generated by replacing V_{GS} by $V_G - I_D R_S$ and V_{DS} by $V_D - I_D(R_S + R_D)$ in the ideal equations. Assume transistor parameters of $V_T = 1$ V and $K_n = 1$ mA/V^2. (*a*) Plot the following curves on the same graph: I_D versus V_D for $V_G = 2$ V and $V_G = 3$ V over the range $0 \le V_D \le 5$ V for (*i*) $R_S = R_D = 0$ and (*ii*) $R_S = R_D = 1$ kΩ. (*b*) Plot the following curves on the same graph: $\sqrt{I_D}$ versus V_G for $V_D = 0.1$ V and $V_D = 5$ V over the range $0 \le I_D \le 1$ mA for (*i*) $R_S = R_D = 0$ and (*ii*) $R_S = R_D = 1$ kΩ.

10.42 An n-channel MOSFET has the same parameters as given in Problem 10.37. The gate terminal is connected to the drain terminal. Plot I_D versus V_{DS} for $0 \le V_{DS} \le 5$ V. Determine the range of V_{DS} over which the transistor is biased in the nonsaturation and saturation regions.

10.43 The channel conductance for a p-channel MOSFET is defined as

$$g_d = \frac{\partial I_D}{\partial V_{SD}}\bigg|_{V_{SD} \to 0}$$

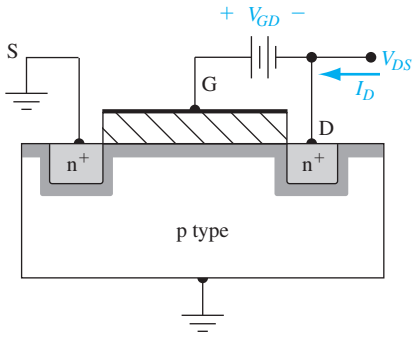


Figure P10.40 | Figure for Problem 10.40.

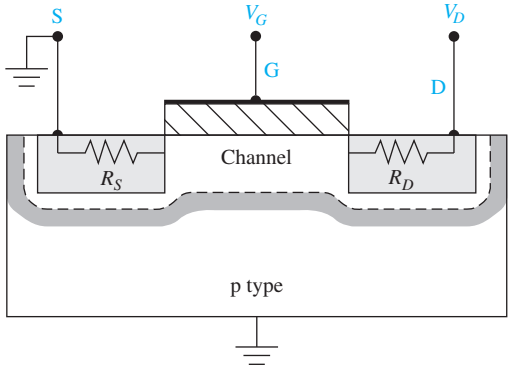


Figure P10.41 | Figure for Problem 10.41.

Plot the channel conductance for the p-channel MOSFET described in Problem 10.38 for $0 \leq V_{SG} \leq 2.4$.

10.44 The transconductance of an n-channel MOSFET is found to be $g_m = \partial I_D / \partial V_{GS} = 1.25$ mA/V when measured at $V_{DS} = 50$ mV. The threshold voltage is $V_T = 0.3$ V. (*a*) Determine the conductance parameter K_n. (*b*) What is the current at $V_{GS} = 0.8$ V and $V_{DS} = 50$ mV? (*c*) Determine the current at $V_{GS} = 0.8$ V and $V_{DS} = 1.5$ V.

10.45 The experimental characteristics of an ideal n-channel MOSFET biased in the saturation region are shown in Figure P10.45. If $W/L = 10$ and $t_{ox} = 425$ Å, determine V_T and μ_n.

10.46 One curve of an n-channel MOSFET is characterized by the following parameters: $I_D(\text{sat}) = 2 \times 10^{-4}$ A, $V_{DS}(\text{sat}) = 4$ V, and $V_T = 0.8$ V.

(*a*) What is the gate voltage?

(*b*) What is the value of the conduction parameter?

(*c*) If $V_G = 2$ V and $V_{DS} = 2$ V, determine I_D.

(*d*) If $V_G = 3$ V and $V_{DS} = 1$ V, determine I_D.

(*e*) For each of the conditions given in (*c*) and (*d*), sketch the inversion charge density and depletion region through the channel.

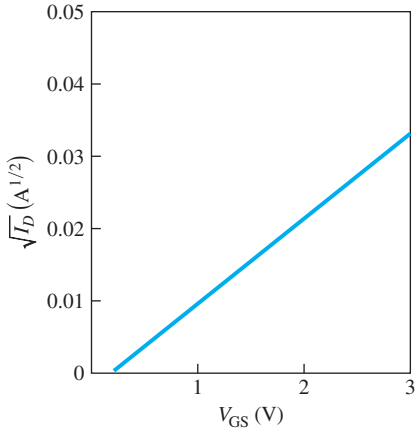


Figure P10.45 | Figure for Problem 10.45.

10.47 (*a*) An ideal n-channel MOSFET has parameters $t_{ox} = 18$ nm $= 180$ Å, $\mu_n = 450$ cm^2/V-s, and $V_T = 0.4$ V. The measured current in the saturation region is $I_D(sat) = 0.8$ mA when biased at $V_{GS} = 2.0$ V. Determine the (*i*) process conduction parameter and (*ii*) width-to-length ratio. (*b*) An ideal p-channel MOSFET has the same oxide thickness as given in part (*a*), a mobility of $\mu_p = 210$ cm^2/V-s, and a threshold voltage of $V_T = -0.4$ V. The measured current in the saturation region is also $I_D(sat) = 0.8$ mA when biased at $V_{SG} = 2.0$ V. Determine the (*i*) process conduction parameter and (*ii*) width-to-length ratio.

10.48 Consider the n-channel MOSFET described in Problem 10.37. (*a*) Calculate g_{mL} for $V_{DS} = 0.10$ V. (*b*) Find g_{ms} for $V_{GS} = 1.5$ V.

10.49 Consider the p-channel MOSFET described in Problem 10.38. (*a*) Calculate g_{mL} for $V_{SD} = 0.10$ V. (*b*) Find g_{ms} for $V_{SG} = 1.5$ V.

10.50 An n-channel MOSFET has the following parameters: $N_a = 5 \times 10^{16}$ cm^{-3}, $t_{ox} = 15$ nm $= 150$ Å, $\mu_n = 450$ cm^2/V-s, $V_{FB} = -0.5$ V, $L = 1.2$ μm, and $W = 8$ μm. (*a*) Determine the body-effect coefficient. (*b*) Plot $\sqrt{I_D(sat)}$ versus V_{GS} over the range $0 \le I_D \le 0.5$ mA for source-to-body voltages of (*i*) $V_{SB} = 0$, (*ii*) $V_{SB} = 1$ V, (*iii*) $V_{SB} = 2$ V, and (*iv*) $V_{SB} = 4$ V. (*c*) What are the threshold voltages for the conditions given in part (*b*)?

10.51 The substrate doping and body-effect coefficient of an n-channel MOSFET are $N_a = 10^{16}$ cm^{-3} and $\gamma = 0.12$ V$^{1/2}$, respectively. The threshold voltage is found to be $V_T = 0.5$ V when biased at $V_{SB} = 2.5$ V. What is the threshold voltage at $V_{SB} = 0$?

10.52 A p-channel MOSFET has an oxide thickness of $t_{ox} = 20$ nm $= 200$ Å and a substrate doping of $N_d = 5 \times 10^{15}$ cm^{-3}. (*a*) Find the body-effect coefficient. (*b*) Determine the body-to-source voltage, V_{BS}, such that the shift in threshold voltage, ΔV_T, from the $V_{BS} = 0$ curve is $\Delta V_T = -0.22$ V.

10.53 An NMOS device has the following parameters: n$^+$ poly gate, $t_{ox} = 400$ Å, $N_a = 10^{15}$ cm^{-3}, and $Q'_{ss} = 5 \times 10^{10}$ cm^{-2}. (*a*) Determine V_T. (*b*) Is it possible to apply a V_{SB} voltage such that $V_T = 0$? If so, what is the value of V_{SB}?

10.54 Investigate the threshold voltage shift due to substrate bias. The threshold shift is given by Equation (10.81). Plot ΔV_T versus V_{SB} over the range $0 \le V_{SB} \le 5$ V for several values of N_a and t_{ox}. Determine the conditions for which ΔV_T is limited to a maximum value of 0.7 V over the range of V_{SB}.

Section 10.4 Frequency Limitations

10.55 Consider an ideal n-channel MOSFET with a width-to-length ratio of $(W/L) = 10$, an electron mobility of $\mu_n = 400$ cm^2/V-s, an oxide thickness of $t_{ox} = 475$ Å, and a threshold voltage of $V_T = +0.65$ V. (*a*) Determine the maximum value of source resistance so that the saturation transconductance g_{ms} is reduced by no more than 20 percent from its ideal value when $V_{GS} = 5$ V. (*b*) Using the value of r_s calculated in part (*a*), how much is g_{ms} reduced from its ideal value when $V_{GS} = 3$ V?

10.56 An n-channel MOSFET has the following parameters:

$$\mu_n = 400 \text{ cm}^2\text{/V-s} \qquad t_{ox} = 500 \text{ Å}$$
$$L = 2 \ \mu\text{m} \qquad W = 20 \ \mu\text{m}$$
$$V_T = +0.75 \text{ V}$$

Assume the transistor is biased in the saturation region at $V_{GS} = 4$ V. (*a*) Calculate the ideal cutoff frequency. (*b*) Assume that the gate oxide overlaps both the source and drain contacts by 0.75 μm. If a load resistance of $R_L = 10$ kΩ is connected to the output, calculate the cutoff frequency.

10.57 Repeat Problem 10.56 for the case when the electrons are traveling at a saturation velocity of $v_{sat} = 4 \times 10^6$ cm/s.

Summary and Review

***10.58** Design an ideal silicon n-channel MOSFET with a polysilicon gate to have a threshold voltage of $V_T = 0.65$ V. Assume an oxide thickness of $t_{ox} = 300$ Å, a channel length of $L = 1.25$ μm, and a nominal value of $Q'_{ss} = 1.5 \times 10^{11}$ cm^{-2}. It is desired to have a drain current of $I_D = 50$ μA at $V_{GS} = 2.5$ V and $V_{DS} = 0.1$ V. Determine the substrate doping concentration, channel width, and type of gate required.

***10.59** Design an ideal silicon n-channel depletion mode MOSFET with a polysilicon gate to have a threshold voltage of $V_T = -0.65$ V. Assume an oxide thickness of $t_{ox} = 300$ Å, a channel length of $L = 1.25$ μm, and a nominal value of $Q'_{ss} = 1.5 \times 10^{11}$ cm^{-2}. It is desired to have a drain current of $I_D(\text{sat}) = 50$ μA at $V_{GS} = 0$. Determine the type of gate, substrate doping concentration, and channel width required.

***10.60** Consider the CMOS inverter circuit shown in Figure 10.59a. Ideal n- and p-channel devices are to be designed with channel lengths of $L = 2.5$ μm and oxide thicknesses of $t_{ox} = 450$ Å. Assume the inversion channel mobilities are one-half the bulk values. The threshold voltages of the n- and p-channel transistors are to be $+0.5$ V and -0.5 V, respectively. The drain current is to be $I_D = 0.256$ mA when the input voltage to the inverter is 1.5 V and 3.5 V with $V_{DD} = 5$ V. The gate material is to be the same in each device. Determine the type of gate, substrate doping concentrations, and channel widths.

***10.61** A complementary pair of ideal n-channel and p-channel MOSFETs is to be designed to produce the same *I–V* characteristics when they are equivalently biased. The devices are to have the same oxide thickness of 250 Å and the same channel length of $L = 2$ μm. Assume the SiO_2 layer is ideal. The n-channel device is to have a channel width of $W = 20$ μm. Assume constant inversion layer mobilities of $\mu_n = 600$ cm^2/V-s and $\mu_p = 220$ cm^2/V-s. (*a*) Determine p-type and n-type substrate doping concentrations. (*b*) What are the threshold voltages? (*c*) What is the width of the p-channel device?

READING LIST

1. Dimitrijev, S. *Principles of Semiconductor Devices*. New York: Oxford University, 2006.
2. Hu, C. C. *Modern Semiconductor Devices for Integrated Cicuits*. Upper Saddle River, NJ: Pearson Prentice Hall, 2010.
3. Kano, K. *Semiconductor Devices*. Upper Saddle River, NJ: Prentice Hall, 1998.
4. Muller, R. S., and T. I. Kamins. *Device Electronics for Integrated Circuits.* 2nd ed. New York: Wiley, 1986.

5. Ng, K. K. *Complete Guide to Semiconductor Devices*. New York: McGraw-Hill, 1995.

6. Nicollian, E. H., and J. R. Brews. *MOS Physics and Technology*. New York: Wiley, 1982.

7. Ong, D. G. *Modern MOS Technology: Processes, Devices, and Design*. New York: McGraw-Hill, 1984.

8. Pierret, R. F. *Semiconductor Device Fundamentals*. Reading, MA: Addison-Wesley, 1996.

9. Roulston, D. J. *An Introduction to the Physics of Semiconductor Devices.* New York: Oxford University Press, 1999.

10. Schroder, D. K. *Advanced MOS Devices, Modular Series on Solid State Devices*. Reading, MA: Addison-Wesley, 1987.

11. Shur, M. *Introduction to Electronic Devices*. New York: John Wiley & Sons, Inc., 1996.

***12.** _____ *Physics of Semiconductor Devices*. Englewood Cliffs, NJ: Prentice Hall, 1990.

13. Singh, J. *Semiconductor Devices: An Introduction*. New York: McGraw-Hill, 1994.

14. _____ *Semiconductor Devices: Basic Principles*. New York: Wiley, 2001.

15. Streetman, B. G., and S. K. Banerjee. *Solid State Electronic Devices.* 6th ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2006.

16. Sze, S. M. *High-Speed Semiconductor Devices*. New York: Wiley, 1990.

17. Sze, S. M. and K. K. Ng. Physics of Semiconductor Devices, 3rd ed. Hoboken, NJ: John Wiley & Sons, Inc., 2007.

18. Taur, Y. and T. H. Ning. *Fundamentals of Modern VLSI Devices,* 2nd ed. Cambridge University Press, 2009.

***19.** Tsividis, Y. *Operation and Modeling of the MOS Transistor.* 2nd ed. Burr Ridge, IL: McGraw-Hill, 1999.

20. Werner, W. M. "The Work Function Difference of the MOS System with Aluminum Field Plates and Polycrystalline Silicon Field Plates." *Solid State Electronics* 17, (1974), pp. 769–775.

21. Yamaguchi, T., S. Morimoto, G. H. Kawamoto, and J. C. DeLacy. "Process and Device Performance of 1 μm-Channel n-Well CMOS Technology." *IEEE Transactions on Electron Devices* ED-31 (February 1984), pp. 205–214.

22. Yang, E. S. *Microelectronic Devices*. New York: McGraw-Hill, 1988.

*Indicates references that are at an advanced level compared to this text.

CHAPTER 11

Metal–Oxide–Semiconductor Field-Effect Transistor: Additional Concepts

In this chapter we present additional concepts that are commonly encountered in metal–oxide–semiconductor field-effect transistors (MOSFETs). These concepts include nonideal effects, small device geometry, breakdown, threshold voltage adjustment by ion implantation, and radiation effects. Although there are a multitude of details that become important when fabricating MOSFETs in ICs, we are able to consider only a few here. Many additional details can be found in more advanced texts. ■

11.0 | PREVIEW

In this chapter, we will:

- Describe and analyze subthreshold conduction, which is the phenomenon whereby current is induced in the channel before the defined threshold voltage is reached.
- Analyze channel length modulation, which is a characteristic of short-channel lengths and leads to a finite output resistance.
- Consider the effects of a decrease in carrier mobility due to increasing gate voltage.
- Analyze the effects of carrier saturation velocity. Carriers can easily reach their saturation velocity in short-channel devices.
- Discuss MOSFET scaling, which describes how various parameters must be changed as device size is decreased.
- Consider the deviations in threshold voltage due to small geometry devices, including short channel length devices and small channel width devices.

- Describe and analyze various voltage breakdown mechanisms in MOSFETs.
- Describe and analyze the technique of threshold voltage adjustment by ion implantation.
- Consider the introduction of trapped oxide charges by ionizing radiation and hot electron effects.

11.1 | NONIDEAL EFFECTS

As with any semiconductor device, the experimental characteristics of MOSFETs deviate to some degree from the ideal relations that have been theoretically derived using the various assumptions and approximations. In this section, we consider five effects that cause deviations from the assumptions used in the ideal derivations. These effects are subthreshold conduction, channel length modulation, mobility variations, velocity saturation, and ballistic transport.

11.1.1 Subthreshold Conduction

The ideal current–voltage relationship predicts zero drain current when the gate-to-source voltage is less than or equal to the threshold voltage. Experimentally, I_D is not zero when $V_{GS} \leq V_T$. Figure 11.1 shows a comparison between the ideal characteristic that was derived, and the experimental results. The drain current, which exists for $V_{GS} \leq V_T$, is known as the *subthreshold current*.

Figure 11.2 shows the energy-band diagram of an MOS structure with a p-type substrate biased so that $\phi_s < 2\phi_{fp}$. At the same time, the Fermi level is closer to the conduction band than the valence band, so the semiconductor surface develops the

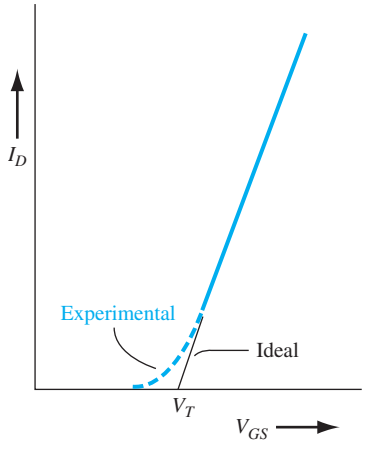


Figure 11.1 | Comparison of ideal and experimental plots of $\sqrt{I_D}$ versus V_{GS}.

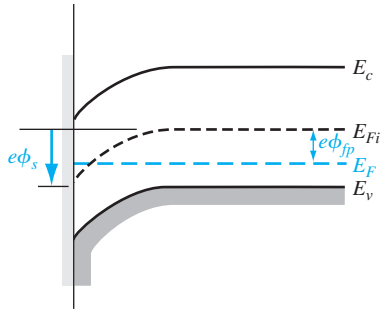


Figure 11.2 | Energy-band diagram when $\phi_{fp} < \phi_s < 2\phi_{fp}$.

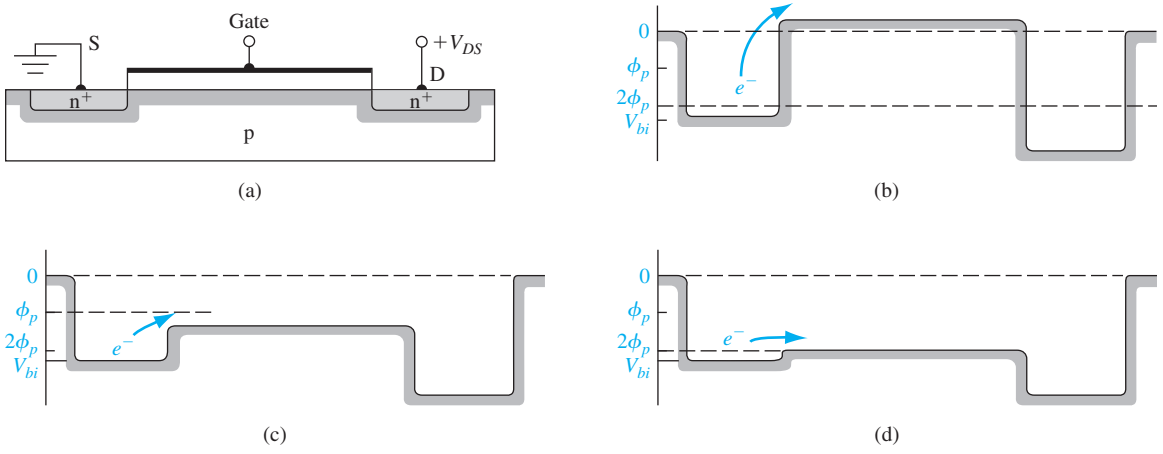


Figure 11.3 | (a) Cross section along channel length of n-channel MOSFET. Energy-band diagrams along channel length at (b) accumulation, (c) weak inversion, and (d) inversion.

characteristics of a lightly doped n-type material. We would expect, then, to observe some conduction between the n^+ source and drain contacts through this weakly inverted channel. The condition for $\phi_{fp} < \phi_s < 2\phi_{fp}$ is known as *weak inversion*.

Figure 11.3 shows the surface potential along the length of the channel at accumulation, weak inversion, and threshold for the case when a small drain voltage is applied. The bulk p-substrate is assumed to be at zero potential. Figure 11.3b, c shows the accumulation and weak inversion cases. There is a potential barrier between the n^+ source and channel region which the electrons must overcome in order to generate a channel current. A comparison of these barriers with those in pn junctions would suggest that the channel current is an exponential function of V_{GS}. In the inversion mode, shown in Figure 11.3d, the barrier is so small that we lose the exponential dependence, since the junction is more like an ohmic contact.

The actual derivation of the subthreshold current is beyond the scope of this chapter. We can write that

$$I_D(\text{sub}) \propto \left[\exp\left(\frac{eV_{GS}}{kT}\right)\right] \cdot \left[1 - \exp\left(\frac{-eV_{DS}}{kT}\right)\right] \tag{11.1}$$

If V_{DS} is larger than a few (kT/e) volts, then the subthreshold current is independent of V_{DS}.

Figure 11.4 shows the exponential behavior of the subthreshold current for several body-to-source voltages. Also shown on the curves are the threshold voltage values. Ideally, a change in gate voltage on approximately 60 mV produces an order of magnitude change in the subthreshold current. A detailed analysis of the subthreshold condition shows that the slope of the I_D versus V_{DS} curve is a function of the semiconductor doping and is also a function of the interface state density. The

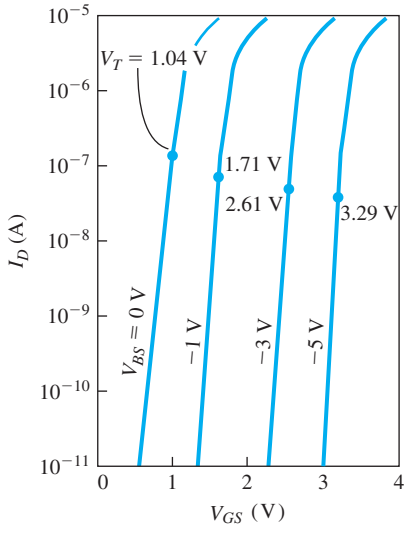


Figure 11.4 | Subthreshold current–voltage characteristics for several values of substrate voltage (the threshold voltage is indicated on each curve). *(From Schroder [17].)*

measurement of the slope of these curves has been used to experimentally determine the oxide–semiconductor interface state density.

If a MOSFET is biased at or even slightly below the threshold voltage, the drain current is not zero. The subthreshold current may add significantly to power dissipation in a large-scale integrated circuit in which hundreds or thousands of MOSFETs are used. The circuit design must include the subthreshold current or ensure that the MOSFET is biased sufficiently below the threshold voltage in the "off" state.

11.1.2 Channel Length Modulation

We assumed in the derivation of the ideal current–voltage relationship that the channel length L was a constant. However, when the MOSFET is biased in the saturation region, the depletion region at the drain terminal extends laterally into the channel, reducing the effective channel length. Since the depletion region width is bias dependent, the effective channel length is also bias dependent and is modulated by the drain-to-source voltage. This channel length modulation effect is shown in Figure 11.5 for an n-channel MOSFET.

The depletion width extending into the p-region of a pn junction under zero bias can be written as

$$x_p = \sqrt{\frac{2\epsilon_s \phi_{fp}}{eN_a}} \tag{11.2}$$

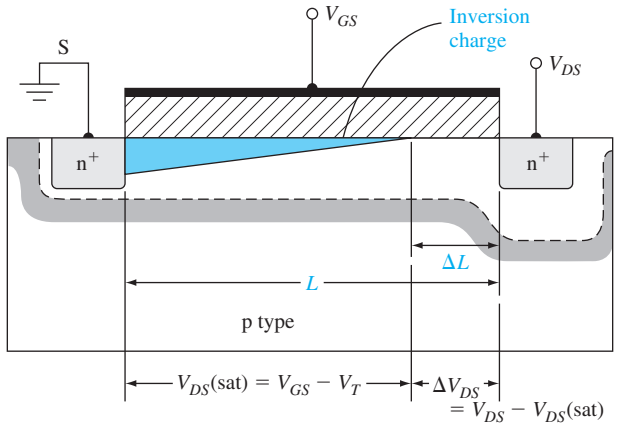


Figure 11.5 | Cross-section of an n-channel MOSFET showing the channel length modulation effect.

For a one-sided n^+p junction, essentially all of the applied reverse-biased voltage is across the low-doped p region. The space charge width of the drain–substrate junction is approximately

$$x_p = \sqrt{\frac{2\epsilon_s}{eN_a}(\phi_{fp} + V_{DS})} \tag{11.3}$$

However, the space charge region defined by ΔL, as shown in Figure 11.5, does not begin to form until $V_{DS} > V_{DS}(\text{sat})$. As a first approximation, we can write that ΔL is the total space charge width minus the space charge width that exists when $V_{DS} = V_{DS}(\text{sat})$, or

$$\Delta L = \sqrt{\frac{2\epsilon_s}{eN_a}}\left[\sqrt{\phi_{fp} + V_{DS}(\text{sat}) + \Delta V_{DS}} - \sqrt{\phi_{fp} + V_{DS}(\text{sat})}\right] \tag{11.4}$$

where

$$\Delta V_{DS} = V_{DS} - V_{DS}(\text{sat}) \tag{11.5}$$

The applied drain-to-source voltage is V_{DS} and we are assuming that $V_{DS} > V_{DS}(\text{sat})$.

As a second approximation at determining ΔL, we can consider Figure 11.6 and revisit the one-dimensional Poisson's equation. The electric field E_{sat} is the lateral electric field at the point where the inversion layer charge is pinched off. Neglecting any charges that exist due to current, we can write

$$\frac{d\mathrm{E}}{dx} = \frac{\rho(x)}{\epsilon_s} \tag{11.6}$$

where $\rho(x) = -eN_a$ and is a constant for a uniformly doped substrate. Integrating Equation (11.6) and applying the boundary conditions give the electric field in the space charge region defined by ΔL:

$$\mathrm{E} = -\frac{eN_a x}{\epsilon_s} - \mathrm{E}_{\text{sat}} \tag{11.7}$$

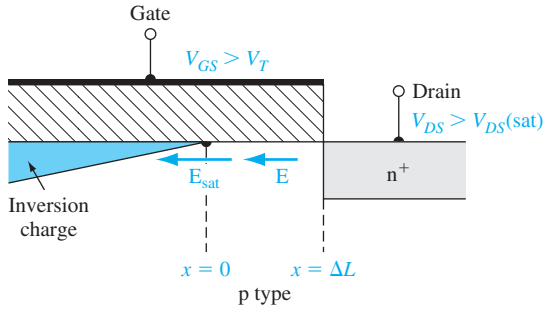


Figure 11.6 | Expanded view of cross section near the drain terminal of an n-channel MOSFET showing the channel length modulation effect.

The potential in this region is

$$\phi(x) = -\int \mathrm{E}dx = \frac{eN_a x^2}{2\epsilon_s} + \mathrm{E}_{\text{sat}} x + C_1 \tag{11.8}$$

where C_1 is a constant of integration. The boundary conditions are $\phi(x = 0) = V_{DS}(\text{sat})$ and $\phi(x = \Delta L) = V_{DS}$. Substituting these boundary conditions into Equation (11.8), we obtain

$$V_{DS} = \frac{eN_a(\Delta L)^2}{2\epsilon_s} + \mathrm{E}_{\text{sat}}(\Delta L) + V_{DS}(\text{sat}) \tag{11.9}$$

Solving for ΔL, we can write

$$\Delta L = \sqrt{\frac{2\epsilon_s}{eN_a}}\left[\sqrt{\phi_{\text{sat}} + [V_{DS} - V_{DS}(\text{sat})]} - \sqrt{\phi_{\text{sat}}}\right] \tag{11.10}$$

where

$$\phi_{\text{sat}} = \frac{2\epsilon_s}{eN_a} \cdot \left(\frac{\mathrm{E}_{\text{sat}}}{2}\right)^2$$

In general, the value of E_{sat} is in the range $10^4 < \mathrm{E}_{\text{sat}} < 2 \times 10^5$ V/cm.

Other models used to determine ΔL include the negative charges due to the drain current and also include two-dimensional effects. These models are not considered here.

Since the drain current is inversely proportional to the channel length, we may write

$$\boxed{I_D' = \left(\frac{L}{L - \Delta L}\right) I_D} \tag{11.11}$$

where I_D' is the actual drain current and I_D is the ideal drain current. Since ΔL is a function of V_{DS}, I_D' is now also a function of V_{DS} even though the transistor is biased in the saturation region.

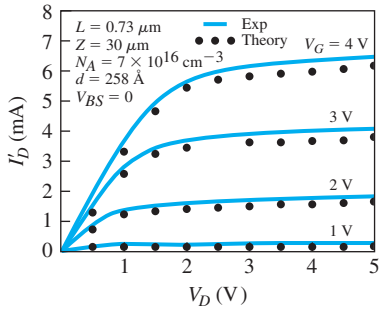


Figure 11.7 | Current–voltage characteristics of a MOSFET showing short-channel effects.
(From Sze [22].)

Since I'_D is now a function of V_{DS}, the output resistance is no longer infinite. The drain current in the saturation region can be written as

$$I'_D = \frac{k'_n}{2} \cdot \frac{W}{L} \cdot [(V_{GS} - V_T)^2 (1 + \lambda V_{DS})] \tag{11.12}$$

where λ is the *channel length modulation parameter.*

The output resistance is given by

$$r_o = \left(\frac{\partial I'_D}{\partial V_{DS}}\right)^{-1} = \left\{\frac{k'_n}{2} \cdot \frac{W}{L} \cdot (V_{GS} - V_T)^2 \cdot \lambda\right\}^{-1} \tag{11.13a}$$

Since λ is normally small, Equation (11.13a) can be written as

$$r_o \cong \frac{1}{\lambda I_D} \tag{11.13b}$$

Figure 11.7 shows some typical I'_D versus V_{DS} curves with positive slopes in the saturation region due to channel length modulation. As the MOSFET dimensions become smaller, the change in the channel length ΔL becomes a larger fraction of the original length L, and the channel length modulation becomes more severe.

EXAMPLE 11.1

Objective: Determine the increase in drain current due to short channel modulation.

Consider an n-channel MOSFET with a substrate doping concentration of $N_a = 2 \times 10^{16}$ cm^{-3}, a threshold voltage of $V_T = 0.4$ V, and a channel length of $L = 1$ μm. The device is biased at $V_{GS} = 1$ V and $V_{DS} = 2.5$ V. Determine the ratio of actual drain current compared to the ideal value.

■ Solution

We find

$$\phi_{fp} = V_t \ln\left(\frac{N_a}{n_i}\right) = (0.0259)\ln\left(\frac{2 \times 10^{16}}{1.5 \times 10^{10}}\right) = 0.3653 \text{ V}$$

$$V_{DS}(\text{sat}) = V_{GS} - V_T = 1.0 - 0.4 = 0.6 \text{ V}$$

and

$$\Delta V_{DS} = V_{DS} - V_{DS}(\text{sat}) = 2.5 - 0.6 = 1.9 \text{ V}$$

Using Equation (11.4), we determine

$$\Delta L = \sqrt{\frac{2\epsilon_s}{eN_a}}\left[\sqrt{\phi_{fp} + V_{DS}(\text{sat}) + \Delta V_{DS}} - \sqrt{\phi_{fp} + V_{DS}(\text{sat})}\right]$$

$$= \sqrt{\frac{2(11.7)(8.85 \times 10^{-14})}{(1.6 \times 10^{-19})(2 \times 10^{16})}}\left[\sqrt{0.3653 + 0.6 + 1.9} - \sqrt{0.3653 + 0.6}\right]$$

$$= 1.807 \times 10^{-5} \text{ cm}$$

or

$$\Delta L = 0.1807 \ \mu\text{m}$$

Then

$$\frac{I'_D}{I_D} = \frac{L}{L - \Delta L} = \frac{1}{1 - 0.1807} = 1.22$$

Comment

The actual drain current increases as the effective channel length decreases when the transistor is biased in the saturation region.

EXERCISE PROBLEM

Ex 11.1 An n-channel MOSFET has the same properties as described in Example 11.1 except for the channel length. The transistor is biased at $V_{GS} = 0.8$ V and $V_{DS} = 2.5$ V. Find the minimum channel length such that the ratio of actual drain current to the ideal drain current due to channel length modulation is no larger than 1.35.

(Ans. $L = 0.698 \ \mu\text{m}$)

11.1.3 Mobility Variation

In the derivation of the ideal *I–V* relationship, we explicitly assumed that the mobility was a constant. However, this assumption must be modified for two reasons. The first effect to be considered is the variation in mobility with gate voltage. The second reason for a mobility variation is that the effective carrier mobility decreases as the carrier approaches the velocity saturation limit. This effect is discussed in the next section.

The inversion layer charge is induced by a vertical electric field, which is shown in Figure 11.8 for an n-channel device. A positive gate voltage produces a force on the electrons in the inversion layer toward the surface. As the electrons travel through the channel toward the drain, they are attracted to the surface, but then are repelled by localized coulombic forces. This effect, schematically shown in Figure 11.9, is called *surface scattering*. The surface scattering effect reduces mobility. If there is a positive fixed oxide charge near the oxide-semiconductor interface, the mobility will be further reduced due to the additional coulomb interaction.

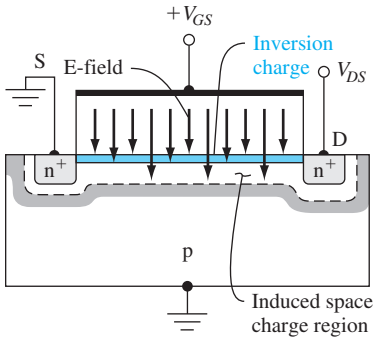


Figure 11.8 | Vertical electric field in an n-channel MOSFET.

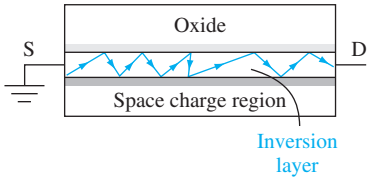


Figure 11.9 | Schematic of carrier surface scattering effects.

The relationship between the inversion charge mobility and the transverse electric field is usually measured experimentally. An effective transverse electric field can be defined as

$$\mathrm{E_{eff}} = \frac{1}{\epsilon_s}\left(|Q'_{SD}\,(\max)| + \frac{1}{2}\,Q'_n\right) \tag{11.14}$$

The effective inversion charge mobility can be determined from the channel conductance as a function of gate voltage. Figure 11.10 shows the effective electron mobility at $T = 300$ K for different doping levels and different oxide thicknesses. The effective mobility is only a function of the electric field at the inversion layer and is independent of oxide thickness. The effective mobility may be represented by

$$\mu_{\mathrm{eff}} = \mu_0\left(\frac{\mathrm{E_{eff}}}{\mathrm{E_0}}\right)^{-1/3} \tag{11.15}$$

where μ_0 and $\mathrm{E_0}$ are constants determined from experimental results.

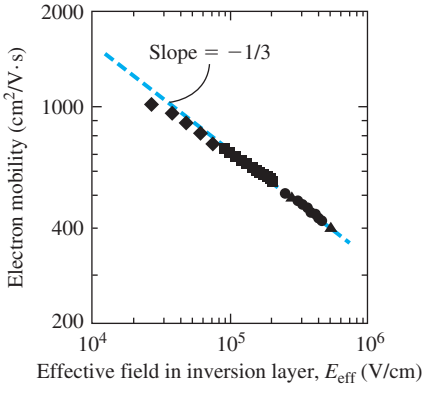


Figure 11.10 | Measured inversion layer electron mobility versus electric field at the inversion layer.
(From Yang [25].)

The effective inversion charge mobility is a strong function of temperature because of lattice scattering. As the temperature is reduced, the mobility increases.

EXAMPLE 11.2

Objective: Calculate the effective electric field at threshold for a given semiconductor doping concentration.

Consider a p-type silicon substrate at $T = 300$ K doped to $N_a = 3 \times 10^{16}$ cm^{-3}.

■ Solution

From the results of Chapter 10, we can calculate

$$\phi_{fp} = V_t \ln\left(\frac{N_a}{n_i}\right) = (0.0259) \ln\left(\frac{3 \times 10^{16}}{1.5 \times 10^{10}}\right) = 0.376 \text{ V}$$

and

$$x_{dT} = \left\{\frac{4\epsilon_s \phi_{fp}}{eN_a}\right\}^{1/2} = \left\{\frac{4(11.7)(8.85 \times 10^{-14})(0.376)}{(1.6 \times 10^{-19})(3 \times 10^{16})}\right\}^{1/2}$$

which is $x_{dT} = 0.18\ \mu$m. Then

$$|Q'_{SD}(\max)| = eN_a x_{dT} = 8.64 \times 10^{-8} \text{ C/cm}^2$$

At the threshold inversion point, we may assume that $Q'_n = 0$, so the effective electric field from Equation (11.14) is found as

$$\text{E}_{\text{eff}} = \frac{1}{\epsilon_s}|Q'_{SD}(\max)| = \frac{8.64 \times 10^{-8}}{(11.7)(8.85 \times 10^{-14})} = 8.34 \times 10^4 \text{ V/cm}$$

■ Comment

We can see, from Figure 11.10, that this value of effective transverse electric field at the surface is sufficient for the effective inversion charge mobility to be significantly less than the bulk semiconductor value.

■ EXERCISE PROBLEM

Ex 11.2 Determine (using Figure 11.10) the effective inversion layer electron mobility for a surface electric field of $\text{E}_{eff} = 2 \times 10^5$ V/cm.

(Ans. $\mu_n \cong 550$ cm^2/V-s)

The effective mobility is a function of gate voltage through the inversion charge density in Equation (11.14). As the gate voltage increases, the carrier mobility decreases even further.

11.1.4 Velocity Saturation

In the analysis of the long-channel MOSFET, we assume the mobility to be constant, which means that the drift velocity increases without limit as the electric field increases. In this ideal case, the carrier velocity increases until the ideal current is attained. However, we have seen that the carrier velocity saturates with increasing electric field. Velocity saturation will become more prominent in shorter-channel devices since the corresponding horizontal electric field is generally larger.

In the ideal *I–V* relationship, current saturation occurs when the inversion charge density becomes zero at the drain terminal, or when

$$V_{DS} = V_{DS}(\text{sat}) = V_{GS} - V_T \tag{11.16}$$

for the n-channel MOSFET. However, velocity saturation can change this saturation condition. Velocity saturation will occur when the horizontal electric field is approximately 10^4 V/cm. If $V_{DS} = 5$ V in a device with a channel length of $L = 1$ μm, the average electric field is 5×10^4 V/cm. Velocity saturation, then, is very likely to occur in short-channel devices.

The modified $I_D(\text{sat})$ characteristics are described approximately by

$$I_D(\text{sat}) = WC_{\text{ox}}(V_{GS} - V_T)v_{\text{sat}} \tag{11.17}$$

where v_{sat} is the saturation velocity (approximately 10^7 cm/s for electrons in bulk silicon) and C_{ox} is the gate oxide capacitance per cm^2. The saturation velocity will decrease somewhat with applied gate voltage because of the vertical electric field and surface scattering. Velocity saturation will yield an $I_D(\text{sat})$ value smaller than that predicted by the ideal relation, and it will yield a smaller $V_{DS}(\text{sat})$ value than predicted. The $I_D(\text{sat})$ current is also approximately linear with V_{GS}, instead of having the ideal square law dependence predicted previously.

There are several models of mobility versus electric field. One particular relation that is commonly used is

$$\mu = \frac{\mu_{\text{eff}}}{\left[1 + \left(\dfrac{\mu_{\text{eff}}\text{E}}{v_{\text{sat}}}\right)^2\right]^{1/2}} \tag{11.18}$$

Figure 11.11 shows a comparison of drain current versus drain-to-source voltage characteristics for constant mobility and for field-dependent mobility. The smaller values of $I_D(\text{sat})$ and the approximate linear dependence on V_{GS} may be noted for the field-dependent mobility curves.

The transconductance is found from

$$g_{ms} = \frac{\partial I_D(\text{sat})}{\partial V_{GS}} = WC_{\text{ox}}\, v_{\text{sat}} \tag{11.19}$$

which is now independent of V_{GS} and V_{DS} when velocity saturation occurs. The drain current is saturated by the velocity saturation effect, which leads to a constant transconductance.

When velocity saturation occurs, the cutoff frequency is given by

$$f_T = \frac{g_m}{2\pi C_G} = \frac{WC_{\text{ox}}v_{\text{sat}}}{2\pi(C_{\text{ox}}WL)} = \frac{v_{\text{sat}}}{2\pi L} \tag{11.20}$$

where the parasitic capacitances are assumed to be negligible.

11.1.5 Ballistic Transport

Scattering events in a semiconductor limit the velocity of carriers to an average drift velocity as discussed in Chapter 5. The average drift velocity is a function of the

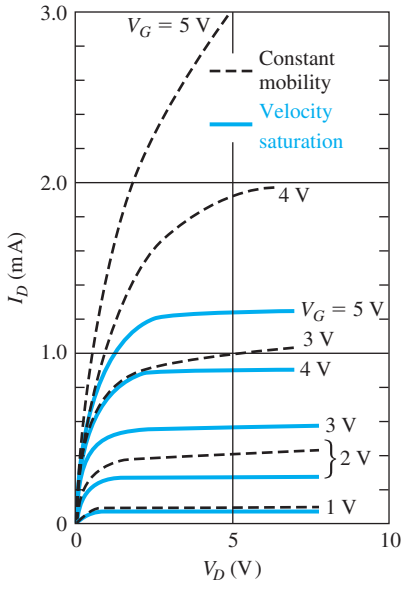


Figure 11.11 | Comparison of I_D versus V_D characteristics for constant mobility (dashed curves) and for field-dependent mobility and velocity saturation effects (solid curves).
(From Sze and Ng [22].)

mean time between collisions or the mean distance between scattering events. In the long-channel device, the channel length L is much longer than the mean distance between collisions l, so that an average carrier drift velocity exists. As the MOSFET channel length is reduced, the mean distance between collisions l may become comparable to L so that the previous analysis may not be valid. If the channel length is further reduced so that $L < l$, then a large fraction of carriers could travel from the source to the drain without experiencing a scattering event. This motion of carriers is called *ballistic transport*.

Ballistic transport means that carriers travel faster than the average drift velocity or the saturation velocity, and this effect can lead to very fast devices. Ballistic transport occurs in submicron ($L < 1$ μm) devices. As the MOSFET technology continues to shrink the channel length toward the 0.1 μm value, the ballistic transport phenomenon will become more important.

TEST YOUR UNDERSTANDING

TYU 11.1 Consider a MOSFET biased in the subthreshold region with $V_D \gg kT/e$. For the ideal relationship given, what change in gate-to-source voltage produces a factor of 10 change in drain current?

(Ans. $\Delta V_{GS} = 59.64$ mV)

TYU 11.2 Consider an NMOS transistor with the following parameters: $\mu_n = 1000$ cm^2/V-s, $C_{ox} = 10^{-8}$ F/cm^2, $W = 10$ μm, $L = 1$ μm, $V_T = 0.4$ V, and $v_{sat} = 5 \times 10^6$ cm/s. Plot on the same graph I_D(sat) versus V_{GS} over the range $0 \leq V_{GS} \leq 4$ V for the case (*a*) of an ideal transistor (Equation (10.45a) and (*b*) when velocity saturation occurs (Equation (11.17).

[Ans. (*a*) I_D(sat) $= 50(V_{GS} - 0.4)^2$ mA; (*b*) I_D(sat) $= 50(V_{GS} - 0.4)$ mA]

11.2 | MOSFET SCALING

As we noted in the previous chapter, the frequency response of MOSFETs increases as the channel length decreases. The driving force in CMOS technology evolution in the last couple of decades has been reduced channel lengths. Channel lengths of 0.13 μm or less are now the norm. One question that must be considered is what other device parameters must be scaled as the channel length is scaled down.

11.2.1 Constant-Field Scaling

The principle of constant-field scaling is that device dimensions and device voltages be scaled such that electric fields (both horizontal and vertical) remain essentially constant. To ensure that the reliability of the scaled device is not compromised, the electric fields in the scaled device must not increase.

Figure 11.12a shows the cross section and parameters of an original NMOS device and Figure 11.12b shows the scaled device, where the scaling parameter is k. Typically, $k \approx 0.7$ per generation of a given technology.

As seen in the figure, the channel length is scaled from L to kL. To maintain a constant horizontal electric field, the drain voltage must also be scaled from V_D to kV_D. The maximum gate voltage will also be scaled from V_G to kV_G so that the gate and drain voltages remain compatible. To maintain a constant vertical electric field, the oxide thickness then must also be scaled from t_{ox} to kt_{ox}.

The maximum depletion width at the drain terminal, for a one-sided pn junction, is

$$x_D = \sqrt{\frac{2\epsilon(V_{bi} + V_D)}{eN_a}} \tag{11.21}$$

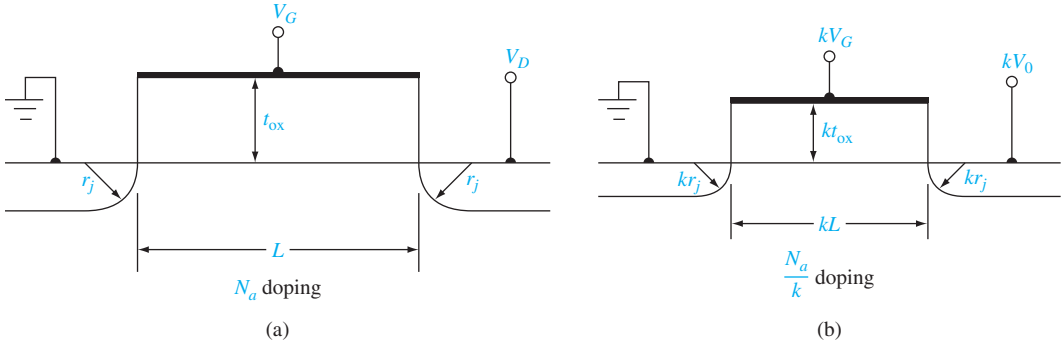


Figure 11.12 | Cross section of (a) original NMOS transistor and (b) scaled NMOS transistor.

Table 11.1 | Summary of constant-field device scaling

	Device and circuit parameters	Scaling factor ($k < 1$)
Scaled parameters	Device dimensions (L, t_{ox}, W, x_j)	k
	Doping concentration (N_a, N_d)	$1/k$
	Voltages	k
Effect on device parameters	Electric field	1
	Carrier velocity	1
	Depletion widths	k
	Capacitance ($C = \epsilon A/t$)	k
	Drift current	k
Effect on circuit parameters	Device density	$1/k^2$
	Power density	1
	Power dissipation per device ($P = IV$)	k^2
	Circuit delay time ($\approx CV/I$)	k
	Power−delay product ($P\tau$)	k^3

Source: Taur and Ning [23].

Since the channel length is being reduced, the depletion widths also need to be reduced. If the substrate doping concentration is increased by the factor ($1/k$), then the depletion width is reduced by approximately the factor k since V_D is reduced by k.

The drain current per channel width, for the transistor biased in the saturation region, can be written as

$$\frac{I_D}{W} = \frac{\mu_n \epsilon_{ox}}{2t_{ox}L}(V_G - V_T)^2 \rightarrow \frac{\mu_n \epsilon_{ox}}{2(kt_{ox})(kL)}(kV_G - V_T)^2 \approx \text{constant} \qquad (11.22)$$

The drift current per channel width remains essentially a constant, so if the channel width is reduced by k, then the drain current is also reduced by k. The area of the device, $A \approx WL$, is then reduced by k^2 and the power, $P = IV$, is also reduced by k^2. The power density in the chip remains unchanged.

Table 11.1 summarizes the device scaling and the effect on circuit parameters. Keep in mind that the width and length of interconnect lines are also assumed to be reduced by the same scaling factor.

11.2.2 Threshold Voltage—First Approximation

In constant-field scaling, the device voltages are reduced by the scaling factor k. It would seem appropriate that the threshold voltage should also be scaled by the same factor. The threshold voltage, for a uniformly doped substrate, can be written as

$$V_T = V_{FB} + 2\phi_{fp} + \frac{\sqrt{2\epsilon e N_a(2\phi_{fp})}}{C_{ox}} \qquad (11.23)$$

The first two terms in Equation (11.23) are functions of material parameters that do not scale and are only very slight functions of doping concentration. The last term is approximately proportional to $\sqrt{k}$, so the threshold voltage does not scale directly with the scaling factor k.

The effect of short channels on the threshold voltage is discussed further in section 11.3 of this chapter.

11.2.3 Generalized Scaling

In constant-field scaling, the applied voltages are scaled with the same scaling factor k as the device dimensions. However, in actual technology evolution, voltages have not been reduced with the same scaling factor. There has been reluctance, for example, to change standardized power supply levels that have been used in circuits earlier. In addition, other factors that do not scale, such as threshold voltage and subthreshold currents, have made the reduction in applied voltages less desirable. As a consequence, electric fields in MOS devices have tended to increase as device dimensions shrinked.

Consequences of increased electric fields are reduced reliability and increased power density. As the power density increases, the device temperature may increase. Increased temperature may affect the device reliability. As the oxide thickness is reduced and the electric field is increased, gate oxides are closer to breakdown and oxide integrity may be more difficult to maintain. In addition, direct tunneling of carriers through the oxide may be more likely to occur. Increased electric fields may also increase the chances of hot-electron effects, which are discussed later in this chapter. Reducing device dimensions, then, can introduce challenging problems that must be solved.

TEST YOUR UNDERSTANDING

TYU 11.3 An NMOS transistor has the following parameters: $L = 1\ \mu$m, $W = 10\ \mu$m, $t_{ox} = 250$ Å, $N_a = 5 \times 10^{15}$ cm^{-3}, and applied voltages of 3 V. If the device is to be scaled using constant-field scaling, determine the new device parameters for a scaling factor of $k = 0.7$. (Ans. $L = 0.7\ \mu$m, $W = 7\ \mu$m, $t_{ox} = 175$ Å, $N_a = 7.14 \times 10^{15}$ cm^{-3}, and applied voltages of 2.1 V)

11.3 | THRESHOLD VOLTAGE MODIFICATIONS

We have derived the ideal MOSFET relations in the previous chapter, including expressions for threshold voltage and for the current–voltage characteristics. We now consider some of the nonideal effects including channel length modulation. Additional effects on threshold voltage occur as the devices shrink in size. A reduction in channel length increases the transconductance and frequency response of the MOSFET, and a reduction in channel width increases the packing density in an integrated circuit. A reduction in either or both the channel length and channel width can affect the threshold voltage.

11.3.1 Short-Channel Effects

For the ideal MOSFET, we have derived the threshold voltage using the concept of charge neutrality in which the sum of charges in the metal oxide inversion layer and

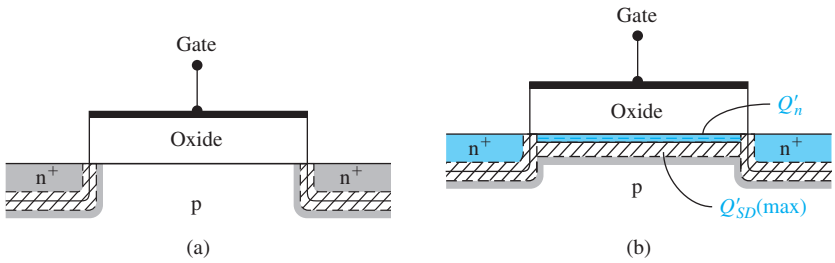


Figure 11.13 | Cross section of a long n-channel MOSFET (a) at flat band and (b) at inversion.

semiconductor space charge region is zero. We assumed that the gate area was the same as the active area in the semiconductor. Using this assumption, we have considered only equivalent surface charge densities and neglected any effects on threshold voltage that may occur due to source and drain space charge regions that extend into the active channel region.

Figure 11.13a shows the cross section of a long n-channel MOSFET at flat band, with zero source and drain voltage applied. The space charge regions at the source and drain extend into the channel region but occupy only a small fraction of the entire channel region. The gate voltage, then, will control essentially all of the space charge induced in the channel region at inversion as shown in Figure 11.13b.

As the channel length decreases, the fraction of charge in the channel region controlled by the gate decreases. This effect can be seen in Figure 11.14 for the flat-band condition. As the drain voltage increases, the reverse-biased space charge region at the drain extends further into the channel area and the gate controls even less bulk charge. The amount of charge in the channel region, $Q'_{SD}(\text{max})$, controlled by the gate, affects the threshold voltage and can be seen from Equation (11.24)

$$V_{TN} = \left(|Q'_{SD}(\text{max})| - Q'_{ss}\right)\left(\frac{t_{\text{ox}}}{\epsilon_{\text{ox}}}\right) + \phi_{ms} + 2\phi_{fp} \tag{11.24}$$

We can quantitatively determine the short-channel effects on the threshold voltage by considering the parameters shown in Figure 11.15. The source and drain junctions are characterized by a diffused junction depth r_j. We will assume that the lateral diffusion distance under the gate is the same as the vertical diffusion distance. This assumption is a reasonably good approximation for diffused junctions but becomes less accurate for ion implanted junctions. We will initially consider the case when the source, drain, and body contacts are all at ground potential.

The basic assumption in this analysis is that the bulk charge in the trapezoidal region under the gate is controlled by the gate. The potential difference across the bulk space charge region is $2\phi_{fp}$ at the threshold inversion point, and the built-in potential barrier height of the source and drain junctions is also approximately $2\,\phi_{fp}$, implying that the three space charge widths are essentially equal. We can then write

$$x_s \approx x_d \approx x_{dT} \equiv x_{dT} \tag{11.25}$$

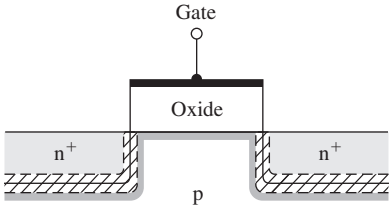


Figure 11.14 | Cross section of a short n-channel MOSFET at flat band.

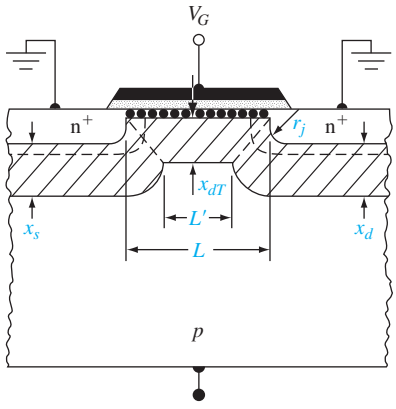


Figure 11.15 | Charge sharing in the short-channel threshold voltage model. *(From Yau [26].)*

Using the geometrical approximation, the average bulk charge per unit area Q'_B in the trapezoid is

$$|Q'_B| \cdot L = eN_a x_{dT}\left(\frac{L + L'}{2}\right) \tag{11.26}$$

From the geometry, we can show that

$$\frac{L + L'}{2L} = \left[1 - \frac{r_j}{L}\left(\sqrt{1 + \frac{2x_{dT}}{r_j}} - 1\right)\right] \tag{11.27}$$

Then

$$|Q'_B| = eN_a x_{dT}\left[1 - \frac{r_j}{L}\left(\sqrt{1 + \frac{2x_{dT}}{r_j}} - 1\right)\right] \tag{11.28}$$

Equation (11.28) is now used in place of $|Q'_{SD}(\text{max})|$ in the expression for the threshold voltage.

Since $|Q'_{SD}(\text{max})| = eN_a x_{dT}$, we can find ΔV_T as

$$\Delta V_T = -\frac{eN_a x_{dT}}{C_{ox}}\left[\frac{r_j}{L}\left(\sqrt{1 + \frac{2x_{dT}}{r_j}} - 1\right)\right] \tag{11.29}$$

where

$$\Delta V_T = V_{T(\text{short channel})} - V_{T(\text{long channel})} \tag{11.30}$$

As the channel length decreases, the threshold voltage shifts in the negative direction so that an n-channel MOSFET shifts toward depletion mode.

EXAMPLE 11.3

Objective: Calculate the threshold voltage shift due to short-channel effects.

Consider an n-channel MOSFET with $N_a = 3 \times 10^{16}\ \text{cm}^{-3}$, $L = 1.0\ \mu\text{m}$, $r_j = 0.3\ \mu\text{m}$, and $t_{ox} = 20\ \text{nm} = 200\ \text{Å}$.

Solution

We can find

$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} = \frac{(3.9)(8.85 \times 10^{-14})}{200 \times 10^{-8}} = 1.726 \times 10^{-7} \text{ F/cm}^2$$

$$\phi_{fp} = V_t \ln\left(\frac{N_a}{n_i}\right) = (0.0259) \ln\left(\frac{3 \times 10^{16}}{1.5 \times 10^{10}}\right) = 0.3758 \text{ V}$$

and

$$x_{dT} = \left[\frac{4\epsilon_s \phi_{fp}}{eN_a}\right]^{1/2} = \left[\frac{4(11.7)(8.85 \times 10^{-14})(0.3758)}{(1.6 \times 10^{-19})(3 \times 10^{16})}\right]^{1/2}$$

$$= 0.18 \times 10^{-4} \text{ cm} = 0.18 \ \mu\text{m}$$

The shift in threshold voltage is now found as

$$\Delta V_T = -\frac{eN_a x_{dT}}{C_{ox}}\left[\frac{r_j}{L}\left(\sqrt{1 + \frac{2x_{dT}}{r_j}} - 1\right)\right]$$

$$= -\frac{(1.6 \times 10^{-19})(3 \times 10^{16})(0.18 \times 10^{-4})}{1.726 \times 10^{-7}}\left[\frac{0.3}{1.0}\left(\sqrt{1 + \frac{2(0.18)}{0.3}} - 1\right)\right]$$

or

$$\Delta V_T = -0.0726 \text{ V}$$

Comment

If the design value of the threshold voltage of an n-channel MOSFET is to be $V_T = 0.35$ V, for example, a shift of $\Delta V_T = -0.0726$ V due to short-channel effects is significant and needs to be taken into account.

EXERCISE PROBLEM

Ex 11.3 Repeat Example 11.3 if the device parameters are $N_a = 10^{16}$ cm^{-3}, $L = 0.75$ μm, $r_j = 0.25$ μm, and $t_{ox} = 12$ nm $= 120$ Å.

(Ans. $\Delta V_T = -0.0769$ V)

The effect of short channels becomes more pronounced as the channel length is reduced further.

The shift in threshold voltage with channel length for an n-channel MOSFET is shown in Figure 11.16. As the substrate doping increases, the initial threshold voltage increases, as we have seen in the previous chapter, and the short-channel threshold shift also becomes larger. The short-channel effects on threshold voltage do not become significant until the channel length becomes less than approximately 2 μm. The threshold voltage shift also becomes smaller as the diffusion depth r_j becomes smaller so that very shallow junctions reduce the threshold voltage dependence on channel length.

Equation (11.29) is derived assuming that the source, channel, and drain space charge widths are all equal. If we now apply a drain voltage, the space charge width at the drain terminal widens, which makes L' smaller, and the amount of bulk charge controlled by the gate voltage decreases. This effect makes the threshold voltage a function of drain voltage. As the drain voltage increases, the threshold voltage of an n-channel MOSFET decreases. The threshold voltage versus channel length is

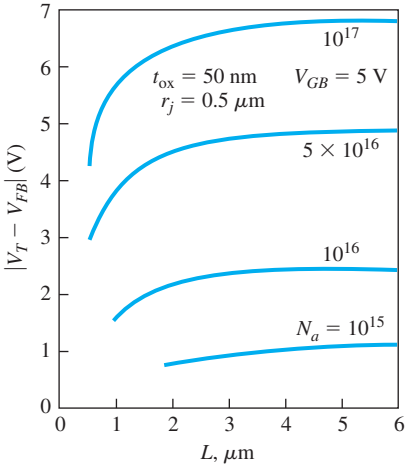


Figure 11.16 | Threshold voltage versus channel length for various substrate dopings. (*From Yau [26].*)

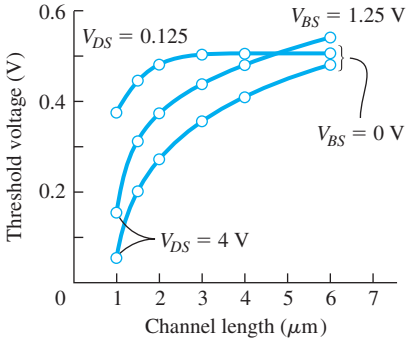


Figure 11.17 | Threshold voltage versus channel length for two values of drain-to-source and body-to-source voltage. (*From Yang [25].*)

plotted in Figure 11.17 for two values of drain-to-source voltage and two values of body-to-source voltage.

11.3.2 Narrow-Channel Effects

Figure 11.18 shows the cross section along the channel width of an n-channel MOSFET biased at inversion. The current is perpendicular to the channel width through the inversion charge. We may note in the figure that there is an additional space charge region at each end of the channel width. This additional charge is controlled by the gate voltage but is not included in the derivation of the ideal threshold voltage relation. The threshold voltage expression must be modified to include this additional charge.

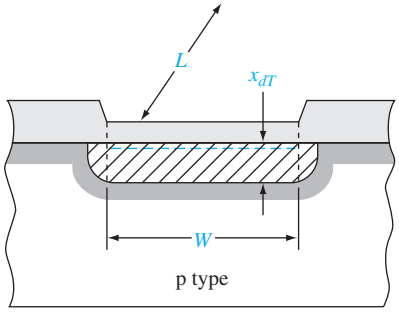


Figure 11.18 | Cross section of an n-channel MOSFET showing the depletion region along the width of the device.

If we neglect short-channel effects, the gate-controlled bulk charge can be written as

$$Q_B = Q_{B0} + \Delta Q_B \tag{11.31}$$

where Q_B is the total bulk charge, Q_{B0} is the ideal bulk charge, and ΔQ_B is the additional bulk charge at the ends of the channel width. For a uniformly doped p-type semiconductor biased at the threshold inversion point, we may write

$$|Q_{B0}| = eN_aWLx_{dT} \tag{11.32}$$

and

$$\Delta Q_B = eN_aLx_{dT}(\xi x_{dT}) \tag{11.33}$$

where ξ is a fitting parameter that accounts for the lateral space charge width. The lateral space charge width may not be the same as the vertical width x_{dT} due to the thicker field oxide at the ends, and/or due to the nonuniform semiconductor doping created by an ion implantation. If the ends were a semicircle, then $\xi = \pi/2$.

We may now write

$$|Q_B| = |Q_{B0}| + |\Delta Q_B| = eN_a\,WL\,x_{dT} + eN_a\,Lx_{dT}\,(\xi x_{dT})$$

$$= eN_aWLx_{dT}\left(1 + \frac{\xi x_{dT}}{W}\right) \tag{11.34}$$

The effect of the end space charge regions becomes significant as the width W decreases and the factor $(\xi\, x_{dT})$ becomes a significant fraction of the width W.

The change in threshold voltage due to the additional space charge is

$$\boxed{\Delta V_T = \frac{eN_a x_{dT}}{C_{\text{ox}}}\left(\frac{\xi x_{dT}}{W}\right)} \tag{11.35}$$

The shift in threshold voltage due to a narrow channel is in the positive direction for the n-channel MOSFET. As the width W becomes smaller, the shift in threshold voltage becomes larger.

EXAMPLE 11.4

Objective: Design the channel width that will limit the threshold voltage shift because of narrow channel effects to a specified value.

Consider a silicon n-channel MOSFET with $N_a = 3 \times 10^{16}$ cm^{-3} and $t_{ox} = 20$ nm $= 200$ Å. Let $\xi = \pi/2$. Assume that the threshold voltage shift is to be limited to $\Delta V_T = 0.2$ V.

■ Solution

We find

$$C_{ox} = 1.726 \times 10^{-7}\ \text{F/cm}^2 \qquad \text{and} \qquad x_{dT} = 0.18\ \mu\text{m}$$

From Equation (11.35), we can express the channel width as

$$W = \frac{eN_a\left(\xi x_{dT}^2\right)}{C_{ox}\,(\Delta V_T)} = \frac{(1.6 \times 10^{-19})(3 \times 10^{16})\left(\frac{\pi}{2}\right)(0.18 \times 10^{-4})^2}{(1.726 \times 10^{-7})(0.2)}$$

$$= 7.08 \times 10^{-5}\ \text{cm}$$

or

$$W = 0.708\ \mu\text{m}$$

Comment

We may note that the threshold voltage shift of $\Delta V_T = 0.2$ V occurs at a channel width of $W = 0.708\ \mu$m, which is approximately four times larger than the induced space charge width x_{dT}.

EXERCISE PROBLEM

Ex 11.4 Repeat Example 11.4 for $N_a = 10^{16}$ cm^{-3} and $t_{ox} = 8$ nm $= 80$ Å. Determine the channel width such that the threshold voltage shift is limited to $\Delta V_T = 0.1$ V.

(Ans. $W = 0.524\ \mu$m)

Figure 11.19 shows the threshold voltage as a function of channel width. We can again note that the threshold voltage shift begins to become apparent for channel widths that are large compared to the induced space charge width.

Figure 11.20a, b shows qualitatively the threshold voltage shifts due to short-channel and narrow-channel effects, respectively, in n-channel MOSFETs. The narrow-channel device produces a larger threshold voltage; the short-channel device produces a smaller threshold voltage. For devices exhibiting both short-channel and narrow-channel effects, the two models need to be combined into a three-dimensional volume approximation of the space charge region controlled by the gate.

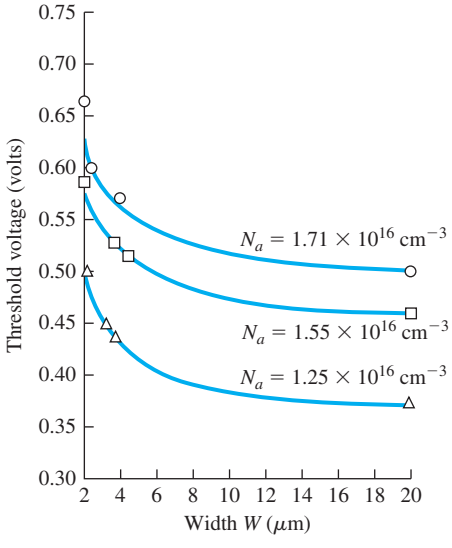


Figure 11.19 | Threshold voltage versus channel width (solid curves, theoretical; points, experimental).
(*From Akers and Sanchez [1].*)

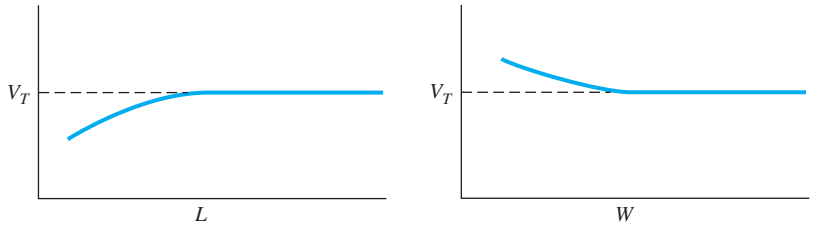


Figure 11.20 | Qualitative variation of threshold voltage (a) with channel length and (b) with channel width.

11.4 | ADDITIONAL ELECTRICAL CHARACTERISTICS

There is a tremendous volume of information on MOSFETs that cannot be included in an introductory text on semiconductor physics and devices. However, two additional topics should be included here: breakdown voltage and threshold adjustment by ion implantation.

11.4.1 Breakdown Voltage

Several voltage breakdown mechanisms in the MOSFET must be considered, including voltage breakdown across the oxide as well as the various voltage breakdown mechanisms in the semiconductor junctions.

Oxide Breakdown We have assumed that the oxide is a perfect insulator. However, if the electric field in the oxide becomes large enough, breakdown can occur, which can lead to a catastrophic failure. In silicon dioxide, the electric field at breakdown is on the order of 6×10^6 V/cm. This breakdown field is larger than that in silicon, but the gate oxides are also quite thin. A gate voltage of approximately 30 V would produce breakdown in an oxide with a thickness of 500 Å. However, a safety margin of a factor of 3 is common, so that the maximum safe gate voltage with $t_{ox} = 500$ Å would be 10 V. A safety margin is necessary since there may be defects in the oxide that lower the breakdown field. Oxide breakdown is normally not a serious problem except in power devices and ultrathin oxide devices. Other oxide degradation problems are discussed later in this chapter.

Avalanche Breakdown Avalanche breakdown may occur by impact ionization in the space charge region near the drain terminal. We have considered avalanche breakdown in pn junctions in Chapter 7. In an ideal planar one-sided pn junction, breakdown is a function primarily of the doping concentration in the low-doped region of the junction. For the MOSFET, the low-doped region corresponds to the semiconductor substrate. If a p-type substrate doping is $N_a = 3 \times 10^{16}$ cm^{-3}, for example, the pn junction breakdown voltage would be approximately 25 V for a planar junction. However, the n^+ drain contact may be a fairly shallow diffused region with

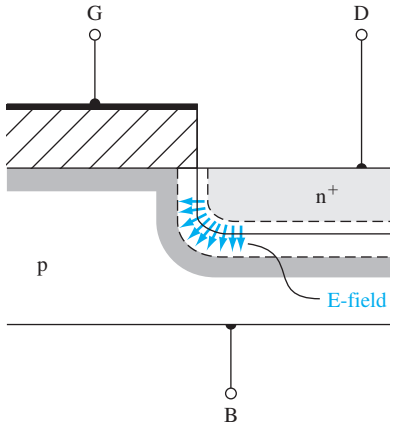


Figure 11.21 | Curvature effect on the electric field in the drain junction.

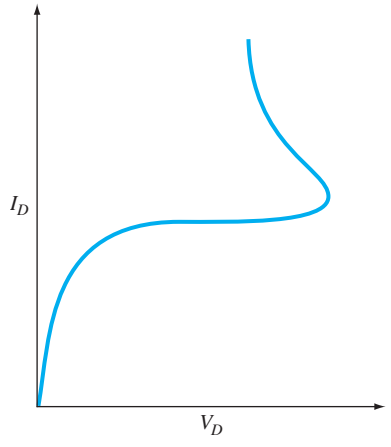


Figure 11.22 | Current–voltage characteristic showing the snapback breakdown effect.

a large curvature. The electric field in the depletion region tends to be concentrated at the curvature, which lowers the breakdown voltage. This curvature effect is shown in Figure 11.21.

Near Avalanche and Snapback Breakdown[1] Another breakdown mechanism results in the S-shaped breakdown curve shown in Figure 11.22. This breakdown process is due to second order effects and can be explained with the aid of Figure 11.23. The n-channel enhancement-mode MOSFET geometry in Figure 11.23a shows the

[1]This section may be postponed until after the bipolar transistor is considered in Chapter 12.

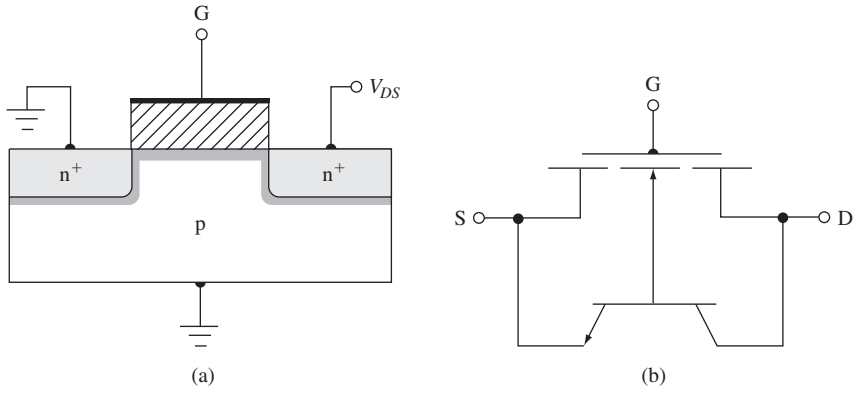


Figure 11.23 | (a) Cross section of the n-channel MOSFET. (b) Equivalent circuit including the parasitic bipolar transistor.

n-type source and drain contacts along with the p-type substrate. The source and body are at ground potential. The n(source)-p(substrate)-n(drain) structure also forms a parasitic bipolar transistor. The equivalent circuit is shown in Figure 11.23b.

Figure 11.24a shows the device when avalanche breakdown is just beginning in the space charge region near the drain. We tend to think of the avalanche breakdown suddenly occurring at a particular voltage. However, avalanche breakdown is a gradual process that starts at low current levels and for electric fields somewhat below the breakdown field. The electrons generated by the avalanche process flow into the drain and contribute to the drain current. The avalanche-generated holes generally flow through the substrate to the body terminal. Since the substrate has a nonzero resistance, a voltage drop is produced as shown. This potential difference drives the source-to-substrate pn junction into forward bias near the source terminal. The source is heavily doped n-type; thus, a large number of electrons can be injected from the source contact into the substrate under forward bias. This process becomes

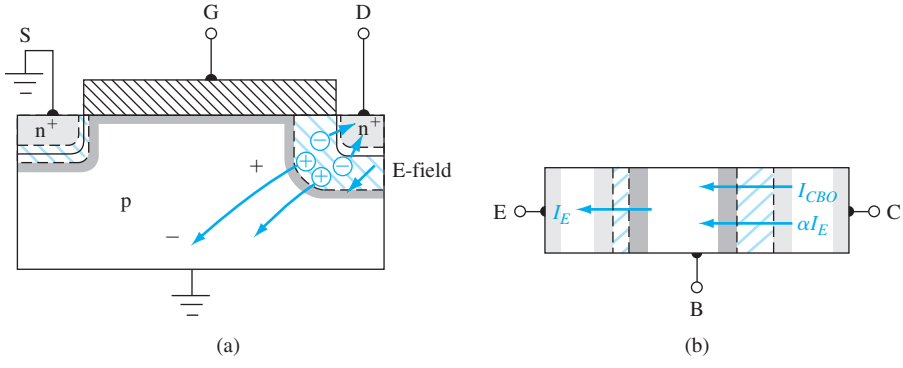


Figure 11.24 | (a) Substrate current–induced voltage drop caused by avalanche multiplication at the drain. (b) Currents in the parasitic bipolar transistor.

severe as the voltage drop in the substrate approaches 0.6 to 0.7 V. A fraction of the injected electrons diffuses across the parasitic base region into the reverse-biased drain space charge region where they also add to the drain current.

The avalanche breakdown process is a function of not only the electric field but the number of carriers involved. The rate of avalanche breakdown increases as the number of carriers in the drain space charge region increases. We now have a regenerative or positive feedback mechanism. Avalanche breakdown near the drain terminal produces the substrate current, which produces the forward-biased source-substrate pn junction voltage. The forward-biased junction injects carriers that can diffuse back to the drain and increase the avalanche process. The positive feedback produces an unstable system.

The snapback or negative resistance portion of the curve shown in Figure 11.22 can now be explained by using the parasitic bipolar transistor. The potential of the base of the bipolar transistor near the emitter (source) is almost floating, since this voltage is determined primarily by the avalanche-generated substrate current rather than an externally applied voltage.

For the open-base bipolar transistor shown in Figure 11.24, we can write

$$I_C = \alpha I_E + I_{CB0} \tag{11.36}$$

where α is the common base current gain and I_{CB0} is the base-collector leakage current. For an open base, $I_C = I_E$, so Equation (11.36) becomes

$$I_C = \alpha I_C + I_{CB0} \tag{11.37}$$

At breakdown, the current in the B–C junction is multiplied by the multiplication factor M, so we have

$$I_C = M(\alpha I_C + I_{CB0}) \tag{11.38}$$

Solving for I_C we obtain

$$I_C = \frac{MI_{CB0}}{1 - \alpha M} \tag{11.39}$$

Breakdown is defined as the condition that produces $I_C \rightarrow \infty$. For a single reverse-biased pn junction, $M \rightarrow \infty$ at breakdown. However, from Equation (11.39), breakdown is now defined to be the condition when $\alpha M \rightarrow 1$ or, for the open-base condition, breakdown occurs when $M \rightarrow 1/\alpha$, which is a much lower multiplication factor than for the simple pn junction.

An empirical relation for the multiplication factor is usually written as

$$M = \frac{1}{1 - (V_{CE}/V_{BD})^m} \tag{11.40}$$

where m is an empirical constant in the range of 3 to 6 and V_{BD} is the junction breakdown voltage.

The common base current gain factor α is a strong function of collector current for small values of collector current. This effect will be discussed in Chapter 12 on bipolar transistors. At low currents, the recombination current in the B-E junction is a significant fraction of the total current so that the common base current gain is small.

As the collector current increases, the value of α increases. As avalanche breakdown begins and I_C is small, particular values of M and V_{CE} are required to produce the condition of $\alpha M = 1$. As the collector current increases, α increases; therefore, smaller values of M and V_{CE} are required to produce the avalanche breakdown condition. The snapback, or negative resistance, breakdown characteristic is then produced.

Only a fraction of the injected electrons from the forward-biased source-substrate junction are collected by the drain terminal. A more exact calculation of the snapback characteristic would necessarily involve taking into account this fraction; thus, the simple model would need to be modified. However, the above discussion qualitatively describes the snapback effect. The snapback effect can be minimized by using a heavily doped substrate that will prevent any significant voltage drop from being developed. A thin epitaxial p-type layer with the proper doping concentration to produce the required threshold voltage can be grown on a heavily doped substrate.

Near Punch-Through Effects Punch-through is the condition at which the drain-to-substrate space charge region extends completely across the channel region to the source-to-substrate space charge region. In this situation, the barrier between the source and drain is completely eliminated and a very large drain current would exist.

However, the drain current will begin to increase rapidly before the actual punch-through condition is reached. This characteristic is referred to as the near punch-through condition, also known as **D**rain-**I**nduced **B**arrier **L**owering (DIBL). Figure 11.25a shows the ideal energy-band diagram from source to drain for a long n-channel MOSFET for the case when $V_{GS} < V_T$ and when the drain-to-source voltage is relatively small. The large potential barriers prevent significant current between the drain and source. Figure 11.25b shows the energy-band diagram when a relatively large drain voltage V_{DS2} is applied. The space charge region near the drain terminal is beginning to interact with the source space charge region and the potential barrier is being lowered. Since the current is an exponential function of barrier height, the current will increase very rapidly with drain voltage once this near punch-through

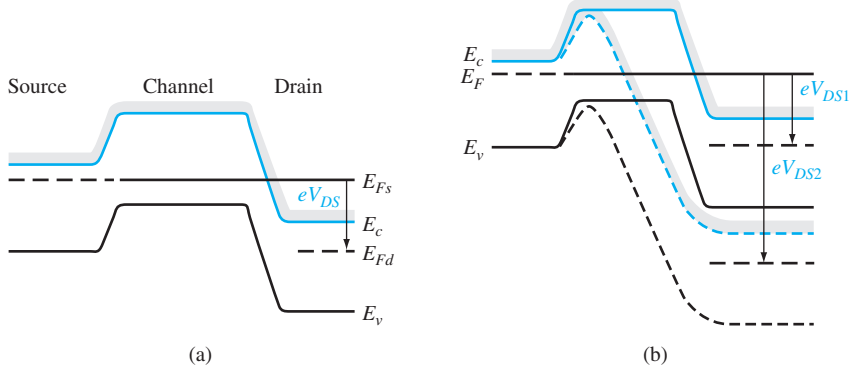


Figure 11.25 | (a) Equipotential plot along the surface of a long-channel MOSFET. (b) Equipotential plot along the surface of a short-channel MOSFET before and after punch-through.

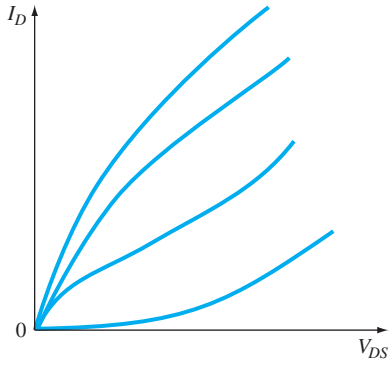


Figure 11.26 | Typical *I–V* characteristics of a MOSFET exhibiting punch-through effects.

condition has been reached. Figure 11.26 shows some typical characteristics of a short-channel device with a near punch-through condition.

EXAMPLE 11.5

Objective: Calculate the theoretical punch-through voltage assuming the abrupt junction approximation.

Consider an n-channel MOSFET with source and drain doping concentrations of $N_d = 10^{19}$ cm^{-3} and a channel region doping of $N_a = 10^{16}$ cm^{-3}. Assume a channel length of $L = 1.2\ \mu$m, and assume the source and body are at ground potential.

■ Solution

The pn junction built-in potential barrier is given by

$$V_{bi} = V_t \ln\left(\frac{N_a N_d}{n_i^2}\right) = (0.0259)\ln\left[\frac{(10^{16})(10^{19})}{(1.5\times 10^{10})^2}\right] = 0.874\ \text{V}$$

The zero-biased source–substrate pn junction width is

$$x_{d0} = \left[\frac{2\epsilon_s V_{bi}}{eN_a}\right]^{1/2} = \left[\frac{2(11.7)(8.85\times 10^{-14})(0.874)}{(1.6\times 10^{-19})(10^{16})}\right]^{1/2} = 0.336\ \mu\text{m}$$

The reverse-biased drain-substrate pn junction width is given by

$$x_d = \left[\frac{2\epsilon_s(V_{bi} + V_{DS})}{eN_a}\right]^{1/2}$$

At punch-through, we will have

$$x_{d0} + x_d = L \qquad \text{or} \qquad 0.336 + x_d = 1.2$$

which gives $x_d = 0.864\ \mu$m at the punch-through condition. We can then find

$$V_{bi} + V_{DS} = \frac{x_d^2 eN_a}{2\epsilon_s} = \frac{(0.864\times 10^{-4})^2(1.6\times 10^{-19})(10^{16})}{2(11.7)(8.85\times 10^{-14})}$$

$$= 5.77\ \text{V}$$

The punch-through voltage is then found as

$$V_{DS} = 5.77 - 0.874 = 4.9 \text{ V}$$

Comment

As the two space charge regions approach punch-through, the abrupt junction approximation is no longer a good assumption. The drain current will begin to increase rapidly before the theoretical punch-through voltage is reached.

EXERCISE PROBLEM

Ex 11.5 Repeat Example 11.5 for a substrate doping concentration of $N_a = 3 \times 10^{16}$ cm^{-3} and a channel length of $L = 0.8$ μm.

(Ans. $V_{pa} = 7.52$ V)

For a doping of 10^{16} cm^{-3}, the two space charge regions will begin to interact when the abrupt depletion layers are approximately 0.25 μm apart. The drain voltage at which this near punch-through condition, also known as drain-induced barrier lowering, occurs is significantly less than the ideal punch-through voltage such as calculated in Example 11.5 (see Problem 11.33).

*11.4.2 The Lightly Doped Drain Transistor

The junction breakdown voltage is a function of the maximum electric field. As the channel length becomes smaller, the bias voltages may not be scaled down accordingly, so the junction electric fields become larger. As the electric field increases, near avalanche breakdown and near punch-through effects become more serious. In addition, as device geometries are scaled down, the parasitic bipolar device becomes more dominant and breakdown effects are enhanced.

One approach that reduces these breakdown effects is to alter the doping profile of the drain contact. The **L**ightly **D**oped **D**rain (LDD) design and doping profiles are shown in Figure 11.27a, the conventional MOSFET and doping profiles are shown in Figure 11.27b for comparison. By introducing the lightly doped region, the peak electric field in the space charge region is reduced and the breakdown effects are minimized. The peak electric field at the drain junction is a function of the semiconductor doping as well as the curvature of the n^+ drain region. Figure 11.28 shows the physical geometries of a conventional n^+ drain contact and an LDD structure superimposed on the same plot. The magnitude of the electric field at the oxide–semiconductor interface in the LDD structure is less than in the conventional structure. The electric field in the conventional device peaks approximately at the metallurgical junction and drops quickly to zero in the drain because no field can exist in the highly conductive n^+ region. On the other hand, the electric field in the LDD device extends across the n-region before dropping to zero at the drain. This effect minimizes breakdown and the hot electron effects, which we discuss in Section 11.5.3.

Two disadvantages of the LDD device are an increase in both fabrication complexity and drain resistance. The added processing steps, however, produce a device with significant improvements in performance. The cross section of the LDD device

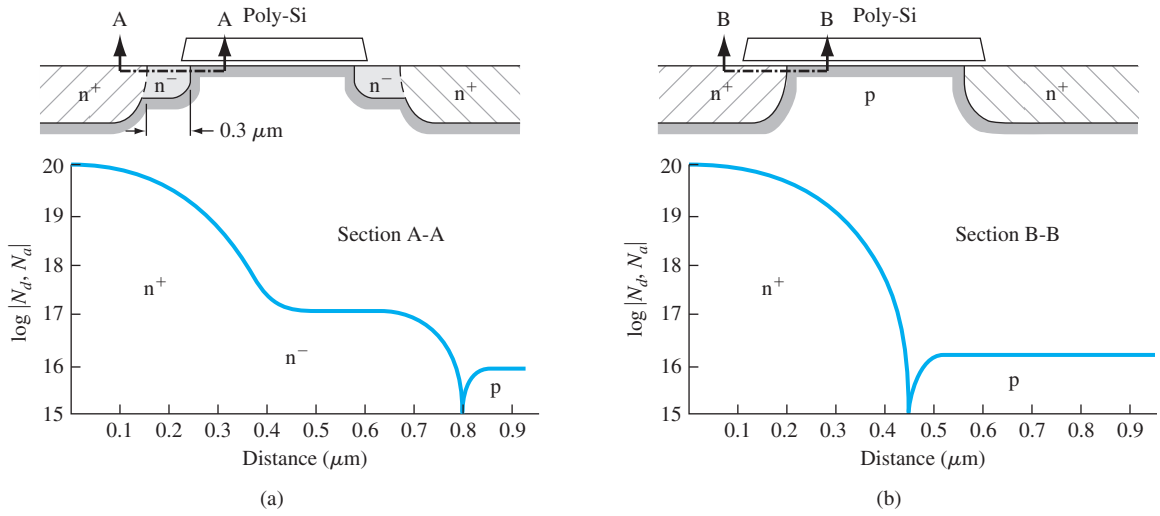


Figure 11.27 | (a) The lightly doped drain (LDD) structure. (b) Conventional structure. (*From Ogura et al. [12].*)

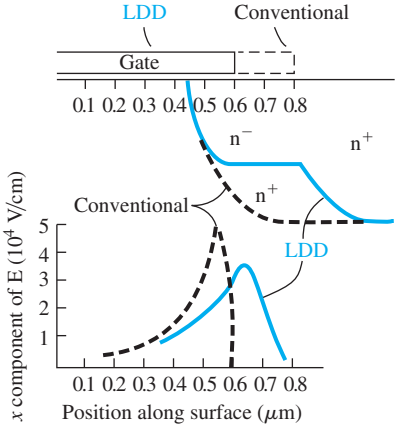


Figure 11.28 | Magnitude of the electric field at the Si–SiO_2 interface as a function of distance; $V_{DS} = 10$ V. $V_{SB} = 2$ V, $V_{GS} = V_T$. (*From Ogura et al. [12].*)

shown in Figure 11.27 indicates a lightly doped n region at the source terminal also. The inclusion of this region does not improve device performance, but does reduce the fabrication complexity as much as possible. The added series resistances increase power dissipation in the device; this must be taken into account in high-power devices.

11.4.3 Threshold Adjustment by Ion Implantation

Several factors, such as fixed oxide charge, metal–semiconductor work function difference, oxide thickness, and semiconductor doping, influence the threshold voltage. Although all of these parameters may be fixed in a particular design and fabrication process, the resulting threshold voltage may not be acceptable for all applications. Ion implantation can be used to change and adjust the substrate doping near the oxide–semiconductor surface to provide the desired threshold voltage. In addition, ion implantation is used for more than doping the channel. It is used extensively as a standard part of device fabrication; for example, it is used to form the source and drain regions of the transistor.

To change the doping and thereby change the threshold voltage, a precise, controlled number of either donor or acceptor ions are implanted into the semiconductor near the oxide surface. When an MOS device is biased in either depletion or inversion and when the implanted dopant atoms are within the induced space charge region, then the ionized dopant charge adds to (or subtracts from) the maximum space charge density, which controls the threshold voltage. An implant of acceptor ions into either a p- or n-type substrate will shift the threshold voltage to more positive values, while an implant of donor ions will shift the threshold voltage to more negative values. Ion implantation can be carried out to change a depletion-mode device to enhancement-mode or an enhancement-mode device to depletion-mode, which is an important application of this technology.

As a first approximation, assume that D_I acceptor atoms per cm^2 are implanted into a p-type substrate directly adjacent to the oxide–semiconductor interface as shown in Figure 11.29a. The shift in threshold voltage due to the implant is

$$\Delta V_T = + \frac{eD_I}{C_{ox}} \tag{11.41}$$

If donor atoms were implanted into the p-type substrate, the space charge density would be reduced; thus, the threshold voltage would shift in the negative voltage direction.

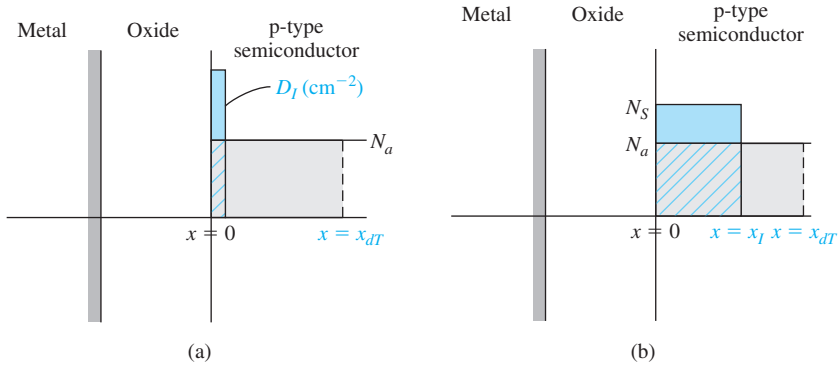


Figure 11.29 | (a) Ion-implanted profile approximated by a delta function. (b) Ion-implanted profile approximated by a step function in which the depth x_i is less than the space charge width x_{dT}.

A second type of implant approximation is the step junction, shown in Figure 11.29b. If the induced space charge width at the threshold inversion point is less than x_I, then the threshold voltage is determined on the basis of a semiconductor with a uniform doping concentration of N_s atoms per cm^3. On the other hand, if the induced space charge width is greater than x_I at the threshold inversion point, then a new expression for x_{dT} must be derived. We can apply Poisson's equation and show that the maximum induced space charge width after the step implant is

$$x_{dT} = \sqrt{\frac{2\epsilon_s}{eN_a}}\left[2\phi_{fp} - \frac{ex_I^2}{2\epsilon_s}(N_s - N_a)\right]^{1/2} \tag{11.42}$$

The threshold voltage after a step implant for the case when $x_{dT} > x_I$ can be written as

$$\boxed{V_T = V_{T0} + \frac{eD_I}{C_{ox}}} \tag{11.43}$$

where V_{T0} is the preimplant threshold voltage. The parameter D_I is given by

$$D_I = (N_s - N_a)x_I \tag{11.44}$$

which is the number per cm^2 of implanted ions. The preimplant threshold voltage is

$$V_{T0} = V_{FB0} + 2\phi_{fp0} + \frac{eN_a x_{dT0}}{C_{ox}} \tag{11.45}$$

where the subscript 0 indicates the preimplant values.

EXAMPLE 11.6

Objective: Design the ion implant dose required to adjust the threshold voltage to a specified value.

Consider an n-channel silicon MOSFET with a doping of $N_a = 5 \times 10^{15}$ cm^{-3}, an oxide thickness of $t_{ox} = 18$ nm $= 180$ Å, and an initial flat-band voltage of $V_{FBO} = -1.25$ V. Determine the ion implantation dose such that a threshold voltage of $V_T = +0.4$ V is obtained.

■ Solution

We find that

$$\phi_{fpO} = V_t \ln\left(\frac{N_a}{n_i}\right) = (0.0259)\ln\left(\frac{5 \times 10^{15}}{1.5 \times 10^{10}}\right) = 0.3294 \text{ V}$$

$$x_{dTO} = \left[\frac{4\epsilon_s \phi_{fpO}}{eN_a}\right]^{1/2} = \left[\frac{4(11.7)(8.85 \times 10^{-14})(0.3294)}{(1.6 \times 10^{-19})(5 \times 10^{15})}\right]^{1/2}$$

$$= 0.4130 \times 10^{-4} \text{ cm}$$

and

$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} = \frac{(3.9)(8.85 \times 10^{-14})}{180 \times 10^{-8}} = 1.9175 \times 10^{-7} \text{ F/cm}^2$$

The initial pre-implant threshold voltage is

$$V_{TO} = V_{FBO} + 2\phi_{fpO} + \frac{eN_a x_{dTO}}{C_{ox}}$$

$$= -1.25 + 2(0.3294) + \frac{(1.6 \times 10^{-19})(5 \times 10^{15})(0.4130 \times 10^{-4})}{1.9175 \times 10^{-7}}$$

$$= -0.419 \text{ V}$$

The threshold voltage after implant, from Equation (11.43), is

$$V_T = V_{TO} + \frac{eD_I}{C_{ox}}$$

so that

$$+0.40 = -0.419 + \frac{(1.6 \times 10^{-19})D_I}{1.9175 \times 10^{-7}}$$

which gives

$$D_I = 9.815 \times 10^{11}\ \text{cm}^{-2}$$

If the uniform step implant extends to a depth of $x_I = 0.15\ \mu$m, for example, then the equivalent acceptor concentration at the surface is

$$N_s - N_a = \frac{D_I}{x_I}$$

or

$$N_s - 5 \times 10^{15} = \frac{9.815 \times 10^{11}}{0.15 \times 10^{-4}}$$

so that

$$N_s = 7.04 \times 10^{16}\ \text{cm}^{-3}$$

Comment

It is assumed in the above calculation that the induced space charge width in the channel region is greater than the ion implant dept x_I. The calculation satisfies our assumptions.

EXERCISE PROBLEM

Ex 11.6 A silicon MOSFET has the following parameters: $N_a = 10^{15}\ \text{cm}^{-3}$, p^+ polysilicon gate with an initial flat-band voltage of $V_{FBO} = +0.95$ V, and an oxide thickness of $t_{ox} = 12$ nm $= 120$ Å. A final threshold voltage of $V_T = +0.40$ V is required. Assume the idealized delta function for the ion implant profile shown in Figure 11.29a. (*a*) What type of ion (acceptor or donor) should be implanted? (*b*) Determine the required ion dose D_I.

[Ans. (*a*) donor ions; (*b*) $D_I = 2.11 \times 10^{12}\ \text{cm}^{-2}$]

The actual implant dose versus distance is neither a delta function nor a step function; it tends to be a Gaussian-type distribution. The threshold shift due to a nonuniform ion implant density may be defined as the shift in curves of N_{inv} versus V_G, where N_{inv} is the inversion carrier density per cm^2. This shift corresponds to an experimental shift of drain current versus V_G when the transistor is biased in the linear mode. The criteria of the threshold inversion point as $\phi_s = 2\phi_{fp}$ in the implanted devices have an uncertain meaning because of the nonuniform doping in the substrate. The determination of the threshold voltage becomes more complicated and will not be made here.

TEST YOUR UNDERSTANDING

TYU 11.4 Repeat Exercise Problem Ex 11.6 for a final threshold voltage of (*a*) $V_T = +0.25$ V and (*b*) $V_T = -0.25$ V.

[Ans. (*a*) $D_I = 8.02 \times 10^{11}\ \text{cm}^{-2}$; (*b*) $D_I = 2.03 \times 10^{11}\ \text{cm}^{-2}$]

*11.5 | RADIATION AND HOT-ELECTRON EFFECTS

We have considered the effects of fixed trapped oxide charge and interface state charge on the capacitance–voltage characteristics of MOS capacitors and on the MOSFET characteristics. These charges can exist because the oxide is essentially a perfect dielectric and a net charge density can exist in a dielectric material. Two processes that generate these charges are ionizing radiation and impact ionization in the drain region of a MOSFET operating near avalanche breakdown.

MOS devices are exposed to ionizing radiation, for example, in communication satellites orbiting through the Van Allen radiation belts. The ionizing radiation can produce additional fixed oxide charge and also additional interface states. In this short discussion of radiation effects in MOSFETs, we are concerned only with the permanent effects that occur in the device characteristics.

Another source can generate oxide charge and interface states: the hot electron effect. Electrons near the drain terminal of a MOSFET operating near avalanche breakdown can have energies that are much larger than the thermal-equilibrium value. These *hot electrons* have energies sufficient to penetrate the oxide–semiconductor barrier.

11.5.1 Radiation-Induced Oxide Charge

Gamma-rays or x-rays incident on semiconductor or oxide materials can interact with valence band electrons. The incident radiation photons can impart enough energy to a valence electron to elevate the electron into the conduction band; an empty state or hole is also produced in the valence band. This process generates electron–hole pairs. These newly generated electrons and holes can move through a material under the influence of an electric field.

Figure 11.30 shows the energy-band diagram of an MOS device with a p-type substrate and a positive gate voltage. The bandgap energy of silicon dioxide is

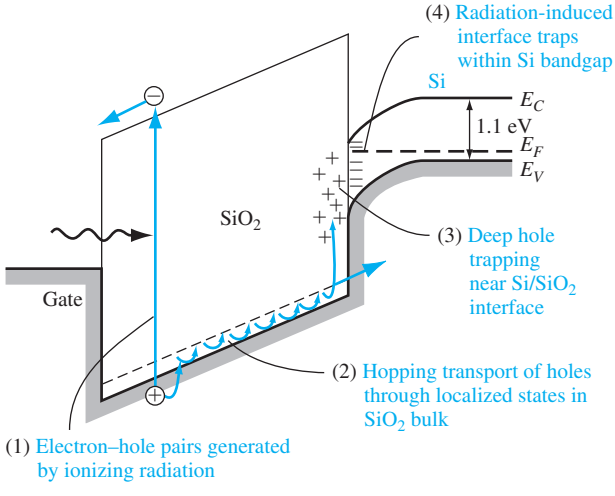


Figure 11.30 | Schematic of ionizing radiation–induced processes in an MOS capacitor with a positive gate bias. (*From Ma and Dressendorfer [7].*)

approximately 9 eV. The figure schematically shows the creation of an electron–hole pair in the oxide by ionizing radiation. The force on the radiation-induced electron is toward the gate and the force on the radiation-induced hole is toward the semiconductor. It has been found that generated electrons in the oxide are fairly mobile, with a mobility value on the order of 20 cm^2/V-s. At high electric fields, the electron velocity in the oxide also saturates at approximately 10^7 cm/s, so that the electron transit time for typical gate oxide thicknesses is on the order of a few picoseconds. For positive gate voltages, the vast majority of radiation-induced electrons flow out through the gate terminal; for this reason, in general these electrons do not play a significant role in the radiation response of MOS devices.

The generated holes, on the other hand, undergo a stochastic hopping transport process through the oxide (shown schematically in Figure 11.30). The hole transport process is dispersive in time and is a function of the electric field, temperature, and oxide thickness. The effective hole mobility in silicon dioxide is typically in the range of 10^{-4} to 10^{-11} cm^2/V-s; thus, holes are relatively immobile when compared with electrons.

When holes reach the silicon–silicon dioxide (Si–SiO_2) interface, a fraction are captured in trapping sites while the remainder flow into the silicon. A net positive radiation-induced charge is then trapped in the oxide due to these captured holes. This trapped charge can last from hours to years. As we have seen, a positive oxide charge causes a negative shift in threshold voltage.

The measured areal hole trap densities are in the range of 10^{12} to 10^{13} cm^{-2} depending upon oxide and device processing. In general, these traps are located within approximately 50 Å of the Si–SiO_2 interface. The hole trap is usually associated with a trivalent silicon defect that has an oxygen vacancy in the SiO_2 structure. The oxygen vacancies are located in a silicon-rich region near the Si–SiO_2 interface.

Since the threshold or flat-band voltage shift is a function of the amount of trapped charge, the voltage shift is a function of applied voltage across the oxide. Figure 11.31 shows the flat-band voltage shift of an MOS capacitor as a function of gate voltage applied during irradiation. For small values of gate voltage, some radiation-generated

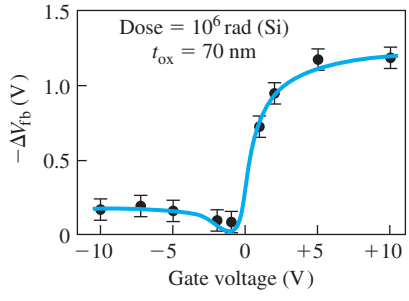


Figure 11.31 | Radiation-induced flat-band voltage shift in an MOS capacitor as a function of applied gate bias during irradiation.
(From Ma and Dressendorfer [7].)

holes and electrons recombine in the oxide. Hence, the amount of charge reaching the Si–SiO_2 interface and being trapped is less than for a large positive gate voltage, where essentially all radiation-generated holes reach the interface without recombining with electrons. If the fraction of generated holes that become trapped is relatively constant, then the voltage shift becomes independent of positive gate bias, as shown in the figure. For negative applied gate voltages, the radiation-induced holes move toward the gate terminal. There can be positive charge trapping in the oxide near the gate, but the effect of this trapped charge on the threshold voltage is small.

EXAMPLE 11.7

Objective: Calculate the threshold voltage shift due to radiation-induced oxide charge trapping.

Consider a MOS device with an oxide thickness of $t_{ox} = 25$ nm $= 250$ Å. Assume that a pulse of ionizing radiation creates 10^{18} electron–hole pairs per cm^3 in the oxide. Also assume that the electrons are swept out through the gate terminal with zero recombination, and that 20 percent of the generated holes are trapped at the oxide–semiconductor interface.

■ Solution

The areal density of holes generated in the oxide is

$$N_h = (10^{18})(250 \times 10^{-8}) = 2.5 \times 10^{12}\ \text{cm}^{-2}$$

The equivalent trapped surface charge is then

$$Q'_{ss} = (2.5 \times 10^{12})(0.2)(1.6 \times 10^{-19}) = 8 \times 10^{-8}\ \text{C/cm}^2$$

We find

$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} = \frac{(3.9)(8.85 \times 10^{-14})}{250 \times 10^{-8}} = 1.381 \times 10^{-7}\ \text{F/cm}^2$$

The threshold voltage shift is then

$$\Delta V_T = -\frac{Q'_{ss}}{C_{ox}} = -\frac{8 \times 10^{-8}}{1.381 \times 10^{-7}} = -0.579\ \text{V}$$

■ Comment

As we have seen previously, a positive fixed oxide charge shifts the threshold voltage in the negative voltage direction. The ionizing radiation may shift an enhancement-mode device into the depletion mode.

■ EXERCISE PROBLEM

Ex 11.7 Repeat Example 11.7 for a MOS device with an oxide thickness of (*a*) $t_{ox} = 12$ nm $= 120$ Å and (*b*) $t_{ox} = 8$ nm $= 80$ Å. (*c*) What can be said about the shift in threshold voltage as the oxide thickness decreases?

[Ans. (*a*) $\Delta V_T = -0.1344$ V; (*b*) $\Delta V_T = -0.0593$ V; (*c*) less shift]

One failure mechanism, therefore, caused by the radiation-induced oxide charge in an n-channel MOSFET in an integrated circuit is a shift from enhancement mode to depletion mode. The device will be turned on rather than off at zero gate voltage; consequently, the circuit function may be disrupted or an excessive power supply current may be generated in the circuit.

The gate voltage in a p-channel MOSFET is normally negative with respect to the substrate. Radiation-generated holes in the oxide are forced to the gate–oxide interface. The trapped charge in this region has less effect on the threshold voltage, so threshold shifts in p-channel MOSFETs are normally smaller if the trap concentrations at the gate–oxide and oxide–semiconductor interfaces are of the same order of magnitude.

11.5.2 Radiation-Induced Interface States

We have considered the effect of interface states on the *C–V* characteristics of an MOS capacitor and on the MOSFET characteristics. The net charge in the interface states of an n-channel MOS device at the threshold inversion point is negative. This negative charge causes a shift in threshold voltage in the positive voltage direction, which is opposite to the shift due to the positive oxide charge. In addition, since the interface states can be charged, they are another source of coulomb interaction with the inversion charge carrier, which means that the inversion carrier mobility is a function of the interface state density through surface-scattering effects. Interface states, then, affect both threshold voltage and carrier mobility.

When MOS devices are exposed to ionizing radiation, additional interface states are generated at the Si–SiO_2 interface. The radiation-induced interface states tend to be donor states in the lower half of the bandgap and acceptor states in the upper half. Figure 11.32 shows the threshold voltage in an n-channel and a p-channel MOSFET as a function of ionizing radiation dose. We initially see the negative threshold voltage shift in both devices due to the radiation-induced positive oxide charge. The reversal in threshold shift at the higher dose levels is attributable to the creation of radiation-induced interface states that tend to compensate the radiation-induced positive oxide charge.

In our discussion of subthreshold conduction, we have mentioned that the slope of the ln I_D versus V_{GS} curves in the subthreshold region is a function of the density of interface states. Figure 11.33 shows the subthreshold current at several total ionizing

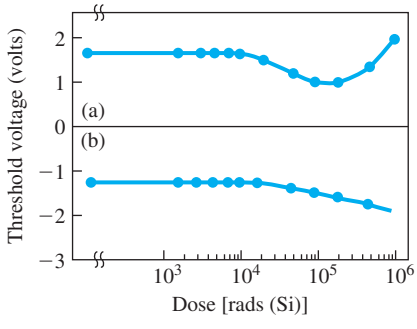


Figure 11.32 | Threshold voltage versus total ionizing radiation dose of (a) an n-channel MOSFET and (b) a p-channel MOSFET.
(From Ma and Dressendorfer [7].)

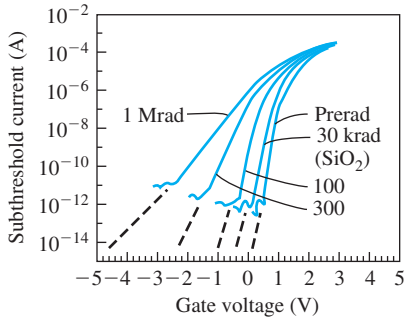


Figure 11.33 | Subthreshold current versus gate voltage of a MOSFET prior to irradiation and at four total radiation dose levels.
(From Ma and Dressendorfer [7].)

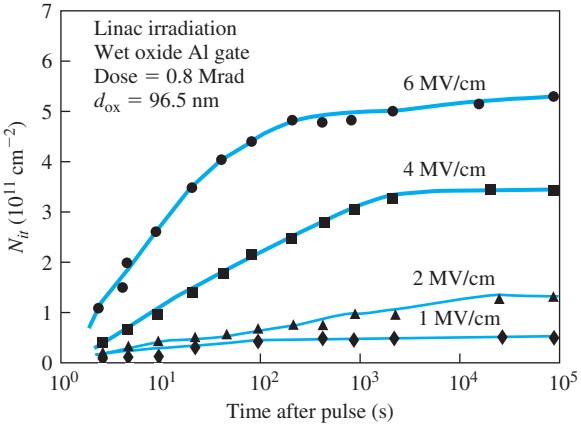


Figure 11.34 | Radiation-induced interface state density versus time after a pulse of ionizing radiation for several values of oxide electric field.
(From Ma and Dressendorfer [7].)

dose levels. The change in slope indicates that the density of interface states is increasing with total dose.

The buildup of radiation-induced interface states occurs over a relatively long time period and is a very strong function of the applied electric field in the oxide. Figure 11.34 shows the radiation-induced interface state density versus time for several applied fields. The final interface state density is reached between 100 to 10,000 seconds after a pulse of ionizing radiation. Almost all models for the generation of radiation-induced interface states depend on the transport or trapping of radiation-generated holes near the Si–SiO_2 interface. This transport and trapping process is time and field dependent, supporting the time and field dependence of the interface state buildup.

The sensitivity of the Si–SiO_2 interface to the buildup of radiation-induced interface states is a strong function of device processing. The interface state buildup in aluminum-gate MOSFETs tends to be smaller than in polysilicon-gate devices. This difference is probably more a result of variations between the two processing technologies than an inherent difference. Hydrogen appears to be important in the radiation-induced interface state buildup—hydrogen tends to passivate dangling silicon bonds at the interface, reducing the preradiation density of interface states. However, devices passivated with hydrogen appear to be more susceptible to the buildup of radiation-induced interface states. The silicon–hydrogen bond at the interface may be broken by the radiation process, which leaves a dangling silicon bond that acts like an interface state trap. These traps at the interface have been identified through electron spin resonance experiments.

Interface states may seriously affect the MOSFET characteristics, which in turn can affect MOSFET circuit performance. Radiation-induced interface states can cause shifts in threshold voltage, affecting circuit performance as we have discussed. A reduction in mobility can affect the speed and output drive capability of a circuit.

11.5.3 Hot-Electron Charging Effects

We have considered breakdown voltage effects in a MOSFET. In particular, as the electric field in the drain junction space charge region increases, electron–hole pairs can be generated by impact ionization. The generated electrons tend to be swept to the drain and generated holes swept into the substrate in an n-channel MOSFET.

Some of the electrons generated in the space charge region are attracted to the oxide due to the electric field induced by a positive gate voltage; this effect is shown in Figure 11.35. These generated electrons have energies far greater than the thermal-equilibrium value and are called hot electrons. If the electrons have energies on the order of 1.5 eV, they may be able to tunnel into the oxide; or in some cases they may be able to overcome the silicon oxide potential barrier and produce a gate current, which may be in the range of femtoamperes (fA) (10^{-15} A) or perhaps picoamperes (pA) (10^{-12} A). A fraction of the electrons traveling through the oxide may be trapped, producing a net negative charge density in the oxide. The probability of electron trapping is usually less than that of hole trapping; but a hot electron–induced gate current may exist over a long period of time, therefore the negative charging effect may build up. The negative oxide charge trapping will cause a local positive shift in the threshold voltage.

The energetic electrons, as they cross the Si–SiO_2 interface, can generate additional interface states. The probable cause of interface state generation is due to the breaking up of silicon-hydrogen bonds—a dangling silicon bond is produced, which acts as an interface state. The charge trapping in interface states causes a shift in threshold voltage, additional surface scattering, and reduced mobility. The hot-electron charging effects are continuous processes, so the device degrades over a period of time. This degradation is obviously an undesirable effect and may tend to limit the useful life of the device. We have discussed the lightly doped drain (LDD) structure in Section 11.4.2. The maximum electric field is reduced in this device, decreasing the probability of impact ionization and hot-electron effects.

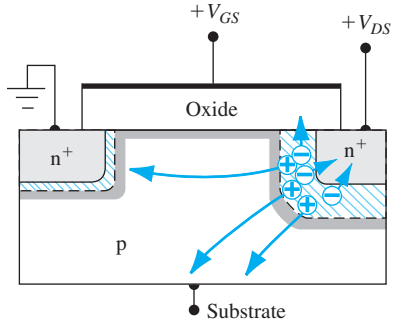


Figure 11.35 | Hot carrier generation, current components, and electron injection into the oxide.

11.6 | SUMMARY

- Advanced concepts of MOSFETs have been discussed in this chapter.
- Subthreshold conduction means that the drain current in a MOSFET is not zero even when the gate-to-source voltage is less than the threshold voltage. In this situation, the transistor is biased in the weak inversion mode and the drain current is dominated by the diffusion rather than the drift mechanism.
- The effective channel length decreases with an increase in drain voltage when the MOSFET is biased in the saturation region, since the depletion region at the drain extends into the channel. This effect is known as channel length modulation.
- The mobility of carriers in the inversion layer is not a constant. As the gate voltage increases, the transverse electric field at the oxide interface increases, causing additional surface scattering. The increased carrier scattering leads to a reduced mobility and a deviation from the ideal current–voltage relation.
- As the channel length decreases, the lateral electric field increases so that carriers flowing through the channel may reach their saturation velocity. In this case, the drain current becomes a linear function of gate-to-source voltage.
- The tendency in MOSFET design is to make devices smaller. The principle of constant-field scaling has been discussed.
- Modification in threshold voltage as device dimensions shrink has been discussed. Because of charge-sharing effects in the substrate, the threshold voltage decreases as channel length decreases and increases as channel width decreases.
- Several voltage breakdown mechanisms have been discussed. These include oxide breakdown, avalanche breakdown, near avalanche or snapback breakdown, and near punch-through effects. These mechanisms are all enhanced as device dimensions shrink. The lightly doped drain transistor tends to minimize the drain breakdown effects.
- Ion implantation can be used as essentially a final process step to change and adjust the substrate doping in the channel region to provide the desired threshold voltage. This process is referred to as threshold voltage adjustment by ion implantation and is used extensively in device fabrication.
- The effect of ionization radiation and hot-electron effects on MOSFET behavior have been briefly discussed.

GLOSSARY OF IMPORTANT TERMS

channel length modulation The change in effective channel length with drain-to-source voltage when the MOSFET is biased in saturation.

drain-induced barrier lowering The near punch-through condition in which the potential barrier between the source and channel region in an off transistor is lowered due to a large applied drain voltage.

hot electrons Electrons with energies far greater than the thermal-equilibrium value caused by acceleration in high electric fields.

lightly doped drain (LDD) A MOSFET with a lightly doped drain region adjacent to the channel to reduce voltage breakdown effects.

narrow-channel effects The shift in threshold voltage as the channel width narrows.

near punch-through The reduction in the potential barrier between source and substrate by the drain-to-substrate voltage, resulting in a rapid increase in drain current.

short-channel effects The shift in threshold voltage as the channel length becomes smaller.

snapback The negative resistance effect during breakdown in a MOSFET caused by the variable current gain in a parasitic bipolar transistor.

subthreshold conduction The process of current conduction in a MOSFET when the transistor is biased below the threshold inversion point.

surface scattering The process of electric field attraction and coulomb repulsion of carriers at the oxide–semiconductor interface as the carriers drift between source and drain.

threshold adjustment The process of altering the threshold voltage by changing the semiconductor doping concentration through ion implantation.

CHECKPOINT

After studying this chapter, the reader should have the ability to:

- Describe the concept and effects of subthreshold conduction.
- Discuss channel length modulation.
- Describe carrier mobility versus gate-to-source voltage and discuss the effects on the current–voltage characteristics of a MOSFET.
- Discuss the effect of velocity saturation on the current–voltage relationship of a MOSFET.
- Define what is meant by constant-field scaling in MOSFET device design, and discuss how device parameters change in constant-field scaling.
- Describe why the threshold voltage changes as the channel length decreases and as the channel width decreases.
- Describe the various voltage breakdown mechanisms in a MOSFET, such as oxide breakdown, avalanche breakdown, snapback breakdown, and near punch-through effects.
- Describe the advantages of the lightly doped drain transistor.
- Discuss the advantages and the process of threshold adjustment by ion implantation.

REVIEW QUESTIONS

1. What is subthreshold conduction? Sketch a drain current versus gate voltage plot that shows the subthreshold current for the transistor biased in the saturation region.
2. What is channel length modulation? Sketch an I–V curve that shows the channel length modulation effect.
3. Why, in general, is the mobility of carriers in the inversion layer not a constant with applied voltage?
4. What is velocity saturation and what is its effect on the I–V relation of a MOSFET?
5. What is constant-field scaling and what parameters in a MOSFET are changed in constant-field scaling?
6. Sketch the space charge region in the channel of a short-channel MOSFET and show the charge-sharing effect. Why does the threshold voltage decrease in a short-channel NMOS device?
7. Sketch the space charge region along the width of an NMOS device. Why does the threshold voltage increase as the channel width of the NMOS device decreases?
8. Sketch I_D versus V_D for an NMOS device, showing the snapback breakdown effect.
9. Sketch the energy bands of an NMOS device between source and drain, showing the near punch-through effect.

10. Sketch the cross section of a lightly doped drain transistor. What are the advantages and disadvantages of this design?

11. What type of ion should be implanted into a MOSFET to increase the threshold voltage? What type of ion should be implanted into a MOSFET to decrease the threshold voltage?

PROBLEMS

(Note: In the following problems, assume the semiconductor and oxide in the MOS system are silicon and silicon dioxide, respectively, and assume the temperature is $T = 300$ K unless otherwise stated.)

Section 11.1 Nonideal Effects

11.1 Assume that the subthreshold current of a MOSFET is given by

$$I_D = 10^{-15} \exp\left(\frac{V_{GS}}{(2.1)V_t}\right)$$

over the range $0 \leq V_{GS} \leq 1$ volt and where the factor 2.1 takes into account the effect of interface states. Assume that 10^6 identical transistors on a chip are all biased at the same V_{GS} and at $V_{DD} = 5$ V. (*a*) Calculate the total current that must be supplied to the chip at $V_{GS} = 0.5$, 0.7, and 0.9 V. (*b*) Calculate the total power dissipated in the chip for the same V_{GS} values.

11.2 The subthreshold current in a MOSFET is given by $I_D = I_S \exp(V_{GS}/nV_t)$. Determine the change in applied V_{GS} for a factor of 10 increase in I_D for (*a*) $n = 1$, (*b*) $n = 1.5$, and (*c*) $n = 2.1$.

11.3 A silicon n-channel MOSFET has a doping concentration of $N_a = 2 \times 10^{16}$ cm^{-3} and a threshold voltage of $V_T = 0.40$ V. Determine the change in channel length, ΔL, for (*a*) $V_{GS} = 1.0$ V, $V_{DS} = 2.0$ V; (*b*) $V_{GS} = 1.0$, $V_{DS} = 4.0$ V; (*c*) $V_{GS} = 2.0$ V, $V_{DS} = 2.0$ V; and (*d*) $V_{GS} = 2.0$, $V_{DS} = 4.0$.

11.4 Consider the MOSFET described in Problem 11.3. (*a*) Determine the minimum channel length so that the incremental change ΔL is no more than 10 percent of the original length L for applied voltages of $V_{GS} = 2$ V and $V_{DS} = 3$ V. (*b*) Repeat part (*a*) for $V_{DS} = 5$ V.

11.5 A silicon MOSFET has parameters $N_a = 4 \times 10^{16}$ cm^{-3}, $t_{ox} = 12$ nm $= 120$ Å, $Q'_{ss} = 4 \times 10^{10}$ cm^{-2}, and $\phi_{ms} = -0.5$ V. The transistor is biased at $V_{GS} = 1.25$ V and $V_{SB} = 0$. (*a*) Calculate ΔL for (*i*) $\Delta V_{DS} = 1$ V, (*ii*) $\Delta V_{DS} = 2$ V, and (*iii*) $\Delta V_{DS} = 4$ V. (*b*) Determine the minimum channel length L such that $\Delta L/L = 0.12$ for $V_{GS} = 1.25$ V and $\Delta V_{DS} = 4$ V.

11.6 The parameters of a silicon MOSFET are $N_a = 3 \times 10^{16}$ cm^{-3}, $V_T = 0.40$ V, $k'_n = 50$ μA/V^2, $L = 0.80$ μm, and $W = 15$ μm. (*a*) Determine the current I'_D at $V_{GS} = 1.0$ V for (*i*) $V_{DS} = 2.0$ V and (*ii*) $V_{DS} = 4.0$ V. (*b*) Defining the output resistance as $r_o = (\Delta I'_D/\Delta V_{DS})^{-1}$, determine r_o for part (*a*). (*c*) Repeat parts (*a*) and (*b*) for $V_{GS} = 2.0$ V.

11.7 Consider an n-channel silicon MOSFET. The parameters are $k'_n = 75$ μA/V^2, $W/L = 10$, and $V_T = 0.35$ V. The applied drain-to-source voltage is $V_{DS} = 1.5$ V. (*a*) For $V_{GS} = 0.8$ V, find (*i*) the ideal drain current, (*ii*) the drain current if $\lambda = 0.02$ V^{-1}, and (*iii*) the output resistance for $\lambda = 0.02$ V^{-1}. (*b*) Repeat part (*a*) for $V_{GS} = 1.25$ V.

11.8 An n-channel MOSFET has a substrate doping concentration of $N_a = 10^{16}$ cm^{-3} and $V_{DS}(\text{sat}) = 2$ V. Using Equation (11.10), plot ΔL versus V_{DS} over the range $2 \leq V_{DS} \leq 5$ V for (*a*) $E_{sat} = 10^4$ V/cm and $E_{sat} = 2 \times 10^5$ V/cm.

11.9 Assume that the lateral electric field at the point where the inversion charge pinches off is given by $E_{sat} = V_{DS}(\text{sat})/L$. (*a*) Determine E_{sat} for $L = 3$, 1.0, 0.50, 0.25, and 0.13 μm. (*b*) For each case given in part (*a*), estimate the carrier drift velocity.

11.10 A silicon n-channel MOSFET has the following parameters: $V_T = 0.45$ V, $\mu_n = 425$ cm^2/V-s, $t_{ox} = 11$ nm $= 110$ Å, $W = 20$ μm, and $L = 1.2$ μm. The substrate doping is $N_a = 3 \times 10^{16}$ cm^{-3}. (*a*) Using Equations (11.4) and (11.11), calculate the output resistance, $r_o = (\partial I_D'/\partial V_{DS})^{-1}$, for $V_{GS} = 0.8$ V and $\Delta V_{DS} = 2$ V. (*b*) Repeat part (*a*) if the channel length is reduced to $L = 0.80$ μm.

11.11 (*a*) Consider an n-channel enhancement mode MOSFET with $(W/L) = 10$, $C_{ox} = 6.9 \times 10^{-8}$ F/cm^2, and $V_T = +1$ V. Assume a constant mobility of $\mu_n = 500$ cm^2/V-s. Plot $\sqrt{I_D}$ versus V_{GS} for $0 \leq V_{GS} \leq 5$ V when the transistor is biased in the saturation region. (*b*) Now assume that the effective mobility in the channel is given by

$$\mu_{\text{eff}} = \mu_0 \left(\frac{E_{\text{eff}}}{E_c}\right)^{-1/3}$$

where $\mu_0 = 1000$ cm^2/V-s and $E_c = 2.5 \times 10^4$ V/cm. As a first approximation, let $E_{\text{eff}} = V_{GS}/t_{ox}$. Using μ_{eff} in place of μ_n in the $\sqrt{I_D}$ versus V_{GS} relation, plot $\sqrt{I_D}$ versus V_{GS} over the same V_{GS} range as in part (*a*). (*c*) Plot the curves from parts (*a*) and (*b*) on the same graph. What can be said about the slopes of the two curves?

11.12 One model used to describe the variation in electron mobility in an NMOS device is

$$\mu_{\text{eff}} = \frac{\mu_0}{1 + \theta(V_{GS} - V_{TN})}$$

where θ is called the mobility degradation parameter. Assume the following parameters: $C_{ox} = 10^{-8}$ F/cm^2, $(W/L) = 25$, $\mu_0 = 800$ cm^2/V-s, and $V_{TN} = 0.5$ V. Plot, on the same graph, $\sqrt{I_D}$ versus V_{GS} for the NMOS device biased in the saturation region over the range $0 \leq V_{GS} \leq 3$ V for (*a*) $\theta = 0$ (ideal case) and (*b*) $\theta = 0.5$ V^{-1}.

11.13 The parameters of an n-channel enhancement-mode MOSFET are $V_T = 0.40$ V, $t_{ox} = 20$ nm $= 200$ Å, $L = 1.0$ μm, and $W = 10$ μm. (*a*) Assuming a constant mobility of $\mu_n = 475$ cm^2/V-s, calculate I_D for $V_{GS} - V_T = 2.0$ V when biased at (*i*) $V_{DS} = 0.5$ V, (*ii*) $V_{DS} = 1.0$ V, (*iii*) $V_{DS} = 1.25$ V, and (*iv*) $V_{DS} = 2.0$ V. (*b*) Consider the piecewise linear model of the carrier velocity versus V_{DS} shown in Figure P11.13. Calculate I_D for the same voltage values given in part (*a*). [See Equation (11.17).] (*c*) Determine the $V_{DS}(\text{sat})$ values for parts (*a*) and (*b*).

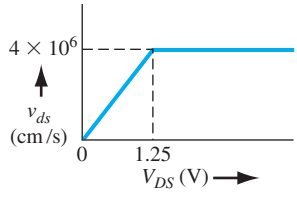


Figure P11.13 | Figure for Problems 11.13 and 11.14.

11.14 Consider an NMOS transistor with a threshold voltage of $V_{TN} = 0.4$ V. Plot, on the same graph, $V_{DS}(\text{sat})$ over the range $0 \leq V_{GS} \leq 3$ V for (*a*) an ideal MOSFET (constant mobility) and (*b*) a device whose drift velocity is given in Figure P11.13.

Section 11.2 MOSFET Scaling

11.15 Apply constant-field scaling to the ideal current–voltage relations in both the saturation and nonsaturation bias regions. (*a*) How does the drain current scale in each bias region? (*b*) How does the power dissipation per device scale in each bias region?

11.16 Consider a MOSFET biased such that carriers are traveling at their saturated velocity in the n channel. If constant-field scaling is applied to the device, how does the drain current scale?

11.17 The initial parameters of an n-channel MOSFET are $k_n' = 0.15$ mA/V^2, $L = 1.2$ μm, $W = 6.0$ μm, and $V_T = 0.45$ V. The device operates over a voltage range of 0 to 3 V. Assume a constant-field scaling factor of $k = 0.65$, but assume the threshold voltage is constant. (*a*) Determine the maximum drain current in the (*i*) original device and (*ii*) scaled device. (*b*) Determine the maximum power dissipation in the (*i*) original device and (*ii*) scaled device.

Section 11.3 Threshold Voltage Modifications

11.18 Consider an n-channel MOSFET with parameters $N_a = 5 \times 10^{16}$ cm^{-3}, $t_{ox} = 12$ nm $=$ 120 Å, and $L = 0.80$ μm. If $r_j = 0.25$ μm, determine the threshold voltage shift due to short-channel effects.

11.19 The parameters of an n-channel MOSFET are $N_a = 2 \times 10^{16}$ cm^{-3}, $L = 0.70$ μm, and $t_{ox} = 8$ nm $= 80$ Å. The diffused junction radius is $r_j = 0.30$ μm. The designed threshold voltage is to be $V_T = 0.35$ V taking into account the shift due to short-channel effects. What is the equivalent long-channel threshold voltage?

11.20 An n-channel MOSFET is doped to $N_a = 3 \times 10^{16}$ cm^{-3} and has an oxide thickness of $t_{ox} = 20$ nm $= 200$ Å. The diffused junction radius is $r_j = 0.30$ μm. Determine the minimum channel length such that the threshold voltage shift due to short-channel effects is limited to $\Delta V_T = -0.15$ V.

***11.21** The shift in threshold voltage due to short-channel effects given by Equation (11.29) assumed all space charge regions were of equal width. If a drain voltage is applied, this condition is no longer valid. Using the same trapezoidal approximation, show that the threshold voltage shift is given by

$$\Delta V_T = -\frac{eN_a x_{dT}}{C_{ox}} \cdot \frac{r_j}{2L}\left[\left(\sqrt{1 + \frac{2x_{ds}}{r_j}} + \alpha^2 - 1\right) + \left(\sqrt{1 + \frac{2x_{dD}}{r_j}} + \beta^2 - 1\right)\right]$$

where

$$\alpha^2 = \frac{x_{ds}^2 - x_{dT}^2}{r_j^2} \qquad \beta^2 = \frac{x_{dD}^2 - x_{dT}^2}{r_j^2}$$

and where x_{ds} and x_{dD} are the source and drain space charge widths, respectively.

***11.22** The threshold voltage shift due to short channel effects, given by Equation (11.29), has been derived assuming L is large enough so that a trapezoidal charge region

*Asterisks next to problems indicate problems that are more difficult.

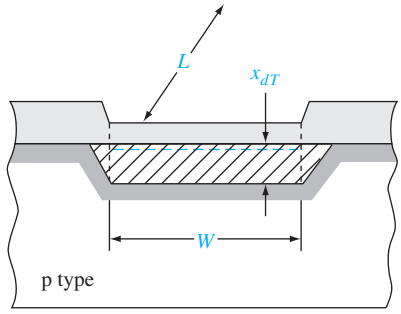


Figure P11.27 | Figure for Problem 11.27.

can be defined as shown in Figure 11.15. Derive the expression for ΔV_T for the case when L becomes small enough such that the trapezoid becomes a triangle. Assume punch-through does not occur.

11.23 Consider the short-channel effect. Plot $V_T - V_{FB}$ versus L as shown in Figure 11.16 over the range $0.5 \leq L \leq 6$ μm. Use the parameters given in the figure and assume $V_{SB} = 0$.

11.24 Repeat Problem 11.23 at $N_a = 10^{16}$ and 10^{17} cm^{-3} for $V_{SB} = 0$, 2, 4, and 6 V.

11.25 Equation (11.29) describes the shift in threshold voltage due to short channel effects. If constant-field scaling is applied, what is the scaling factor in ΔV_T?

11.26 An n-channel MOSFET has parameters of $N_a = 3 \times 10^{16}$ cm^{-3}, $W = 2.2$ μm, and $t_{ox} = 8$ nm $= 80$ Å. Neglecting short-channel effects, calculate the shift in threshold voltage due to narrow-channel effects. Assume the fitting parameter is $\xi = \pi/2$.

11.27 Consider an n-channel MOSFET with $N_a = 10^{16}$ cm^{-3} and $t_{ox} = 12$ nm $= 120$ Å. The depletion region at each end of the channel can be approximated by a triangular region, as shown in Figure P11.27. Assume both the lateral and vertical depletion widths are equal to x_{dT}. The threshold voltage shift due to narrow-channel effects is to be limited to $\Delta V_T = +0.045$ V. Determine the minimum channel width W. Neglect short-channel effects.

11.28 Consider the narrow-channel effect. Use the transistor parameters described in Example 11.4. Plot $V_T - V_{FB}$ over the range $0.5 \leq W \leq 5$ μm for a long-channel device.

11.29 Equation (11.35) describes the shift in threshold voltage due to narrow-channel effects. If constant-field scaling is applied, what is the scaling factor in ΔV_T?

Section 11.4 Additional Electrical Characteristics

11.30 (*a*) A MOS device has a silicon dioxide insulator with a thickness of $t_{ox} = 20$ nm $= 200$ Å. (*i*) Determine the ideal breakdown voltage. (*ii*) If a safety factor of 3 is required, determine the maximum safe gate voltage that may be applied. (*b*) Repeat part (*a*) for an oxide thickness of $t_{ox} = 8$ nm $= 80$ Å.

11.31 (*a*) In a power MOSFET, a maximum gate voltage of 8 V is to be applied. Determine the minimum oxide thickness necessary if a safety factor of 3 is specified. (*b*) Repeat part (*a*) if the maximum gate voltage is 12 V.

***11.32** The snapback breakdown condition is defined to be when $\alpha M = 1$, where α is the common base current gain and M is the multiplication constant given by Equation (11.40). Let $m = 3$ and let $V_{BD} = 15$ V. The common base current gain is a very strong function of junction current I_D. Assume α is described by the relation

$$\alpha = (0.18)\log_{10}\left(\frac{I_D}{3 \times 10^{-9}}\right)$$

where I_D is given in amperes. Plot the curve of I_D versus V_{CE} which satisfies the snapback condition over the range $10^{-8} \leq I_D \leq 10^{-3}$ A. (Use a log scale for the current.)

11.33 Near punch-through occurs when the two depletion regions are within approximately six Debye lengths of each other. The extrinsic Debye length L_D is defined as

$$L_D = \left[\frac{\epsilon_s(kT/e)}{eN_a}\right]^{1/2}$$

Consider the n-channel MOSFET in Example 11.5. Calculate the near punch-through voltage. How does this voltage compare to the ideal punch-through voltage determined in the example?

11.34 The near punch-through voltage (see Problem 11.33) of an n-channel MOSFET is to be no less than $V_{DS} = 5$ V. The source and drain regions are doped $N_d = 10^{19}$ cm^{-3}, and the channel region is doped $N_a = 3 \times 10^{16}$ cm^{-3}. The source and body are at ground potential. Determine the minimum channel length.

11.35 Repeat Problem 11.34 if a source–substrate voltage $V_{SB} = 2$ V is applied.

11.36 The threshold voltage of an n-channel MOSFET, with an oxide thickness of $t_{ox} = 12$ nm $= 120$ Å, needs to be shifted in the positive direction by 0.80 V. Determine the type of ion implant and the implant dose required.

11.37 The threshold voltage of a p-channel MOSFET, with an oxide thickness of $t_{ox} = 18$ nm $= 180$ Å, needs to be shifted in the negative direction by 0.60 V. Determine the type of ion implant and the implant dose required.

11.38 An n-channel MOSFET has an n^+ polysilicon gate, a substrate doping of $N_a = 6 \times 10^{15}$ cm^{-3}, an oxide thickness of $t_{ox} = 15$ nm $= 150$ Å, and an oxide trapped charge density of $Q'_{ss} = 5 \times 10^{10}$ cm^{-2}. (*a*) Determine the threshold voltage. (*b*) The required threshold voltage is $V_T = +0.50$ V. Determine the type and ion implant density necessary to achieve this specification. Assume the implant is directly adjacent to the oxide-semiconductor interface.

11.39 A p-channel MOSFET has a p^+ polysilicon gate, a substrate doping of $N_d = 2 \times 10^{16}$ cm^{-3}, an oxide thickness of $t_{ox} = 18$ nm $= 180$ Å, and an oxide trapped charge density of $Q'_{ss} = 10^{11}$ cm^{-2}. (*a*) Calculate the threshold voltage. (*b*) The required threshold voltage is $V_T = -0.40$ V. Determine the type and ion implant density necessary to achieve this specification. Assume the implant is directly adjacent to the oxide-semiconductor interface.

11.40 Consider an n-channel MOSFET with a substrate doping of $N_a = 4 \times 10^{15}$ cm^{-3}, an oxide thickness of $t_{ox} = 8$ nm $= 80$ Å, and an initial flat-band voltage of $V_{FB} = -1.25$ V. (*a*) Determine the threshold voltage. (*b*) For an enhancement-mode device, the required threshold voltage is $V_T = +0.40$ V. Determine the type and ion implant density that is necessary to achieve this specification. (*c*) Repeat part (*b*) for a depletion-mode device with a required threshold voltage of $V_T = -0.40$ V.

11.41 The channel of a device with $t_{ox} = 500$ Å and a p-type substrate with $N_a = 10^{14}$ cm^{-3} is implanted with acceptors using an effective dose of $D_I = 10^{12}$ cm^{-2}. The

implant is approximated as a step function with $x_I = 0.2\ \mu$m. Calculate the shift in threshold voltage due to back bias effects for $V_{SB} = 1.3$, and 5 volts.

11.42 A MOSFET has the following parameters: n^+ poly gate, $t_{ox} = 80$ Å, $N_d = 10^{17}$ cm^{-3}, and $Q'_{ss} = 5 \times 10^{10}$ cm^{-2}. (*a*) What is the threshold voltage of this MOSFET? Is the device enhancement- or depletion-mode? (*b*) What type of implant and dose are required such that $V_T = 0$?

Section 11.5 Radiation and Hot-Electron Effects

11.43 One rad (Si) produces on the average 8×10^{12} electron–hole pairs/cm^3 in silicon dioxide.[1] Assume that a pulse of ionizing radiation with a total dose of 10^5 rads (Si) is incident on an MOS device with a 750 Å oxide. Assume that there is no electron-hole recombination and that the electrons are swept out through the gate terminal. If 10 percent of the generated holes are trapped at the oxide–semiconductor interface, calculate the threshold voltage shift.

11.44 Reconsider Problem 11.43. If the threshold voltage shift is to be no more than $\Delta V_T = -0.50$ V, calculate the maximum percentage of holes that can be trapped.

11.45 Show that, for the simple model for radiation-induced hole trapping we have considered, the threshold voltage shift is proportional to $\Delta V_T \propto -t_{ox}^2$. Thin oxides are one requirement for radiation-tolerant MOS devices.

Summary and Review

***11.46** Design a silicon n-channel MOSFET with a polysilicon gate to have a threshold voltage of $V_T = +0.30$ V. The oxide thickness is to be $t_{ox} = 12$ nm $= 120$ Å and the channel length is to be $L = 0.80\ \mu$m. Assume $Q'_{ss} = 0$. It is desired to have a drain current of $I_D = 80\ \mu$A at $V_{GS} = 1.25$ V and $V_{DS} = 0.25$ V. Determine the substrate doping concentration, channel width, and type of gate required. Specify any possible ion implantation process and take short channel effects into account.

***11.47** A particular process produces an n-channel MOSFET with the following properties:

$t_{ox} = 325$ Å	$L = 0.8\ \mu$m
$N_a = 10^{16}$ cm^3	$W = 20\ \mu$m
n^+ polysilicon gate	$r_j = 0.35\ \mu$m
$Q'_{ss} = 10^{11}$ cm^{-2}	

The desired threshold voltage is $V_T = 0.35$ V at $T = 300$ K. Design an additional process to achieve this objective by using ion implantation, which produces a step function profile that is 0.35 μm deep.

***11.48** A CMOS inverter is to be designed in which both the n-channel and p-channel devices have the same magnitude of doping concentration equal to 10^{16} cm^{-3}, equal oxide thickness of $t_{ox} = 150$ Å, equal oxide trapped charge of $Q'_{ss} = +8 \times 10^{10}$ cm^{-2}. The gate of the n channel is p^+ poly and the gate of the p channel is n^+ poly. Determine the type of ion implant and the implant dose in each device such that the final threshold voltages are $V_{TN} = +0.5$ V and $V_{TP} = -0.5$ V.

---

[1]One rad (Si) is equivalent to 100 ergs of energy deposited per cm^3 in silicon. We normally use this same total dose notation for the total dose effects in silicon dioxide.

READING LIST

1. Akers, L. A., and J. J. Sanchez. "Threshold Voltage Models of Short, Narrow, and Small Geometry MOSFETs: A Review." *Solid State Electronics* 25 (July 1982), pp. 621–641.
2. Baliga, B. J. *Fundamentals of Power Semiconductor Devices.* Springer, Berlin, Germany 2008.
3. Brews, J. R. "Threshold Shifts Due to Nonuniform Doping Profiles in Surface Channel MOSFETs." *IEEE Transactions on Electron Devices* ED-26 (November 1979), pp. 1696–1710.
4. Dimitrijev, S. *Principles of Semiconductor Devices*. New York: Oxford University Press, 2006.
5. Kano, K. *Semiconductor Devices*. Upper Saddle River, NJ: Prentice Hall, 1998.
6. Klaassen, F. M., and W. Hes. "On the Temperature Coefficient of the MOSFET Threshold Voltage." *Solid State Electronics* 29 (August 1986), pp. 787–789.
7. Ma, T. P., and P. V. Dressendorfer. *Ionizing Radiation Effects in MOS Devices and Circuits*. New York: John Wiley and Sons, 1989.
8. Muller, R. S., and T. I. Kamins. *Device Electronics for Integrated Circuits*. 2nd ed. New York: John Wiley and Sons, 1986.
9. Neamen, D. A., B. Buchanan, and W. Shedd. "Ionizing Radiation Effects in SOS Structures." *IEEE Transactions on Nuclear Science* NS–22 (December 1975), pp. 2179–2202.

*10. Nicollian, E. H., and J. R. Brews. *MOS Physics and Technology*. New York: John Wiley and Sons, 1982.

11. Ning, T. H., P. W. Cook, R. H. Dennard, C. M. Osburn, S. E. Schuster, and H. N. Yu. "1 μm MOSFET VLSI Technology: Part IV—Hot Electron Design Constraints." *IEEE Transactions on Electron Devices* ED-26 (April 1979), pp. 346–353.
12. Ogura, S., P. J. Tsang, W. W. Walker, D. L. Critchlow, and J. F. Shepard. "Design and Characteristics of the Lightly Doped Drain-Source (LDD) Insulated Gate Field-Effect Transistor." *IEEE Transactions on Electron Devices* ED-27 (August 1980), pp. 1359–1367.
13. Ong, D. G. *Modern MOS Technology: Processes*, *Devices*, *and Design*. New York: McGraw-Hill, 1984.
14. Pierret, R. F. *Semiconductor Device Fundamentals*. Reading, MA: Addison-Wesley, 1996.
15. Roulston, D. J. *An Introduction to the Physics of Semiconductor Devices*. New York: Oxford University Press, 1999.
16. Sanchez, J. J., K. K. Hsueh, and T. A. DeMassa. "Drain-Engineered Hot-Electron-Resistant Device Structures: A Review." *IEEE Transactions on Electron Devices* ED-36 (June 1989), pp. 1125–1132.
17. Schroder, D. K. *Advanced MOS Devices*, *Modular Series on Solid State Devices*. Reading, MA: Addison-Wesley, 1987.
18. Shur, M. *Introduction to Electronic Devices*. New York: John Wiley and Sons, 1996.

*19. _____. *Physics of Semiconductor Devices*. Englewood Cliffs, NJ: Prentice Hall, 1990.

20. Singh, J. *Semiconductor Devices: Basic Principles*. New York: John Wiley and Sons, 2001.

21. Streetman, B. G., and S. Banerjee. *Solid State Electronic Devices*. 6th ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2006.

22. Sze, S. M., and K. K. Ng. Physics of Semiconductor Devices, 3rd ed. Hoboken, NJ: John Wiley and Sons, 2007.

23. Taur, Y. and T. H. Ning. *Fundamentals of Modern VLSI Devices*, 2nd ed. New York Cambridge University Press, 2009.

***24.** Tsividis, Y. *Operation and Modeling of the MOS Transistor*, 2nd ed. Burr Ridge, IL: McGraw-Hill, 1999.

25. Yang, E. S. *Microelectronic Devices*. New York: McGraw-Hill, 1988.

26. Yau, L. D. "A Simple Theory to Predict the Threshold Voltage of Short-Channel IGFETs." *Solid-State Electronics* 17 (October 1974), pp. 1059–1063.

*Indicates references that are at an advanced level compared to this text.

CHAPTER 12

The Bipolar Transistor

The transistor is a multijunction semiconductor device that, in conjunction with other circuit elements, is capable of current gain, voltage gain, and signal power gain. The transistor is therefore referred to as an active device, whereas the diode is passive. The basic transistor action is the control of current at one terminal by the voltage applied across the other two terminals of the device.

The **B**ipolar **J**unction **T**ransistor (BJT) is one of two major types of transistors. The fundamental physics of the BJT is developed in this chapter. The bipolar transistor is used extensively in analog electronic circuits because of its high current gain.

Two complementary configurations of BJTs, the npn and pnp devices, can be fabricated. Electronic circuit design becomes very versatile when the two types of devices are used in the same circuit. ■

12.0 | PREVIEW

In this chapter, we will:

- Discuss the physical structure of the bipolar transistor, which has three separately doped regions and two pn junctions that are sufficiently close together so interactions occur between the two junctions.
- Discuss the basic principle of operation of the bipolar transistor, including the various possible modes of operation.
- Derive expressions for the minority carrier concentrations through the device for various operating modes.
- Derive expressions for the various current components in the bipolar transistor.
- Define common-base and common-emitter current gains.
- Define the limiting factors and derive expressions for the current gain.
- Discuss several nonideal effects in bipolar transistors, including base width modulation and high-level injection effects.

- Develop the small-signal equivalent circuit of the bipolar transistor. This circuit is used to relate small-signal currents and voltages in analog circuits.
- Define and derive expressions for the frequency limiting factors.
- Present the geometries and characteristics of a few specialized bipolar transistor designs.

12.1 | THE BIPOLAR TRANSISTOR ACTION

The bipolar transistor has three separately doped regions and two pn junctions. Figure 12.1 shows the basic structure of an npn bipolar transistor and a pnp bipolar transistor, along with the circuit symbols. The three terminal connections are called the emitter, base, and collector. The width of the base region is small compared to the minority carrier diffusion length. The (++) and (+) notation indicates the relative magnitudes of the impurity doping concentrations normally used in the bipolar transistor, with (++) meaning very heavily doped and (+) meaning moderately doped. The emitter region has the largest doping concentration; the collector region has the smallest. The reasons for using these relative impurity concentrations, and for the narrow base width, will become clear as we develop the theory of the bipolar transistor. The concepts developed for the pn junction apply directly to the bipolar transistor.

The block diagrams of Figure 12.1 show the basic structure of the transistor, but in very simplified sketches. Figure 12.2a shows a cross section of a classic npn bipolar transistor fabricated in an integrated circuit configuration, and Figure 12.2b shows the cross section of an npn bipolar transistor fabricated by a more modern technology. One can immediately observe that the actual structure of the bipolar transistor is not nearly as simple as the block diagrams of Figure 12.1 might suggest. A reason for the complexity is that terminal connections are made at the surface; in order to minimize semiconductor resistances, heavily doped n^+ buried layers must be included. Another reason for complexity arises out of the desire to fabricate more than one bipolar transistor on a single piece of semiconductor material. Individual transistors must be isolated from each other since all collectors, for example, will not be at the same potential. This isolation is accomplished by adding p^+ regions so that devices are separated by reverse-biased pn junctions as shown in Figure 12.2a, or they are isolated by large oxide regions as shown in Figure 12.2b.

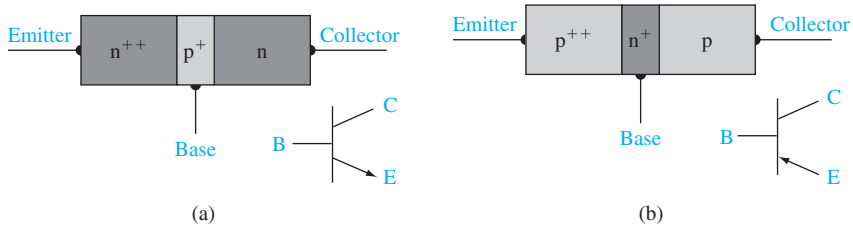


Figure 12.1 | Simplified block diagrams and circuit symbols of (a) npn and (b) pnp bipolar transistors.

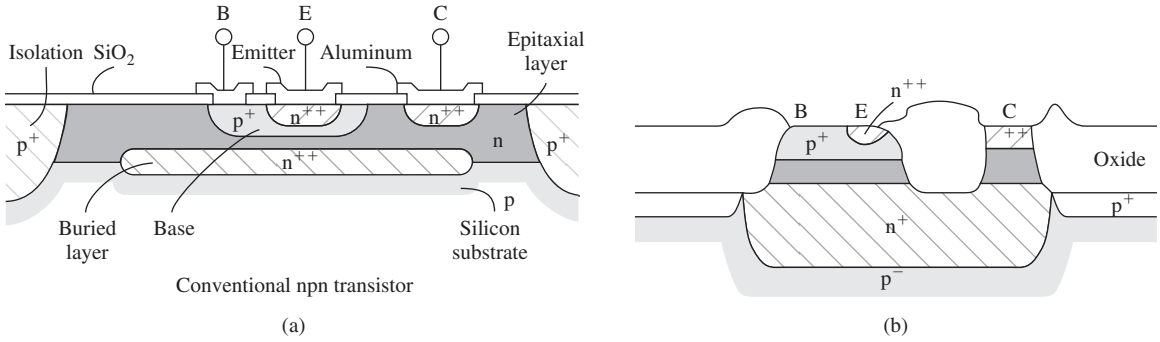


Figure 12.2 | Cross section of (a) a conventional integrated circuit npn bipolar transistor and (b) an oxide-isolated npn bipolar transistor.
(From Muller and Kamins [4].)

An important point to note from the devices shown in Figure 12.2 is that the bipolar transistor is not a symmetrical device. Although the transistor may contain two n regions or two p regions, the impurity doping concentrations in the emitter and collector are different and the geometry of these regions can be vastly different. The block diagrams of Figure 12.1 are highly simplified, but useful, concepts in the development of the basic transistor theory.

12.1.1 The Basic Principle of Operation

The npn and pnp transistors are complementary devices. We develop the bipolar transistor theory using the npn transistor, but the same basic principles and equations also apply to the pnp device. Figure 12.3 shows an idealized impurity doping profile in an npn bipolar transistor for the case when each region is uniformly doped. Typical impurity doping concentrations in the emitter, base, and collector may be on the order of 10^{19}, 10^{17}, and 10^{15} cm^{-3}, respectively.

The base–emitter (B–E) pn junction is forward biased and the base–collector (B–C) pn junction is reverse biased in the normal bias configuration as shown in

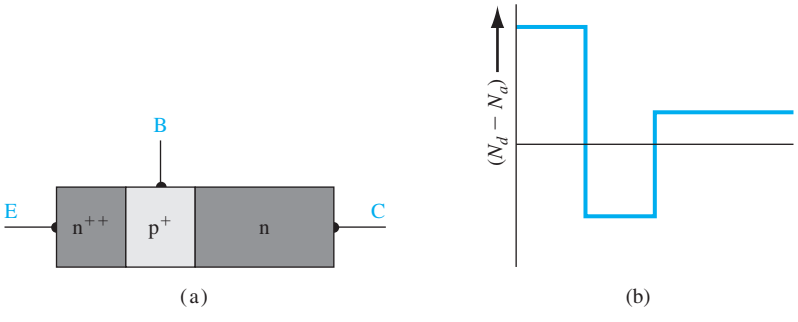


Figure 12.3 | Idealized doping profile of a uniformly doped npn bipolar transistor.

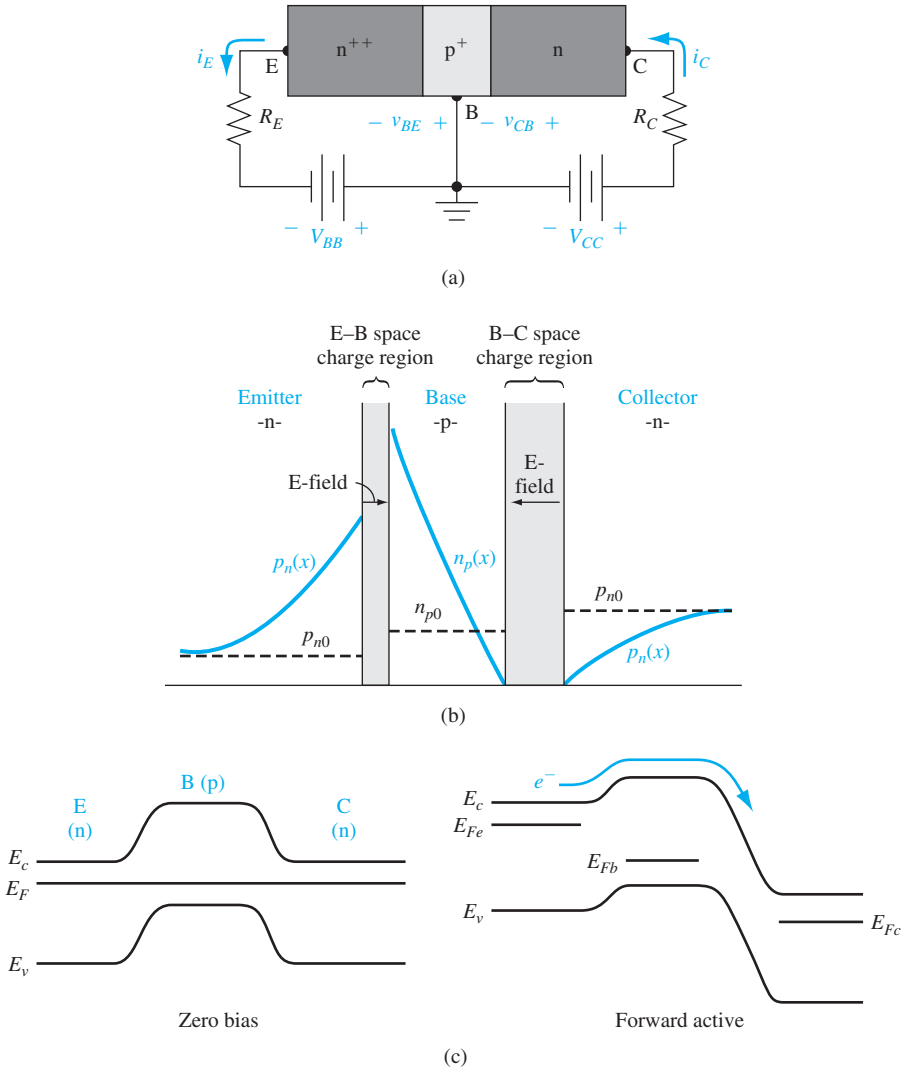


Figure 12.4 | (a) Biasing of an npn bipolar transistor in the forward-active mode, (b) minority carrier distribution in an npn bipolar transistor operating in the forward-active mode, and (c) energy-band diagram of the npn bipolar transistor under zero bias and under a forward-active mode bias.

Figure 12.4a. This configuration is called the *forward-active* operating mode: The B–E junction is forward biased so electrons from the emitter are injected across the B–E junction into the base. These injected electrons create an excess concentration of minority carriers in the base. The B–C junction is reverse biased, so the minority carrier electron concentration at the edge of the B–C junction is ideally zero. We expect the electron concentration in the base to be like that shown in

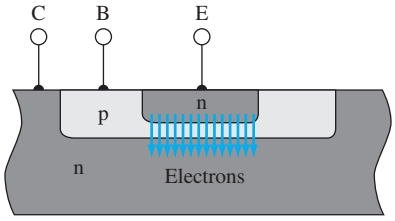


Figure 12.5 | Cross section of an npn bipolar transistor showing the injection and collection of electrons in the forward-active mode.

Figure 12.4b. The large gradient in the electron concentration means that electrons injected from the emitter will diffuse across the base region into the B–C space charge region, where the electric field will sweep the electrons into the collector. We want as many electrons as possible to reach the collector without recombining with any majority carrier holes in the base. For this reason, the width of the base needs to be small compared with the minority carrier diffusion length. If the base width is small, then the minority carrier electron concentration is a function of both the B–E and B–C junction voltages. The two junctions are close enough to be called *interacting* pn junctions.

Figure 12.5 shows a cross section of an npn transistor with the injection of electrons from the n-type emitter (hence the name emitter) and the collection of the electrons in the collector (hence the name collector).

12.1.2 Simplified Transistor Current Relation—Qualitative Discussion

We can gain a basic understanding of the operation of the transistor and the relations between the various currents and voltages by considering a simplified analysis. After this discussion, we delve into a more detailed analysis of the physics of the bipolar transistor.

The minority carrier concentrations are again shown in Figure 12.6 for an npn bipolar transistor biased in the forward-active mode. Ideally, the minority carrier electron concentration in the base is a linear function of distance, which implies no recombination. The electrons diffuse across the base and are swept into the collector by the electric field in the B–C space charge region.

Collector Current Assuming the ideal linear electron distribution in the base, the collector current can be written as a diffusion current given by

$$i_C = eD_n A_{BE}\frac{dn(x)}{dx} = eD_n A_{BE}\left[\frac{n_B(0) - 0}{0 - x_B}\right] = \frac{-eD_n A_{BE}}{x_B} \cdot n_{B0}\exp\left(\frac{v_{BE}}{V_t}\right) \quad (12.1)$$

where A_{BE} is the cross-sectional area of the B–E junction, n_{B0} is the thermal-equilibrium electron concentration in the base, and V_t is the thermal voltage. The

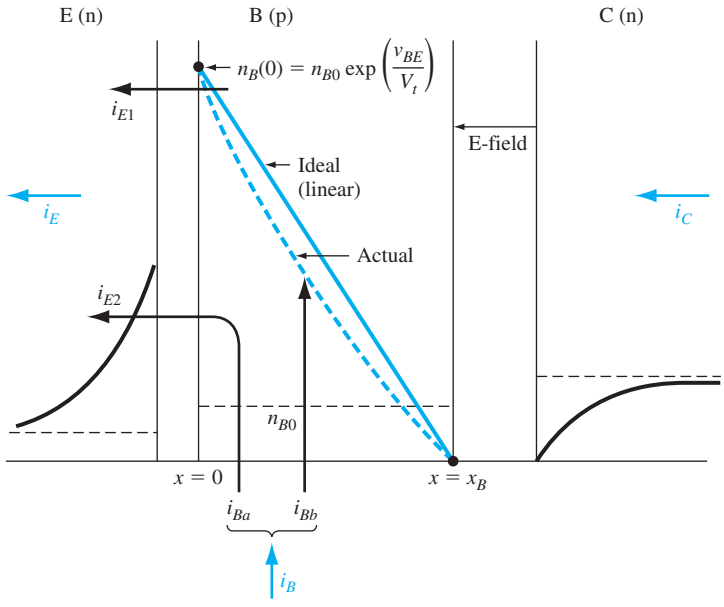


Figure 12.6 | Minority carrier distributions and basic currents in a forward-biased npn bipolar transistor.

diffusion of electrons is in the $+x$ direction so that the conventional current is in the $-x$ direction. Considering magnitudes only, Equation (12.1) can be written as

$$i_C = I_S \exp\left(\frac{v_{BE}}{V_t}\right) \tag{12.2}$$

The collector current is controlled by the base–emitter voltage; that is, the current at one terminal of the device is controlled by the voltage applied to the other two terminals of the device. As we have mentioned, this is the basic transistor action.

Emitter Current One component of emitter current, i_{E1}, shown in Figure 12.6 is due to the flow of electrons injected from the emitter into the base. This current, then, is equal to the collector current given by Equation (12.1).

Since the base–emitter junction is forward biased, majority carrier holes in the base are injected across the B–E junction into the emitter. These injected holes produce a pn junction current i_{E2} as indicated in Figure 12.6. This current is only a B–E junction current so this component of emitter current is not part of the collector current. Since i_{E2} is a forward-biased pn junction current, we can write (considering magnitude only)

$$i_{E2} = I_{S2} \exp\left(\frac{v_{BE}}{V_t}\right) \tag{12.3}$$

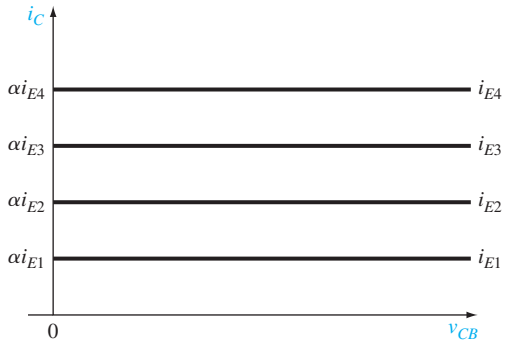


Figure 12.7 | Ideal bipolar transistor common-base current–voltage characteristics.

where I_{S2} involves the minority carrier hole parameters in the emitter. The total emitter current is the sum of the two components, or

$$i_E = i_{E1} + i_{E2} = i_C + i_{E2} = I_{SE}\exp\left(\frac{v_{BE}}{V_t}\right) \tag{12.4}$$

Since all current components in Equation (12.4) are functions of $\exp(v_{BE}/V_t)$, the ratio of collector current to emitter current is a constant. We can write

$$\frac{i_C}{i_E} \equiv \alpha \tag{12.5}$$

where α is called the *common-base current gain*. By considering Equation (12.4), we see that $i_C < i_E$ or $\alpha < 1$. Since i_{E2} is not part of the basic transistor action, we would like this component of current to be as small as possible. We would then like the common-base current gain to be as close to unity as possible.

Referring to Figure 12.4a and Equation (12.4), note that the emitter current is an exponential function of the base–emitter voltage and the collector current is $i_C = \alpha i_E$. To a first approximation, the collector current is independent of the base–collector voltage as long as the B–C junction is reverse biased. We can sketch the common-base transistor characteristics as shown in Figure 12.7. The bipolar transistor acts like a constant current source.

Base Current As shown in Figure 12.6, the component of emitter current i_{E2} is a B–E junction current so that this current is also a component of base current shown as i_{Ba}. This component of base current is proportional to $\exp(v_{BE}/V_t)$.

There is also a second component of base current. We have considered the ideal case in which there is no recombination of minority carrier electrons with majority carrier holes in the base. However, in reality, there will be some recombination. Since majority carrier holes in the base are disappearing, they must be resupplied by a flow of positive charge into the base terminal. This flow of charge is indicated as a current i_{Bb} in Figure 12.6. The number of holes per unit time recombining in the base is directly related to the number of minority carrier electrons in the base

[see Equation (6.13)]. Therefore, the current i_{Bb} is also proportional to $\exp(v_{BE}/V_t)$. The total base current is the sum of i_{Ba} and i_{Bb} and is proportional to $\exp(v_{BE}/V_t)$.

The ratio of collector current to base current is a constant since both currents are directly proportional to $\exp(v_{BE}/V_t)$. We can then write

$$\frac{i_C}{i_B} \equiv \beta \tag{12.6}$$

where β is called the *common-emitter current gain.* Normally, the base current will be relatively small so that, in general, the common-emitter current gain is much larger than unity (on the order of 100 or larger).

12.1.3 The Modes of Operation

Figure 12.8 shows the npn transistor in a simple circuit. In this configuration, the transistor may be biased in one of three modes of operation. If the B–E voltage is zero or reverse biased ($V_{BE} \leq 0$), then majority carrier electrons from the emitter will not be injected into the base. The B–C junction is also reverse biased; thus, the emitter and collector currents will be zero for this case. This condition is referred to as *cutoff*—all currents in the transistor are zero.

When the B–E junction becomes forward biased, an emitter current will be generated as we have discussed, and the injection of electrons into the base results in a collector current. We may write the KVL equations around the collector–emitter loop as

$$V_{CC} = I_C R_C + V_{CB} + V_{BE} = V_R + V_{CE} \tag{12.7}$$

If V_{CC} is large enough and if V_R is small enough, then $V_{CB} > 0$, which means that the B–C junction is reverse biased for this npn transistor. Again, this condition is the forward-active region of operation.

As the forward-biased B–E voltage increases, the collector current and hence V_R will also increase. The increase in V_R means that the reverse-biased C–B voltage decreases, or $|V_{CB}|$ decreases. At some point, the collector current may become large

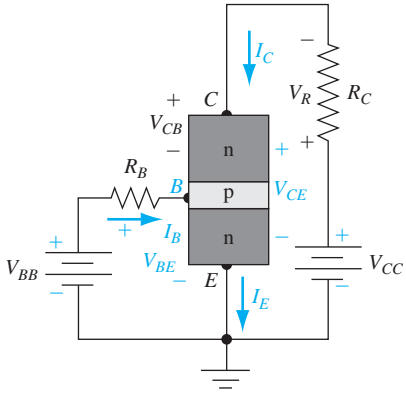


Figure 12.8 | An npn bipolar transistor in a common-emitter circuit configuration.

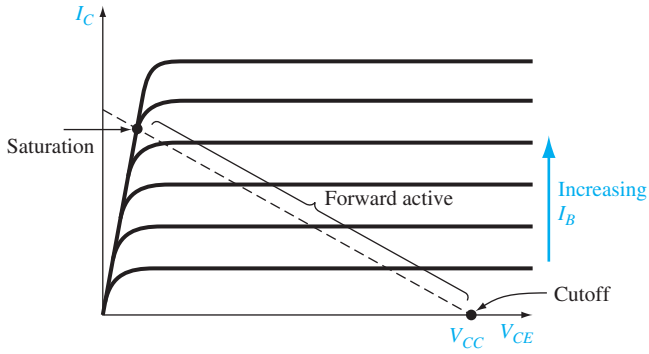


Figure 12.9 | Bipolar transistor common-emitter current–voltage characteristics with load line superimposed.

enough that the combination of V_R and V_{CC} produces 0 V across the B–C junction. A slight increase in I_C beyond this point will cause a slight increase in V_R and the B–C junction will become forward biased ($V_{CB} < 0$). This condition is called *saturation.*[1] In the saturation mode of operation, both B–E and B–C junctions are forward biased and the collector current is no longer controlled by the B–E voltage.

Figure 12.9 shows the transistor current characteristics, I_C versus V_{CE}, for constant base currents when the transistor is connected in the common-emitter configuration (Figure 12.8). When the collector–emitter voltage is large enough so that the base–collector junction is reverse biased, the collector current is a constant in this first-order theory. For small values of C–E voltage, the base–collector junction becomes forward biased and the collector current decreases to zero for a constant base current.

Writing a Kirchhoff's voltage equation around the C–E loop, we find

$$V_{CE} = V_{CC} - I_C R_C \tag{12.8}$$

Equation (12.8) shows a linear relation between collector current and collector–emitter voltage. This linear relation is called a *load line* and is plotted in Figure 12.9. The load line, superimposed on the transistor characteristics, can be used to visualize the bias condition and operating mode of the transistor. The cutoff mode occurs when $I_C = 0$, saturation occurs when there is no longer a change in collector current for a change in base current, and the forward-active mode occurs when the relation $I_C = \beta I_B$ is valid. These three operating modes are indicated on the figure.

A fourth mode of operation for the bipolar transistor is possible, although not with the circuit configuration shown in Figure 12.8. This fourth mode, known as

[1]The concept of "saturation" for the bipolar transistor is not the same as the principle of the "saturation region" for the MOSFET described in Chapter 10. The term "saturation" as applied to the BJT means that the output current and output voltage do not change as the base–emitter voltage changes. The term "saturation region" as applied to the MOSFET means that the output current does not change (ideally) with a change in the drain-to-source voltage.

inverse active, occurs when the B–E junction is reverse biased and the B–C junction is forward biased. In this case the transistor is operating "upside down," and the roles of the emitter and collector are reversed. We have argued that the transistor is not a symmetrical device; therefore, the inverse-active characteristics will not be the same as the forward-active characteristics.

The junction voltage conditions for the four operating modes are shown in Figure 12.10.

12.1.4 Amplification with Bipolar Transistors

Voltages and currents can be amplified by bipolar transistors in conjunction with other elements. We demonstrate this amplification qualitatively in the following discussion. Figure 12.11 shows an npn bipolar transistor in a common-emitter configuration. The dc voltage sources, V_{BB} and V_{CC}, are used to bias the transistor in the forward-active mode. The voltage source v_i represents a time-varying input voltage (such as a signal from a satellite) that needs to be amplified.

Figure 12.12 shows the various voltages and currents that are generated in the circuit assuming that v_i is a sinusoidal voltage. The sinusoidal voltage v_i induces a sinusoidal component of base current superimposed on a dc quiescent value. Since $i_C = \beta i_B$, then a relatively large sinusoidal collector current is superimposed on a dc value of collector current. The time-varying collector current induces a time-varying voltage across the R_C resistor which, by Kirchhoff's voltage law, means that a sinusoidal voltage, superimposed on a dc value, exists between the collector and emitter of the bipolar transistor. The sinusoidal voltages in the collector–emitter portion of the circuit are larger than the signal input voltage v_i, so that the circuit has produced a *voltage gain* in the time-varying signals. Hence, the circuit is known as a *voltage amplifier.*

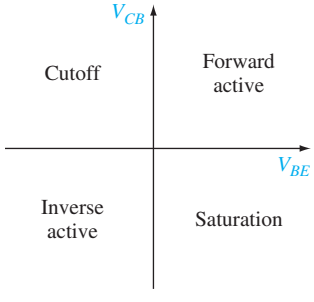


Figure 12.10 | Junction voltage conditions for the four operating modes of a bipolar transistor.

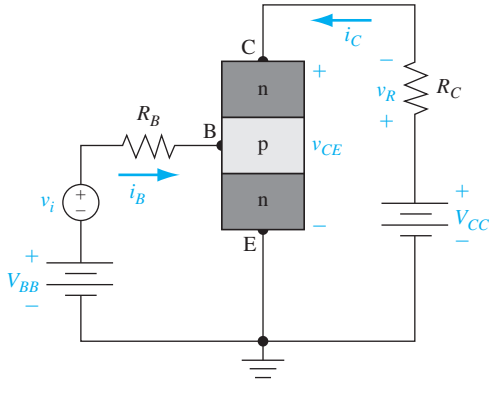


Figure 12.11 | Common-emitter npn bipolar circuit configuration with a time-varying signal voltage v_i included in the base–emitter loop.

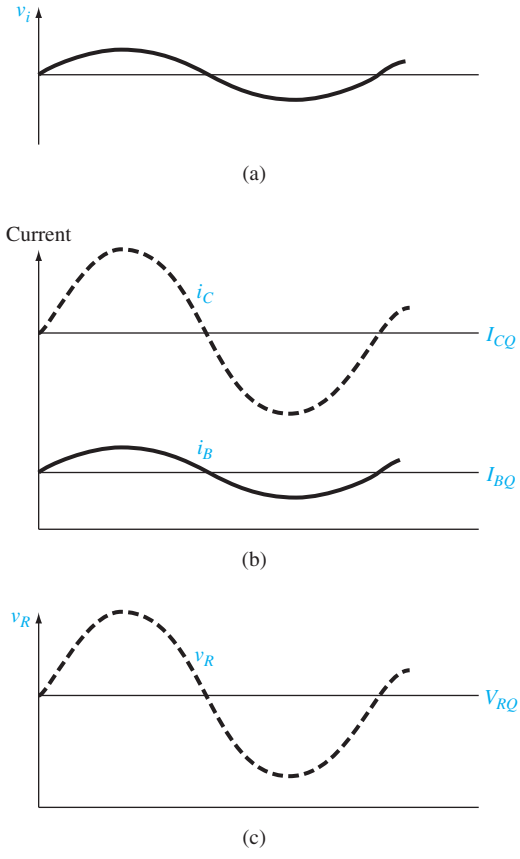


Figure 12.12 | Currents and voltages existing in the circuit shown in Figure 12.11. (a) Input sinusoidal signal voltage. (b) Sinusoidal base and collector currents superimposed on the quiescent dc values. (c) Sinusoidal voltage across the R_C resistor superimposed on the quiescent dc value.

In the remainder of the chapter, we consider the operation and characteristics of the bipolar transistor in more detail.

12.2 | MINORITY CARRIER DISTRIBUTION

We are interested in calculating currents in the bipolar transistor that, as in the simple pn junction, are determined by minority carrier diffusion. Since diffusion currents are produced by minority carrier gradients, we must determine the steady-state minority carrier distribution in each of the three transistor regions. Let us first consider the forward-active mode, and then the other modes of operation. Table 12.1 summarizes the notation used in the following analysis.

Table 12.1 | Notation used in the analysis of the bipolar transistor

Notation	Definition
For both the npn and pnp transistors	
N_E, N_B, N_C	Doping concentrations in the emitter, base, and collector
x_E, x_B, x_C	Widths of neutral emitter, base, and collector regions
D_E, D_B, D_C	*Minority carrier* diffusion coefficients in emitter, base, and collector regions
L_E, L_B, L_C	*Minority carrier* diffusion lengths in emitter, base, and collector regions
τ_{E0}, τ_{B0}, τ_{C0}	*Minority carrier* lifetimes in emitter, base, and collector regions
For the npn	
p_{E0}, n_{B0}, p_{C0}	Thermal-equilibrium *minority carrier* hole, electron, and hole concentrations in the emitter, base, and collector
$p_E(x')$, $n_B(x)$, $p_C(x'')$	Total *minority carrier* hole, electron, and hole concentrations in the emitter, base, and collector
$\delta p_E(x')$, $\delta n_B(x)$, $\delta p_C(x'')$	Excess *minority carrier* hole, electron, and hole concentrations in the emitter, base, and collector
For the pnp	
n_{E0}, p_{B0}, n_{C0}	Thermal-equilibrium *minority carrier* electron, hole, and electron concentrations in the emitter, base, and collector
$n_E(x')$, $p_B(x)$, $n_C(x'')$	Total *minority carrier* electron, hole, and electron concentrations in the emitter, base, and collector
$\delta n_E(x')$, $\delta p_B(x)$, $\delta n_C(x'')$	Excess *minority carrier* electron, hole, and electron concentrations in the emitter, base, and collector

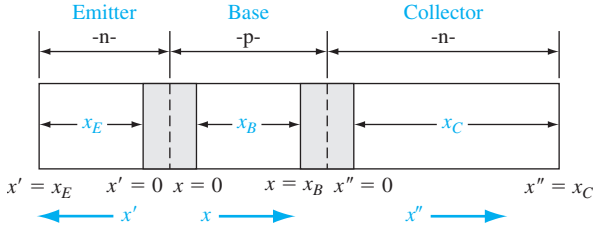


Figure 12.13 | Geometry of the npn bipolar transistor used to calculate the minority carrier distribution.

12.2.1 Forward-Active Mode

Consider a uniformly doped npn bipolar transistor with the geometry shown in Figure 12.13. When we consider the individual emitter, base, and collector regions, we shift the origin to the edge of the space charge region and consider a positive x, x', or x'' coordinate as shown in the figure.

In the forward-active mode, the B–E junction is forward biased and the B–C is reverse biased. We expect the minority carrier distributions to look like those shown

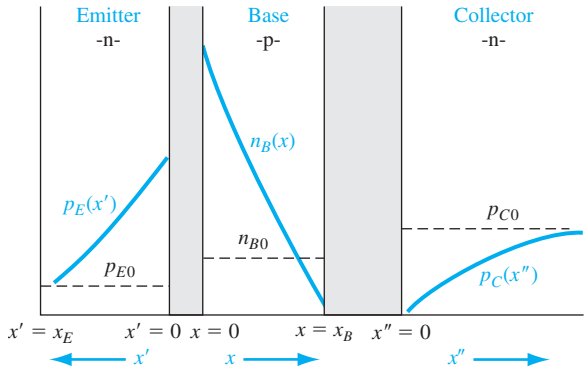


Figure 12.14 | Minority carrier distribution in an npn bipolar transistor operating in the forward-active mode.

in Figure 12.14. As there are two n regions, we have minority carrier holes in both emitter and collector. To distinguish between these two minority carrier hole distributions, we use the notation shown in the figure. Keep in mind that we are dealing only with minority carriers. The parameters p_{E0}, n_{B0}, and p_{C0} denote the thermal-equilibrium minority carrier concentrations in the emitter, base, and collector, respectively. The functions $p_E(x')$, $n_B(x)$, and $p_C(x'')$ denote the steady-state minority carrier concentrations in the emitter, base, and collector, respectively. We assume that the neutral collector length x_C is long compared to the minority carrier diffusion length L_C in the collector, but we take into account a finite emitter length x_E. If we assume that the surface recombination velocity at $x' = x_E$ is infinite, then the excess minority carrier concentration at $x' = x_E$ is zero, or $p_E(x' = x_E) = p_{E0}$. An infinite surface recombination velocity is a good approximation when an ohmic contact is fabricated at $x' = x_E$.

Base Region The steady-state excess minority carrier electron concentration is found from the ambipolar transport equation, which we discussed in detail in Chapter 6. For a zero electric field in the neutral base region, the ambipolar transport equation in steady state reduces to

$$D_B \frac{\partial^2(\delta n_B(x))}{\partial x^2} - \frac{\delta n_B(x)}{\tau_{B0}} = 0 \tag{12.9}$$

where δn_B is the excess minority carrier electron concentration, and D_B and τ_{B0} are the minority carrier diffusion coefficient and lifetime in the base region, respectively. The excess electron concentration is defined as

$$\delta n_B(x) = n_B(x) - n_{B0} \tag{12.10}$$

The general solution to Equation (12.9) can be written as

$$\delta n_B(x) = A \exp\left(\frac{+x}{L_B}\right) + B \exp\left(\frac{-x}{L_B}\right) \tag{12.11}$$

where L_B is the minority carrier diffusion length in the base, given by $L_B = \sqrt{D_B \tau_{B0}}$. The base is of finite width so both exponential terms in Equation (12.11) must be retained.

The excess minority carrier electron concentrations at the two boundaries become

$$\delta n_B(x = 0) \equiv \delta n_B(0) = A + B \tag{12.12a}$$

and

$$\delta n_B(x = x_B) \equiv \delta n_B(x_B) = A \exp\left(\frac{+x_B}{L_B}\right) + B \exp\left(\frac{-x_B}{L_B}\right) \tag{12.12b}$$

The B–E junction is forward biased, so the boundary condition at $x = 0$ is

$$\delta n_B(0) = n_B(x = 0) - n_{B0} = n_{B0}\left[\exp\left(\frac{eV_{BE}}{kT}\right) - 1\right] \tag{12.13a}$$

The B–C junction is reverse biased, so the second boundary condition at $x = x_B$ is

$$\delta n_B(x_B) = n_B(x = x_B) - n_{B0} = 0 - n_{B0} = -n_{B0} \tag{12.13b}$$

From the boundary conditions given by Equations (12.13a) and (12.13b), the coefficients A and B from Equations (12.12a) and (12.12b) can be determined. The results are

$$A = \frac{-n_{B0} - n_{B0}\left[\exp\left(\frac{eV_{BE}}{kT}\right) - 1\right]\exp\left(\frac{-x_B}{L_B}\right)}{2\sinh\left(\frac{x_B}{L_B}\right)} \tag{12.14a}$$

and

$$B = \frac{n_{B0}\left[\exp\left(\frac{eV_{BE}}{kT}\right) - 1\right]\exp\left(\frac{x_B}{L_B}\right) + n_{B0}}{2\sinh\left(\frac{x_B}{L_B}\right)} \tag{12.14b}$$

Then, substituting Equations (12.14a) and (12.14b) into Equation (12.9), we can write the excess minority carrier electron concentration in the base region as

$$\boxed{\delta n_B(x) = \frac{n_{B0}\left\{\left[\exp\left(\frac{eV_{BE}}{kT}\right) - 1\right]\sinh\left(\frac{x_B - x}{L_B}\right) - \sinh\left(\frac{x}{L_B}\right)\right\}}{\sinh\left(\frac{x_B}{L_B}\right)}} \tag{12.15a}$$

Equation (12.15a) may look formidable with the sinh functions. We have stressed that we want the base width x_B to be small compared to the minority carrier diffusion length L_B. This condition may seem somewhat arbitrary at this point, but the reason becomes clear as we proceed through all of the calculations. Since we want $x_B < L_B$, the argument in the sinh functions is always less than unity and in most cases will be much less than unity. Figure 12.15 shows a plot of sinh (y) for $0 \leq y \leq 1$ and also shows the linear approximation for small values of y. If $y < 0.4$, the sinh (y) function differs from its linear approximation by less than 3 percent. All of this leads to the *conclusion that the excess electron concentration* δn_B *in Equation (12.15a) is approximately a linear function of* x *through the neutral base region.*

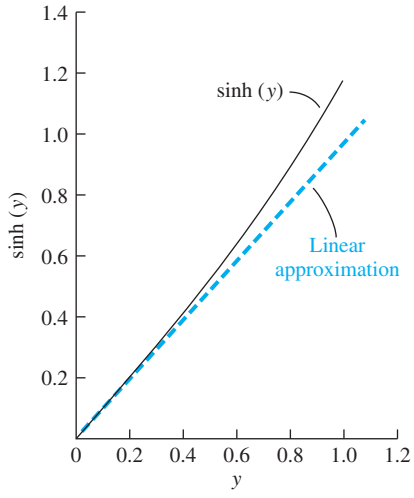


Figure 12.15 | Hyperbolic sine function and its linear approximation.

Using the approximation that $\sinh(x) \approx x$ for $x \ll 1$, the excess electron concentration in the base is given by

$$\delta n_B(x) \approx \frac{n_{B0}}{x_B}\left\{\left[\exp\left(\frac{eV_{BE}}{kT}\right) - 1\right](x_B - x) - x\right\} \tag{12.15b}$$

We use this linear approximation later in some of the example calculations. The difference in the excess carrier concentrations determined from Equations (12.15a) and (12.15b) is demonstrated in the following exercise.

TEST YOUR UNDERSTANDING

TYU 12.1 The emitter and base of a silicon npn bipolar transistor are uniformly doped at impurity concentrations of 10^{18} cm^{-3} and 10^{16} cm^{-3}, respectively. A forward-bias B–E voltage of $V_{BE} = 0.610$ V is applied. The neutral base width is $x_B = 2\ \mu$m and the minority carrier diffusion length in the base is $L_B = 10\ \mu$m. Calculate the excess minority carrier concentration in the base at (*a*) $x = 0$ and (*b*) $x = x_B/2$. (*c*) Determine the ratio of the actual minority carrier concentration at $x = x_B/2$ [Equation (12.15a)] to that in the ideal case of a linear minority carrier distribution [Equation (12.15b)].

[Ans. (*a*) $n_B(0) = 3.81 \times 10^{14}$ cm^{-3}; (*b*) $\delta n_B(x_B/2) = n_B(x_B/2) \approx$ 1.8947×10^{14} cm^{-3}; (*c*) Ratio $= (1.8947 \times 10^{14}/1.9042 \times 10^{14}) = 0.9950$]

Table 12.2 shows the Taylor expansions of some of the hyperbolic functions that are encountered in this section of the chapter. In most cases, we consider only the linear terms when expanding these functions.

Table 12.2 | Taylor expansions of hyperbolic functions

Function	Taylor expansion
$\sinh(x)$	$x + \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots$
$\cosh(x)$	$1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \cdots$
$\tanh(x)$	$x - \frac{x^3}{3} + \frac{2x^5}{15} + \cdots$

Emitter Region Consider, now, the minority carrier hole concentration in the emitter. The steady-state excess hole concentration is determined from the equation

$$D_E \frac{\partial^2[\delta p_E(x')]}{\partial x'^2} - \frac{\delta p_E(x')}{\tau_{E0}} = 0 \tag{12.16}$$

where D_E and τ_{E0} are the minority carrier diffusion coefficient and minority carrier lifetime, respectively, in the emitter. The excess hole concentration is given by

$$\delta p_E(x') = p_E(x') - p_{E0} \tag{12.17}$$

The general solution to Equation (12.16) can be written as

$$\delta p_E(x') = C \exp\left(\frac{+x'}{L_E}\right) + D \exp\left(\frac{-x'}{L_E}\right) \tag{12.18}$$

where $L_E = \sqrt{D_E \tau_{E0}}$. If we assume the neutral emitter length x_E is not necessarily long compared to L_E, then both exponential terms in Equation (12.18) must be retained.

The excess minority carrier hole concentrations at the two boundaries are

$$\delta p_E(x' = 0) \equiv \delta p_E(0) = C + D \tag{12.19a}$$

and

$$\delta p_E(x' = x_E) \equiv \delta p_E(x_E) = C \exp\left(\frac{x_E}{L_E}\right) + D \exp\left(\frac{-x_E}{L_E}\right) \tag{12.19b}$$

Again, the B–E junction is forward biased, so

$$\delta p_E(0) = p_E(x' = 0) - p_{E0} = p_{E0}\left[\exp\left(\frac{eV_{BE}}{kT}\right) - 1\right] \tag{12.20a}$$

An infinite surface recombination velocity at $x' = x_E$ implies that

$$\delta p_E(x_E) = 0 \tag{12.20b}$$

Solving for C and D using Equations (12.19) and (12.20) yields the excess minority carrier hole concentration in Equation (12.18):

$$\delta p_E(x') = \frac{p_{E0}\left[\exp\left(\frac{eV_{BE}}{kT}\right) - 1\right] \sinh\left(\frac{x_E - x'}{L_E}\right)}{\sinh\left(\frac{x_E}{L_E}\right)} \tag{12.21a}$$

This excess concentration will also vary approximately linearly with distance if x_E is small. We find

$$\delta p_E(x') \approx \frac{p_{E0}}{x_E}\left[\exp\left(\frac{eV_{BE}}{kT}\right) - 1\right](x_E - x') \tag{12.21b}$$

If x_E is comparable to L_E, then $\delta p_E(x')$ shows an exponential dependence on x_E.

TEST YOUR UNDERSTANDING

TYU 12.2 Consider a silicon npn bipolar transistor with emitter and base regions uniformly doped at concentrations of $10^{18}\,\text{cm}^{-3}$ and $10^{16}\,\text{cm}^{-3}$, respectively. A forward-bias B–E voltage of $V_{BE} = 0.610$ V is applied. The neutral emitter width is $x_E = 4\ \mu$m and the minority carrier diffusion length in the emitter is $L_E = 4\ \mu$m. Calculate the excess minority carrier concentration in the emitter at (*a*) $x' = 0$ and (*b*) $x' = x_E/2$.

[Ans. (*a*) $3.808 \times 10^{12}\,\text{cm}^{-3}$; (*b*) $1.689 \times 10^{12}\,\text{cm}^{-3}$]

Collector Region The excess minority carrier hole concentration in the collector can be determined from the equation

$$D_C \frac{\partial^2[\delta p_C(x'')]}{\partial x''^2} - \frac{\delta p_C(x'')}{\tau_{C0}} = 0 \tag{12.22}$$

where D_C and τ_{C0} are the minority carrier diffusion coefficient and minority carrier lifetime, respectively, in the collector. We can express the excess minority carrier hole concentration in the collector as

$$\delta p_C(x'') = p_C(x'') - p_{C0} \tag{12.23}$$

The general solution to Equation (12.22) can be written as

$$\delta p_C(x'') = G\exp\left(\frac{x''}{L_C}\right) + H\exp\left(\frac{-x''}{L_C}\right) \tag{12.24}$$

where $L_C = \sqrt{D_C\tau_{C0}}$. If we assume that the collector is long, then the coefficient G must be zero since the excess concentration must remain finite. The second boundary condition gives

$$\delta p_C(x'' = 0) \equiv \delta p_C(0) = p_C(x'' = 0) - p_{C0} = 0 - p_{C0} = -p_{C0} \tag{12.25}$$

The excess minority carrier hole concentration in the collector is then given as

$$\delta p_C(x'') = -p_{C0}\exp\left(\frac{-x''}{L_C}\right) \tag{12.26}$$

This result is exactly what we expect from the results of a reverse-biased pn junction.

TEST YOUR UNDERSTANDING

TYU 12.3 Consider the collector region of an npn bipolar transistor biased in the forward-active region. At what value of x'', compared to L_C, does the magnitude of the minority carrier concentration reach 95 percent of the thermal-equilibrium value? ($x'' \approx 3L_C$ Ans.)

12.2.2 Other Modes of Operation

The bipolar transistor can also operate in the cutoff, saturation, or inverse-active mode. We qualitatively discuss the minority carrier distributions for these operating conditions and treat the actual calculations as problems at the end of the chapter.

Figure 12.16a shows the minority carrier distribution in an npn bipolar transistor in cutoff. In cutoff, both the B–E and B–C junctions are reverse biased; thus, the minority carrier concentrations are zero at each space charge edge. The emitter and collector regions are assumed to be "long" in this figure, while the base is narrow compared with the minority carrier diffusion length. Since $x_B \ll L_B$, essentially all minority carriers are swept out of the base region.

Figure 12.16b shows the minority carrier distribution in the npn bipolar transistor operating in saturation. Both the B–E and B–C junctions are forward biased; thus, excess minority carriers exist at the edge of each space charge region. However, since a collector current still exists when the transistor is in saturation, a gradient will still exist in the minority carrier electron concentration in the base.

Finally, Figure 12.17a shows the minority carrier distribution in the npn transistor for the inverse-active mode. In this case, the B–E is reverse biased and the B–C is forward biased. Electrons from the collector are now injected into the base. The gradient in the minority carrier electron concentration in the base is in the opposite

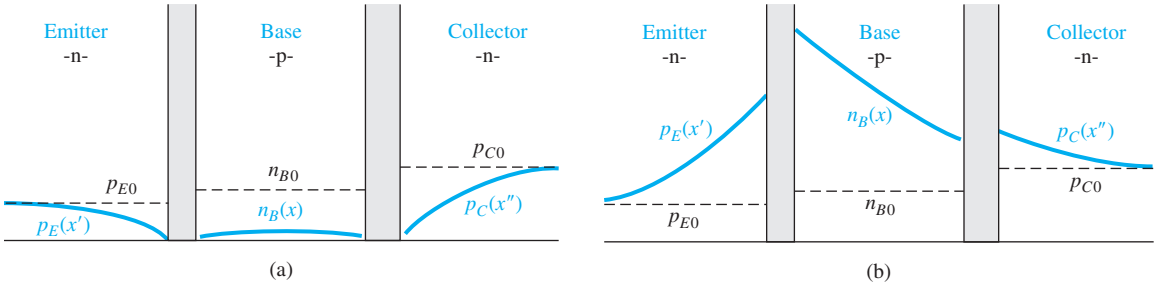


Figure 12.16 | Minority carrier distribution in an npn bipolar transistor operating in (a) cutoff and (b) saturation.

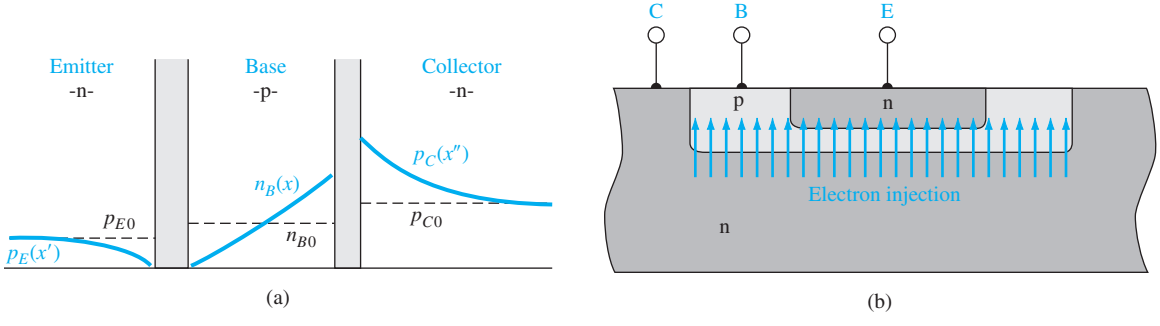


Figure 12.17 | (a) Minority carrier distribution in an npn bipolar transistor operating in the inverse-active mode. (b) Cross section of an npn bipolar transistor showing the injection and collection of electrons in the inverse-active mode.

direction compared with the forward-active mode, so the emitter and collector currents will change direction. Figure 12.17b shows the injection of electrons from the collector into the base. Since the B–C area is normally much larger than the B–E area, not all of the injected electrons will be collected by the emitter. The relative doping concentrations in the base and collector are also different compared with those in the base and emitter; thus, we see that the transistor is not symmetrical. We then expect the characteristics to be significantly different between the forward-active and inverse-active modes of operation.

12.3 | TRANSISTOR CURRENTS AND LOW-FREQUENCY COMMON-BASE CURRENT GAIN

The basic principle of operation of the bipolar transistor is the control of the collector current by the B–E voltage. The collector current is a function of the number of majority carriers reaching the collector after being injected from the emitter across the B–E junction. The *common-base current gain* is defined as the ratio of collector current to emitter current. The flow of various charged carriers leads to definitions of particular currents in the device. We can use these definitions to define the current gain of the transistor in terms of several factors.

12.3.1 Current Gain—Contributing Factors

Figure 12.18 shows the various particle flux components in the npn bipolar transistor. We define the various flux components and then consider the resulting currents. Although there seems to be a large number of flux components, we may help clarify the situation by correlating each factor with the minority carrier distributions shown in Figure 12.14.

The factor J_{nE}^{-} is the electron flux injected from the emitter into the base. As the electrons diffuse across the base, a few will recombine with majority carrier holes. The majority carrier holes that are lost by recombination must be replenished from the base terminal. This replacement hole flux is denoted by J_{RB}^{+}. The electron flux that reaches the collector is J_{nC}^{-}. The majority carrier holes from the base that are injected back into the emitter result in a hole flux denoted by J_{pE}^{+}. Some electrons and holes

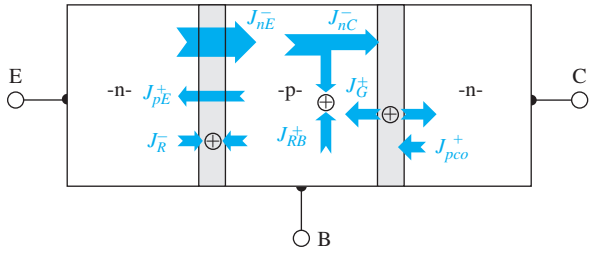


Figure 12.18 | Particle current density or flux components in an npn bipolar transistor operating in the forward-active mode.

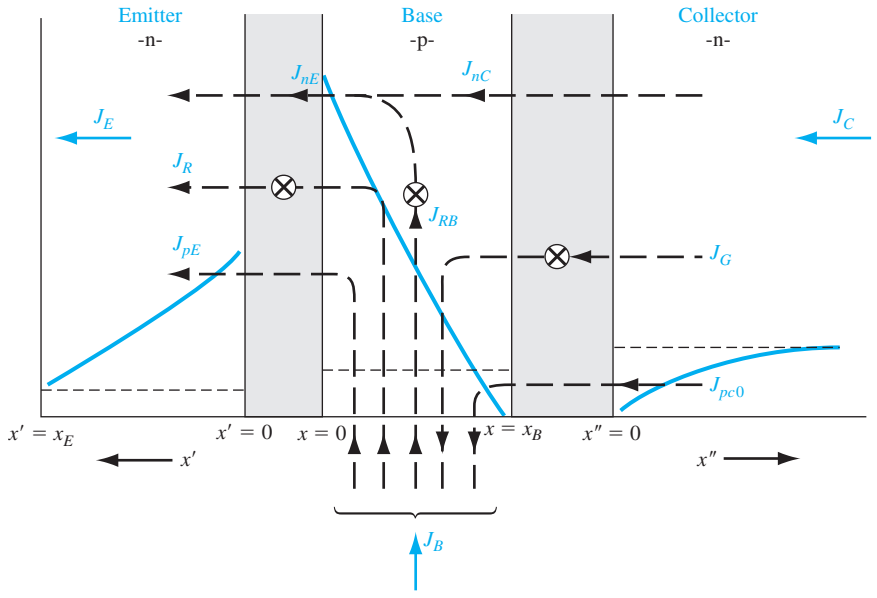


Figure 12.19 | Current density components in an npn bipolar transistor operating in the forward-active mode.

that are injected into the forward-biased B–E space charge region will recombine in this region. This recombination leads to the electron flux J_R^-. Generation of electrons and holes occurs in the reverse-biased B–C junction. This generation yields a hole flux J_G^+. Finally, the ideal reverse-saturation current in the B–C junction is denoted by the hole flux J_{pc0}^-.

The corresponding electric current density components in the npn transistor are shown in Figure 12.19 along with the minority carrier distributions for the forward-active mode. The curves are the same as in Figure 12.14. As in the pn junction, the currents in the bipolar transistor are defined in terms of minority carrier diffusion currents. The current densities are defined as follows:

J_{nE}: Due to the diffusion of minority carrier electrons in the base at $x = 0$.

J_{nC}: Due to the diffusion of minority carrier electrons in the base at $x = x_B$.

J_{RB}: The difference between J_{nE} and J_{nC}, which is due to the recombination of excess minority carrier electrons with majority carrier holes in the base. The J_{RB} current is the flow of holes into the base to replace the holes lost by recombination.

J_{pE}: Due to the diffusion of minority carrier holes in the emitter at $x' = 0$.

J_R: Due to the recombination of carriers in the forward-biased B–E junction.

J_{pc0}: Due to the diffusion of minority carrier holes in the collector at $x'' = 0$.

J_G: Due to the generation of carriers in the reverse-biased B–C junction.

The currents J_{RB}, J_{pE}, and J_R are B–E junction currents only and do not contribute to the collector current. The currents J_{pc0} and J_G are B–C junction currents only. These current components do not contribute to the transistor action or the current gain.

The dc common-base current gain is defined as

$$\alpha_0 = \frac{I_C}{I_E} \tag{12.27}$$

If we assume that the active cross-sectional area is the same for the collector and emitter, then we can write the current gain in terms of the current densities, or

$$\alpha_0 = \frac{J_C}{J_E} = \frac{J_{nC} + J_G + J_{pc0}}{J_{nE} + J_R + J_{pE}} \tag{12.28}$$

We are primarily interested in determining how the collector current will change with a change in emitter current. The small-signal, or sinusoidal, common-base current gain is defined as

$$\alpha = \frac{\partial J_C}{\partial J_E} = \frac{J_{nC}}{J_{nE} + J_R + J_{pE}} \tag{12.29}$$

The reverse-biased B–C currents, J_G and J_{pc0}, are not functions of the emitter current.

We can rewrite Equation (12.29) in the form

$$\alpha = \left(\frac{J_{nE}}{J_{nE} + J_{pE}}\right)\left(\frac{J_{nC}}{J_{nE}}\right)\left(\frac{J_{nE} + J_{pE}}{J_{nE} + J_R + J_{pE}}\right) \tag{12.30a}$$

or

$$\alpha = \gamma\alpha_T\delta \tag{12.30b}$$

The factors in Equation (12.30b) are defined as:

$$\gamma = \left(\frac{J_{nE}}{J_{nE} + J_{pE}}\right) \quad \equiv \text{emitter injection efficiency factor} \tag{12.31a}$$

$$\alpha_T = \left(\frac{J_{nC}}{J_{nE}}\right) \quad \equiv \text{base transport factor} \tag{12.31b}$$

$$\delta = \frac{J_{nE} + J_{pE}}{J_{nE} + J_R + J_{pE}} \equiv \text{recombination factor} \tag{12.31c}$$

We would like to have the change in collector current be exactly the same as the change in emitter current or, ideally, to have $\alpha = 1$. However, a consideration of Equation (12.29) shows that α will always be less than unity. The goal is to make α as close to unity as possible. To achieve this goal, we must make each term in Equation (12.30b) as close to unity as possible, since each factor is less than unity.

The *emitter injection efficiency factor* γ takes into account the minority carrier hole diffusion current in the emitter. This current is part of the emitter current, but does not contribute to the transistor action in that J_{pE} is not part of the collector current. The *base transport factor* α_T takes into account any recombination of excess minority carrier electrons in the base. Ideally, we want no recombination in the base. The *recombination factor* δ takes into account the recombination in the

forward-biased B–E junction. The current J_R contributes to the emitter current, but does not contribute to collector current.

12.3.2 Derivation of Transistor Current Components and Current Gain Factors

We now wish to determine the various transistor current components and each of the gain factors in terms of the electrical and geometrical parameters of the transistor. The results of these derivations show how the various parameters in the transistor influence the electrical properties of the device and point the way to the design of a "good" bipolar transistor.

Emitter Injection Efficiency Factor Consider, initially, the emitter injection efficiency factor. We have from Equation (12.31a)

$$\gamma = \left(\frac{J_{nE}}{J_{nE} + J_{pE}}\right) = \frac{1}{\left(1 + \dfrac{J_{pE}}{J_{nE}}\right)} \tag{12.32}$$

We derived the minority carrier distribution functions for the forward-active mode in Section 12.2.1. Noting that J_{nE}, as defined in Figure 12.19, is in the negative x direction, we can write the current densities as

$$J_{pE} = -eD_E \frac{d[\delta p_E(x')]}{dx'}\bigg|_{x'=0} \tag{12.33a}$$

and

$$J_{nE} = (-)eD_B \frac{d[\delta n_B(x)]}{dx}\bigg|_{x=0} \tag{12.33b}$$

where $\delta p_E(x')$ and $\delta n_B(x)$ are given by Equations (12.21) and (12.15), respectively.

Taking the appropriate derivatives of $\delta p_E(x')$ and $\delta n_B(x)$, we obtain

$$J_{pE} = \frac{eD_E p_{E0}}{L_E}\left[\exp\left(\frac{eV_{BE}}{kT}\right) - 1\right] \cdot \frac{1}{\tanh\,(x_E/L_E)} \tag{12.34a}$$

and

$$J_{nE} = \frac{eD_B n_{B0}}{L_B}\left\{\frac{1}{\sinh\,(x_B/L_B)} + \frac{[\exp\,(eV_{BE}/kT) - 1]}{\tanh\,(x_B/L_B)}\right\} \tag{12.34b}$$

Positive J_{pE} and J_{nE} values imply that the currents are in the directions shown in Figure 12.19. If we assume that the B–E junction is biased sufficiently far in the forward bias so that $V_{BE} \gg kT/e$, then

$$\exp\left(\frac{eV_{BE}}{kT}\right) \gg 1$$

and also

$$\frac{\exp\,(eV_{BE}/kT)}{\tanh\,(x_B/L_B)} \gg \frac{1}{\sinh\,(x_B/L_B)}$$

The emitter injection efficiency, from Equation (12.32), then becomes

$$\gamma = \frac{1}{1 + \dfrac{p_{E0} D_E L_B}{n_{B0} D_B L_E} \cdot \dfrac{\tanh(x_B/L_B)}{\tanh(x_E/L_E)}} \tag{12.35a}$$

If we assume that all the parameters in Equation (12.35a) except p_{E0} and n_{B0} are fixed, then in order for $\gamma \approx 1$, we must have $p_{E0} \ll n_{B0}$. We can write

$$p_{E0} = \frac{n_i^2}{N_E} \qquad \text{and} \qquad n_{B0} = \frac{n_i^2}{N_B}$$

where N_E and N_B are the impurity doping concentrations in the emitter and base, respectively. Then the condition that $p_{E0} \ll n_{B0}$ implies that $N_E \gg N_B$. For the emitter injection efficiency to be close to unity, the emitter doping must be large compared to the base doping. This condition means that many more electrons from the n-type emitter than holes from the p-type base will be injected across the B–E space charge region. If both $x_B \ll L_B$ and $x_E \ll L_E$, then the emitter injection efficiency can be written as

$$\gamma \approx \frac{1}{1 + \dfrac{N_B}{N_E} \cdot \dfrac{D_E}{D_B} \cdot \dfrac{x_B}{x_E}} \tag{12.35b}$$

EXAMPLE 12.1

Objective: Calculate the emitter injection efficiency.

Assume the following transistor parameters: $N_B = 10^{15}\,\text{cm}^{-3}$, $N_E = 10^{17}\,\text{cm}^{-3}$, $D_E = 10\,\text{cm}^2/\text{s}$, $D_B = 20\,\text{cm}^2/\text{s}$, $x_B = 0.80\,\mu\text{m}$, and $x_E = 0.60\,\mu\text{m}$.

■ Solution

From Equation (12.35b), we find

$$\gamma \cong \frac{1}{1 + \left(\dfrac{N_B}{N_E}\right)\left(\dfrac{D_E}{D_B}\right)\left(\dfrac{x_B}{x_E}\right)} = \frac{1}{1 + \left(\dfrac{10^{15}}{10^{17}}\right)\left(\dfrac{10}{20}\right)\left(\dfrac{0.80}{0.60}\right)} = 0.9934$$

■ Comment

This simple example shows a typical magnitude of the emitter injection efficiency.

■ EXERCISE PROBLEM

Ex 12.1 Repeat Example 12.1 if the base and emitter doping concentrations are $N_B = 5 \times 10^{15}\,\text{cm}^{-3}$ and $N_E = 10^{18}\,\text{cm}^{-3}$, respectively.
(Ans. $\gamma = 0.9966$)

Base Transport Factor The next term to consider is the base transport factor, given by Equation (12.31b) as $\alpha_T = J_{nC}/J_{nE}$. From the definitions of the current directions shown in Figure 12.19, we can write

$$J_{nC} = (-)eD_B \left.\frac{d[\delta n_B(x)]}{dx}\right|_{x=x_B} \tag{12.36a}$$

and

$$J_{nE} = (-)eD_B \left.\frac{d[\delta n_B(x)]}{dx}\right|_{x=0} \tag{12.36b}$$

Using the expression for $\delta n_B(x)$ given in Equation (12.15), we find that

$$J_{nC} = \frac{eD_B n_{B0}}{L_B}\left\{\frac{[\exp(eV_{BE}/kT) - 1]}{\sinh(x_B/L_B)} + \frac{1}{\tanh(x_B/L_B)}\right\} \tag{12.37}$$

The expression for J_{nE} is given in Equation (12.34a).

If we again assume that the B–E junction is biased sufficiently far in the forward bias so that $V_{BE} \gg kT/e$, then $\exp(eV_{BE}/kT) \gg 1$. Substituting Equations (12.37) and (12.34b) into Equation (12.31b), we have

$$\alpha_T = \frac{J_{nC}}{J_{nE}} \approx \frac{\exp(eV_{BE}/kT) + \cosh(x_B/L_B)}{1 + \exp(eV_{BE}/kT)\cosh(x_B/L_B)} \tag{12.38}$$

In order for α_T to be close to unity, the neutral base width x_B must be much smaller than the minority carrier diffusion length in the base L_B. If $x_B \ll L_B$, then $\cosh(x_B/L_B)$ will be just slightly greater than unity. In addition, if $\exp(eV_{BE}/kT) \gg 1$, then the base transport factor is approximately

$$\boxed{\alpha_T \approx \frac{1}{\cosh(x_B/L_B)}} \tag{12.39a}$$

For $x_B \ll L_B$, we may expand the cosh function in a Taylor series, so that

$$\boxed{\alpha_T \approx \frac{1}{\cosh(x_B/L_B)} \approx \frac{1}{1 + \frac{1}{2}(x_B/L_B)^2} \approx 1 - \frac{1}{2}(x_B/L_B)^2} \tag{12.39b}$$

The base transport factor α_T will be close to one if $x_B \ll L_B$. We can now see why we indicated earlier that the neutral base width x_B would be less than L_B.

EXAMPLE 12.2

Objective: Calculate the base transport factor.

Assume transistor parameters of $x_B = 0.80\ \mu$m and $L_B = 10.0\ \mu$m.

■ **Solution**

From Equation (12.39a), we find

$$\alpha_T \cong \frac{1}{\cosh\left(\frac{x_B}{L_B}\right)} = \frac{1}{\cosh\left(\frac{0.80}{10.0}\right)} = 0.9968$$

■ **Comment**

This simple example shows a typical magnitude of the base transport factor.

■ **EXERCISE PROBLEM**

Ex 12.2 Repeat Example 12.2 for $x_B = 1.2\ \mu$m and $L_B = 10.0\ \mu$m. (Ans. $\alpha_T = 0.9928$)

Recombination Factor The recombination factor is given by Equation (12.31c). We can write

$$\delta = \frac{J_{nE} + J_{pE}}{J_{nE} + J_R + J_{pE}} \approx \frac{J_{nE}}{J_{nE} + J_R} = \frac{1}{1 + J_R/J_{nE}} \tag{12.40}$$

We have assumed in Equation (12.40) that $J_{pE} \ll J_{nE}$. The recombination current density, due to the recombination in a forward-biased pn junction, was discussed in Chapter 8 and can be written as

$$J_R = \frac{ex_{BE}n_i}{2\tau_0}\exp\left(\frac{eV_{BE}}{2kT}\right) = J_{r0}\exp\left(\frac{eV_{BE}}{2kT}\right) \tag{12.41}$$

where x_{BE} is the B–E space charge width.

The current J_{nE} from Equation (12.34b) can be approximated as

$$J_{nE} = J_{s0}\exp\left(\frac{eV_{BE}}{kT}\right) \tag{12.42}$$

where

$$J_{s0} = \frac{eD_B n_{B0}}{L_B \tanh(x_B/L_B)} \tag{12.43}$$

The recombination factor, from Equation (12.40), can then be written as

$$\boxed{\delta = \frac{1}{1 + \dfrac{J_{r0}}{J_{s0}}\exp\left(\dfrac{-eV_{BE}}{2kT}\right)}} \tag{12.44}$$

The recombination factor is a function of the B–E voltage. As V_{BE} increases, the recombination current becomes less dominant and the recombination factor approaches unity.

EXAMPLE 12.3

Objective: Calculate the recombination factor.

Assume the following transistor parameters: $x_{BE} = 0.10\ \mu$m, $\tau_o = 10^{-7}$ s, $N_B = 5 \times 10^{15}$ cm^{-3}, $D_B = 20$ cm^2/s, $L_B = 10\ \mu$m, and $x_B = 0.80\ \mu$m. Assume $V_{BE} = 0.50$ V.

■ Solution

From Equation (12.41), we find

$$J_{r0} = \frac{ex_{BE}n_i}{2\tau_o} = \frac{(1.6 \times 10^{-19})(0.10 \times 10^{-4})(1.5 \times 10^{10})}{2(10^{-7})} = 1.2 \times 10^{-7}\ \text{A/cm}^2$$

and from Equation (12.43), we find

$$J_{s0} = \frac{eD_B n_{B0}}{L_B \tanh(x_B/L_B)} = \frac{eD_B(n_i^2/N_B)}{L_B \tanh(x_B/L_B)}$$

$$= \frac{(1.6 \times 10^{-19})(20)[(1.5 \times 10^{10})^2/5 \times 10^{15}]}{(10 \times 10^{-4})\tanh(0.80/10.0)} = 1.804 \times 10^{-9}\ \text{A/cm}^2$$

Then from Equation (12.44), the recombination factor is found as

$$\delta = \frac{1}{1 + \dfrac{J_{r0}}{J_{s0}} \cdot \exp\left(\dfrac{-V_{BE}}{2V_t}\right)} = \frac{1}{1 + \left(\dfrac{1.2 \times 10^{-7}}{1.804 \times 10^{-9}}\right) \cdot \exp\left(\dfrac{-0.50}{2(0.0259)}\right)}$$

$$= 0.99574$$

Comment

This simple example shows a typical magnitude of the recombination factor.

EXERCISE PROBLEM

Ex 12.3 Repeat Example 12.3 for $V_{BE} = 0.65$ V. (Ans. $\delta = 0.99976$)

The recombination factor must also include surface effects. The surface effects can be described by the surface recombination velocity as we discussed in Chapter 6. Figure 12.20a shows the B–E junction of an npn transistor near the semiconductor surface. We assume that the B–E junction is forward biased. Figure 12.20b shows the excess minority carrier electron concentration in the base along the cross section A-A′. This curve is the usual forward-biased junction minority carrier concentration. Figure 12.20c shows the excess minority carrier electron concentration along the cross section C-C′ from the surface. We have showed earlier that the excess concentration at a surface is smaller than the excess concentration in the bulk material. With this electron distribution, there is a diffusion of electrons from the bulk toward the surface where the electrons recombine with the majority carrier holes. Figure 12.20d shows the injection of electrons from the emitter into the base and the diffusion of

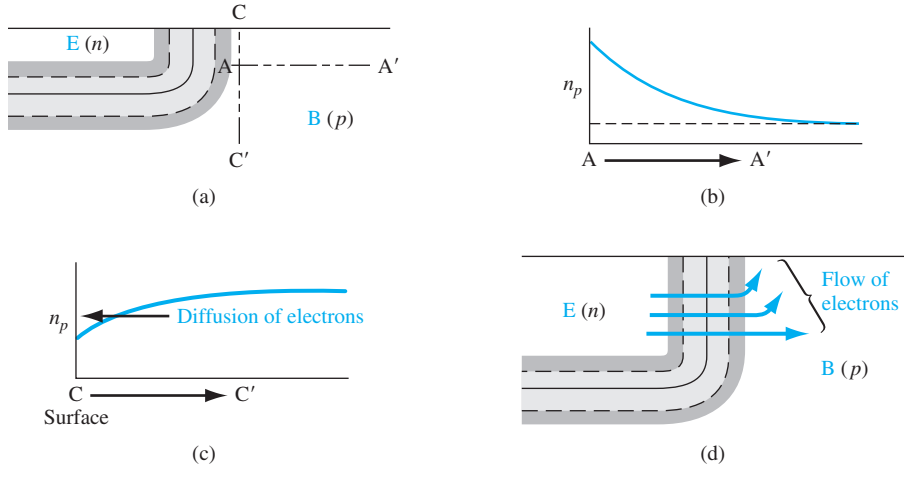


Figure 12.20 | The surface at the E–B junction showing the diffusion of carriers toward the surface.

electrons toward the surface. This diffusion generates another component of recombination current and this component of recombination current must be included in the recombination factor δ. Although the actual calculation is difficult because of the two-dimensional analysis required, the form of the recombination current is the same as that of Equation (12.41).

12.3.3 Summary

Although we have considered an npn transistor in all of the derivations, exactly the same analysis applies to a pnp transistor; the same minority carrier distributions are obtained except that the electron concentrations become hole concentrations and vice versa. The current directions and voltage polarities also change.

We have been considering the common-base current gain, defined in Equation (12.27) as $\alpha_0 = I_C/I_E$. The common-emitter current gain is defined as $\beta_0 = I_C/I_B$. From Figure 12.8 we see that $I_E = I_B + I_C$. We can determine the relation between common-emitter and common-base current gains from the KCL equation. We can write

$$\frac{I_E}{I_C} = \frac{I_B}{I_C} + 1$$

Substituting the definitions of current gains, we have

$$\frac{1}{\alpha_0} = \frac{1}{\beta_0} + 1$$

Since this relation actually holds for both dc and small-signal conditions, we can drop the subscript. The common-emitter current gain can now be written in terms of the common-base current gain as

$$\beta = \frac{\alpha}{1 - \alpha}$$

The common-base current gain, in terms of the common-emitter current gain, is found to be

$$\alpha = \frac{\beta}{1 + \beta}$$

Table 12.3 summarizes the expressions for the limiting factors in the common-base current gain assuming that $x_B \ll L_B$ and $x_E \ll L_E$. Also given are the approximate expressions for the common-base current gain and the common-emitter current gain.

12.3.4 Example Calculations of the Gain Factors

If we assume a typical value of β to be 100, then $\alpha = 0.99$. If we also assume that $\gamma = \alpha_T = \delta$, then each factor would have to be equal to 0.9967 in order that $\beta = 100$. This calculation gives an indication of how close to unity each factor must be in order to achieve a reasonable current gain.

Table 12.3 | Summary of limiting factors

Emitter injection efficiency

$$\gamma \approx \frac{1}{1 + \frac{N_B}{N_E} \cdot \frac{D_E}{D_B} \cdot \frac{x_B}{x_E}} \qquad (x_B \ll L_B), (x_E \ll L_E)$$

Base transport factor

$$\alpha_T \approx \frac{1}{1 + \frac{1}{2}\left(\frac{x_B}{L_B}\right)^2} \qquad (x_B \ll L_B)$$

Recombination factor

$$\delta = \frac{1}{1 + \frac{J_{r0}}{J_{s0}} \exp\left(\frac{-eV_{BE}}{2kT}\right)}$$

Common-base current gain

$$\alpha = \gamma \alpha_T \delta \approx \frac{1}{1 + \frac{N_B}{N_E} \cdot \frac{D_E}{D_B} \cdot \frac{x_B}{x_E} + \frac{1}{2}\left(\frac{x_B}{L_B}\right)^2 + \frac{J_{r0}}{J_{s0}} \exp\left(\frac{-eV_{BE}}{2kT}\right)}$$

Common-emitter current gain

$$\beta = \frac{\alpha}{1 - \alpha} \approx \frac{1}{\frac{N_B}{N_E} \cdot \frac{D_E}{D_B} \cdot \frac{x_B}{x_E} + \frac{1}{2}\left(\frac{x_B}{L_B}\right)^2 + \frac{J_{r0}}{J_{s0}} \exp\left(\frac{-eV_{BE}}{2kT}\right)}$$

DESIGN EXAMPLE 12.4

Objective: Design the ratio of emitter doping to base doping in order to achieve an emitter injection efficiency factor of $\gamma = 0.9967$.

Consider an npn bipolar transistor. Assume, for simplicity, that $D_E = D_B$, $L_E = L_B$, and $x_E = x_B$.

Solution

Equation (12.35b) reduces to

$$\gamma = \frac{1}{1 + \frac{p_{E0}}{n_{B0}}} = \frac{1}{1 + \frac{n_i^2/N_E}{n_i^2/N_B}}$$

so

$$\gamma = \frac{1}{1 + \frac{N_B}{N_E}} = 0.9967$$

Then

$$\frac{N_B}{N_E} = 0.00331 \qquad \text{or} \qquad \frac{N_E}{N_B} = 302$$

Comment

The emitter doping concentration must be much larger than the base doping concentration to achieve a high emitter injection efficiency.

EXERCISE PROBLEM

Ex 12.4 Assume that transistor parameters are the same as described in Example 12.4. In addition, let $N_E = 6 \times 10^{18}$ cm^{-3}. Determine the base doping concentration such that the emitter injection efficiency is $\gamma = 0.9950$. (Ans. $N_a = 3.02 \times 10^{16}$ cm^{-3})

DESIGN EXAMPLE 12.5

Objective: Design the base width required to achieve a base transport factor of $\alpha_T = 0.9967$. Consider a pnp bipolar transistor. Assume that $D_B = 10$ cm^2/s and $\tau_{B0} = 10^{-7}$s.

Solution

The base transport factor applies to both pnp and npn transistors and is given by

$$\alpha_T = \frac{1}{\cosh\,(x_B/L_B)} = 0.9967$$

Then

$$x_B/L_B = 0.0814$$

We have

$$L_B = \sqrt{D_B\tau_{B0}} = \sqrt{(10)(10^{-7})} = 10^{-3}\text{ cm}$$

so that the base width must then be

$$x_B = 0.814 \times 10^{-4}\text{ cm} = 0.814\ \mu\text{m}$$

Comment

If the base width is less than approximately 0.8 μm, then the required base transport factor will be achieved. In most cases, the base transport factor will not be the limiting factor in the bipolar transistor current gain.

EXERCISE PROBLEM

Ex 12.5 Assume that transistor parameters are the same as described in Example 12.5. Determine the minimum base width x_B such that the base transport factor is $\alpha_T = 0.9980$. (Ans. $x_B = 0.6349$ μm)

DESIGN EXAMPLE 12.6

Objective: Determine the forward-biased B–E voltage required to achieve a recombination factor equal to $\delta = 0.9967$.

Consider an npn bipolar transistor at $T = 300$ K. Assume that $J_{r0} = 10^{-8}$ A/cm^2 and that $J_{s0} = 10^{-11}$ A/cm^2.

Solution

The recombination factor, from Equation (12.44), is

$$\delta = \frac{1}{1 + \dfrac{J_{r0}}{J_{s0}}\exp\left(\dfrac{-eV_{BE}}{2kT}\right)}$$

We then have

$$0.9967 = \frac{1}{1 + \dfrac{10^{-8}}{10^{-11}} \exp\left(\dfrac{-eV_{BE}}{2kT}\right)}$$

We can rearrange this equation and write

$$\exp\left(\frac{+eV_{BE}}{2kT}\right) = \frac{0.9967 \times 10^3}{1 - 0.9967} = 3.02 \times 10^5$$

Then

$$V_{BE} = 2(0.0259) \ln (3.02 \times 10^5) = 0.654 \text{ V}$$

Comment

This example demonstrates that the recombination factor may be an important limiting factor in the bipolar current gain. In this example, if V_{BE} is smaller than 0.654 V, then the recombination factor δ will fall below the desired 0.9967 value.

EXERCISE PROBLEM

Ex 12.6 If $J_{r0} = 10^{-8}$ A/cm^2 and $J_{s0} = 10^{-11}$ A/cm^2, determine the value of V_{BE} such that $\delta = 0.9950$. (Ans. $V_{BE} = 0.6320$ V)

EXAMPLE 12.7

Objective: Calculate the common-emitter current gain of a silicon npn bipolar transistor at $T = 300$ K given a set of parameters.

Assume the following parameters:

$D_E = 10$ cm^2/s	$x_B = 0.70\ \mu$m
$D_B = 25$ cm^2/s	$x_E = 0.50\ \mu$m
$\tau_{E0} = 1 \times 10^{-7}$ s	$N_E = 1 \times 10^{18}$ cm^{-3}
$\tau_{B0} = 5 \times 10^{-7}$ s	$N_B = 1 \times 10^{16}$ cm^{-3}
$J_{r0} = 5 \times 10^{-8}$ A/cm^2	$V_{BE} = 0.65$ V

The following parameters are calculated:

$$p_{E0} = \frac{(1.5 \times 10^{10})^2}{1 \times 10^{18}} = 2.25 \times 10^2 \text{ cm}^{-3}$$

$$n_{B0} = \frac{(1.5 \times 10^{10})^2}{1 \times 10^{16}} = 2.25 \times 10^4 \text{ cm}^{-3}$$

$$L_E = \sqrt{D_E \tau_{E0}} = 10^{-3} \text{ cm}$$

$$L_B = \sqrt{D_B \tau_{B0}} = 3.54 \times 10^{-3} \text{ cm}$$

Solution

The emitter injection efficiency factor, from Equation (12.35a), is

$$\gamma = \frac{1}{1 + \dfrac{(2.25 \times 10^2)(10)(3.54 \times 10^{-3})}{(2.25 \times 10^4)(25)(10^{-3})} \cdot \dfrac{\tanh (0.0198)}{\tanh (0.050)}} = 0.9944$$

The base transport factor, from Equation (12.39a), is

$$\alpha_T = \frac{1}{\cosh\left(\dfrac{0.70 \times 10^{-4}}{3.54 \times 10^{-3}}\right)} = 0.9998$$

The recombination factor, from Equation (12.44), is

$$\delta = \frac{1}{1 + \dfrac{5 \times 10^{-8}}{J_{s0}} \exp\left(\dfrac{-0.65}{2(0.0259)}\right)}$$

where

$$J_{s0} = \frac{eD_B n_{B0}}{L_B \tanh\left(\dfrac{x_B}{L_B}\right)} = \frac{(1.6 \times 10^{-19})(25)(2.25 \times 10^4)}{3.54 \times 10^{-3} \tanh\,(1.977 \times 10^{-2})} = 1.29 \times 10^{-9}\ \text{A/cm}^2$$

We can now calculate $\delta = 0.99986$. The common-base current gain is then

$$\alpha = \gamma\alpha_T\delta = (0.9944)(0.9998)(0.99986) = 0.99406$$

which gives a common-emitter current gain of

$$\beta = \frac{\alpha}{1 - a} = \frac{0.99406}{1 - 0.99406} = 167$$

Comment

In this example, the emitter injection efficiency is the limiting factor in the current gain.

EXERCISE PROBLEM

Ex 12.7 Assume that $\gamma = \alpha_T = 0.9980$, $J_{r0} = 5 \times 10^{-9}$ A/cm^2, and $J_{s0} = 2 \times 10^{-11}$ A/cm^2. Determine the common-emitter current gain β for (*a*) $V_{BE} = 0.550$ V and (*b*) $V_{BE} = 0.650$ V. [Ans. (*a*) $\beta = 98.5$; (*b*) $\beta = 204$]

TEST YOUR UNDERSTANDING

NOTE: In the following Test Your Understanding questions, assume a silicon npn bipolar transistor at $T = 300$ K has the following minority carrier parameters: $D_E = 8$ cm^2/s, $D_B = 20$ cm^2/s, $D_C = 12$ cm^2/s, $\tau_{E0} = 10^{-8}$ s, $\tau_{B0} = 10^{-7}$ s, and $\tau_{C0} = 10^{-6}$ s.

TYU 12.4 If the emitter doping concentration is $N_E = 5 \times 10^{18}$ cm^{-3}, find the base doping concentration such that the emitter injection efficiency is $\gamma = 0.9950$. Assume $x_E = 2x_B = 2\ \mu$m.
(Ans. $N_B = 1.08 \times 10^{17}$ cm^{-3})

TYU 12.5 Assume that $\alpha_T = \delta = 0.9967$, $x_B = x_E = 1\ \mu$m, $N_B = 5 \times 10^{16}$ cm^{-3}, and $N_E = 5 \times 10^{18}$ cm^{-3}. Determine the common-emitter current gain β.
(Ans. $\beta = 92.4$)

TYU 12.6 Assume that $\gamma = \delta = 0.9967$ and $x_B = 0.80\ \mu$m. Determine the common-emitter current gain β.
(Ans. $\beta = 121$)

12.4 | NONIDEAL EFFECTS

In all previous discussions, we have considered a transistor with uniformly doped regions, low injection, constant emitter and base widths, an ideal constant energy bandgap, uniform current densities, and junctions that are not in breakdown. If any of these ideal conditions is not present, then the transistor properties will deviate from the ideal characteristics we have derived.

12.4.1 Base Width Modulation

We have implicitly assumed that the neutral base width x_B is constant. This base width, however, is a function of the B–C voltage, since the width of the space charge region extending into the base region varies with B–C voltage. As the B–C reverse-biased voltage increases, the B–C space charge region width increases, which reduces x_B. A change in the neutral base width will change the collector current as can be observed in Figure 12.21. A reduction in base width will cause the gradient in the minority carrier concentration to increase, which in turn causes an increase in the diffusion current. This effect is known as *base width modulation;* it is also called the *Early effect.*

The Early effect can be seen in the current–voltage characteristics shown in Figure 12.22. In most cases, a constant base current is equivalent to a constant B–E voltage. Ideally the collector current is independent of the B–C voltage so that the slope of the curves would be zero; thus, the output conductance of the transistor would be zero. However, the base width modulation, or Early effect, produces a nonzero slope and gives rise to a finite output conductance. If the collector current characteristics are extrapolated to zero collector current, the curves intersect the voltage axis at a point that is defined as the Early voltage. The Early voltage is considered to be a positive value. It is a common parameter given in transistor specifications; typical values of Early voltage are in the 100- to 300-V range.

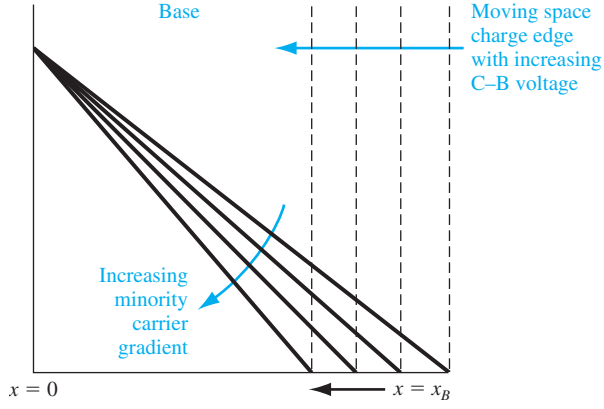


Figure 12.21 | The change in the base width and the change in the minority carrier gradient as the B–C space charge width changes.

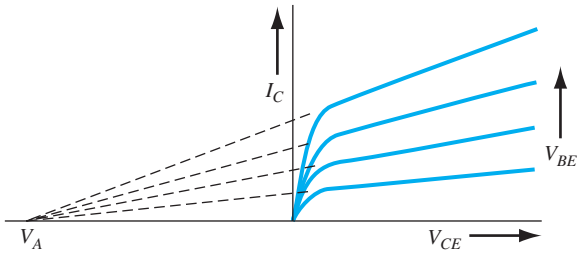


Figure 12.22 | The collector current versus collector–emitter voltage showing the Early effect and Early voltage.

From Figure 12.22, we can write that

$$\frac{dI_C}{dV_{CE}} \equiv g_o = \frac{I_C}{V_{CE} + V_A} = \frac{1}{r_o} \tag{12.45a}$$

where V_A and V_{CE} are defined as positive quantities, g_o is defined as the output conductance, and r_o is defined as the output resistance. Equation (12.45a) can be rewritten in the form

$$I_C = g_o\,(V_{CE} + V_A) = \frac{1}{r_o}(V_{CE} + V_A) \tag{12.45b}$$

showing that the collector current is now an explicit function of the collector–emitter voltage or the collector–base voltage.

EXAMPLE 12.8

Objective: Calculate the change in collector current with a change in neutral base width, and estimate the Early voltage.

Consider a uniformly doped silicon npn bipolar transistor with the following parameters: $N_B = 5 \times 10^{16}$ cm^{-3}, $N_C = 2 \times 10^{15}$ cm^{-3}, $x_{B0} = 0.70$ μm, and $D_B = 25$ cm^2/s. Assume that $x_{B0} \ll L_B$ and that $V_{BE} = 0.60$ V. The collector–base voltage is in the range $2 \le V_{CB} \le 10$ V.

■ Solution

Assuming $x_{B0} \ll L_B$, the excess minority carrier electron concentration in the base can be approximated by Equation (12.15b), which is

$$\delta n_B(x) \cong \frac{n_{B0}}{x_B}\left\{\left[\exp\left(\frac{V_{BE}}{V_t}\right) - 1\right](x_B - x) - x\right\}$$

The collector current is

$$|J_C| = eD_B\frac{d[\delta n_B(x)]}{dx} \cong \frac{eD_B n_{B0}}{x_B}\exp\left(\frac{V_{BE}}{V_t}\right)$$

The value of n_{B0} is found as

$$n_{B0} = \frac{n_i^2}{N_B} = \frac{(1.5 \times 10^{10})^2}{5 \times 10^{16}} = 4.5 \times 10^3 \text{ cm}^{-3}$$

For $V_{CB} = 2$ V, we find (see the following Exercise Problem Ex 12.8)

$$x_B = x_{B0} - x_{dB} = 0.70 - 0.0518 = 0.6482\ \mu\text{m}$$

and

$$|J_C| = \frac{(1.6 \times 10^{-19})(25)(4.5 \times 10^3)}{0.6482 \times 10^{-4}} \exp\left(\frac{0.60}{0.0259}\right) = 3.195 \text{ A/cm}^2$$

For $V_{CB} = 10$ V, we find (see the following Exercise Problem Ex 12.8)

$$x_B = 0.70 - 0.103 = 0.597 \ \mu\text{m}$$

and

$$|J_C| = \frac{(1.6 \times 10^{-19})(25)(4.5 \times 10^3)}{0.597 \times 10^{-4}} \exp\left(\frac{0.60}{0.0259}\right) = 3.469 \text{ A/cm}^2$$

We now can find, from Equation (12.45a)

$$\frac{dJ_C}{dV_{CE}} = \frac{\Delta J_C}{\Delta V_{CB}} = \frac{J_C}{V_{CE} + V_A} = \frac{J_C}{V_{BE} + V_{CB} + V_A}$$

or

$$\frac{3.469 - 3.195}{8} = \frac{3.195}{0.60 + 2 + V_A}$$

The Early voltage is then determined to be

$$V_A = 90.7 \text{ V}$$

Comment

This example indicates how much the collector current can change as the neutral base width changes with a change in the B–C space charge width, and it also indicates the magnitude of the Early voltage.

EXERCISE PROBLEM

Ex 12.8 Consider a silicon npn bipolar transistor with parameters described in Example 12.8. Determine the neutral base width for a C–B voltage of (*a*) $V_{CB} = 2$ V and (*b*) $V_{CB} = 10$ V. Neglect the B–E space charge width.

[Ans. (*a*) $x_B = 0.597\ \mu\text{m}$; (*b*) $x_B = 0.482\ \mu\text{m}$]

The previous example and exercise problem demonstrate, too, that we can expect variations in transistor properties due to tolerances in transistor-fabrication processes. There will be variations, in particular, in the base width of narrow-base transistors that will cause variations in the collector current characteristics simply due to the tolerances in processing.

12.4.2 High Injection

The ambipolar transport equation that we have used to determine the minority carrier distributions assumed low injection. As V_{BE} increases, the injected minority carrier concentration may approach, or even become larger than, the majority carrier concentration. If we assume quasi–charge neutrality, then the majority carrier hole concentration in the p-type base at $x = 0$ will increase as shown in Figure 12.23 because of the excess holes.

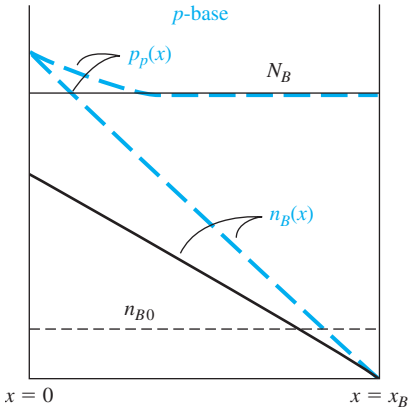


Figure 12.23 | Minority and majority carrier concentrations in the base under low and high injection (solid line: low injection; dashed line: high injection).

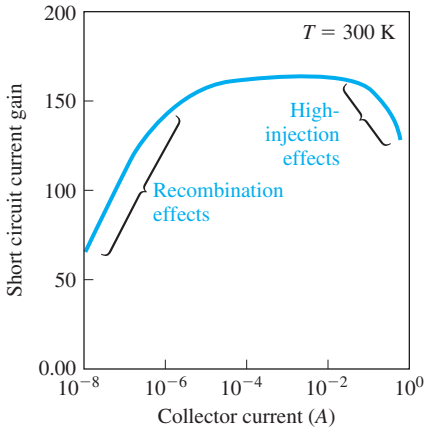


Figure 12.24 | Common-emitter current gain versus collector current. *(From Shur [14].)*

Two effects occur in the transistor at high injection. The first effect is a reduction in emitter injection efficiency. Since the majority carrier hole concentration at $x = 0$ increases with high injection, more holes are injected back into the emitter because of the forward-biased B–E voltage. An increase in the hole injection causes an increase in the J_{pE} current and an increase in J_{pE} reduces the emitter injection efficiency. The common-emitter current gain decreases, then, with high injection. Figure 12.24 shows a typical common-emitter current gain versus collector current curve. The low gain at low currents is due to the small recombination factor and the drop-off at the high current is due to the high-injection effect.

We now consider the second high-injection effect. At low injection, the majority carrier hole concentration at $x = 0$ for the npn transistor is

$$p_p(0) = p_{p0} = N_a \tag{12.46a}$$

and the minority carrier electron concentration is

$$n_p(0) = n_{p0}\exp\left(\frac{eV_{BE}}{kT}\right) \tag{12.46b}$$

The pn product is

$$p_p(0)n_p(0) = p_{p0}n_{p0}\exp\left(\frac{eV_{BE}}{kT}\right) \tag{12.46c}$$

At high injection, Equation (12.46c) still applies. However, $p_p(0)$ will also increase, and for very high injection it will increase at nearly the same rate as $n_p(0)$. The increase in $n_p(0)$ will asymptotically approach the function

$$n_p(0) \approx n_{p0}\exp\left(\frac{eV_{BE}}{2kT}\right) \tag{12.47}$$

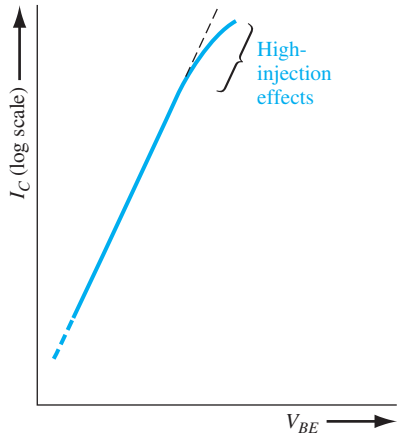


Figure 12.25 | Collector current versus base–emitter voltage showing high-injection effects.

The excess minority carrier concentration in the base, and hence the collector current, will increase at a slower rate with B–E voltage in high injection than low injection. This effect is shown in Figure 12.25. The high-injection effect is very similar to the effect of a series resistance in a pn junction diode.

12.4.3 Emitter Bandgap Narrowing

Another phenomenon affecting the emitter injection efficiency is bandgap narrowing. We have implied from our previous discussion that the emitter injection efficiency factor will continue to increase and approach unity as the ratio of emitter doping to base doping continues to increase. As silicon becomes heavily doped, the discrete donor energy level in an n-type emitter splits into a band of energies. The distance between donor atoms decreases as the concentration of impurity donor atoms increases, and the splitting of the donor level is caused by the interaction of donor atoms with each other. As the doping continues to increase, the donor band widens, becomes skewed, and moves up toward the conduction band, eventually merging with it. At this point, the effective bandgap energy has decreased. Figure 12.26 shows a plot of the change in the bandgap energy with impurity doping concentration.

A reduction in the bandgap energy increases the intrinsic carrier concentration. The intrinsic carrier concentration is given by

$$n_i^2 = N_c N_v \exp\left(\frac{-E_g}{kT}\right) \tag{12.48}$$

In a heavily doped emitter, the intrinsic carrier concentration can be written as

$$n_{iE}^2 = N_c N_v \exp\left[\frac{-(E_{g0} - \Delta E_g)}{kT}\right] = n_i^2 \exp\left(\frac{\Delta E_g}{kT}\right) \tag{12.49}$$

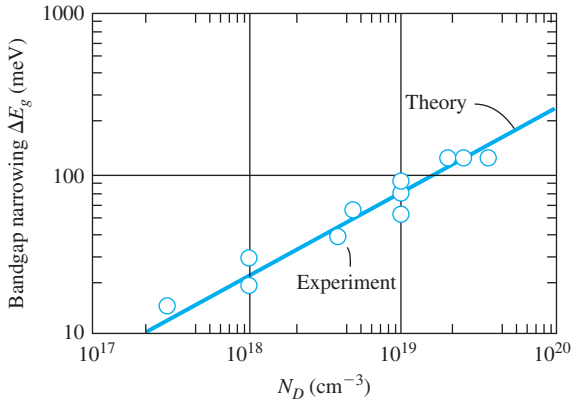


Figure 12.26 | Bandgap narrowing factor versus donor impurity concentration in silicon.
(From Sze [19].)

where E_{g0} is the bandgap energy at a low doping concentration and ΔE_g is the bandgap narrowing factor.

The emitter injection efficiency factor is given by Equation (12.35) as

$$\gamma = \frac{1}{1 + \dfrac{p_{E0} D_E L_B}{n_{B0} D_B L_E} \cdot \dfrac{\tanh\,(x_B/L_B)}{\tanh\,(x_E/L_E)}}$$

The term p'_{E0} is the thermal-equilibrium minority carrier concentration in the emitter, taking into account bandgap narrowing, and can be written as

$$p'_{E0} = \frac{n_{iE}^2}{N_E} = \frac{n_i^2}{N_E} \exp\left(\frac{\Delta E_g}{kT}\right) \tag{12.50}$$

As the emitter doping concentration increases, ΔE_g increases; thus, p'_{E0} does not continue to decrease with increasing emitter doping N_E. If p'_{E0} starts to increase because of the bandgap narrowing, the emitter injection efficiency begins to fall off instead of continuing to increase with increased emitter doping.

EXAMPLE 12.9

Objective: Determine the increase in p_{E0} in the emitter due to bandgap narrowing.

Consider a silicon emitter at $T = 300$ K. Assume the emitter doping increases from 10^{18} to 10^{19} cm^{-3}. Determine the new value of p'_{E0} and determine the ratio p'_{E0}/p_{E0}.

■ Solution

For emitter doping concentrations of $N_E = 10^{18}$ and 10^{19} cm^{-3}, we have, neglecting bandgap narrowing,

$$p_{E0} = \frac{n_i^2}{N_E} = \frac{(1.5 \times 10^{10})^2}{10^{18}} = 2.25 \times 10^2 \text{ cm}^{-3}$$

and

$$p_{E0} = \frac{n_i^2}{N_E} = \frac{(1.5 \times 10^{10})^2}{10^{19}} = 2.25 \times 10^1 \text{ cm}^{-3}$$

Taking into account the bandgap narrowing effect shown in Figure 12.26, we obtain, respectively, for $N_E = 10^{18}$ and 10^{19} cm^{-3},

$$p'_{E0} = \frac{n_i^2}{N_E}\exp\left(\frac{\Delta E_g}{kT}\right) = \frac{(1.5 \times 10^{10})^2}{10^{18}}\exp\left(\frac{0.020}{0.0259}\right) = 4.87 \times 10^2 \text{ cm}^{-3}$$

and

$$p'_{E0} = \frac{(1.5 \times 10^{10})^2}{10^{19}}\exp\left(\frac{0.080}{0.0259}\right) = 4.94 \times 10^2 \text{ cm}^{-3}$$

Taking the ratio of p'_{E0}/p_{E0} for $N_E = 10^{18}$ cm^{-3}, we find

$$\frac{p'_{E0}}{p_{E0}} = \exp\left(\frac{\Delta E_g}{kT}\right) = \exp\left(\frac{0.020}{0.0259}\right) = 2.16$$

and for $N_E = 10^{19}$ cm^{-3}, we find

$$\frac{p'_{E0}}{p_{E0}} = \exp\left(\frac{0.080}{0.0259}\right) = 21.95$$

Comment

If the emitter doping concentration increases from 10^{18} to 10^{19} cm^{-3}, the thermal-equilibrium minority carrier concentration actually increases rather than decreasing by a factor of 10 as would be expected.

EXERCISE PROBLEM

Ex 12.9 Determine the thermal-equilibrium minority carrier concentration for an emitter doping concentration of $N_E = 10^{20}$ cm^{-3} taking into account bandgap narrowing.

(Ans. $p_{E0} = 2.25$ cm^{-3}, $p'_{E0} = 1.618 \times 10^4$ cm^{-3})

As the emitter doping increases, the bandgap narrowing factor, ΔE_g, will increase; this can actually cause p_{E0} to increase. As p_{E0} increases, the emitter injection efficiency decreases; this then causes the transistor gain to decrease, as shown in Figure 12.24. A very high emitter doping may result in a smaller current gain than we anticipate because of the bandgap narrowing effect.

12.4.4 Current Crowding

It is tempting to neglect the effects of base current in a transistor since the base current is usually much smaller than either the collector or the emitter current. Figure 12.27 is a cross section of an npn transistor showing the lateral distribution of base current. The base region is typically less than a micrometer thick, so there can be a sizable base resistance. The nonzero base resistance results in a lateral potential difference under the emitter region. For the npn transistor, the potential decreases from the edge of the emitter toward the center. The emitter is highly doped, so as a first approximation the emitter can be considered an equipotential region.

The number of electrons from the emitter injected into the base is exponentially dependent on the B–E voltage. With the lateral voltage drop in the base between the edge and center of the emitter, more electrons will be injected near the emitter edges than in the center, causing the emitter current to be crowded toward the edges. This

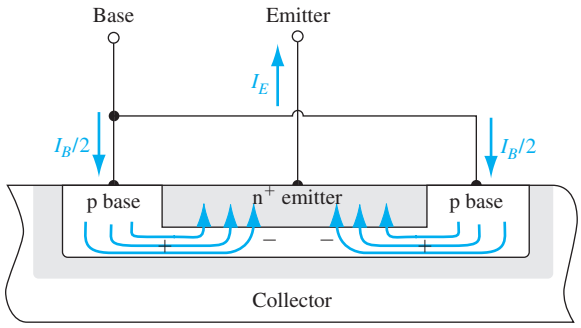


Figure 12.27 | Cross section of an npn bipolar transistor showing the base current distribution and the lateral potential drop in the base region.

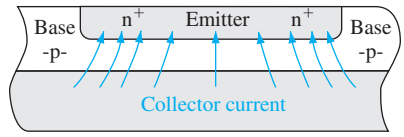


Figure 12.28 | Cross section of an npn bipolar transistor showing the emitter current crowding effect.

current crowding effect is schematically shown in Figure 12.28. The larger current density near the emitter edge may cause localized heating effects as well as localized high-injection effects. The nonuniform emitter current also results in a nonuniform lateral base current under the emitter. A two-dimensional analysis would be required to calculate the actual potential drop versus distance because of the nonuniform base current. Another approach is to slice the transistor into a number of smaller parallel transistors and to lump the resistance of each base section into an equivalent external resistance.

Power transistors, designed to handle large currents, require large emitter areas to maintain reasonable current densities. To avoid the current crowding effect, these transistors are usually designed with narrow emitter widths and fabricated with an interdigitated design. Figure 12.29 shows the basic geometry. In effect, many narrow emitters are connected in parallel to achieve the required emitter area.

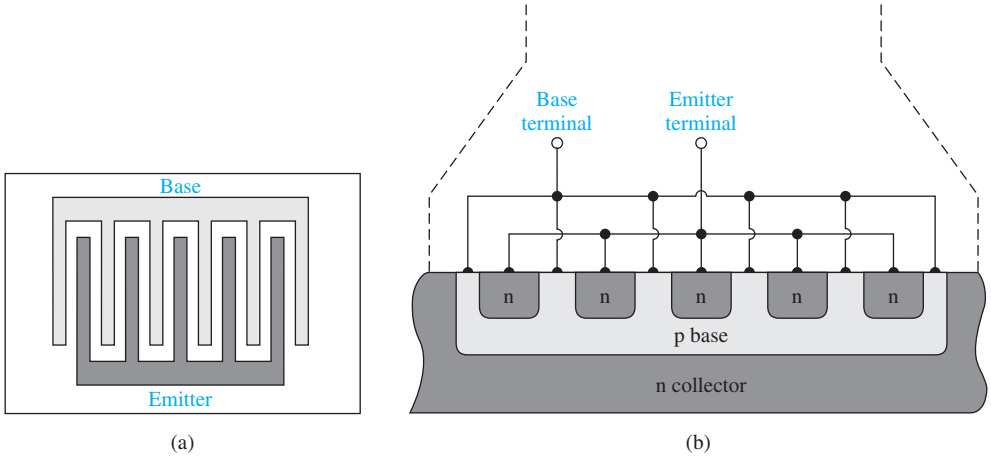


Figure 12.29 | (a) Top view and (b) cross section of an interdigitated npn bipolar transistor structure.

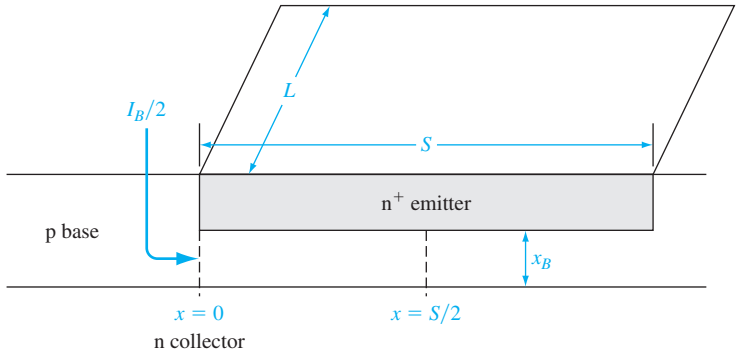


Figure 12.30 | Figure for TYU 12.7.

TEST YOUR UNDERSTANDING

TYU 12.7 Consider the geometry shown in Figure 12.30. The base doping concentration is $N_B = 10^{16}$ cm^{-3}, the neutral base width is $x_B = 0.80$ μm, the emitter width is $S = 10$ μm, and the emitter length is $L = 10$ μm. (*a*) Determine the resistance of the base between $x = 0$ and $x = S/2$. Assume a hole mobility of $\mu_p = 400$ cm^2/V-s. (*b*) If the base current in this region is uniform and given by $I_B/2 = 5$ μA, determine the potential difference between $x = 0$ and $x = S/2$. (*c*) Using the results of part (*b*), what is the ratio of emitter current density at $x = 0$ to that at $x = S/2$? [Ans. (*a*) 9.77 kΩ; (*b*) 48.83 mV; (*c*) 6.59]

*12.4.5 Nonuniform Base Doping

In the analysis of the bipolar transistor, we have assumed uniformly doped regions. However, uniform doping rarely occurs. Figure 12.31 shows a doping profile in a doubly diffused npn transistor. We can start with a uniformly doped n-type substrate, diffuse acceptor atoms from the surface to form a compensated p-type base, and then diffuse donor atoms from the surface to form a doubly compensated n-type emitter. The diffusion process results in a nonuniform doping profile.

We determined in Chapter 5 that a graded impurity concentration leads to an induced electric field. For the p-type base region in thermal equilibrium, we can write

$$J_p = e\mu_p N_a \mathrm{E} - eD_p \frac{dN_a}{dx} = 0 \tag{12.51}$$

Then

$$\mathrm{E} = +\left(\frac{kT}{e}\right)\frac{1}{N_a}\frac{dN_a}{dx} \tag{12.52}$$

According to the example of Figure 12.31, dN_a/dx is negative; hence, the induced electric field is in the negative x direction.

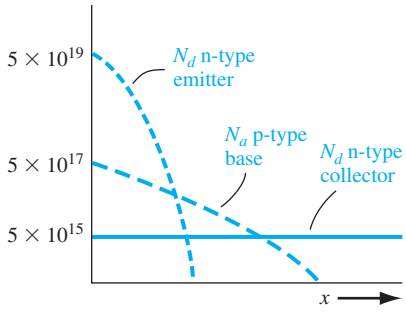


Figure 12.31 | Impurity concentration profiles of a double-diffused npn bipolar transistor.

Electrons are injected from the n-type emitter into the base, and the minority carrier base electrons begin diffusing toward the collector region. The induced electric field in the base, because of the nonuniform doping, produces a force on the electrons in the direction toward the collector. The induced electric field, then, aids the flow of minority carriers across the base region. This electric field is called an *accelerating field.*

The accelerating field will produce a drift component of current that is in addition to the existing diffusion current. Since the minority carrier electron concentration varies across the base, the drift current density will not be constant. The total current across the base, however, is nearly constant. The induced electric field in the base due to nonuniform base doping will alter the minority carrier distribution through the base so that the sum of drift current and diffusion current will be a constant. Calculations have shown that the uniformly doped base theory is very useful for estimating the base characteristics.

12.4.6 Breakdown Voltage

There are two breakdown mechanisms to consider in a bipolar transistor. The first is called punch-through. As the reverse-biased B–C voltage increases, the B–C space charge region widens and extends farther into the neutral base. It is possible for the B–C depletion region to penetrate completely through the base and reach the B–E space charge region, the effect called *punch-through*. Figure 12.32a shows the energy-band diagram of an npn bipolar transistor in thermal equilibrium, and Figure 12.32b shows the energy-band diagram for two values of reverse-biased B–C junction voltage. When a small C–B voltage, V_{R1}, is applied, the B–E potential barrier is not affected; thus, the transistor current is still essentially zero. When a large reverse-biased voltage, V_{R2}, is applied, the depletion region extends through the base region and the B–E potential barrier is lowered because of the C–B voltage. The lowering of the potential barrier at the B–E junction produces a large increase in current with a very small increase in C–B voltage. This effect is the punch-through breakdown phenomenon.

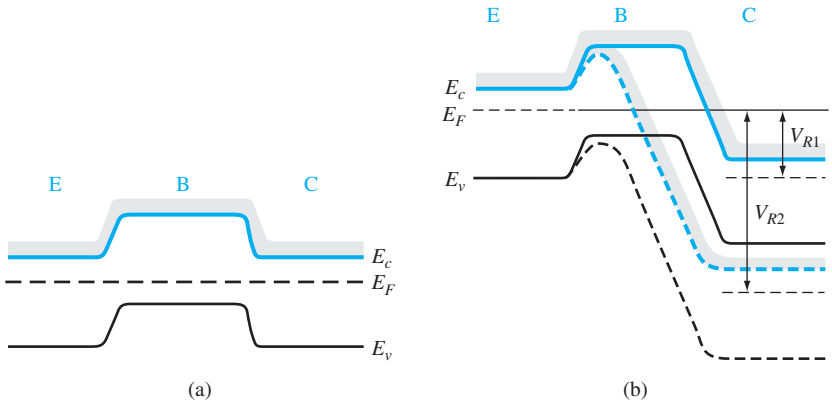


Figure 12.32 | Energy-band diagram of an npn bipolar transistor (a) in thermal equilibrium, and (b) with a reverse-biased B–C voltage before punch-through, V_{R1}, and after punch-through, V_{R2}.

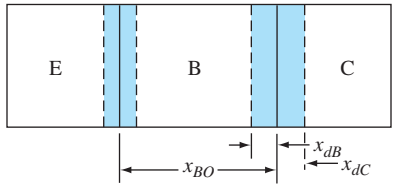


Figure 12.33 | Geometry of a bipolar transistor to calculate the punch-through voltage.

Figure 12.33 shows the geometry for calculating the punch-through voltage. Assume that N_B and N_C are the uniform impurity doping concentrations in the base and collector, respectively. Let x_{BO} be the metallurgical width of the base and let x_{dB} be the space charge width extending into the base from the B–C junction. If we neglect the narrow space charge width of a zero-biased or forward-biased B–E junction, then punch-through, assuming the abrupt junction approximation, occurs when $x_{dB} = x_{BO}$. We can write that

$$x_{dB} = x_{BO} = \left\{\frac{2\epsilon_s(V_{bi} + V_{pt})}{e} \cdot \frac{N_C}{N_B} \cdot \frac{1}{N_C + N_B}\right\}^{1/2} \tag{12.53}$$

where V_{pt} is the reverse-biased B–C voltage at punch-through. Neglecting V_{bi} compared to V_{pt}, we can solve for V_{pt} as

$$\boxed{V_{pt} = \frac{ex_{BO}^2}{2\epsilon_s} \cdot \frac{N_B(N_C + N_B)}{N_C}} \tag{12.54}$$

DESIGN EXAMPLE 12.10

Objective: Design the collector doping concentration and collector width to meet a punch-through voltage specification.

Consider a uniformly doped silicon bipolar transistor with a metallurgical base width of 0.5 μm and a base doping of $N_B = 10^{16}$ cm^{-3}. The punch-through voltage is to be $V_{pt} = 25$ V.

■ Solution

The maximum collector doping concentration can be determined from Equation (12.54) as

$$25 = \frac{(1.6 \times 10^{-19})(0.5 \times 10^{-4})^2(10^{16})(N_C + 10^{16})}{2(11.7)(8.85 \times 10^{-14})N_C}$$

or

$$12.94 = 1 + \frac{10^{16}}{N_C}$$

which yields

$$N_C = 8.38 \times 10^{14} \text{ cm}^{-3}$$

Using this n-type doping concentration for the collector, we can determine the minimum width of the collector region such that the depletion region extending into the collector will not reach the substrate and cause breakdown in the collector region. We have, using the results of Chapter 7,

$$x_{dC} = x_C = 5.97 \ \mu\text{m}$$

■ Comment

From Figure 7.15, the expected avalanche breakdown voltage for this junction is greater than 300 V. Obviously punch-through will occur before the normal breakdown voltage in this case. For a larger punch-through voltage, a larger metallurgical base width will be required, since a lower collector doping concentration is becoming impractical. A larger punch-through voltage will also require a larger collector width in order to avoid premature breakdown in this region.

■ EXERCISE PROBLEM

Ex 12.10 The metallurgical base width of a silicon npn bipolar transistor is $x_{B0} = 0.80$ μm. The base and collector doping concentrations are $N_B = 5 \times 10^{16}$ cm^{-3} and $N_C = 2 \times 10^{15}$ cm^{-3}, respectively. (*a*) Determine the punch-through voltage. (*b*) What is the expected avalanche breakdown voltage?

[Ans. (*a*) $V_{pt} = 643$ V; (*b*) $BV = 180$ V]

The second breakdown mechanism to consider is avalanche breakdown, but taking into account the gain of the transistor.[2] Figure 12.34a is an npn transistor with a reverse-biased voltage applied to the B–C junction and with the emitter left open. The current I_{CBO} is the reverse-biased junction current. Figure 12.34b shows the transistor with an applied C–E voltage and with the base terminal left open. This bias

[2]The doping concentrations in the base and collector of the transistor are small enough that Zener breakdown is not a factor to be considered.

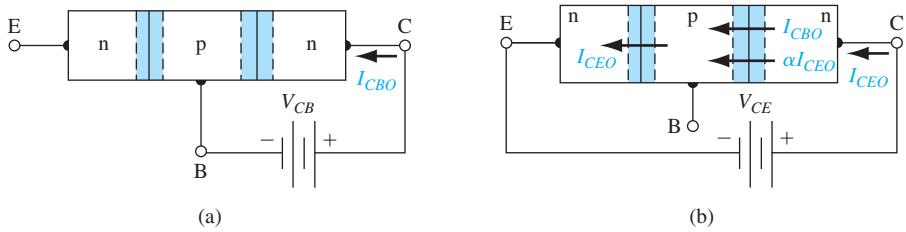


Figure 12.34 | (a) Open-emitter configuration with saturation current I_{CBO}: (b) Open-base configuration with saturation current I_{CEO}.

condition also makes the B–C junction reverse biased. The current in the transistor for this bias configuration is denoted as I_{CEO}.

The current I_{CBO} shown in Figure 12.34b is the normal reverse-biased B–C junction current. Part of this current is due to the flow of minority carrier holes from the collector across the B–C space charge region into the base. The flow of holes into the base makes the base positive with respect to the emitter, and the B–E junction becomes forward biased. The forward-biased B–E junction produces the current I_{CEO}, due primarily to the injection of electrons from the emitter into the base. The injected electrons diffuse across the base toward the B–C junction. These electrons are subject to all of the recombination processes in the bipolar transistor. When the electrons reach the B–C junction, this current component is αI_{CEO} where α is the common-base current gain. We therefore have

$$I_{CEO} = \alpha I_{CEO} + I_{CBO} \tag{12.55a}$$

or

$$I_{CEO} = \frac{I_{CBO}}{1 - \alpha} \approx \beta I_{CBO} \tag{12.55b}$$

where β is the common-emitter current gain. The reverse-biased junction current I_{CBO} is multiplied by the current gain β when the transistor is biased in the open-base configuration.

When the transistor is biased in the open-emitter configuration as in Figure 12.34a, the current I_{CBO} at breakdown becomes $I_{CBO} \rightarrow MI_{CBO}$, where M is the multiplication factor. An empirical approximation for the multiplication factor is usually written as

$$M = \frac{1}{1 - (V_{CB}/BV_{CBO})^n} \tag{12.56}$$

where n is an empirical constant, usually between 3 and 6, and BV_{CBO} is the B–C breakdown voltage with the emitter left open.

When the transistor is biased with the base open circuited as shown in Figure 12.34b, the currents in the B–C junction at breakdown are multiplied, so that

$$I_{CEO} = M(\alpha I_{CEO} + I_{CBO}) \tag{12.57}$$

Solving for I_{CEO}, we obtain

$$I_{CEO} = \frac{MI_{CBO}}{1 - \alpha M} \tag{12.58}$$

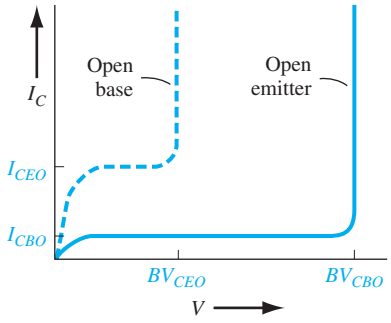


Figure 12.35 | Relative breakdown voltages and saturation currents of the open-base and open-emitter configurations.

The condition for breakdown corresponds to

$$\alpha M = 1 \tag{12.59}$$

Using Equation (12.56) and assuming that $V_{CB} \approx V_{CE}$, Equation (12.59) becomes

$$\frac{\alpha}{1 - (BV_{CEO}/BV_{CBO})^n} = 1 \tag{12.60}$$

where BV_{CEO} is the C–E voltage at breakdown in the open-base configuration. Solving for BV_{CEO}, we find

$$BV_{CEO} = BV_{CBO}\sqrt[n]{1 - \alpha} \tag{12.61}$$

where, again, α is the common-base current gain. The common-emitter and common-base current gains are related by

$$\beta = \frac{\alpha}{1 - \alpha} \tag{12.62a}$$

Normally $\alpha \approx 1$, so that

$$1 - \alpha \approx \frac{1}{\beta} \tag{12.62b}$$

Then Equation (12.61) can be written as

$$\boxed{BV_{CEO} = \frac{BV_{CBO}}{\sqrt[n]{\beta}}} \tag{12.63}$$

The breakdown voltage in the open-base configuration is smaller, by the factor $\sqrt[n]{\beta}$, than the actual avalanche junction breakdown voltage. This characteristic is shown in Figure 12.35.

DESIGN EXAMPLE 12.11

Objective: Design a bipolar transistor to meet a breakdown voltage specification.

Consider a silicon bipolar transistor with a common-emitter current gain of $\beta = 100$ and a base doping concentration of $N_B = 10^{17}$ cm^{-3}. The minimum open-base breakdown voltage is to be 15 V.

■ Solution

From Equation (12.63), the minimum open-emitter junction breakdown voltage must be

$$BV_{CBO} = \sqrt[n]{\beta} BV_{CEO}$$

Assuming the empirical constant n is 3, we find

$$BV_{CBO} = \sqrt[3]{100}(15) = 69.6 \text{ V}$$

From Figure 7.15, the maximum collector doping concentration should be approximately 7×10^{15} cm^{-3} to achieve this breakdown voltage.

■ Comment

In a transistor circuit, the transistor must be designed to operate under a worst-case situation. In this example, the transistor must be able to operate in an open-base configuration without going into breakdown. As we have determined previously, an increase in breakdown voltage can be achieved by decreasing the collector doping concentration.

■ EXERCISE PROBLEM

Ex 12.11 A uniformly doped silicon bipolar transistor has base and collector doping concentrations of $N_B = 7 \times 10^{16}$ cm^{-3} and $N_C = 3 \times 10^{15}$ cm^{-3}, respectively. The common-emitter current gain is $\beta = 125$. Assuming an empirical constant of $n = 3$, determine (*a*) BV_{CBO} and (*b*) BV_{CEO}.

[Ans. (*a*) $BV_{CBO} \cong 125$ V; (*b*) $BV_{CEO} = 25$ V]

TEST YOUR UNDERSTANDING

TYU 12.8 A particular transistor has an output resistance of 200 kΩ and an Early voltage of $V_A = 125$ V. Determine the change in collector current when V_{CE} increases from 2 V to 8 V. (Ans. $\Delta I_C = 30\ \mu$A)

TYU 12.9 (*a*) If, because of fabrication tolerances, the neutral base width for a set of transistors varies over the range of $0.800 \le x_B \le 1.00\ \mu$m, determine the variation in the base transport factor α_T. Assume $L_B = 1.414 \times 10^{-3}$ cm. (*b*) Using the results of part (*a*) and assuming $\gamma = \delta = 0.9967$, what is the variation in common-emitter current gain? [Ans. (*a*) $0.9975 \le \alpha_T \le 0.9984$; (*b*) $109 \le \beta \le 121$]

TYU 12.10 The base impurity doping concentration is $N_B = 3 \times 10^{16}$ cm^{-3} and the metallurgical base width is $x_B = 0.70\ \mu$m. The minimum required punch-through breakdown voltage is specified to be $V_{pt} = 70$ V. What is the maximum allowed collector doping concentration? (Ans. $N_C = 5.81 \times 10^{15}$ cm^{-3})

12.5 | EQUIVALENT CIRCUIT MODELS

In order to analyze a transistor circuit either by hand calculations or using computer codes, one needs a mathematical model, or equivalent circuit, of the transistor. There are several possible models, each one having certain advantages and disadvantages.

A detailed study of all possible models is beyond the scope of this chapter. However, we will consider three equivalent circuit models. Each of these follows directly from the work we have done on the pn junction diode and on the bipolar transistor. Computer analysis of electronic circuits is more commonly used than hand calculations, but it is instructive to consider the types of transistor model used in computer codes.

It is useful to divide bipolar transistors into two categories—switching and amplification—defined by their use in electronic circuits. Switching usually involves turning a transistor from its "off" state, or cutoff, to its "on" state, either forward-active or saturation, and then back to its "off" state. Amplification usually involves superimposing sinusoidal signals on dc values so that bias voltages and currents are only perturbed. The *Ebers–Moll model* is used in switching applications; the *hybrid-pi* model is used in amplification applications.

*12.5.1 Ebers–Moll Model

The Ebers–Moll model, or equivalent circuit, is one of the classic models of the bipolar transistor. This particular model is based on the interacting diode junctions and is applicable in any of the transistor operating modes. Figure 12.36 shows the current directions and voltage polarities used in the Ebers–Moll model. The currents are defined as all entering the terminals so that

$$I_E + I_B + I_C = 0 \tag{12.64}$$

The direction of the emitter current is opposite to what we have considered up to this point, but as long as we are consistent in the analysis, the defined direction does not matter.

The collector current can be written in general as

$$I_C = \alpha_F I_F - I_R \tag{12.65a}$$

where α_F is the common-base current gain in the forward-active mode. In this mode, Equation (12.65a) becomes

$$I_C = \alpha_F I_F + I_{CS} \tag{12.65b}$$

where the current I_{CS} is the reverse-biased B–C junction current. The current I_F is given by

$$I_F = I_{ES}\left[\exp\left(\frac{eV_{BE}}{kT}\right) - 1\right] \tag{12.66}$$

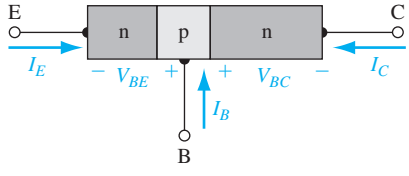


Figure 12.36 | Current direction and voltage polarity definitions for the Ebers–Moll model.

If the B–C junction becomes forward biased, such as in saturation, then we can write the current I_R as

$$I_R = I_{CS}\left[\exp\left(\frac{eV_{BC}}{kT}\right) - 1\right] \tag{12.67}$$

Using Equations (12.66) and (12.67), the collector current from Equation (12.65a) can be written as

$$I_C = \alpha_F I_{ES}\left[\exp\left(\frac{eV_{BE}}{kT}\right) - 1\right] - I_{CS}\left[\exp\left(\frac{eV_{BC}}{kT}\right) - 1\right] \tag{12.68}$$

We can also write the emitter current as

$$I_E = \alpha_R I_R - I_F \tag{12.69}$$

or

$$I_E = \alpha_R I_{CS}\left[\exp\left(\frac{eV_{BC}}{kT}\right) - 1\right] - I_{ES}\left[\exp\left(\frac{eV_{BE}}{kT}\right) - 1\right] \tag{12.70}$$

The current I_{ES} is the reverse-biased B–E junction current and α_R is the common-base current gain for the inverse-active mode. Equations (12.68) and (12.70) are the classic Ebers–Moll equations.

Figure 12.37 shows the equivalent circuit corresponding to Equations (12.68) and (12.70). The current sources in the equivalent circuit represent current components that depend on voltages across other junctions. The Ebers–Moll model has four parameters: α_F, α_R, I_{ES}, and I_{CS}. However, only three parameters are independent. The reciprocity relationship states that

$$\alpha_F I_{ES} = \alpha_R I_{CS} \tag{12.71}$$

Since the Ebers–Moll model is valid in each of the four operating modes, we can, for example, use the model for the transistor in saturation. In the saturation mode, both B–E and B–C junctions are forward biased, so that $V_{BE} > 0$ and $V_{BC} > 0$. The B–E voltage will be a known parameter since we will apply a voltage across this junction. The forward-biased B–C voltage is a result of driving the transistor into saturation and is the unknown to be determined from the Ebers–Moll equations.

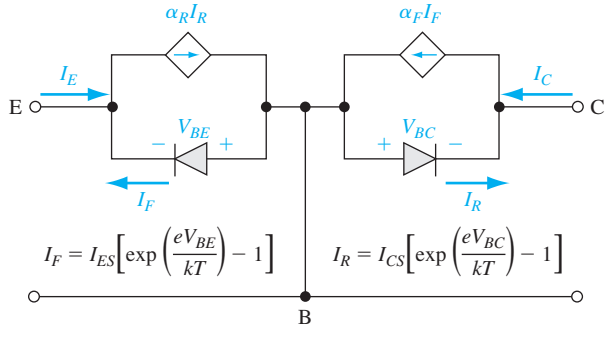


Figure 12.37 | Basic Ebers–Moll equivalent circuit.

Normally in electronic circuit applications, the collector–emitter voltage at saturation is of interest. We can define the C–E saturation voltage as

$$V_{CE}(\text{sat}) = V_{BE} - V_{BC} \tag{12.72}$$

We find an expression for $V_{CE}(\text{sat})$ by combining the Ebers–Moll equations. In the following example, we see how the Ebers–Moll equations can be used in a hand calculation, and we may also see how a computer analysis would make the calculations easier.

Combining Equations (12.64) and (12.70), we have

$$-(I_B + I_C) = \alpha_R I_{CS}\left[\exp\left(\frac{eV_{BC}}{kT}\right) - 1\right] - I_{ES}\left[\exp\left(\frac{eV_{BE}}{kT}\right) - 1\right] \tag{12.73}$$

If we solve for $[\exp(eV_{BC}/kT) - 1]$ from Equation (12.73), and substitute the resulting expression into Equation (12.68), we can then find V_{BE} as

$$V_{BE} = V_t \ln\left[\frac{I_C(1-\alpha_R) + I_B + I_{ES}(1-\alpha_F\alpha_R)}{I_{ES}(1-\alpha_F\alpha_R)}\right] \tag{12.74}$$

where V_t is the thermal voltage. Similarly, if we solve for $[\exp(eV_{BE}/kT) - 1]$ from Equation (12.68), and substitute this expression into Equation (12.73), we can find V_{BC} as

$$V_{BC} = V_t \ln\left[\frac{\alpha_F I_B - (1-\alpha_F)I_C + I_{CS}(1-\alpha_F\alpha_R)}{I_{CS}(1-\alpha_F\alpha_R)}\right] \tag{12.75}$$

We may neglect the I_{ES} and I_{CS} terms in the numerators of Equations (12.74) and (12.75). Solving for $V_{CE}(\text{sat})$, we have

$$V_{CE}(\text{sat}) = V_{BE} - V_{CB} = V_t \ln\left[\frac{I_C(1-\alpha_R) + I_B}{\alpha_F I_B - (1-\alpha_F)I_C} \cdot \frac{I_{CS}}{I_{ES}}\right] \tag{12.76}$$

The ratio of I_{CS} to I_{ES} can be written in terms of α_F and α_R from Equation (12.71). We can finally write

$$V_{CE}(\text{sat}) = V_t \ln\left[\frac{I_C(1-\alpha_R) + I_B}{\alpha_F I_B - (1-\alpha_F)I_C} \cdot \frac{\alpha_F}{\alpha_R}\right] \tag{12.77}$$

EXAMPLE 12.12

Objective: Calculate the collector–emitter saturation voltage of a bipolar transistor at $T = 300$ K.

Assume that $\alpha_F = 0.99$, $\alpha_R = 0.20$, $I_C = 1$ mA, and $I_B = 50\ \mu$A.

■ Solution

Substituting the parameters into Equation (12.77), we have

$$V_{CE}(\text{sat}) = (0.0259)\ln\left[\frac{(1)(1-0.2) + (0.05)}{(0.99)(0.05) - (1-0.99)(1)}\left(\frac{0.99}{0.20}\right)\right] = 0.121 \text{ V}$$

■ Comment

This $V_{CE}(\text{sat})$ value is typical of collector–emitter saturation voltages. Because of the log function, $V_{CE}(\text{sat})$ is not a strong function of I_C or I_B.

EXERCISE PROBLEM

Ex 12.12 Repeat Example 12.12 for transistor parameters of $\alpha_F = 0.992$, $\alpha_R = 0.05$, $I_C = 0.5$ mA, and $I_B = 50\ \mu$A. (Ans. $V_{CE} = 0.141$ V)

12.5.2 Gummel–Poon Model

The Gummel–Poon model of the BJT considers more physics of the transistor than the Ebers–Moll model. This model can be used if, for example, there is a nonuniform doping concentration in the base.

The electron current density in the base of an npn transistor can be written as

$$J_n = e\mu_n n(x)\mathrm{E} + eD_n\frac{dn(x)}{dx} \tag{12.78}$$

An electric field will occur in the base if nonuniform doping exists in the base. This is discussed in Section 12.4.5. The electric field, from Equation (12.52), can be written in the form

$$\mathrm{E} = \frac{kT}{e}\cdot\frac{1}{p(x)}\cdot\frac{dp(x)}{dx} \tag{12.79}$$

where $p(x)$ is the majority carrier hole concentration in the base. Under low injection, the hole concentration is just the acceptor impurity concentration. With the doping profile shown in Figure 12.31, the electric field is negative (from the collector to the emitter). The direction of this electric field aids the flow of electrons across the base.

Substituting Equation (12.79) into Equation (12.78), we obtain

$$J_n = e\mu_n n(x)\cdot\frac{kT}{e}\cdot\frac{1}{p(x)}\cdot\frac{dp(x)}{dx} + eD_n\frac{dn(x)}{dx} \tag{12.80}$$

Using Einstein's relation, we can write Equation (12.80) in the form

$$J_n = \frac{eD_n}{p(x)}\left[n(x)\frac{dp(x)}{dx} + p(x)\frac{dn(x)}{dx}\right] = \frac{eD_n}{p(x)}\cdot\frac{d(pn)}{dx} \tag{12.81}$$

Equation (12.81) can be written in the form

$$\frac{J_n p(x)}{eD_n} = \frac{d(pn)}{dx} \tag{12.82}$$

Integrating Equation (12.82) through the base region while assuming that the electron current density is essentially a constant and the diffusion coefficient is a constant, we find

$$\frac{J_n}{eD_n}\int_0^{x_B} p(x)\,dx = \int_0^{x_B}\frac{dp(x)}{dx}\,dx = p(x_B)n(x_B) - p(0)n(0) \tag{12.83}$$

Assuming that the B–E junction is forward biased and the B–C junction is reverse biased, we have $n(0) = n_{B0}\exp(V_{BE}/V_t)$ and $n(x_B) = 0$. We may note that $n_{B0}p = n_i^2$ so that Equation (12.83) can be written as

$$J_n = \frac{-eD_n n_i^2\exp(V_{BE}/V_t)}{\displaystyle\int_0^{x_B} p(x)\,dx} \tag{12.84}$$

The integral in the denominator is the total majority carrier charge in the base and is known as the *base Gummel number,* defined as Q_B.

If we perform the same analysis in the emitter, we find that the hole current density in the emitter of an npn transistor can be expressed as

$$J_p = \frac{-eD_p n_i^2 \exp(V_{BE}/V_t)}{\displaystyle\int_0^{x_E} n(x')\,dx'} \tag{12.85}$$

The integral in the denominator is the total majority carrier charge in the emitter and is known as the emitter Gummel number, defined as Q_E.

Since the currents in the Gummel–Poon model are functions of the total integrated charges in the base and emitter, these currents can easily be determined for nonuniformly doped transistors.

The Gummel–Poon model can also take into account nonideal effects, such as the Early effect and high-level injection. As the B–C voltage changes, the neutral base width changes so that the base Gummel number Q_B changes. The change in Q_B with B–C voltage then makes the electron current density given by Equation (12.84) a function of the B–C voltage. This is the base width modulation effect or Early effect as discussed previously in Section 12.4.1.

If the B–E voltage becomes too large, low injection no longer applies, which leads to high-level injection. In this case, the total hole concentration in the base increases because of the increased excess hole concentration. This means that the base Gummel number will increase. The change in base Gummel number implies, from Equation (12.84), that the electron current density will also change. High-level injection has also been previously discussed in Section 12.4.2.

The Gummel–Poon model can then be used to describe the basic operation of the transistor as well as to describe nonideal effects.

12.5.3 Hybrid-Pi Model

Bipolar transistors are commonly used in circuits that amplify time-varying or sinusoidal signals. In these linear amplifier circuits, the transistor is biased in the forward-active region and small sinusoidal voltages and currents are superimposed on dc voltages and currents. In these applications, the sinusoidal parameters are of interest, so it is convenient to develop a small-signal equivalent circuit of the bipolar transistor using the small-signal admittance parameters of the pn junction developed in Chapter 8.

Figure 12.38a shows an npn bipolar transistor in a common-emitter configuration with the small-signal terminal voltages and currents. Figure 12.38b shows the cross section of the npn transistor. The C, B, and E terminals are the external connections to the transistor, while the C′, B′, and E′ points are the idealized internal collector, base, and emitter regions.

We can begin constructing the equivalent circuit of the transistor by considering the various terminals individually. Figure 12.39a shows the equivalent circuit between the external input base terminal and the external emitter terminal. The resistance r_b is the series resistance in the base between the external base terminal B and the internal base region B′. The B′−E′ junction is forward biased, so C_π is

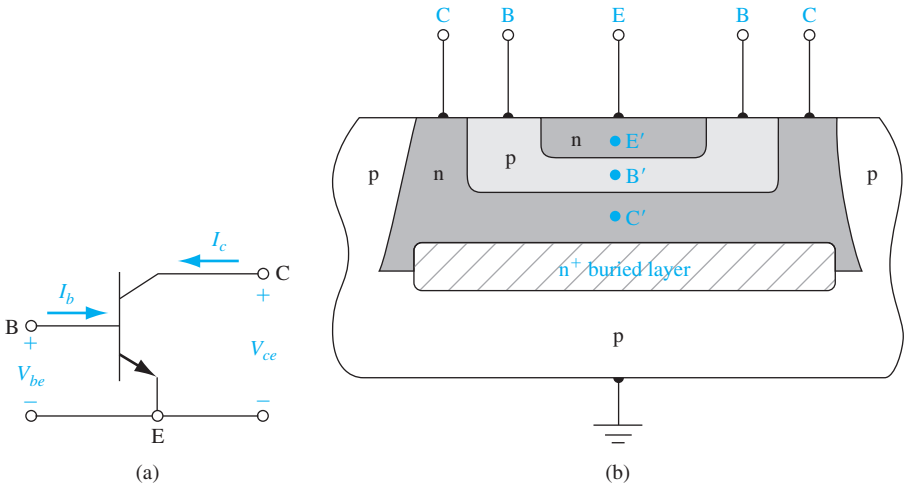


Figure 12.38 | (a) Common-emitter npn bipolar transistor with small-signal current and voltages. (b) Cross section of an npn bipolar transistor for the hybrid-pi model.

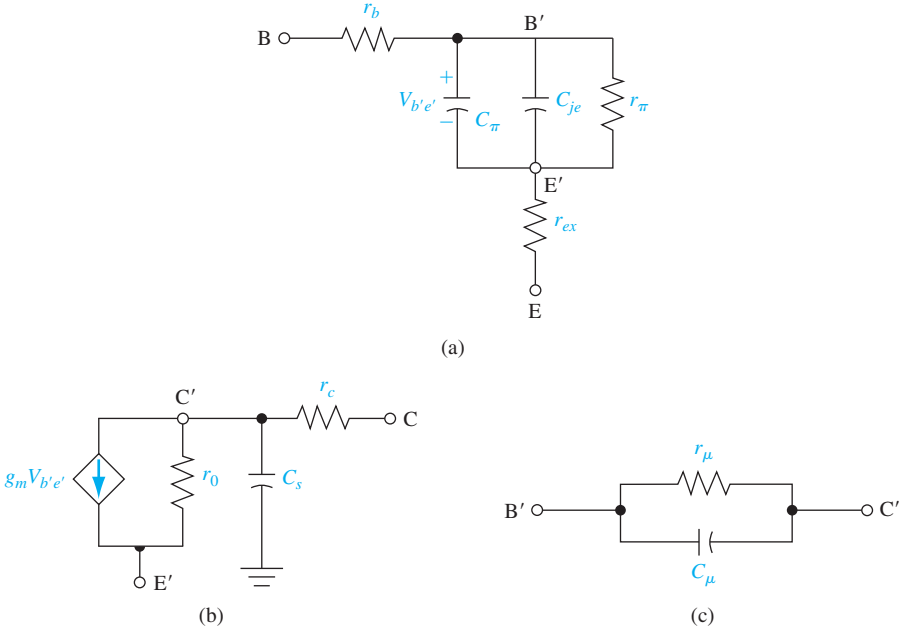


Figure 12.39 | Components of the hybrid-pi equivalent circuit between (a) the base and emitter, (b) the collector and emitter, and (c) the base and collector.

the junction diffusion capacitance and r_π is the junction diffusion resistance. The diffusion capacitance C_π is the same as the diffusion capacitance C_d given by Equation (8.105), and the diffusion resistance r_π is the same as the diffusion resistance r_d given by Equation (8.68). The values of both parameters are functions of the junction

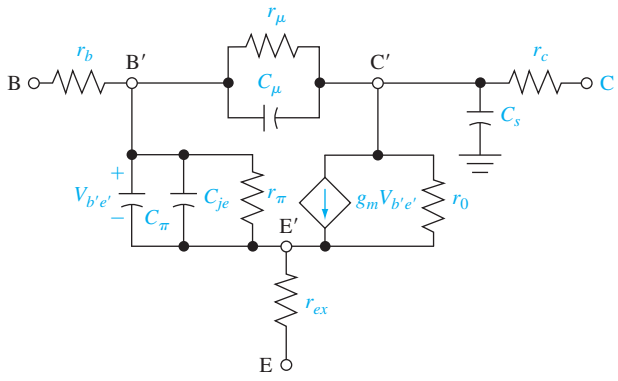


Figure 12.40 | Hybrid-pi equivalent circuit.

current. These two elements are in parallel with the junction capacitance, which is C_{je}. Finally, r_{ex} is the series resistance between the external emitter terminal and the internal emitter region. This resistance is usually very small and may be on the order of 1 to 2 Ω.

Figure 12.39b shows the equivalent circuit looking into the collector terminal. The r_c resistance is the series resistance between the external and internal collector connections and the capacitance C_s is the junction capacitance of the reverse-biased collector-substrate junction. The dependent current source, $g_m V_{b'e'}$, is the collector current in the transistor, which is controlled by the internal base–emitter voltage. The resistance r_0 is the inverse of the output conductance g_0 and is primarily due to the Early effect.

Finally, Figure 12.39c shows the equivalent circuit of the reverse-biased B′−C′ junction. The C_μ parameter is the reverse-biased junction capacitance and r_μ is the reverse-biased diffusion resistance. Normally, r_μ is on the order of megohms and can be neglected. The value of C_μ is usually much smaller than C_π but, because of the feedback effect that leads to the Miller effect and Miller capacitance, C_μ cannot be ignored in most cases. The Miller capacitance is the equivalent capacitance between B′ and E′ due to C_μ and the feedback effect, which includes the gain of the transistor. The Miller effect also reflects C_μ between the C′ and E′ terminals at the output. However, the effect on the output characteristics can usually be ignored.

Figure 12.40 shows the complete hybrid-pi equivalent circuit. A computer simulation is usually required for this complete model because of the large number of elements. However, some simplifications can be made in order to gain an appreciation for the frequency effects of the bipolar transistor. The capacitances lead to frequency effects in the transistor, which means that the gain, for example, is a function of the input signal frequency.

EXAMPLE 12.13

Objective: Determine, to a first approximation, the frequency at which the small-signal current gain decreases to $1/\sqrt{2}$ of its low-frequency value.

Consider the simplified hybrid-pi circuit shown in Figure 12.41. We are ignoring C_μ, C_s, r_μ, C_{je}, r_0, and the series resistances. We must emphasize that this is a first-order calculation and that C_μ normally cannot be neglected.

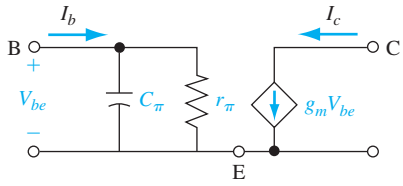


Figure 12.41 | Simplified hybrid-pi equivalent circuit.

■ Solution

At very low frequency, we may neglect C_π so that

$$V_{be} = I_b\, r_\pi \qquad \text{and} \qquad I_c = g_m V_{be} = g_m\, r_\pi\, I_b$$

We can then write

$$h_{fe0} = \frac{I_c}{I_b} = g_m r_\pi$$

where h_{fe0} is the low-frequency, small-signal common-emitter current gain.

Taking into account C_π, we have

$$V_{be} = I_b \left(\frac{r_\pi}{1 + j\omega r_\pi C_\pi} \right)$$

Then

$$I_c = g_m V_{be} = I_b \left(\frac{h_{fe0}}{1 + j\omega r_\pi C_\pi} \right)$$

or the small-signal current gain can be written as

$$A_i = \frac{I_c}{I_b} = \left(\frac{h_{fe0}}{1 + j\omega r_\pi C_\pi} \right)$$

The magnitude of the current gain is then

$$|A_i| = \left| \frac{I_c}{I_b} \right| = \frac{h_{fe0}}{\sqrt{1 + (\omega r_\pi\, C_\pi)^2}} = \frac{h_{fe0}}{\sqrt{1 + (2\pi f r_\pi C_\pi)^2}}$$

The magnitude of the current gain drops to $1/\sqrt{2}$ of its low-frequency value at $f = 1/2\pi r_\pi C_\pi$.

If, for example, $r_\pi = 2.6\ \text{k}\Omega$ and $C_\pi = 4$ pF, then

$$f = 15.3\ \text{MHz}$$

■ Comment

High-frequency transistors must have small-diffusion capacitances, implying the use of small devices.

■ EXERCISE PROBLEM

Ex 12.13 Using the results of Example 12.13, determine the maximum value of C_π such that the frequency at which $|A_i| = h_{fe0}/\sqrt{2}$ is $f = 35$ MHz.

(Ans. $C_\pi = 1.75$ pF)

12.6 | FREQUENCY LIMITATIONS

The hybrid-pi equivalent circuit, developed in the last section, introduces frequency effects through the capacitor–resistor circuits. We now discuss the various physical factors in the bipolar transistor affecting the frequency limitations of the device and then define the transistor cutoff frequency, which is a figure of merit for a transistor.

12.6.1 Time-Delay Factors

The bipolar transistor is a transit-time device. When the voltage across the B–E junction increases, for example, additional carriers from the emitter are injected into the base, diffuse across the base, and are collected in the collector region. As the frequency increases, this transit time can become comparable to the period of the input signal. At this point, the output response will no longer be in phase with the input and the magnitude of the current gain will decrease.

The total emitter-to-collector time constant or delay time is composed of four separate time constants. We can write

$$\tau_{ec} = \tau_e + \tau_b + \tau_d + \tau_c \tag{12.86}$$

where

τ_{ec} = emitter-to-collector time delay
τ_e = emitter–base junction capacitance charging time
τ_b = base transit time
τ_d = collector depletion region transit time
τ_c = collector capacitance charging time

The equivalent circuit of the forward-biased B–E junction is given in Figure 12.39a. The capacitance C_{je} is the junction capacitance. If we ignore the series resistance, then the emitter–base junction capacitance charging time is

$$\boxed{\tau_e = r'_e\,(C_{je} + C_p)} \tag{12.87}$$

where r'_e is the emitter junction or diffusion resistance. The capacitance C_p includes any parasitic capacitance between the base and emitter. The resistance r'_e is found as the inverse of the slope of the I_E versus V_{BE} curve. We obtain

$$r'_e = \frac{kT}{e} \cdot \frac{1}{I_E} \tag{12.88}$$

where I_E is the dc emitter current.

The second term, τ_b, is the base transit time, the time required for the minority carriers to diffuse across the neutral base region. The base transit time is related to the diffusion capacitance C_π of the B–E junction. For the npn transistor, the electron current density in the base can be written as

$$J_n = -en_B(x)v(x) \tag{12.89}$$

where $v(x)$ is an average velocity. We can write

$$v(x) = dx/dt \quad \text{or} \quad dt = dx/v(x) \tag{12.90}$$

The transit time can then be found by integrating, or

$$\tau_b = \int_0^{x_B} dt = \int_0^{x_B} \frac{dx}{v(x)} = \int_0^{x_B} \frac{en_B(x)\,dx}{(-J_n)} \tag{12.91}$$

The electron concentration in the base is approximately linear (see Equation (12.15b)) so we can write

$$n_B(x) \cong n_{B0}\left[\exp\left(\frac{eV_{BE}}{kT}\right)\right]\left(1 - \frac{x}{x_B}\right) \tag{12.92}$$

and the electron current density is given by

$$J_n = eD_n \frac{dn_B(x)}{dx} \tag{12.93}$$

The base transit time is then found by combining Equations (12.92) and (12.93) with Equation (12.91). We find that

$$\boxed{\tau_b = \frac{x_B^2}{2D_n}} \tag{12.94}$$

The third time-delay factor is τ_d, the collector depletion region transit time. Assuming that the electrons in the npn device travel across the B–C space charge region at their saturation velocity, we have

$$\boxed{\tau_d = \frac{x_{dc}}{v_s}} \tag{12.95}$$

where x_{dc} is the B–C space charge width and v_s is the electron saturation velocity.

The fourth time-delay factor, τ_c, is the collector capacitance charging time. The B–C is reverse biased so that the diffusion resistance in parallel with the junction capacitance is very large. The charging time constant is then a function of the collector series resistance r_c. We can write

$$\boxed{\tau_c = r_c(C_\mu + C_s)} \tag{12.96}$$

where C_μ is the B–C junction capacitance and C_s is the collector-to-substrate capacitance. The series resistance in small epitaxial transistors is usually small; thus, the time delay τ_c may be neglected in some cases.

Example calculations of the various time-delay factors are given in the next section as part of the cutoff frequency discussion.

12.6.2 Transistor Cutoff Frequency

The current gain as a function of frequency is developed in Example 12.13 so that we can also write the common-base current gain as

$$\alpha = \frac{\alpha_0}{1 + j\dfrac{f}{f_\alpha}} \tag{12.97}$$

where α_0 is the low-frequency common-base current gain and f_α is defined as the *alpha cutoff frequency*. The frequency f_α is related to the emitter-to-collector time delay τ_{ec} as

$$f_\alpha = \frac{1}{2\pi\tau_{ec}} \tag{12.98}$$

When the frequency is equal to the alpha cutoff frequency, the magnitude of the common-base current gain is $1/\sqrt{2}$ of its low-frequency value.

We can relate the alpha cutoff frequency to the common-emitter current gain by considering

$$\beta = \frac{\alpha}{1-\alpha} \tag{12.99}$$

We may replace α in Equation (12.99) with the expression given by Equation (12.97). When the frequency f is of the same order of magnitude as f_α, then

$$|\beta| = \left|\frac{\alpha}{1-\alpha}\right| \approx \frac{f_\alpha}{f} \tag{12.100}$$

where we have assumed that $\alpha_0 \approx 1$. When the signal frequency is equal to the alpha cutoff frequency, the magnitude of the common-emitter current gain is equal to unity. The usual notation is to define this *cutoff frequency* as f_T, so we have

$$\boxed{f_T = \frac{1}{2\pi\tau_{ec}}} \tag{12.101}$$

From the analysis in Example 12.13, we may also write the common-emitter current gain as

$$\beta = \frac{\beta_0}{1 + j(f/f_\beta)} \tag{12.102}$$

where f_β is called the *beta cutoff frequency* and is the frequency at which the magnitude of the common-emitter current gain β drops to $1/\sqrt{2}$ of its low-frequency value.

Combining Equations (12.99) and (12.97), we can write

$$\beta = \frac{\alpha}{1-\alpha} = \frac{\dfrac{\alpha_0}{1 + j(f/f_T)}}{1 - \dfrac{\alpha_0}{1 + j(f/f_T)}} = \frac{\alpha_0}{1 - \alpha_0 + j(f/f_T)} \tag{12.103}$$

or

$$\beta = \frac{\alpha_0}{(1-\alpha_0)\left[1 + j\dfrac{f}{(1-\alpha_0)f_T}\right]} \approx \frac{\beta_0}{1 + j\dfrac{\beta_0 f}{f_T}} \tag{12.104}$$

where

$$\beta_0 = \frac{\alpha_0}{1-\alpha_0} \approx \frac{1}{1-\alpha_0}$$

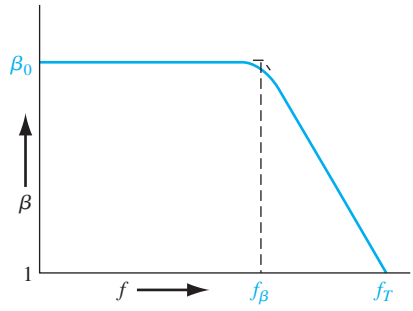


Figure 12.42 | Bode plot of common-emitter current gain versus frequency.

Comparing Equations (12.104) and (12.102), the beta cutoff frequency is related to the cutoff frequency by

$$f_\beta \cong \frac{f_T}{\beta_0} \tag{12.105}$$

Figure 12.42 shows a Bode plot of the common-emitter current gain as a function of frequency and shows the relative values of the beta and cutoff frequencies. Keep in mind that the frequency is plotted on a log scale, so f_β and f_T usually have significantly different values.

EXAMPLE 12.14

Objective: Calculate the emitter-to-collector transit time and the cutoff frequency of a bipolar transistor, with the following parameters.

Consider a silicon npn transistor at $T = 300$ K. Assume the following parameters:

$$\begin{aligned} I_E &= 1 \text{ mA} & C_{je} &= 1 \text{ pF} \\ x_B &= 0.5\ \mu\text{m} & D_n &= 25 \text{ cm}^2/\text{s} \\ x_{dc} &= 2.4\ \mu\text{m} & r_c &= 20\ \Omega \\ C_\mu &= 0.1 \text{ pF} & C_s &= 0.1 \text{ pF} \end{aligned}$$

■ Solution

We will initially calculate the various time-delay factors. If we neglect the parasitic capacitance, the emitter–base junction charging time is

$$\tau_e = r'_e C_{je}$$

where

$$r'_e = \frac{kT}{e} \cdot \frac{1}{I_E} = \frac{0.0259}{1 \times 10^{-3}} = 25.9\ \Omega$$

Then

$$\tau_e = (25.9)(10^{-12}) = 25.9 \text{ ps}$$

The base transit time is

$$\tau_b = \frac{x_B^2}{2D_n} = \frac{(0.5 \times 10^{-4})^2}{2(25)} = 50 \text{ ps}$$

The collector depletion region transit time is

$$\tau_d = \frac{x_{dc}}{v_s} = \frac{2.4 \times 10^{-4}}{10^7} = 24 \text{ ps}$$

The collector capacitance charging time is

$$\tau_c = r_c\,(C_\mu + C_s) = (20)(0.2 \times 10^{-12}) = 4 \text{ ps}$$

The total emitter-to-collector time delay is then

$$\tau_{ec} = 25.9 + 50 + 24 + 4 = 103.9 \text{ ps}$$

so that the cutoff frequency is calculated as

$$f_T = \frac{1}{2\pi\,\tau_{ec}} = \frac{1}{2\pi\,(103.9 \times 10^{-12})} = 1.53 \text{ GHz}$$

If we assume a low-frequency common-emitter current gain of $\beta = 100$, then the beta cutoff frequency is

$$f_\beta = \frac{f_T}{\beta_0} = \frac{1.53 \times 10^9}{100} = 15.3 \text{ MHz}$$

■ Comment

The design of high-frequency transistors requires small device geometries in order to reduce capacitances, and narrow base widths in order to reduce the base transit time.

■ EXERCISE PROBLEM

Ex 12.14 Consider a bipolar transistor with the same parameters as described in Example 12.14 except that $I_E = 50\ \mu$A, $C_{je} = 0.40$ pF, and $C_\mu = 0.05$ pF. Determine the emitter-to-collector transit time, the cutoff frequency, and the beta cutoff frequency.

(Ans. $\tau_{ec} = 282.2$ ps, $f_T = 564$ MHz, $f_\beta = 5.64$ MHz)

12.7 | LARGE-SIGNAL SWITCHING

Switching a transistor from one state to another is strongly related to the frequency characteristics just discussed. However, switching is considered to be a large-signal change, whereas the frequency effects assumed only small changes in the magnitude of the signal.

12.7.1 Switching Characteristics

Consider an npn transistor in the circuit shown in Figure 12.43a switching from cutoff to saturation, and then switching back from saturation to cutoff. We describe the physical processes taking place in the transistor during the switching cycle.

Consider, initially, the case of switching from cutoff to saturation. Assume that in cutoff $V_{BE} \approx V_{BB} < 0$, thus the B–E junction is reverse biased. At $t = 0$, assume that V_{BB} switches to a value of V_{BB0} as shown in Figure 12.43b. We assume that V_{BB0} is sufficiently positive to eventually drive the transistor into saturation. For $0 \le t \le t_1$, the base current supplies charge to bring the B–E junction from reverse bias to a slight forward bias. The space charge width of the B–E junction is narrowing, and ionized donors and acceptors are being neutralized. A small amount of charge is also

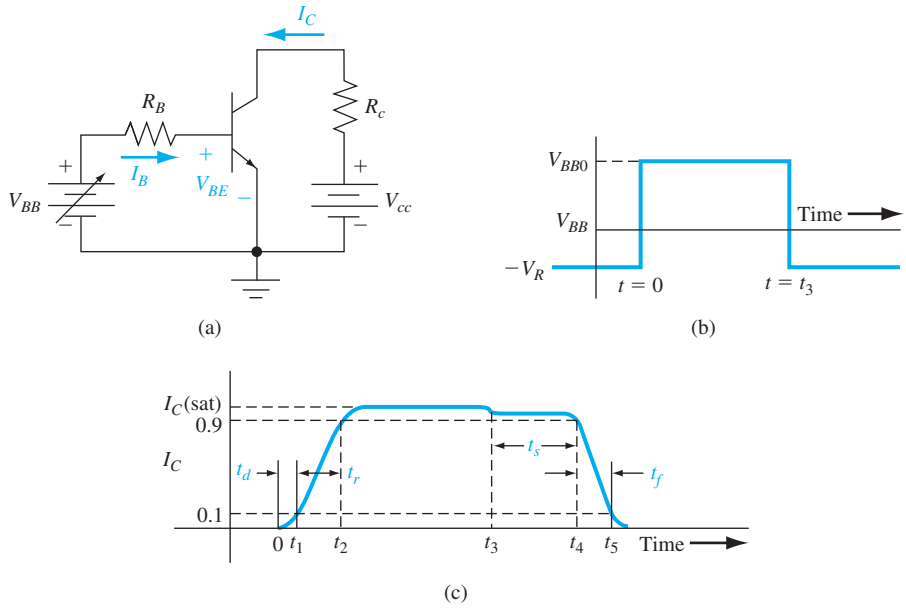


Figure 12.43 | (a) Circuit used for transistor switching. (b) Input base drive for transistor switching. (c) Collector current versus time during transistor switching.

injected into the base during this time. The collector current increases from zero to 10 percent of its final value during this time period, referred to as the delay time.

During the next time period, $t_1 \leq t \leq t_2$, the base current is supplying charge, which increases the B–E junction voltage from near cutoff to near saturation. During this time, additional carriers are being injected into the base so that the gradient of the minority carrier electron concentration in the base increases, causing the collector current to increase. We refer to this time period as the rise time, during which the collector current increases from 10 to 90 percent of the final value. For $t > t_2$, the base drive continues to supply base current, driving the transistor into saturation and establishing the final minority carrier distribution in the device.

The switching of the transistor from saturation to cutoff involves removing all of the excess minority carriers stored in the emitter, base, and collector regions. Figure 12.44 shows the charge storage in the base and collector when the transistor is in saturation. The charge Q_B is the excess charge stored in a forward-active transistor, and Q_{BX} and Q_C are the extra charges stored when the transistor is biased in saturation. At $t = t_3$, the base voltage V_{BB} switches to a negative value of $(-V_R)$. The base current in the transistor reverses direction as was the case in switching a pn junction diode from forward to reverse bias. The reverse base current pulls the excess stored carriers from the emitter and base regions. Initially, the collector current does not change significantly, since the gradient of the minority carrier concentration in the base does not

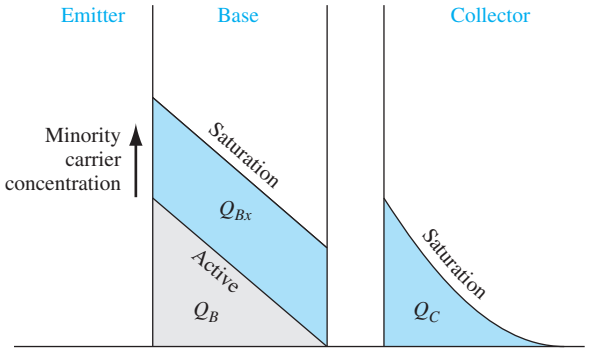


Figure 12.44 | Charge storage in the base and collector at saturation and in the active mode.

change instantaneously. Recall that when the transistor is biased in saturation, both the B–E and B–C junctions are forward biased. The charge Q_{BX} in the base must be removed to reduce the forward-biased B–C voltage to 0 V before the collector current can change. This time delay is called the *storage time* and is denoted by t_s. The storage time is the time between the point at which V_{BB} switches to the time when the collector current is reduced to 90 percent of its maximum saturation value. The storage time is usually the most important parameter in the switching speed of the bipolar transistor.

The final switching delay time is the fall time t_f during which the collector current decreases from the 90 percent to the 10 percent value. During this time, the B–C junction is reverse biased but excess carriers in the base are still being removed, and the B–E junction voltage is decreasing.

The switching-time response of the transistor can be determined by using the Ebers–Moll model. The frequency-dependent gain parameters must be used, and normally the Laplace transform technique is used to obtain the time response. The details of this analysis are quite tedious and are presented here.

12.7.2 The Schottky-Clamped Transistor

One method frequently employed to reduce the storage time and increase the switching speed is the use of a Schottky-clamped transistor. This is a normal npn bipolar device with a Schottky diode connected between base and collector, as shown in Figure 12.45a. The circuit symbol for the Schottky-clamped transistor is shown in Figure 12.45b. When the transistor is biased in the forward-active mode, the B–C junction is reverse biased; hence, the Schottky diode is reverse biased and effectively out of the circuit. The characteristics of the Schottky-clamped transistor—or simply the Schottky transistor—are those of the normal npn bipolar device.

When the transistor is driven into saturation, the B–C junction becomes forward biased; hence, the Schottky diode also becomes forward biased. We may recall from our discussion in Chapter 9 that the effective turn-on voltage of the Schottky diode is approximately half that of the pn junction. The difference in turn-on voltage means that most of the excess base current is shunted through the Schottky diode and away

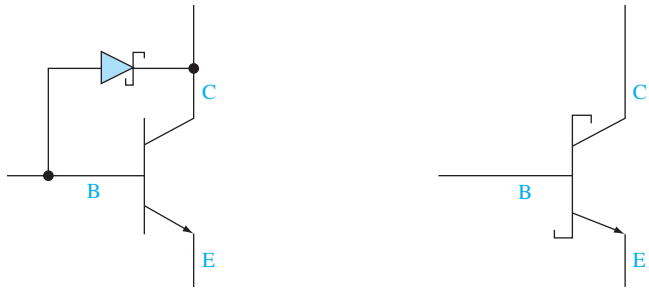


Figure 12.45 | (a) The Schottky-clamped transistor. (b) Circuit symbol of the Schottky-clamped transistor.

from the base so that the amount of excess stored charge in the base and collector is drastically reduced. The excess minority carrier concentration in the base and collector at the B–C junction is an exponential function of V_{BC}. If V_{BC} is reduced from 0.5 to 0.3 V, for example, the excess minority carrier concentration is reduced by over three orders of magnitude. The reduced excess stored charge in the base of the Schottky transistor greatly reduces the storage time—storage times on the order of 1 ns or less are common in Schottky transistors.

*12.8 | OTHER BIPOLAR TRANSISTOR STRUCTURES

This section is intended to briefly introduce three specialized bipolar transistor structures. The first structure is the polysilicon emitter bipolar junction transistor (BJT), the second is the SiGe-base transistor, and the third is the heterojunction bipolar transistor (HBT). The polysilicon emitter BJT is being used in some recent integrated circuits, and the SiGe-base transistor and HBT are intended for high-frequency/high-speed applications.

12.8.1 Polysilicon Emitter BJT

The emitter injection efficiency is degraded by the carriers injected from the base back into the emitter. The emitter width, in general, is thin, which increases speed and reduces parasitic resistance. However, a thin emitter increases the gradient in the minority carrier concentration, as indicated in Figure 12.19. The increase in the gradient increases the B–E junction current, which in turn decreases the emitter injection efficiency and decreases the common-emitter current gain. This effect is also shown in the summary of Table 12.3.

Figure 12.46 shows the idealized cross section of an npn bipolar transistor with a polysilicon emitter. As shown in the figure, there is a very thin n^+ single-crystal silicon region between the p-type base and the n-type polysilicon. As a first approximation to the analysis, we may treat the polysilicon portion of the emitter as low-mobility silicon, which means that the corresponding diffusion coefficient is small.

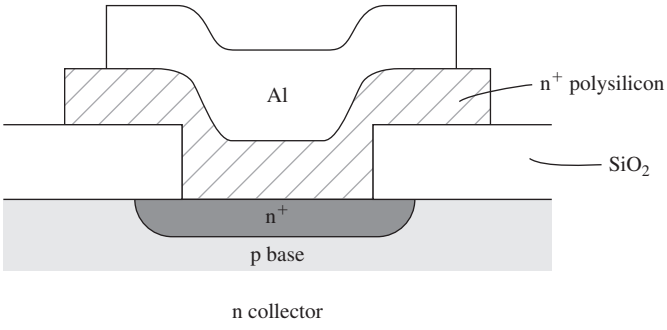


Figure 12.46 | Simplified cross section of an npn polysilicon emitter BJT.

If the neutral widths of both the polysilicon and single-crystal portions of the emitter are much smaller than the respective diffusion lengths, then the minority carrier distribution functions will be linear in each region. Both the minority carrier concentration and diffusion current must be continuous across the polysilicon/silicon interface. We can therefore write

$$eD_{E(\text{poly})}\frac{d(\delta p_{E(\text{poly})})}{dx} = eD_{E(\text{n}^+)}\frac{d(\delta p_{E(\text{n}^+)})}{dx} \tag{12.106a}$$

or

$$\frac{d(\delta p_{E(\text{n}^+)})}{dx} = \frac{D_{E(\text{poly})}}{D_{E(\text{n}^+)}}\cdot\frac{d(\delta p_{E(\text{poly})})}{dx} \tag{12.106b}$$

Since $D_{E(\text{poly})} < D_{E(\text{n}^+)}$, then the gradient of the minority carrier concentration at the emitter edge of the B–E depletion region in the n^+ region is reduced as Figure 12.47 shows. This implies that the current back-injected from the base into the emitter is reduced so that the common-emitter current gain is increased.

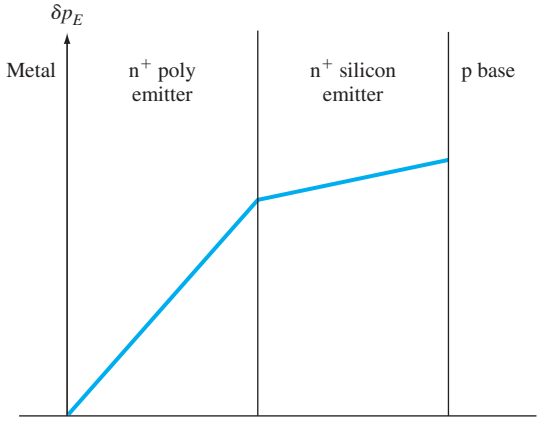


Figure 12.47 | Excess minority carrier hole concentrations in n^+ polysilicon and n^+ silicon emitter.

12.8.2 Silicon–Germanium Base Transistor

The bandgap energy of germanium (Ge) (~0.67 eV) is significantly smaller than the bandgap energy of silicon (Si) (~1.12 eV). By incorporating Ge into Si, the bandgap energy will decrease compared to pure Si. If Ge is incorporated into the base region of a Si bipolar transistor, the decrease in bandgap energy will influence the device characteristics. The desired Ge concentration profile is to have the largest amount of Ge near the base–collector junction and the least amount of Ge near the base–emitter junction. Figure 12.48a shows an ideal uniform boron doping concentration in the p-type base and a linear Ge concentration profile.

The energy bands of a SiGe-base npn transistor compared to a Si-base npn transistor, assuming the boron and Ge concentrations given in Figure 12.48a, are shown in Figure 12.48b. The emitter–base junctions of the two transistors are essentially identical, since the Ge concentration is very small in this region. However, the bandgap energy of the SiGe-base transistor near the base–collector junction is smaller than that of the Si-base transistor. The base current is determined by the base–emitter junction parameters and hence will be essentially the same in the two transistors. This change in bandgap energy will influence the collector current.

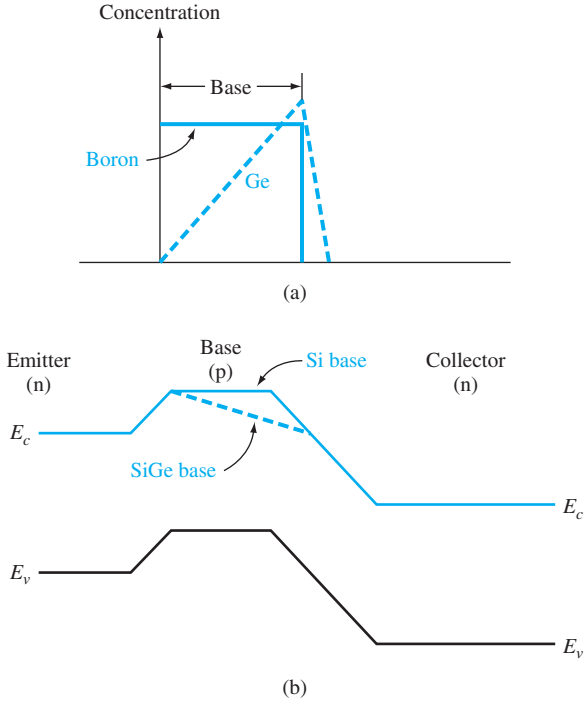


Figure 12.48 | (a) Assumed boron and germanium concentrations in the base of the SiGe-base transistor. (b) Energy-band diagram of the Si- and SiGe-base transistors.

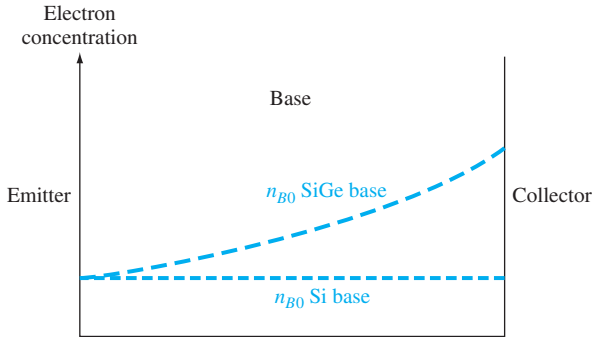


Figure 12.49 | Thermal-equilibrium minority carrier electron concentration through the base of the Si- and SiGe-base transistors.

Collector Current and Current Gain Effects Figure 12.49 shows the thermal-equilibrium minority carrier electron concentration through the base region of the SiGe and Si transistors. This concentration is given by

$$n_{B0} = \frac{n_i^2}{N_B} \tag{12.107}$$

where N_B is assumed to be constant. The intrinsic concentration, however, is a function of the bandgap energy. We may write

$$\frac{n_i^2(\text{SiGe})}{n_i^2(\text{Si})} = \exp\left(\frac{\Delta E_g}{kT}\right) \tag{12.108}$$

where n_i (SiGe) is the intrinsic carrier concentration in the SiGe material, n_i(Si) is the intrinsic carrier concentration in the Si material, and ΔE_g is the change in the bandgap energy of the SiGe material compared to that of Si.

The collector current in a SiGe-base transistor will increase. As a first approximation, we can see this from the previous analysis. The collector current is found from Equation (12.36a), in which the derivative is evaluated at the base–collector junction. This means that the value of n_{B0} in the collector current expression in Equation (12.37) is the value at the base–collector junction. Since this value is larger for the SiGe-base transistor (Figure 12.49), the collector current will be larger compared to the Si-base transistor. Since the base currents are the same in the two transistors, the increase in collector current then implies that the current gain in the SiGe-base transistor is larger. If the bandgap narrowing is 100 meV, then the increase in the collector current and current gain will be approximately a factor of 4.

Early Voltage Effects The Early voltage in a SiGe-base transistor is larger than that of the Si-base transistor. The explanation for this effect is less obvious than the explanation for the increase in collector current and current gain. For a bandgap narrowing of 100 meV, the Early voltage is increased by approximately a factor of 12. Incorporating Ge into the base region can increase the Early voltage by a large factor.

Base Transit Time and Emitter–Base Charging Time Effects The decrease in bandgap energy from the base–emitter junction to the base–collector junction induces an electric field in the base that helps accelerate electrons across the p-type base region. For a bandgap narrowing of 100 meV, the induced electric field can be on the order of 10^3 to 10^4 V/cm. This electric field reduces the base transit time by approximately a factor of 2.5.

The emitter–base junction charging time constant, given by Equation (12.87), is directly proportional to the emitter diffusion resistance r'_e. This parameter is inversely proportional to the emitter current, as seen in Equation (12.88). For a given base current, the emitter current in the SiGe-base transistor is larger, since the current gain is larger. The emitter–base junction charging time is then smaller in a SiGe-base transistor than that in a Si-base transistor.

The reduction in both the base transit time and the emitter–base charging time increases the cutoff frequency of the SiGe-base transistor. The cutoff frequency of these devices can be substantially higher than that of the Si-base device.

12.8.3 Heterojunction Bipolar Transistors

As mentioned previously, one of the basic limitations of the current gain in the bipolar transistor is the emitter injection efficiency. The emitter injection efficiency γ can be increased by reducing the value of the thermal-equilibrium minority carrier concentration p_{E0} in the emitter. However, as the emitter doping increases, the bandgap narrowing effect offsets any improvement in the emitter injection efficiency. One possible solution is to use a wide-bandgap material for the emitter, which will minimize the injection of carriers from the base back into the emitter.

Figure 12.50a shows a discrete aluminum gallium arsenide (AlGaAs)/gallium arsenide (GaAs) heterojunction bipolar transistor, and Figure 12.50b shows the energy-band diagram of the n-AlGaAs emitter to p-GaAs base junction. The large potential barrier V_h limits the number of holes that will be injected back from the base into the emitter.

The intrinsic carrier concentration is a function of bandgap energy as

$$n_i^2 \propto \exp\left(\frac{-E_g}{kT}\right)$$

For a given emitter doping, the number of minority carrier holes injected into the emitter is reduced by a factor of

$$\exp\left(\frac{\Delta E_g}{kT}\right)$$

in changing from a narrow- to wide-bandgap emitter. If $\Delta E_g = 0.30$ eV, for example, n_i^2 would be reduced by approximately 10^5 at $T = 300$ K. The drastic reduction in n_i^2 for the wide-bandgap emitter means that the requirements of a very high emitter doping can be relaxed and a high emitter injection efficiency can still be obtained. A lower emitter doping reduces the bandgap narrowing effect.

The heterojunction GaAs bipolar transistor has the potential of being a very high-frequency device. A lower emitter doping in the wide-bandgap emitter leads to

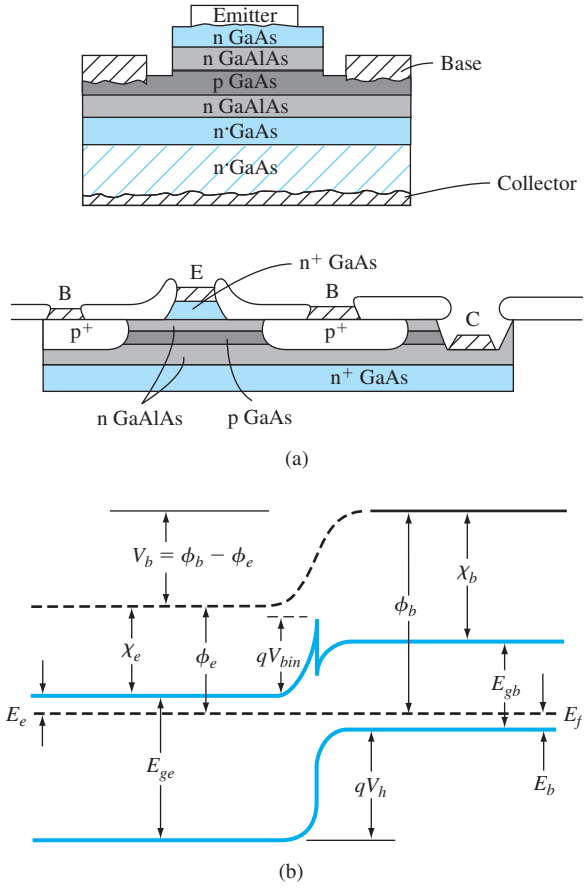


Figure 12.50 | (a) Cross section of AlGaAs/GaAs heterojunction bipolar transistor showing a discrete and integrated structure. (b) Energy-band diagram of the n-AlGaAs emitter and p-GaAs base junction.
(From Tiwari et al. [20].)

a smaller junction capacitance, increasing the speed of the device. Also, for the GaAs npn device, the minority carriers in the base are electrons with a high mobility. The electron mobility in GaAs is approximately five times that in silicon; thus, the base transit time in the GaAs base is very short. Experimental AlGaAs/GaAs heterojunction transistors with base widths on the order of 0.1 μm have shown cutoff frequencies on the order of 40 GHz.

One disadvantage of GaAs is the low minority carrier lifetime. The small lifetime is not a factor in the base of a narrow-base device, but results in a larger B–E recombination current, which decreases the recombination factor and reduces the current gain. A current gain of 150 has been reported.

12.9 | SUMMARY

- There are two complementary bipolar transistors—npn and pnp. Each transistor has three separately doped regions and two pn junctions. The center region (base) is very narrow, so the two pn junctions are said to be interacting junctions.
- In the forward-active mode, the B–E junction is forward biased and the B–C junction is reverse biased. Majority carriers from the emitter are injected into the base where they become minority carriers. These minority carriers diffuse across the base into the B–C space charge region where they are swept into the collector.
- When a transistor is biased in the forward-active mode of operation, the current at one terminal of the transistor (collector current) is controlled by the voltage applied across the other two terminals of the transistor (base–emitter voltage). This is the basic transistor action.
- The minority carrier concentrations are determined in each region of the transistor. The principal currents in the device are determined by the diffusion of these minority carriers.
- The common-base current gain, which leads to the common-emitter current gain, is a function of three factors—emitter injection efficiency, base transport factor, and recombination factor. The emitter injection efficiency takes into account carriers from the base that are injected back into the emitter, the base transport factor takes into account recombination in the base region, and the recombination factor takes into account carriers that recombine within the forward-biased B–E junction.
- Several nonideal effects are considered:

  1. Base width modulation, or Early effect—the change in the neutral base width with a change in B–C voltage, producing a change in collector current with a change in B–C or C–E voltage.
  2. High-injection effects that cause the collector current to increase at a slower rate with base–emitter voltage.
  3. Emitter bandgap narrowing that produces a smaller emitter injection efficiency because of a very large emitter region doping concentration.
  4. Current crowding effects that produce a larger current density at the emitter edge than in the center of the emitter.
  5. A nonuniform base doping concentration that induces an electric field in the base region, which aids the flow of minority carriers across the base.
  6. Two breakdown voltage mechanisms—punch-through and avalanche.

- Three equivalent circuits or mathematical models of the transistor are considered. The Ebers–Moll model and equivalent circuit are applicable in any of the transistor operating modes. The Gummel–Poon model is convenient to use when nonuniform doping exists in the transistor. The small-signal hybrid-pi model applies to transistors operating in the forward-active mode in linear amplifier circuits.
- The cutoff frequency of a transistor, a figure of merit for the transistor, is the frequency at which the magnitude of the common-emitter current gain becomes equal to unity. The frequency response is a function of the emitter–base junction capacitance charging time, the base transit time, the collector depletion region transit time, and the collector capacitance charging time.
- The switching characteristics are closely related to the frequency limitations although switching involves large changes in currents and voltages. An important parameter in switching is the charge storage time, which applies to a transistor switching from saturation to cutoff.

GLOSSARY OF IMPORTANT TERMS

alpha cutoff frequency The frequency at which the magnitude of the common-base current is $1/\sqrt{2}$ of its low-frequency value; also equal to the cutoff frequency.

bandgap narrowing The reduction in the forbidden energy bandgap with high emitter doping concentration.

base transit time The time that it takes a minority carrier to cross the neutral base region.

base transport factor The factor in the common-base current gain that accounts for recombination in the neutral base width.

base width modulation The change in the neutral base width with C–E or C–B voltage.

beta cutoff frequency The frequency at which the magnitude of the common-emitter current gain is $1/\sqrt{2}$ of its low-frequency value.

collector capacitance charging time The time constant that describes the time required for the B–C and collector–substrate space charge widths to change with a change in emitter current.

collector depletion region transit time The time that it takes a carrier to be swept across the B–C space charge region.

common-base current gain The ratio of collector current to emitter current.

common-emitter current gain The ratio of collector current to base current.

current crowding The nonuniform current density across the emitter junction area created by a lateral voltage drop in the base region due to a finite base current and base resistance.

cutoff The bias condition in which zero- or reverse-biased voltages are applied to both transistor junctions, resulting in zero transistor currents.

cutoff frequency The frequency at which the magnitude of the common-emitter current gain is unity.

early effect Another term for base width modulation.

early voltage The value of voltage (magnitude) at the intercept on the voltage axis obtained by extrapolating the I_C versus V_{CE} curves to zero current.

emitter–base junction capacitance charging time The time constant describing the time for the B–E space charge width to change with a change in emitter current.

emitter injection efficiency factor The factor in the common-base current gain that takes into account the injection of carriers from the base into the emitter.

forward active The bias condition in which the B–E junction is forward biased and the B–C junction is reverse biased.

inverse active The bias condition in which the B–E junction is reverse biased and the B–C junction is forward biased.

output conductance The ratio of a differential change in collector current to the corresponding differential change in C–E voltage.

CHECKPOINT

After studying this chapter, the reader should have the ability to:

- Describe the basic operation of the transistor.
- Sketch the energy bands of the transistor in thermal equilibrium and when biased in the various operating modes.
- Calculate, to a good first approximation, the collector current as a function of base–emitter voltage.

- Sketch the minority carrier concentrations throughout the transistor under the various operating modes.
- Define the various diffusion and other current components in the transistor from the minority carrier distribution curves.
- Explain the physical mechanisms of the current gain limiting factors.
- Define the current-limiting factors from the current components in the transistor.
- Describe the physical mechanism of base width modulation and its effect on the current–voltage characteristics of the transistor.
- Describe the voltage breakdown mechanisms in a bipolar transistor.
- Sketch the simplified small-signal hybrid-pi equivalent circuit of the transistor biased in the forward-active mode.
- Describe qualitatively the four time-delay or time-constant components in the frequency response of the bipolar transistor.

REVIEW QUESTIONS

1. Describe the charge flow in an npn bipolar transistor biased in the forward-active mode. Is the current by drift or diffusion?
2. Define the common-emitter current gain and explain why, to a first approximation, the current gain is a constant. What is the relation between the common-emitter and common-base current gains?
3. Explain the conditions of the cutoff, saturation, and inverse-active modes.
4. Sketch the minority carrier concentrations in a pnp bipolar transistor biased in the forward-active mode.
5. Define and describe the three limiting factors in the common-base current gain. Why does the base doping concentration affect the emitter injection efficiency?
6. Describe base width modulation. Sketch an I–V curve that shows the base width modulation effect.
7. What is meant by high injection?
8. Explain emitter current crowding.
9. Define I_{CBO} and I_{CEO}, and explain why $I_{CEO} > I_{CBO}$.
10. Sketch a simplified hybrid-pi model for an npn bipolar transistor and explain when this equivalent circuit is used.
11. Describe the time-delay factors in the frequency limitation of the bipolar transistor.
12. What is the cutoff frequency of a bipolar transistor?
13. Describe the response of a bipolar transistor when it is switching between saturation and cutoff.

PROBLEMS

(Note: In the following problems, use the transistor geometry shown in Figure 12.13. Assume $T = 300$ K unless otherwise stated.)

Section 12.1 The Bipolar Transistor Action

12.1 For a uniformly doped $n^{++}p^{+}n$ bipolar transistor in thermal equilibrium, (*a*) sketch the energy-band diagram, (*b*) sketch the electric field through the device, and (*c*) repeat parts (*a*) and (*b*) for the transistor biased in the forward-active region.

12.2 Consider a $p^{++}n^{+}p$ bipolar transistor, uniformly doped in each region. Sketch the energy-band diagram for the case when the transistor is (*a*) in thermal equilibrium, (*b*) biased in the forward-active mode, (*c*) biased in the inverse-active region, and (*d*) biased in cutoff with both the B–E and B–C junctions reverse biased.

12.3 The parameters of the base region in a silicon npn bipolar transistor are $D_n = 18\ \text{cm}^2/\text{s}$, $n_{B0} = 4 \times 10^3\ \text{cm}^{-3}$, $x_B = 0.80\ \mu\text{m}$, and $A_{BE} = 5 \times 10^{-5}\ \text{cm}^2$. (*a*) Comparing Equations (12.1) and (12.2), calculate the magnitude of I_S. (*b*) Calculate the collector current for (*i*) $v_{BE} = 0.58$ V, (*ii*) $v_{BE} = 0.65$ V, and (*iii*) $v_{BE} = 0.72$ V.

12.4 An npn silicon bipolar transistor has the following base parameters: $D_n = 22\ \text{cm}^2/\text{s}$, $x_B = 0.80\ \mu\text{m}$, and $n_{B0} = 2 \times 10^4\ \text{cm}^{-3}$. (*a*) The collector current is to be $|i_C| = 2$ mA when biased at $v_{BE} = 0.60$ V. What is the required cross-sectional area A_{BE}? (*b*) Using the results of part (*a*), what is the value of v_{BE} such that $|i_C| = 5$ mA?

12.5 Consider the transistor described in Problem 12.3. (*a*) For a common-base current gain of $\alpha = 0.9850$, determine the common-emitter current gain [note: $\beta = \alpha/(1 - \alpha)$]. (*b*) Determine the emitter and base currents corresponding to the collector currents determined in Problem 12.3. (*c*) Repeats parts (*a*) and (*b*) for a common-base current gain of $\alpha = 0.9940$.

12.6 A bipolar transistor is biased in the forward-active region. (*a*) For a base current of $I_B = 4.2\ \mu\text{A}$ and a collector current of $I_C = 0.625$ mA, determine (*i*) β, (*ii*) α, and (*iii*) I_E. (*b*) For a collector current of $I_C = 1.254$ mA and an emitter current of $I_E = 1.273$ mA, determine (*i*) β, (*ii*) α, and (*iii*) I_B. (*c*) For a base current of $I_B = .065\ \mu\text{A}$ and a common-emitter current gain of $\beta = 150$, determine (*i*) α, (*ii*) I_C, and (*iii*) I_E.

12.7 Assume that an npn bipolar transistor has a common-emitter current gain of $\beta = 100$. (*a*) Sketch the ideal current–voltage characteristics (i_C versus v_{CE}), like those in Figure 12.9, as i_B varies from zero to 0.1 mA in 0.01-mA increments. Let v_{CE} vary over the range $0 \leq v_{CE} \leq 10$ V. (*b*) Assuming $V_{CC} = 10$ V and $R_C = 1\ \text{k}\Omega$ in the circuit in Figure 12.8, superimpose the load line on the transistor characteristics in part (*a*). (*c*) Plot, on the resulting graph, the value of i_C and v_{CE} corresponding to $i_B = 0.05$ mA.

12.8 Consider Figure 12.8. Assume $V_{CC} = 3$ V and $V_{BE} = 0.65$ V. (*a*) For $R_C = 25\ \text{k}\Omega$, (*i*) plot I_C versus V_{CE} over the range $0.20 \leq V_{CE} \leq 3$ V. (*ii*) At what value of I_C does $V_{CB} = 0$? (*b*) Repeat part (*a*) for $R_C = 10\ \text{k}\Omega$.

Section 12.2 Minority Carrier Distribution

12.9 A uniformly doped silicon npn bipolar transistor at $T = 300$ K is biased in the forward-active mode. The doping concentrations are $N_E = 8 \times 10^{17}\ \text{cm}^{-3}$, $N_B = 2 \times 10^{16}\ \text{cm}^{-3}$, and $N_C = 10^{15}\ \text{cm}^{-3}$. (*a*) Determine the thermal-equilibrium values p_{E0}, n_{B0}, and p_{C0}. (*b*) For $V_{BE} = 0.640$ V, calculate the values of n_B at $x = 0$ and p_E at $x' = 0$. (*c*) Sketch the minority carrier concentrations through the device and label each curve.

12.10 A silicon pnp bipolar transistor at $T = 300$ K is uniformly doped and is biased in the forward-active mode. The doping concentrations are $N_E = 5 \times 10^{17}\ \text{cm}^{-3}$, $N_B = 10^{16}\ \text{cm}^{-3}$, and $N_C = 10^{15}\ \text{cm}^{-3}$. (*a*) Find the thermal-equilibrium values n_{E0}, p_{B0}, and n_{C0}. (*b*) Determine the values of p_B at $x = 0$ and n_E at $x' = 0$ for $V_{EB} = 0.615$ V. (*c*) Sketch the minority carrier concentrations through the device and label each curve.

12.11 Consider a uniformly doped silicon npn bipolar transistor at $T = 300$ K. The device is biased in the forward-active mode with $V_{CB} = 2.5$ V. The metallurgical base width is $x_{B0} = 1.0\ \mu$m. The doping concentrations are $N_E = 8 \times 10^{17}$ cm^{-3}, $N_B = 2 \times 10^{16}$ cm^{-3}, and $N_C = 10^{15}$ cm^{-3}. (*a*) Determine the B–E voltage such that the minority carrier electron concentration, n_B, at $x = 0$ is 10 percent of the majority carrier hole concentration. (*b*) At this bias, determine the minority carrier hole concentration at $x' = 0$.

12.12 Consider the minority carrier electron concentration in the base of an npn bipolar transistor as given by Equation (12.15a). In this problem, we want to compare the gradient of the electron concentration evaluated at the B–C junction to that evaluated at the B–E junction. In particular, calculate the ratio of $d(\delta n_B)/dx$ at $x = x_B$ to $d(\delta n_B)/dx$ at $x = 0$ for (*a*) $x_B/L_B = 0.1$, (*b*) $x_B/L_B = 1.0$, and (*c*) $x_B/L_B = 10$.

12.13 Derive the expressions for the coefficients given by Equations (12.14a) and (12.14b).

***12.14** Derive the expression for the excess minority carrier hole concentration in the base region of a uniformly doped pnp bipolar transistor operating in the forward-active region.

12.15 The excess electron concentration in the base of an npn bipolar transistor is given by Equation (12.15a). The linear approximation is given by Equation (12.15b). If $\delta n_{B0}(x)$ is the linear approximation given by Equation (12.15b) and $\delta n_B(x)$ is the actual distribution given by Equation (12.15a), determine

$$\frac{\delta n_{B0}(x) - \delta n_B(x)}{\delta n_{B0}(x)} \times 100\%$$

at $x = x_B/2$ for (*a*) $x_B/L_B = 0.1$ and (*b*) $x_B/L_B = 1.0$. Assume $V_{BE} \gg kT/e$.

12.16 Consider a uniformly doped silicon pnp bipolar transistor biased in the forward-active mode at low injection. The excess minority carrier hole concentration at $x = 0$ is $\delta p_B(0) = 10^{15}$ cm^{-3} and the excess minority carrier hole concentration at $x = x_B$ is $\delta p_B(x_B) = -5 \times 10^3$ cm^{-3}. (*a*) What is the majority carrier electron concentration in the base region and what is the E–B voltage? (*b*) Assuming $x_B = 0.80\ \mu$m and $D_B = 10$ cm^2/s, calculate the magnitude of diffusion current density at (*i*) $x = 0$ and (*ii*) $x = x_B$ for the case when $x_B \ll L_B$. (See Equation (12.15b).) (*c*) Repeat part (*b*) for the case when $x_B = L_B = 12\ \mu$m. (See Equation (12.15a).) (*d*) Determine the ratio $J(x = x_B)/J(x = 0)$ for parts (*b*) and (*c*).

***12.17** (*a*) A uniformly doped npn bipolar transistor at $T = 300$ K is biased in saturation. Starting with the continuity equation for minority carriers, show that the excess electron concentration in the base region can be expressed as

$$\delta n_B(x) = n_{B0}\left\{\left[\exp\left(\frac{eV_{BE}}{kT}\right) - 1\right]\left(1 - \frac{x}{x_B}\right) + \left[\exp\left(\frac{eV_{BC}}{kT}\right) - 1\right]\left(\frac{x}{x_B}\right)\right\}$$

for $x_B/L_B \ll 1$ where x_B is the neutral base width. (*b*) Show that the minority carrier diffusion current in the base is then given by

$$J_n = -\frac{eD_B n_{B0}}{x_B}\left[\exp\left(\frac{eV_{BE}}{kT}\right) - \exp\left(\frac{eV_{BC}}{kT}\right)\right]$$

(*c*) Show that the total excess minority carrier charge (C/cm^2) in the base region is given by

$$\delta Q_{nB} = \frac{-en_{B0}x_B}{2}\left\{\left[\exp\left(\frac{eV_{BE}}{kT}\right) - 1\right] + \left[\exp\left(\frac{eV_{BC}}{kT}\right) - 1\right]\right\}$$

*Asterisks next to problems indicate problems that are more difficult.

12.18 Consider a silicon npn bipolar transistor at $T = 300$ K with uniform doping concentrations of $N_E = 10^{18}$ cm^{-3}, $N_B = 5 \times 10^{16}$ cm^{-3}, and $N_C = 10^{15}$ cm^{-3}. Let $D_B = 25$ cm^2/s, $x_B = 0.70$ μm, and assume $x_B \ll L_B$. The transistor is operating in saturation with $|J_n| = 125$ A/cm^2 and $V_{BE} = 0.70$ V. Determine (*a*) V_{BC}, (*b*) $V_{CE}(sat)$, (*c*) the #/cm^2 of excess minority carrier electrons in the base region, and (*d*) the #/cm^2 of excess minority carrier holes in the long collector. Let $L_C = 35$ μm.

12.19 An npn silicon bipolar transistor at $T = 300$ K has uniform dopings of $N_E = 10^{19}$ cm^{-3}, $N_B = 10^{17}$ cm^{-3}, and $N_C = 7 \times 10^{15}$ cm^{-3}. The transistor is operating in the inverse-active mode with $V_{BE} = -2$ V and $V_{BC} = 0.565$ V. (*a*) Sketch the minority carrier distribution through the device. (*b*) Determine the minority carrier concentrations at $x = x_B$ and $x' = 0$. (*c*) If the metallurgical base width is 1.2 μm, determine the neutral base width.

12.20 A uniformly doped silicon pnp bipolar transistor at $T = 300$ K with dopings of $N_E = 5 \times 10^{17}$ cm^{-3}, $N_B = 10^{16}$ cm^{-3}, and $N_C = 5 \times 10^{14}$ cm^{-3} is biased in the inverse-active mode. What is the maximum B–C voltage so that the low-injection condition applies?

Section 12.3 Transistor Currents and Low-Frequency Common-Base Current Gain

12.21 (*a*) The following currents are measured in a uniformly doped npn bipolar transistor.

$$I_{nE} = 0.50 \text{ mA} \qquad I_{pE} = 3.5 \ \mu\text{A}$$
$$I_{nC} = 0.495 \text{ mA} \qquad I_R = 5.0 \ \mu\text{A}$$
$$I_G = 0.50 \ \mu A \qquad I_{pc0} = 0.50 \ \mu\text{A}$$

Determine the following current gain parameters: (*i*) γ, (*ii*) α_T, (*iii*) δ, (*iv*) α, and (*v*) β. (*b*) If the required value of common-emitter current gain is $\beta = 120$, determine new values of I_{nC}, I_{pE}, and I_R to meet this specification assuming $\gamma = \alpha_T = \delta$.

12.22 A silicon pnp bipolar transistor at $T = 300$ K has a B–E cross-sectional area of $A_{BE} = 5 \times 10^{-4}$ cm^2, neutral base width of $x_B = 0.70$ μm, a neutral emitter width of $x_E = 0.50$ μm, and uniform doping concentrations of $N_E = 5 \times 10^{17}$ cm^{-3}, $N_B = 10^{16}$ cm^{-3}, and $N_C = 10^{15}$ cm^{-3}. Other transistor parameters are $D_B = 10$ cm^2/s, $D_E = 15$ cm^2/s, $\tau_{E0} = \tau_{B0} = 5 \times 10^{-7}$ s, and $\tau_{C0} = 2 \times 10^{-6}$ s. The transistor is biased in the forward-active mode and the recombination factor is $\delta = 0.995$. Determine the collector current for (*a*) $V_{EB} = 0.550$ V, (*b*) $I_B = 0.80$ μA, and (*c*) $I_E = 125$ μA.

12.23 Consider a uniformly doped npn bipolar transistor at $T = 300$ K with the following parameters:

$N_E = 10^{18}$ cm^{-3}	$N_B = 5 \times 10^{16}$ cm^{-3}	$N_C = 10^{15}$ cm^{-3}
$D_E = 8$ cm^2/s	$D_B = 15$ cm^2/s	$D_C = 12$ cm^2/s
$\tau_{E0} = 10^{-8}$ s	$\tau_{B0} = 5 \times 10^{-8}$ s	$\tau_{C0} = 10^{-7}$ s
$x_E = 0.8$ μm	$x_B = 0.7$ μm	$J_{r0} = 3 \times 10^{-8}$ A/cm^2

For $V_{BE} = 0.60$ V and $V_{CE} = 5$ V, calculate (*a*) the currents J_{nE}, J_{pE}, J_{nC}, and J_R and (*b*) the current gain factors γ, α_T, δ, α, and β.

12.24 Three npn bipolar transistors have identical parameters except for the base doping concentrations and neutral base widths. The base parameters for the three devices are as follows:

Device	Base doping	Base width
A	$N_B = N_{B0}$	$x_B = x_{B0}$
B	$N_B = 2N_{B0}$	$x_B = x_{B0}$
C	$N_B = N_{B0}$	$x_B = x_{B0}/2$

(The base doping concentration for the B device is twice that of A and C, and the neutral base width for the C device is half that of A and B.)

(*a*) Determine the ratio of the emitter injection efficiency of (*i*) device B to device A and (*ii*) device C to device A.

(*b*) Repeat part (*a*) for the base transport factor.

(*c*) Repeat part (*a*) for the recombination factor.

(*d*) Which device has the largest common-emitter current gain β?

12.25 Repeat Problem 12.24 for three devices in which the emitter parameters vary. The emitter parameters for the three devices are as follows:

Device	Emitter doping	Emitter width
A	$N_E = N_{E0}$	$x_E = x_{E0}$
B	$N_E = 2N_{E0}$	$x_E = x_{E0}$
C	$N_E = N_{E0}$	$x_E = x_{E0}/2$

12.26 An npn silicon transistor is biased in the inverse-active mode with $V_{BE} = -3$ V and $V_{BC} = 0.6$ V. The doping concentrations are $N_E = 10^{18}$ cm^{-3}, $N_B = 10^{17}$ cm^{-3}, and $N_C = 10^{16}$ cm^{-3}. Other parameters are $x_B = 1$ μm, $\tau_{E0} = \tau_{B0} = \tau_{C0} = 2 \times 10^{-7}$ s, $D_E = 10$ cm^2/s, $D_B = 20$ cm^2/s, $D_C = 15$ cm^2/s, and $A = 10^{-3}$ cm^2. (*a*) Calculate and plot the minority carrier distribution in the device. (*b*) Calculate the collector and emitter currents. (Neglect geometry factors and assume the recombination factor is unity.)

12.27 (*a*) Calculate the base transport factor, α_T, for $x_B/L_B = 0.01$, 0.10, 1.0, and 10. Assuming that γ and δ are unity, determine β for each case. (*b*) Calculate the emitter injection efficiency, γ, for $N_B/N_E = 0.01$, 0.10, 1.0, and 10. Assuming that α_T and δ are unity, determine α for each case. (*c*) Considering the results of parts (*a*) and (*b*), what conclusions can be made concerning when the base transport factor or when the emitter injection efficiency are the limiting factors for the common-emitter current gain?

12.28 (*a*) Calculate the recombination factor for $V_{BE} = 0.2$, 0.4, and 0.6 V. Assume the following parameters:

$$D_B = 25 \text{ cm}^2/\text{s} \qquad D_E = 10 \text{ cm}^2/\text{s}$$
$$N_E = 5 \times 10^{18} \text{ cm}^{-3} \qquad N_B = 1 \times 10^{17} \text{ cm}^{-3}$$
$$N_C = 5 \times 10^{15} \text{ cm}^{-3} \qquad x_B = 0.7 \ \mu\text{m}$$
$$\tau_{B0} = \tau_{E0} = 10^{-7} \text{ s} \qquad J_{r0} = 2 \times 10^{-9} \text{ A/cm}^2$$
$$n_i = 1.5 \times 10^{10} \text{ cm}^{-3}$$

(*b*) Assuming the base transport and emitter injection efficiency factors are unity, calculate the common-emitter current gain for the conditions in part (*a*).

(*c*) Considering the results of part (*b*), what can be said about the recombination factor being the limiting factor in the common-emitter current gain.

12.29 Consider a uniformly dope silicon npn bipolar transistor at $T = 300$ K with the following parameters: $D_B = 23$ cm^2/s, $D_E = 8$ cm^2/s, $\tau_{B0} = 2 \times 10^{-7}$ s, $\tau_{E0} = 8 \times 10^{-8}$ s, $N_B = 2 \times 10^{16}$ cm^{-3}, and $x_E = 0.35$ μm. The recombination factor has been determined to be $\delta = 0.9975$. The required common-emitter current gain is $\beta = 150$. A minimum neutral base width of $x_B = 0.80$ μm can be fabricated. (*a*) Determine an appropriate neutral base width and the minimum emitter doping concentration, N_E, to meet this specification. (*b*) Using the results of part (*a*), what are the values of α_T and γ?

***12.30** (*a*) The recombination current density, J_{r0}, in an npn silicon bipolar transistor at $T = 300$ K is $J_{r0} = 5 \times 10^{-8}$ A/cm^2. The uniform dopings are $N_E = 10^{18}$ cm^{-3}, $N_B = 5 \times 10^{16}$ cm^{-3}, and $N_C = 10^{15}$ cm^{-3}. Other parameters are $D_E = 10$ cm^2/s, $D_B = 25$ cm^2/s, $\tau_{E0} = 10^{-8}$ s, and $\tau_{B0} = 10^{-7}$ s. Determine the neutral base width so that the recombination factor is $\delta = 0.995$ when $V_{BE} = 0.55$ V (*b*) If J_{r0} remains constant with temperature, what is the value of δ when $V_{BE} = 0.55$ V for the case when the temperature is $T = 400$ K? Use the value of x_B determined in part (*a*).

12.31 (*a*) Plot, for a bipolar transistor, the base transport factor, α_T, as a function of (x_B/L_B) over the range $0.01 \leq (x_B/L_B) \leq 10$. (Use a log scale on the horizontal axis.) (*b*) Assuming that the emitter injection efficiency and recombination factors are unity, plot the common-emitter gain for the conditions in part (*a*). (*c*) Considering the results of part (*b*), what can be said about the base transport factor being the limiting factor in the common-emitter current gain?

12.32 (*a*) Plot the emitter injection efficiency as a function of the doping ratio, N_B/N_E, over the range $0.01 \leq N_B/N_E \leq 10$. Assume that $D_E = D_B$, $L_B = L_E$, and $x_B = x_E$. (Use a log scale on the horizontal axis.) Neglect bandgap narrowing effects. (*b*) Assuming that the base transport factor and recombination factors are unity, plot the common-emitter current gain for the conditions in part (*a*). (*c*) Considering the results of part (*b*), what can be said about the emitter injection efficiency being the limiting factor in the common-emitter current gain.

12.33 (*a*) Plot the recombination factor as a function of the forward-bias B–E voltage for $0.1 \leq V_{BE} \leq 0.6$. Assume the following parameters:

$$\begin{aligned} D_B &= 25 \text{ cm}^2/\text{s} & D_E &= 10 \text{ cm}^2/\text{s} \\ N_E &= 5 \times 10^{18} \text{ cm}^{-3} & N_B &= 1 \times 10^{17} \text{ cm}^{-3} \\ N_C &= 5 \times 10^{15} \text{ cm}^{-3} & x_B &= 0.7 \ \mu\text{m} \\ \tau_{B0} &= \tau_{E0} = 10^{-7} \text{ s} & J_{r0} &= 2 \times 10^{-9} \text{ A/cm}^2 \\ n_i &= 1.5 \times 10^{10} \text{ cm}^{-3} \end{aligned}$$

(*b*) Assuming the base transport and emitter injection efficiency factors are unity, plot the common-emitter current gain for the conditions in part (*a*). (*c*) Considering the results of part (*b*), what can be said about the recombination factor being the limiting factor in the common-emitter current gain.

12.34 The emitter in a BJT is often made very thin to achieve high operating speed. In this problem, we investigate the effect of emitter width on current gain. Consider the emitter injection efficiency given by Equation (12.35a). Assume that $N_E = 100$ N_B, $D_E = D_B$, and $L_E = L_B$. Also let $x_B = 0.1$ L_B. Plot the emitter injection efficiency for $0.01L_E \leq x_E \leq 10L_E$. From these results, discuss the effect of emitter width on the current gain.

Section 12.4 Nonideal Effects

12.35 An npn bipolar transistor is biased in the forward-active mode. (*a*) The collector current is $I_C = 1.2$ mA when biased at $V_{CE} = 2$ V. The Early voltage is $V_A = 120$ V. Determine (*i*) the output resistance r_o, (*ii*) the output conductance g_o, and (*iii*) the collector current when biased at $V_{CE} = 4$ V. (*b*) Repeat part (*a*) if the collector current is $I_C = 0.25$ mA when biased at $V_{CE} = 2$ V and the Early voltage is $V_A = 160$ V.

12.36 The output resistance of a pnp bipolar transistor is $r_o = 180$ kΩ. The Early voltage is $V_A = 80$ V. Determine the change in collector current if V_{EC} increases from 2 to 5 V.

12.37 A uniformly doped silicon npn bipolar transistor at $T = 300$ K has parameters $N_E = 2 \times 10^{18}$ cm^{-3}, $N_B = 2 \times 10^{16}$ cm^{-3}, $N_C = 2 \times 10^{15}$ cm^{-3}, $x_{BO} = 0.85$ μm, and $D_B = 25$ cm^2/s. Assume $x_{BO} \ll L_B$ and let $V_{BE} = 0.650$ V. (*a*) Determine the electron diffusion current density in the base for (*i*) $V_{CB} = 4$ V, (*ii*) $V_{CB} = 8$ V, and (*iii*) $V_{CB} = 12$ V. (*b*) Estimate the Early voltage.

***12.38** The base width of a bipolar transistor is normally small to provide a large current gain and increased speed. The base width also affects the Early voltage. In a silicon npn bipolar transistor at $T = 300$ K, the doping concentrations are $N_E = 10^{18}$ cm^{-3}, $N_B = 3 \times 10^{16}$ cm^{-3}, and $N_C = 5 \times 10^{15}$ cm^{-3}. Assume $D_B = 20$ cm^2/s and $\tau_{B0} = 5 \times 10^{-7}$ s, and let $V_{BE} = 0.70$ V. Using voltages $V_{CB} = 5$ V and $V_{CB} = 10$ V as two data points, estimate the Early voltage for metallurgical base widths of (*a*) 1.0 μm, (*b*) 0.80 μm, and (*c*) 0.60 μm.

12.39 A uniformly doped pnp silicon bipolar transistor has a base doping of $N_B = 10^{16}$ cm^{-3}, a collector doping of $N_C = 10^{15}$ cm^{-3}, a metallurgical base width of $x_{B0} = 0.70$ μm, a base minority carrier diffusion coefficient of $D_B = 10$ cm^2/s, and a B–E cross-sectional area of $A_{BE} = 10^{-4}$ cm^2. The transistor is biased in the forward-active mode with $V_{EB} = 0.625$ V. Neglecting the B–E space charge width and assuming $x_B \ll L_B$, (*a*) determine the change in neutral base width as V_{BC} changes from 1 to 5 V, (*b*) find the corresponding change in collector current, (*c*) estimate the Early voltage, and (*d*) find the output resistance.

12.40 Consider a uniformly doped silicon npn bipolar transistor in which $x_E = x_B$, $L_E = L_B$, and $D_E = D_B$. Assume that $\alpha_T = \delta = 0.995$ and let $N_B = 10^{17}$ cm^{-3}, Calculate and plot the common-emitter current gain β for $N_E = 10^{17}$, 10^{18}, 10^{19}, and 10^{20} cm^{-3}, and for the case (*a*) when the bandgap narrowing effect is neglected, and (*b*) when the bandgap narrowing effect is taken into account.

12.41 A silicon pnp bipolar transistor at $T = 300$ K is to be designed so that the emitter injection efficiency is $\gamma = 0.996$. Assume that $x_E = x_B$, $L_E = L_B$, $D_E = D_B$, and let $N_E = 10^{19}$ cm^{-3}. (*a*) Determine the maximum base doping, taking into account bandgap narrowing. (*b*) If bandgap narrowing were neglected, what would be the maximum base doping required?

12.42 The current crowding effect, to a first approximation, can be determined by using the geometry shown in Figure P12.42. Assume that one-half of the base current enters from each side of the emitter strip and flows uniformly to the center of the emitter. The base is p type with the following parameters: $N_B = 2 \times 10^{16}$ cm^{-3}, $x_B = 0.65$ μm, $\mu_p = 250$ cm^2/V-s, and $L = 25$ μm. (*a*) Assume $S = 10$ μm. (*i*) Calculate the resistance between $x = 0$ and $x = S/2$. (*ii*) If $I_B/2 = 5$ μA, determine the voltage drop between $x = 0$ and $x = S/2$. (*iii*) For $V_{BE} = 0.60$ V at $x = S/2$, determine the ratio of electrons being injected into the base at $x = S/2$ compared to $x = 0$. (*b*) Repeat part (*a*) for $S = 3$ μm.

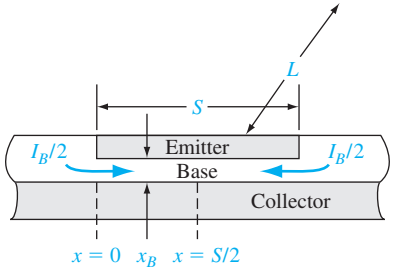


Figure P12.42 | Figure for Problems 12.42 and 12.43.

12.43 Consider the geometry shown in Figure P12.42 and the device parameters given in Problem 12.42 except for the emitter width S. Determine the maximum value of S such that the ratio of electrons being injected into the base at $x = S/2$ compared to $x = 0$ is no less than 0.90.

***12.44** The base doping in a diffused n^+pn bipolar transistor can be approximated by an exponential as

$$N_B = N_B(0) \exp\left(\frac{-ax}{x_B}\right)$$

where a is a constant and is given by

$$a = \ln\left(\frac{N_B(0)}{N_B(x_B)}\right)$$

(a) Show that, in thermal equilibrium, the electric field in the neutral base region is a constant. (b) Indicate the direction of the electric field. Does this electric field aid or retard the flow of minority carrier electrons across the base? (c) Derive an expression for the steady-state minority carrier electron concentration in the base under forward bias. Assume no recombination occurs in the base. (Express the electron concentration in terms of the electron current density.)

12.45 Consider a uniformly doped pnp silicon bipolar transistor with doping concentrations of $N_E = 10^{18}$ cm^{-3}, $N_B = 5 \times 10^{16}$ cm^{-3}, and $N_C = 2 \times 10^{15}$ cm^{-3}. The common-base current gain is $\alpha = 0.9930$. Determine (a) BV_{BC0}, (b) BV_{EC0}, and (c) the emitter–base breakdown voltage. (Assume $N = 3$ for the empirical constant.)

12.46 A high-voltage silicon npn bipolar transistor is to be designed such that the uniform base doping is $N_B = 10^{16}$ cm^{-3} and the common-emitter current grain is $\beta = 50$. The breakdown voltage BV_{CEO} is to be at least 60 V. Determine the maximum collector doping and the minimum collector length to support this voltage. (Assume $n = 3$.)

12.47 A silicon npn bipolar transistor is uniformly doped with $N_B = 5 \times 10^{16}$ cm^{-3} and $N_C = 8 \times 10^{15}$ cm^{-3}. The metallurgical base width is $x_{B0} = 0.50$ μm with $V_{BE} = 0$ and $V_{CB} = 0$. (a) Determine the expected avalanche B–C breakdown voltage. (b) Calculate the value of V_{CB} at which punch-through occurs.

12.48 Consider an npn silicon bipolar transistor with doping concentrations of $N_B = 2 \times 10^{16}$ cm^{-3} and $N_C = 5 \times 10^{15}$ cm^{-3}, and with a metallurgical base width of $x_{B0} = 0.65$ μm. Let $V_{BE} = 0.625$ V. (a) Determine V_{CE} at punch-through. (b) Calculate the magnitude of the maximum electric field in the B–C space charge region at punch-through.

12.49 A uniformly doped silicon pnp bipolar transistor has doping concentrations of $N_E = 10^{18}$ cm^{-3}, $N_B = 5 \times 10^{16}$ cm^{-3}, and $N_C = 3 \times 10^{15}$ cm^{-3}. Determine the minimum metallurgical base width such that the punch-through voltage is $V_{pt} = 15$ V.

Section 12.5 Equivalent Circuit Models

12.50 The V_{CE}(sat) voltage of an npn transistor in saturation continues to decrease slowly as the base current increases. In the Ebers–Moll model, assume $\alpha_F = 0.99$, $\alpha_R = 0.20$, and $I_C = 1$ mA. For $T = 300$ K, determine the base current, I_B, necessary to give (*a*) V_{CE}(sat) = 0.30 V, (*b*) V_{CE}(sat) = 0.20 V, and (*c*) V_{CE}(sat) = 0.10 V.

12.51 Consider an npn bipolar transistor biased in the active mode. Using the Ebers–Moll model, derive the equation for the base current, I_B, in terms of α_F, α_R, I_{ES}, I_{CS}, and V_{BE}.

12.52 Consider the Ebers–Moll model and let the base terminal be open so $I_B = 0$. Show that, when a collector–emitter voltage is applied, we have

$$I_C \equiv I_{CEO} = I_{CS}\frac{(1 - \alpha_F\alpha_R)}{(1 - \alpha_F)}$$

12.53 The parameters in the Ebers-Moll model are $\alpha_F = 0.9920$, $I_{ES} = 5 \times 10^{-14}$ A, and $I_{CS} = 10^{-13}$ A. Let $T = 300$ K. Plot I_C versus V_{CB} for $-0.5 < V_{CB} < 2$ V and for (*a*) $V_{BE} = 0.2$ V, (*b*) $V_{BE} = 0.4$ V, and (*c*) $V_{BE} = 0.6$ V. (Note that $V_{CB} = -V_{BC}$.)

12.54 The collector–emitter saturation voltage, from the Ebers–Moll model, is given by Equation (12.77). Consider a power BJT in which $\alpha_F = 0.975$, $\alpha_R = 0.150$, and $I_C = 5$ A. Plot V_{CE}(sat) versus I_B over the range $0.15 \leq I_B \leq 1$ A.

Section 12.6 Frequency Limitations

12.55 Consider a uniformly doped silicon bipolar transistor at $T = 300$ K with the following parameters:

$I_E = 0.25$ mA	$C_{je} = 0.35$ pF
$x_B = 0.65\ \mu$m	$D_n = 25$ cm^2/s
$x_{dc} = 2.2\ \mu$m	$r_c = 18\ \Omega$
$C_s = C_\mu = 0.020$ pF	$\beta = 125$

(*a*) Determine the transit time factors (*i*) τ_e, (*ii*) τ_b, (*iii*) τ_d, and (*iv*) τ_c.

(*b*) Find the total transit time τ_{ec}. (*c*) Calculate the cutoff frequency f_T.

(*c*) Find the beta cutoff frequency f_β.

12.56 In a particular bipolar transistor, the base transit time is 20 percent of the total delay time. The base width is 0.5 μm and the base diffusion coefficient is $D_B = 20$ cm^2/s. Determine the cutoff frequency.

12.57 Assume the base transit time of a BJT is 100 ps and carriers cross the 1.2 μm B–C space charge region at a speed of 10^7 cm/s. The emitter–base junction charging time is 25 ps and the collector capacitance and resistance are 0.10 pF and 10 Ω, respectively. Determine the cutoff frequency.

Summary and Review

***12.58** (*a*) A silicon npn bipolar transistor at $T = 300$ K is to be designed such that the common-emitter current gain is at least $\beta = 120$ and the Early voltage is at least $V_A = 140$ V. (*b*) Repeat part (*a*) for a pnp silicon bipolar transistor.

***12.59** Design a uniformly doped silicon npn bipolar transistor so that $\beta = 100$ at $T = 300$ K. The maximum CE voltage is to be 15 V and any breakdown voltage is to be at least three times this value. Assume the recombination factor is constant at $\delta = 0.995$. The transistor is to be operated in low injection with a maximum collector current of $I_C = 5$ mA. Bandgap narrowing effects and base width modulation effects are to be minimized. Let $D_E = 6$ cm^2/s, D_B=25 cm^2/s, $\tau_{E0} = 10^{-8}$ s, and $\tau_{B0} = 10^{-7}$ s. Determine doping concentrations, the metallurgical base width, the active area, and the maximum allowable V_{BE}.

***12.60** Design a pair of complementary npn and pnp bipolar transistors. The transistors are to have the same metallurgical base and emitter widths of $W_B = 0.75$ μm and $x_E = 0.5$ μm. Assume that the following minority carrier parameters apply to each device.

$$D_n = 23 \text{ cm}^2/\text{s} \qquad \tau_{n0} = 10^{-7} \text{ s}$$
$$D_p = 8 \text{ cm}^2/\text{s} \qquad \tau_{p0} = 5 \times 10^{-8} \text{ s}$$

The collector doping concentration in each device is 5×10^{15} cm^{-3} and the recombination factor in each device is constant at $\delta = 0.9950$. (*a*) Design, if possible, the devices so that $\beta = 100$ in each device. If this is not possible, how close a match can be obtained? (*b*) With equal forward-bias base–emitter voltages applied, the collector currents are to be $I_C = 5$ mA with each device operating in low injection. Determine the active cross-sectional areas.

READING LIST

1. Dimitrijev, S. *Principles of Semiconductor Devices*. New York: Oxford University, 2006.
2. Hu, C. C. *Modern Semiconductor Devices for Integrated Circuits*. Upper Saddle River, NJ: Pearson Prentice Hall, 2010.
3. Kano, K. *Semiconductor Devices.* Upper Saddle River, NJ: Prentice Hall, 1998.
4. Muller, R. S., and T. I. Kamins. *Device Electronics for Integrated Circuits.* 2nd ed. New York: John Wiley & Sons, 1986.
5. Navon, D. H. *Semiconductor Microdevices and Materials.* New York: Holt, Rinehart, & Winston, 1986.
6. Neudeck, G. W. *The Bipolar Junction Transistor.* Vol. 3 of the *Modular Series on Solid State Devices.* 2nd ed. Reading, MA: Addison-Wesley, 1989.
7. Ng, K. K. *Complete Guide to Semiconductor Devices.* New York: McGraw-Hill, 1995.
8. Ning, T. H., and R. D. lsaac. "Effect of Emitter Contact on Current Gain of Silicon Bipolar Devices." *Polysilicon Emitter Bipolar Transistors.* eds. A. K. Kapoor and D. J. Roulston. New York: IEEE Press, 1989.
9. Pierret, R. F. *Semiconductor Device Fundamentals.* Reading, MA: Addison-Wesley, 1996.
10. Roulston, D. J. *Bipolar Semiconductor Devices*. New York: McGraw-Hill, 1990.
11. _____. *An Introduction to the Physics of Semiconductor Devices.* New York: Oxford University Press, 1999.

*12. Shur, M. *GaAs Devices and Circuits.* New York: Plenum Press, 1987.

13. _____. *Introduction to Electronic Devices.* New York: John Wiley & Sons, Inc.,1996.

*14. _____. *Physics of Semiconductor Devices.* Englewood Cliffs, NJ: Prentice Hall, 1990.

15. Singh, J. *Semiconductor Devices: An Introduction*. New York: McGraw-Hill, 1994.

16. _____. *Semiconductor Devices: Basic Principles.* New York: John Wiley & Sons, Inc., 2001.

17. Streetman, B. G., and S. K. Banerjee. *Solid State Electronic Devices,* 6th ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2006.

18. Sze, S. M. *High-Speed Semiconductor Devices.* New York: John Wiley & Sons, 1990.

19. _____. and K. K. Ng. *Physics of Semiconductor Devices,* 3rd ed. Hoboken, NJ: John Wiley & Sons, Inc., 2007.

20. Tiwari, S., S. L. Wright, and A. W. Kleinsasser. "Transport and Related Properties of (Ga, Al)As/GaAs Double Heterojunction Bipolar Junction Transistors." *IEEE Transactions on Electron Devices,* ED-34 (February 1987), pp. 185–187.

***21.** Taur, Y. and T. H. Ning. *Fundamentals of Modern VLSI Devices,* 2nd ed. Cambridge University Press, 2009.

***22.** Wang, S. *Fundamentals of Semiconductor Theory and Device Physics.* Englewood Cliffs, NJ: Prentice Hall, 1989.

***23.** Warner, R. M. Jr., and B. L. Grung. *Transistors: Fundamentals for the Integrated-Circuit Engineer.* New York: John Wiley & Sons, 1983.

24. Yang, E. S. *Microelectronic Devices.* New York: McGraw-Hill, 1988.

***25.** Yuan, J. S. *SiGe, GaAs, and InP Heterojunction Bipolar Transistors.* New York: John Wiley & Sons, Inc., 1999.

*Indicates references that are at an advanced level compared to this text.

CHAPTER 13

The Junction Field-Effect Transistor

The Junction Field-Effect Transistor (JFET) is a separate class of field-effect transistors. The MOSFET has been considered in Chapters 10 and 11. In this chapter, we cover the physics and properties of the JFET. Although we have discussed the MOS and bipolar transistors in previous chapters, the material in this chapter only presumes a knowledge of semiconductor material properties and the characteristics of pn and Schottky barrier junctions.

As with the transistors considered in previous chapters, the JFET, in conjunction with other circuit elements, is capable of voltage gain and signal power gain. Again, the basic transistor action is the control of current at one terminal by the voltage across the other two terminals of the device.

There are two general categories of JFETs. The first is the pn junction FET, or pn JFET, and the second is the **ME**tal-**S**emiconductor **F**ield-**E**ffect **T**ransistor, or MESFET. The pn JFET is fabricated with a pn junction and the MESFET is fabricated with a Schottky barrier rectifying junction. ■

13.0 | PREVIEW

In this chapter, we will:

- Present the geometry and discuss the basic operation of the pn JFET and MESFET devices.
- Analyze the modulation of the channel conductance of the JFET by an electric field perpendicular to the channel. The modulating electric field is induced in the space charge region of a reverse-biased pn junction or reverse-biased Schottky barrier junction.
- Derive the ideal current–voltage characteristics of the JFET in terms of the semiconductor material and geometrical properties of the device.

- Consider the transistor gain, or transconductance, of the JFET.
- Discuss a few nonideal effects in JFETs, including channel-length modulation and velocity saturation effects.
- Develop a small-signal equivalent circuit of the JFET that is used to relate small-signal currents and voltages in the device.
- Examine various physical factors affecting the frequency response and limitations of JFETs, and derive an expression for the cutoff frequency.
- Present the geometry and characteristics of a specialized JFET called HEMT.

13.1 | JFET CONCEPTS

The concept of the field-effect phenomenon was the basis for the first proposed solid-state transistor. Patents filed in the 1920s and 1930s conceived and investigated the transistor shown in Figure 13.1. A voltage applied to the metal plate modulated the conductance of the semiconductor under the metal and controlled the current between the ohmic contacts. Good semiconductor materials and processing technology were not available at that time, so the device was not seriously considered again until the 1950s.

The phenomenon of modulating the conductance of a semiconductor by an electric field applied perpendicular to the surface of a semiconductor is called field effect. This type of transistor has also been called the unipolar transistor, to emphasize that only one type of carrier, the majority carrier, is involved in the operation. We will qualitatively discuss the basic operation of the two types of JFETs in this section, and introduce some of the JFET terminology.

13.1.1 Basic pn JFET Operation

The first type of field-effect transistor is the pn junction field-effect transistor, or pn JFET. A simplified cross section of a symmetrical device is shown in Figure 13.2. The n region between the two p regions is known as the channel and, in this n-channel

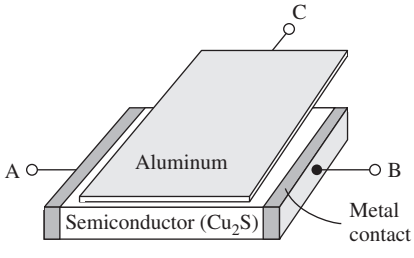


Figure 13.1 | Idealization of the Lilienfeld transistor.
(From Pierret [10].)

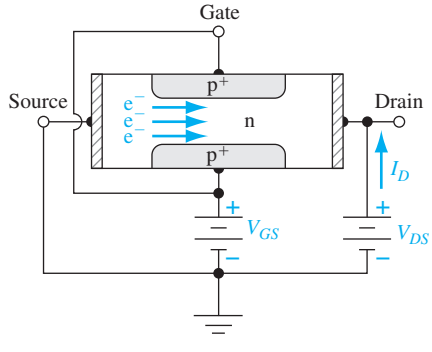


Figure 13.2 | Cross section of a symmetrical n-channel pn junction FET.

device, majority carrier electrons flow between the source and drain terminals. The source is the terminal from which carriers enter the channel from the external circuit, the drain is the terminal where carriers leave, or are drained from, the device, and the gate is the control terminal. The two gate terminals shown in Figure 13.2 are tied together to form a single gate connection. Since majority carrier electrons are primarily involved in the conduction in this n-channel transistor, the JFET is a majority-carrier device.

A complementary p-channel JFET can also be fabricated in which the p and n regions are reversed from those of the n-channel device. Holes will flow in the p-type channel between source and drain and the source terminal will now be the source of the holes. The current direction and voltage polarities in the p-channel JFET are the reverse of those in the n-channel device. The p-channel JFET is generally a lower frequency device than the n-channel JFET due to the lower hole mobility.

Figure 13.3a shows an n-channel pn JFET with zero volts applied to the gate. If the source is at ground potential, and if a small positive drain voltage is applied, a drain current I_D is produced between the source and drain terminals. The n channel is essentially a resistance so the I_D versus V_{DS} characteristic, for small V_{DS} values, is approximately linear, as shown in the figure.

When we apply a voltage to the gate of a pn JFET with respect to the source and drain, we alter the channel conductance. If a negative voltage is applied to the gate of the n-channel pn JFET shown in Figure 13.3, the gate-to-channel pn junction becomes reverse biased. The space charge region now widens so the channel region becomes narrower and the resistance of the n channel increases. The slope of the I_D versus V_{DS} curve, for small V_{DS}, decreases. These effects are shown in Figure 13.3b. If a larger negative gate voltage is applied, the condition shown in Figure 13.3c can be achieved. The reverse-biased gate-to-channel space charge region has completely filled the channel region. This condition is known as *pinchoff*. The drain current at pinchoff is essentially zero, since the depletion region isolates the source and drain terminals. Figure 13.3c shows the I_D versus V_{DS} curve for this case, as well as the other two cases.

The current in the channel is controlled by the gate voltage. The control of the current in one part of the device by a voltage in another part of the device is the basic transistor action. This device is a normally on or *depletion mode* device, which means that a voltage must be applied to the gate terminal to turn the device off.

Now consider the situation in which the gate voltage is held at zero volts, $V_{GS} = 0$, and the drain voltage changes. Figure 13.4a is a replica of Figure 13.3a for zero gate voltage and a small drain voltage. As the drain voltage increases (positive), the gate-to-channel pn junction becomes reverse biased near the drain terminal so that the space charge region extends further into the channel. The channel is essentially a resistor, and the effective channel resistance increases as the space charge region widens; therefore, the slope of the I_D versus V_{DS} characteristic decreases as shown in Figure 13.4b. The effective channel resistance now varies along the channel length and, since the channel current must be constant, the voltage drop through the channel becomes dependent on position.

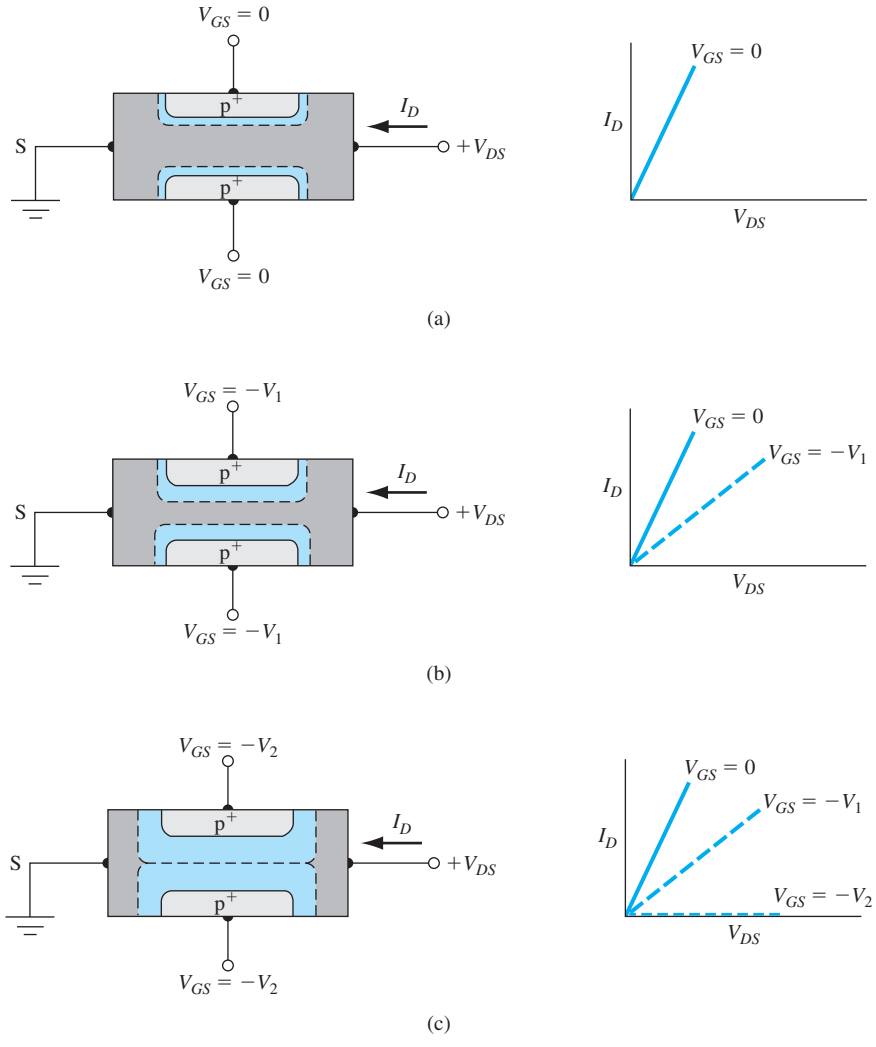


Figure 13.3 | Gate-to-channel space charge regions and *I–V* characteristics for small V_{DS} values and for (a) zero gate voltage, (b) small reverse-biased gate voltage, and (c) a gate voltage to achieve pinchoff.

If the drain voltage increases further, the condition shown in Figure 13.4c can result. The channel has been pinched off at the drain terminal. Any further increase in drain voltage will not cause an increase in drain current. The *I–V* characteristic for this condition is also shown in this figure. The drain voltage at pinchoff is referred to as $V_{DS}(\text{sat})$. For $V_{DS} > V_{DS}(\text{sat})$, the transistor is said to be in the saturation region and the drain current, for this ideal case, is independent of V_{DS}. At first glance, we might expect the drain current to go to zero when the channel becomes pinched off at the drain terminal, but we will show why this does not happen.

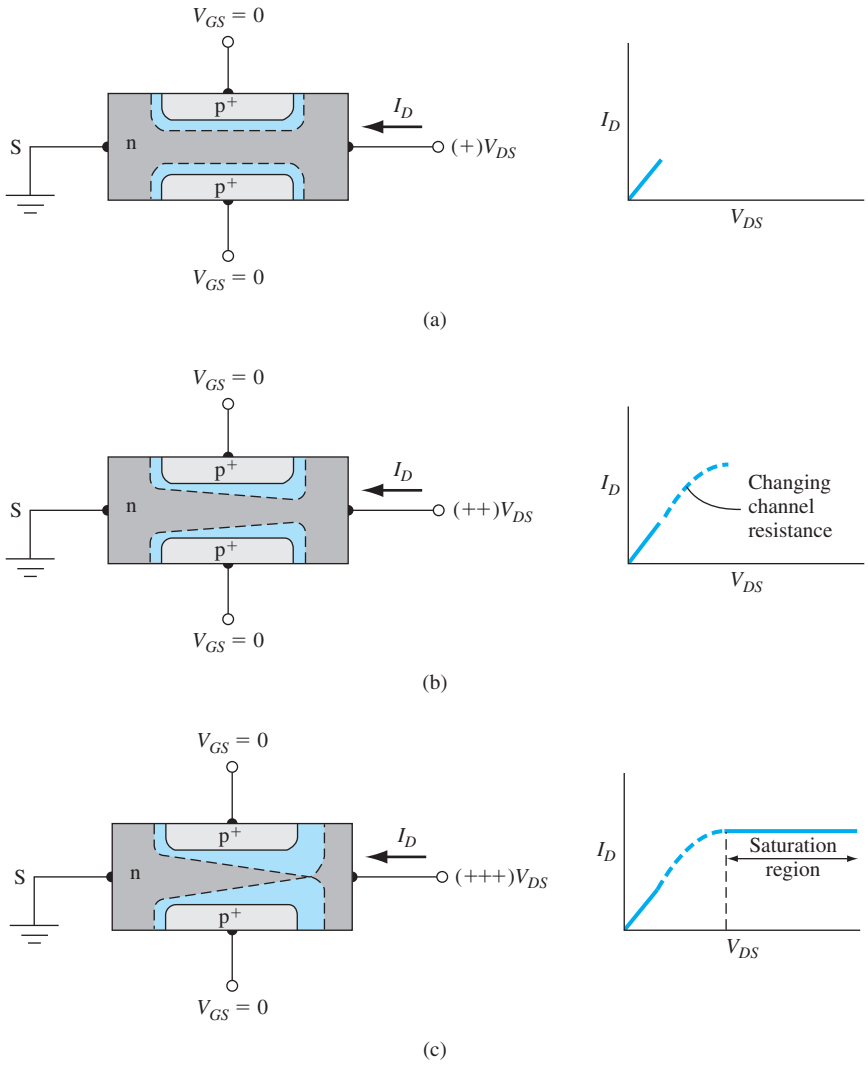


Figure 13.4 | Gate-to-channel space charge regions and *I–V* characteristics for zero gate voltage and for (a) a small drain voltage, (b) a larger drain voltage, and (c) a drain voltage to achieve pinchoff at the drain terminal.

Figure 13.5 shows an expanded view of the pinchoff region in the channel. The n channel and drain terminal are now separated by a space charge region which has a length ΔL. The electrons move through the n channel from the source and are injected into the space charge region where, subjected to the E-field force, they are swept through into the drain contact area. If we assume that $\Delta L \ll L$, then the electric field in the n-channel region remains unchanged from the V_{DS}(sat) case; the drain current will remain constant as V_{DS} changes. Once the carriers are in the drain region,

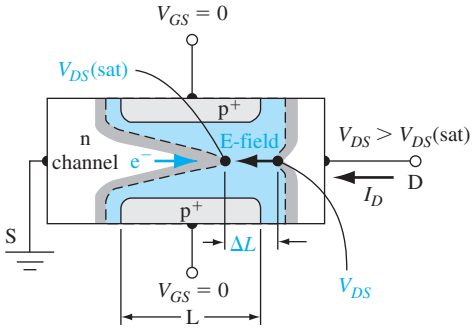


Figure 13.5 | Expanded view of the space charge region in the channel for $V_{DS} > V_{DS}$(sat).

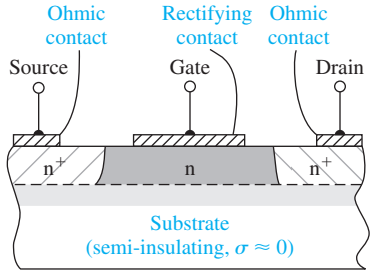


Figure 13.6 | Cross section of an n-channel MESFET with a semi-insulating substrate.

the drain current will be independent of V_{DS}; thus, the device looks like a constant current source.

13.1.2 Basic MESFET Operation

The second type of junction field-effect transistor is the MESFET. The gate junction in the pn junction FET is replaced by a Schottky barrier rectifying contact. Although MESFETs can be fabricated in silicon, they are usually associated with gallium arsenide or other compound semiconductor materials. A simplified cross section of a GaAs MESFET is shown in Figure 13.6. A thin epitaxial layer of GaAs is used for the active region; the substrate is a very high resistivity GaAs material referred to as a semi-insulating substrate. GaAs is intentionally doped with chromium, which behaves as a single acceptor close to the center of the energy bandgap, to make it semi-insulating with a resistivity as high as 10^9 Ω-cm. The advantages of these devices include higher electron mobility, hence smaller transit time and faster response; and decreased parasitic capacitance and a simplified fabrication process, resulting from the semi-insulating GaAs substrate.

In the MESFET shown in Figure 13.6, a reverse-biased gate-to-source voltage induces a space charge region under the metal gate that modulates the channel conductance as in the case of the pn JFET. The space charge region will eventually reach the substrate if the applied negative gate voltage is sufficiently large. This condition, again, is known as pinchoff. The device shown in this figure is also a depletion mode device, since a gate voltage must be applied to pinch off the channel.

If we treat the semi-insulating substrate as an intrinsic material, then the energy-band diagram of the substrate–channel–metal structure is as shown in Figure 13.7 for the case of zero bias applied to the gate. Because there is a potential barrier between the channel and substrate and between the channel and metal, the majority carrier electrons are confined to the channel region.

Consider, now, another type of MESFET in which the channel is pinched off even at $V_{GS} = 0$. Figure 13.8a shows this condition, in which the channel thickness is smaller than the zero-biased space charge width. To open a channel, the depletion region must be reduced: A forward-bias voltage must be applied to the gate–semiconductor

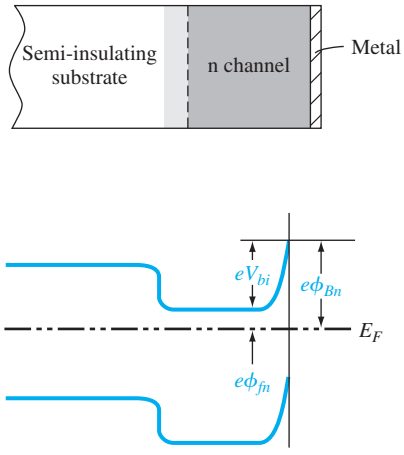


Figure 13.7 | Idealized energy-band diagram of the substrate–channel–metal in the n-channel MESFET.

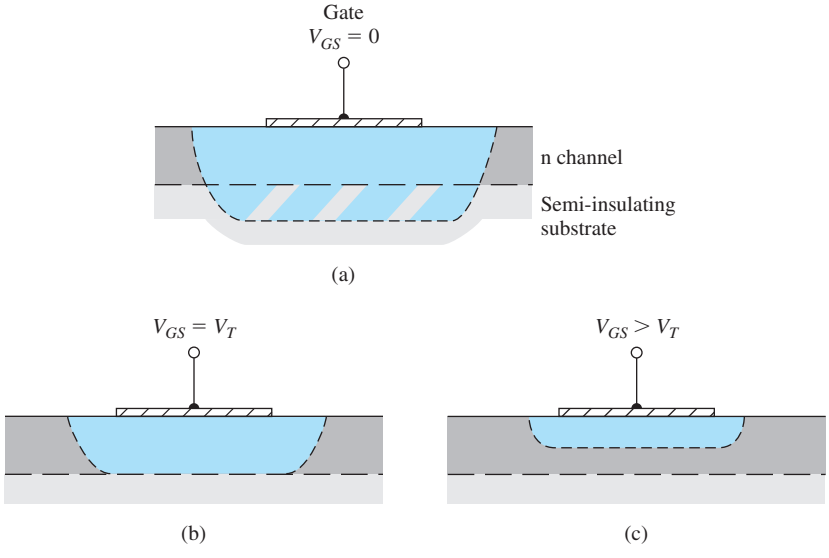


Figure 13.8 | Channel space charge region of an enhancement mode MESFET for (a) $V_{GS} = 0$, (b) $V_{GS} = V_T$, and (c) $V_{GS} > V_T$.

junction. When a slightly forward-bias voltage is applied, the depletion region just extends through the channel—a condition known as *threshold,* shown in Figure 13.8b. The threshold voltage is the gate-to-source voltage that must be applied to create the pinchoff condition. The threshold voltage for this n-channel MESFET is positive, in contrast to the negative voltage for the n-channel depletion mode device. If a larger forward bias is applied, the channel region opens as shown in Figure 13.8c. The

applied forward-bias gate voltage is limited to a few tenths of a volt before there is significant gate current. This device is known as an n-channel enhancement mode MESFET. Enhancement mode p-channel MESFETs and enhancement mode pn junction FETs have also been fabricated. The advantage of enhancement mode MESFETs is that circuits can be designed in which the voltage polarity on the gate and drain is the same. However, the output voltage swing will be quite small with these devices.

13.2 | THE DEVICE CHARACTERISTICS

To describe the basic electrical characteristics of the JFET, we initially consider a uniformly doped depletion mode pn JFET and then later discuss the enhancement mode device. The pinchoff voltage and drain-to-source saturation voltage are defined and expressions for these parameters derived in terms of geometry and electrical properties. The ideal current–voltage relationship is developed, and then the transconductance, or transistor gain is determined.

Figure 13.9a shows a symmetrical, two-sided pn JFET and Figure 13.9b shows a MESFET with the semi-insulating substrate. One can derive the ideal DC current–voltage relationship for both devices by simply considering the two-sided device to be two JFETs in parallel. We derive the *I–V* characteristics in terms of I_{D1} so that the drain current in the two-sided device becomes $I_{D2} = 2I_{D1}$. We ignore any depletion region at the substrate of the one-sided device in the ideal case.

13.2.1 Internal Pinchoff Voltage, Pinchoff Voltage, and Drain-to-Source Saturation Voltage

n-channel pn JFET Figure 13.10a shows a simplified one-sided n-channel pn JFET. The metallurgical channel thickness between the p^+ gate region and the substrate is a, and the induced depletion region width for the one-sided p^+n junction is h. Assume the drain-to-source voltage is zero. If we assume the abrupt depletion approximation, then the space charge width is given by

$$h = \left[\frac{2\epsilon_s(V_{bi} - V_{GS})}{eN_d}\right]^{1/2} \tag{13.1}$$

where V_{GS} is the gate-to-source voltage and V_{bi} is the built-in potential barrier. For a reverse-biased p^+n junction, V_{GS} must be a negative voltage.

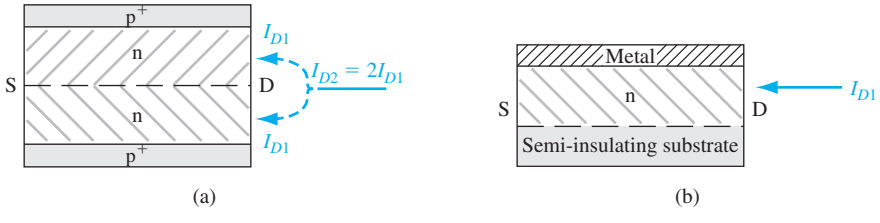


Figure 13.9 | Drain currents of (a) a symmetrical, two-sided pn JFET, and (b) a one-sided MESFET.

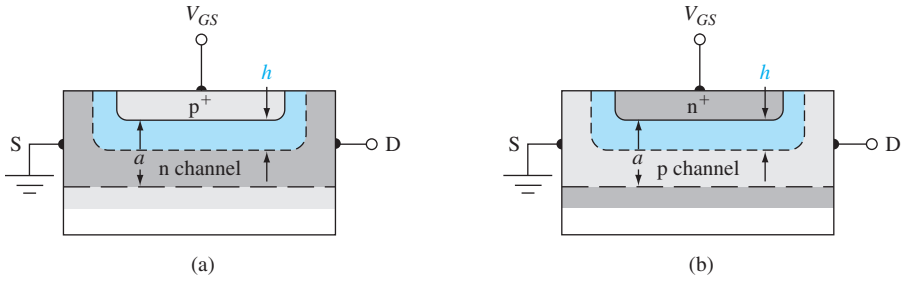


Figure 13.10 | Geometries of simplified (a) n-channel and (b) p-channel pn JFETs.

At pinchoff, $h = a$ and the total potential across the p^+n junction is called the *internal pinchoff voltage,* denoted by V_{p0}. We now have

$$a = \left[\frac{2\epsilon_s V_{p0}}{eN_d}\right]^{1/2} \tag{13.2}$$

or

$$\boxed{V_{p0} = \frac{ea^2 N_d}{2\epsilon_s}} \tag{13.3}$$

Note that the internal pinchoff voltage is defined as a positive quantity.

The internal pinchoff voltage V_{p0} is not the gate-to-source voltage to achieve pinchoff. The gate-to-source voltage that must be applied to achieve pinchoff is described as the *pinchoff voltage* and is also variously called the *turn-off voltage* or *threshold voltage*. The pinchoff voltage is denoted by V_p and is defined from Equations (13.1) and (13.2) as

$$V_{bi} - V_p = V_{p0} \qquad \text{or} \qquad \boxed{V_p = V_{bi} - V_{p0}} \tag{13.4}$$

The gate-to-source voltage to achieve pinchoff in an n-channel depletion mode JFET is negative; thus, $V_{p0} > V_{bi}$.

EXAMPLE 13.1

Objective: Calculate the internal pinchoff voltage and pinchoff voltage of an n-channel JFET.

Assume that the p^+n junction of a uniformly doped silicon n-channel JFET at $T = 300$ K has doping concentrations of $N_a = 10^{18}$ cm^{-3} and $N_d = 10^{16}$ cm^{-3}. Assume that the metallurgical channel thickness, a, is 0.75 μm $= 0.75 \times 10^{-4}$ cm.

■ Solution

The internal pinchoff voltage is given by Equation (13.3), so we have

$$V_{p0} = \frac{ea^2 N_d}{2\epsilon_s} = \frac{(1.6 \times 10^{-19})(0.75 \times 10^{-4})^2(10^{16})}{2(11.7)(8.85 \times 10^{-14})} = 4.35 \text{ V}$$

The built-in potential barrier is

$$V_{bi} = V_t \ln\left(\frac{N_a N_d}{n_i^2}\right) = (0.0259)\ln\left[\frac{(10^{18})(10^{16})}{(1.5 \times 10^{10})^2}\right] = 0.814 \text{ V}$$

The pinchoff voltage, from Equation (13.4), is then found as

$$V_p = V_{bi} - V_{p0} = 0.814 - 4.35 = -3.54 \text{ V}$$

■ Comment

The pinchoff voltage, or gate-to-source voltage to achieve pinchoff, for the n-channel depletion mode device is a negative quantity as we have said.

■ EXERCISE PROBLEM

Ex 13.1 A silicon n-channel JFET at $T = 300$ K has a gate doping concentration of $N_a = 10^{18}$ cm^{-3} and a channel doping concentration of $N_d = 2 \times 10^{16}$ cm^{-3}. Determine the metallurgical channel thickness, a, such that the pinchoff voltage is $V_p = -2.50$ V.

(Ans. $a = 0.464\ \mu$m)

The pinchoff voltage is the gate-to-source voltage that must be applied to turn the JFET off and so must be within the voltage range of the circuit design. The magnitude of the pinchoff voltage must also be less than the breakdown voltage of the junction.

p-channel pn JFET Figure 13.10b shows a p-channel JFET with the same basic geometry as the n-channel JFET we considered. The induced depletion region for the one-sided n^+p junction is again denoted by h and is given by

$$h = \left[\frac{2\epsilon_s(V_{bi} + V_{GS})}{eN_a}\right]^{1/2} \tag{13.5}$$

For a reverse-biased n^+p junction, V_{GS} must be positive. The internal pinchoff voltage is again defined to be the total pn junction voltage to achieve pinchoff, so that when $h = a$ we have

$$a = \left[\frac{2\epsilon_s V_{p0}}{eN_a}\right]^{1/2} \tag{13.6}$$

or

$$\boxed{V_{p0} = \frac{ea^2 N_q}{2\epsilon_s}} \tag{13.7}$$

The internal pinchoff voltage for the p-channel device is also defined to be a positive quantity.

The pinchoff voltage is again defined as the gate-to-source voltage to achieve the pinchoff condition. For the p-channel depletion mode device, we have, from Equation (13.5), at pinchoff

$$V_{bi} + V_p = V_{p0} \qquad \text{or} \qquad \boxed{V_p = V_{p0} - V_{bi}} \tag{13.8}$$

The pinchoff voltage for a p-channel depletion mode JFET is a positive quantity.

DESIGN EXAMPLE 13.2

Objective: Design the channel doping concentration and metallurgical channel thickness to achieve a given pinchoff voltage.

Consider a silicon p-channel pn JFET at $T = 300$ K. Assume that the gate doping concentration is $N_d = 10^{18}$ cm^{-3}. Determine the channel doping concentration and channel thickness so that the pinchoff voltage is $V_p = 2.25$ V.

■ Solution

There is not a unique solution to this design problem. We will pick a channel doping concentration of $N_a = 2 \times 10^{16}$ cm^{-3} and determine the channel thickness. The built-in potential barrier is

$$V_{bi} = V_t \ln\left(\frac{N_a N_d}{n_i^2}\right) = (0.0259) \ln\left[\frac{(2 \times 10^{16})(10^{18})}{(1.5 \times 10^{10})^2}\right] = 0.832 \text{ V}$$

From Equation (13.8), the internal pinchoff voltage must be

$$V_{p0} = V_{bi} + V_p = 0.832 + 2.25 = 3.08 \text{ V}$$

and from Equation (13.6), the channel thickness can be determined as

$$a = \left[\frac{2\epsilon_s V_{p0}}{eN_a}\right]^{1/2} = \left[\frac{2(11.7)(8.85 \times 10^{-14})(3.08)}{(1.6 \times 10^{-19})(2 \times 10^{16})}\right]^{1/2} = 0.446 \ \mu\text{m}$$

■ Comment

If the channel doping concentration chosen were larger, the required channel thickness would decrease; a very small value of channel thickness would be difficult to fabricate within reasonable tolerance limits.

■ EXERCISE PROBLEM

Ex 13.2 The n$^+$p junction of a uniformly doped silicon p-channel JFET at $T = 300$ K has doping concentrations of $N_d = 10^{18}$ cm^{-3} and $N_a = 10^{16}$ cm^{-3}. The metallurgical channel thickness is $a = 0.40$ μm. Determine the internal pinchoff voltage and the pinchoff voltage of the JFET.

(Ans. $V_{od} = 1.236$ V, $V_p = 0.422$ V)

Also, we will see later that if the channel doping concentration were smaller the current capability of the device would decrease. There are definite tradeoffs to be considered in any design problem.

We have determined the pinchoff voltage for both n-channel and p-channel JFETs when the drain-to-source voltage is zero. Now consider the case when both gate and drain voltages are applied. The depletion region width will vary with distance through the channel. Figure 13.11 shows the simplified geometry for an n-channel device. The depletion width h_1 at the source end is a function of V_{bi} and V_{GS} but is not a function of drain voltage. The depletion width at the drain terminal is given by

$$h_2 = \left[\frac{2\epsilon_s(V_{bi} + V_{DS} - V_{GS})}{eN_d}\right]^{1/2} \tag{13.9}$$

Again, we must keep in mind that V_{GS} is a negative quantity for the n-channel device.

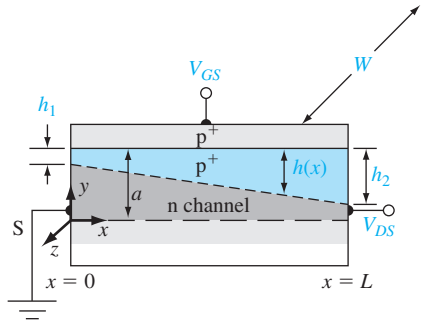


Figure 13.11 | Simplified geometry of an n-channel pn JFET.

Pinchoff at the drain terminal occurs when $h_2 = a$. At this point we reach what is known as the saturation condition; thus, we can write that $V_{DS} = V_{DS}(\text{sat})$. Then

$$a = \left[\frac{2\epsilon_s(V_{bi} + V_{DS}(\text{sat}) - V_{GS})}{eN_d}\right]^{1/2} \tag{13.10}$$

This can be rewritten as

$$V_{bi} + V_{DS}(\text{sat}) - V_{GS} = \frac{ea^2N_d}{2\epsilon_s} = V_{p0} \tag{13.11}$$

or

$$\boxed{V_{DS}(\text{sat}) = V_{p0} - (V_{bi} - V_{GS})} \tag{13.12}$$

Equation (13.12) gives the drain-to-source voltage to cause pinchoff at the drain terminal. The drain-to-source saturation voltage decreases with increasing reverse-biased gate-to-source voltage. We may note that Equation (13.12) has no meaning if $|V_{GS}| > |V_p|$.

In a p-channel JFET, the voltage polarities are the reverse of those in the n-channel device. We can show that, in the p-channel JFET at saturation,

$$\boxed{V_{SD}(\text{sat}) = V_{p0} - (V_{bi} + V_{GS})} \tag{13.13}$$

where now the source is positive with respect to the drain.

13.2.2 Ideal DC Current–Voltage Relationship—Depletion Mode JFET

The derivation of the ideal current–voltage relation of the JFET is somewhat tedious, and the resulting equations are cumbersome in hand calculations. Before we go through this derivation, consider the following expression, which is a good approximation

to the *I–V* characteristics when the JFET is biased in the saturation region. This equation is used extensively in JFET applications and is given by

$$I_D = I_{DSS}\left(1 - \frac{V_{GS}}{V_p}\right)^2 \tag{13.14}$$

where I_{DSS} is the saturation current when $V_{GS} = 0$. At the end of this section, we compare the approximation given by Equation (13.14) and the ideal current–voltage equation that we have derived.

***I–V* Derivation** The ideal current–voltage relationship of the JFET is derived by starting with Ohm's law. Consider an n-channel JFET with the geometry shown in Figure 13.11. We are considering half of the two-sided symmetrical geometry. The differential resistance of the channel at a point x in the channel is

$$dR = \frac{\rho dx}{A(x)} \tag{13.15}$$

where ρ is the resistivity and $A(x)$ is the cross-sectional area. If we neglect the minority carrier holes in the n channel, the channel resistivity is

$$\rho = \frac{1}{e\mu_n N_d} \tag{13.16}$$

The cross-sectional area is given by

$$A(x) = [a - h(x)]\,W \tag{13.17}$$

where W is the channel width. Equation (13.15) can now be written as

$$dR = \frac{dx}{e\mu_n N_d[a - h(x)]\,W} \tag{13.18}$$

The differential voltage across a differential length dx can be written as

$$dV(x) = I_{D1}\,dR(x) \tag{13.19}$$

where the drain current I_{D1} is constant through the channel. Substituting Equation (13.18) into Equation (13.19), we have

$$dV(x) = \frac{I_{D1}\,dx}{e\mu_n N_d W[a - h(x)]} \tag{13.20a}$$

or

$$I_{D1}\,dx = e\mu_n N_d\,W[a - h(x)]\,dV(x) \tag{13.20b}$$

The depletion width $h(x)$ is given by

$$h(x) = \left\{\frac{2\epsilon_s[V(x) + V_{bi} - V_{GS}]}{eN_d}\right\}^{1/2} \tag{13.21}$$

where $V(x)$ is the potential in the channel due to the drain-to-source voltage. Solving for $V(x)$ in Equation (13.21) and taking the differential, we have

$$dV(x) = \frac{eN_d h(x)\,dh(x)}{\epsilon_s} \tag{13.22}$$

Then Equation (13.20b) becomes

$$I_{D1}\,dx = \frac{\mu_n (eN_d)^2 W}{\epsilon_s}\,[ah(x)\,dh(x) - h(x)^2\,dh(x)] \tag{13.23}$$

The drain current I_{D1} is found by integrating Equation (13.23) along the channel length. Assuming the current and mobility are constant through the channel, we obtain

$$I_{D1} = \frac{\mu_n (eN_d)^2 W}{\epsilon_s L}\left[\int_{h_1}^{h_2} ah\,dh - \int_{h_1}^{h_2} h^2\,dh\right] \tag{13.24}$$

or

$$I_{D1} = \frac{\mu_n (eN_d)^2 W}{\epsilon_s L}\left[\frac{a}{2}(h_2^2 - h_1^2) - \frac{1}{3}(h_2^3 - h_1^3)\right] \tag{13.25}$$

Noting that

$$h_2^2 = \frac{2\epsilon_s (V_{DS} + V_{bi} - V_{GS})}{eN_d} \tag{13.26a}$$

$$h_1^2 = \frac{2\epsilon_s (V_{bi} - V_{GS})}{eN_d} \tag{13.26b}$$

and

$$V_{p0} = \frac{ea^2 N_d}{2\epsilon_s} \tag{13.26c}$$

Equation (13.25) can be written as

$$I_{D1} = \frac{\mu_n (eN_d)^2 W a^3}{2\epsilon_s L}\left[\frac{V_{DS}}{V_{p0}} - \frac{2}{3}\left(\frac{V_{DS} + V_{bi} - V_{GS}}{V_{p0}}\right)^{3/2} + \frac{2}{3}\left(\frac{V_{bi} - V_{GS}}{V_{p0}}\right)^{3/2}\right] \tag{13.27}$$

We may define

$$\boxed{I_{P1} \equiv \frac{\mu_n (eN_d)^2 W a^3}{6\epsilon_s L}} \tag{13.28}$$

where I_{P1} is called the pinchoff current. Equation (13.27) becomes

$$\boxed{I_{D1} = I_{P1}\left[3\left(\frac{V_{DS}}{V_{p0}}\right) - 2\left(\frac{V_{DS} + V_{bi} - V_{GS}}{V_{p0}}\right)^{3/2} + 2\left(\frac{V_{bi} - V_{GS}}{V_{p0}}\right)^{3/2}\right]} \tag{13.29}$$

Equation (13.29) is valid for $0 \le |V_{GS}| \le |V_p|$ and $0 \le V_{DS} \le V_{DS}(\text{sat})$. The pinchoff current I_{P1} would be the maximum drain current in the JFET if the zero-biased depletion regions could be ignored or if V_{GS} and V_{bi} were both zero.

Equation (13.29) is the current–voltage relationship for the one-sided n-channel JFET in the nonsaturation region. For the two-sided symmetrical JFET shown in Figure 13.9a, the total drain current would be $I_{D2} = 2I_{D1}$.

Equation (13.27) can also be written as

$$I_{D1} = G_{01}\left\{V_{DS} - \frac{2}{3}\sqrt{\frac{1}{V_{p0}}}\left[(V_{DS} + V_{bi} - V_{GS})^{3/2} - (V_{bi} - V_{GS})^{3/2}\right]\right\} \tag{13.30}$$

where

$$G_{01} = \frac{\mu_n (eN_d)^2\, Wa^3}{2\epsilon_s L V_{p0}} = \frac{e\mu_n N_d Wa}{L} = \frac{3I_{P1}}{V_{p0}} \tag{13.31}$$

The channel conductance is defined as

$$g_d = \left.\frac{\partial I_{D1}}{\partial V_{DS}}\right|_{V_{DS}\to 0} \tag{13.32}$$

Taking the derivative of Equation (13.30) with respect to V_{DS}, we obtain

$$g_d = \left.\frac{\partial I_{D1}}{\partial V_{DS}}\right|_{V_{DS}\to 0} = G_{01}\left[1 - \left(\frac{V_{bi} - V_{GS}}{V_{p0}}\right)^{1/2}\right] \tag{13.33}$$

We may note from Equation (13.33) that G_{01} would be the conductance of the channel if both V_{bi} and V_{GS} were zero. This condition would exist if no space charge regions existed in the channel. We may also note, from Equation (13.33), that the channel conductance is modulated or controlled by the gate voltage. This channel conductance modulation is the basis of the field-effect phenomenon.

We have shown that the drain becomes pinched off, for the n-channel JFET, when

$$V_{DS} = V_{DS}(\text{sat}) = V_{p0} - (V_{bi} - V_{GS}) \tag{13.34}$$

In the saturation region, the saturation drain current is determined by setting $V_{DS} = V_{DS}(\text{sat})$ in Equation (13.29) so that

$$\boxed{I_{D1} = I_{D1}(\text{sat}) = I_{P1}\left\{1 - 3\left(\frac{V_{bi} - V_{GS}}{V_{p0}}\right)\left[1 - \frac{2}{3}\sqrt{\frac{V_{bi} - V_{GS}}{V_{p0}}}\right]\right\}} \tag{13.35}$$

The ideal saturation drain current is independent of the drain-to-source voltage. Figure 13.12 shows the ideal current–voltage characteristics of a silicon n-channel JFET.

EXAMPLE 13.3

Objective: Calculate the maximum current in an n-channel JFET.

Consider a silicon n-channel JFET at $T = 300$ K with the following parameters: $N_a = 10^{18}$ cm^{-3}, $N_d = 10^{16}$ cm^{-3}, $a = 0.75$ μm, $L = 10$ μm, $W = 30$ μm, and $\mu_n = 1000$ cm^2/V-s.

■ Solution

The pinchoff current from Equation (13.28) becomes

$$I_{P1} = \frac{(1000)[(1.6 \times 10^{-19})(10^{16})]^2\,(30 \times 10^{-4})(0.75 \times 10^{-4})^3}{6(11.7)(8.85 \times 10^{-14})(10 \times 10^{-4})} = 0.522 \text{ mA}$$

We also have from Example 13.1 that $V_{bi} = 0.814$ V and $V_{p0} = 4.35$ V. The maximum current occurs when $V_{GS} = 0$, so from Equation (13.35)

$$I_{D1}(\text{max}) = I_{P1}\left\{1 - 3\left(\frac{V_{bi}}{V_{p0}}\right)\left[1 - \frac{2}{3}\sqrt{\frac{V_{bi}}{V_{p0}}}\right]\right\} \tag{13.36}$$

or

$$I_{D1}(\text{max}) = (0.522)\left\{1 - 3\left(\frac{0.814}{4.35}\right)\left[1 - \frac{2}{3}\sqrt{\frac{0.814}{4.35}}\right]\right\} = 0.313 \text{ mA}$$

Comment

The maximum current through the JFET is less than the pinchoff current I_{P1}.

EXERCISE PROBLEM

Ex 13.3 Consider an n-channel silicon pn JFET with parameters $N_a = 10^{18}$ cm^{-3}, $N_d = 10^{16}$ cm^{-3}, $a = 0.40$ μm, $L = 5$ μm, $W = 50$ μm, and $\mu_n = 900$ cm^2/V-s. Calculate the pinchoff current I_{P1} and the maximum drain current I_{D1}(sat) for $V_{GS} = 0$.

[Ans. $I_{P1} = 0.237$ mA, I_{D1}(sat) $= 22.13$ μA]

The maximum saturation current calculated in this example is considerably less than that shown in Figure 13.12 because of the big difference in the width-to-length ratios. Once the pinchoff voltage of JFET has been designed, the channel width W is the primary design variable for determining the current capability of a device.

Summary Equations (13.29) and (13.35) are rather cumbersome to use in any hand calculations. We may show that, in the saturation region, the drain current is given to a good approximation by Equation (13.14), stated at the beginning of this section as

$$I_D = I_{DSS}\left(1 - \frac{V_{GS}}{V_p}\right)^2$$

The current I_{DSS} is the maximum drain current and is the same as I_{D1}(max) in Equation (13.36). The parameter V_{GS} is the gate-to-source voltage and V_p is the pinchoff

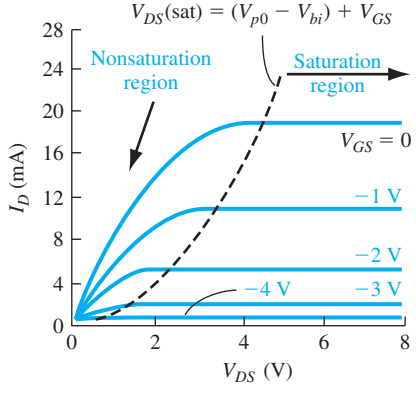


Figure 13.12 | Ideal current–voltage characteristics of a silicon n-channel JFET with $a = 1.5$ μm, $W/L = 170$, and $N_d = 2.5 \times 10^{15}$ cm^{-3}.
(From Yang [22].)

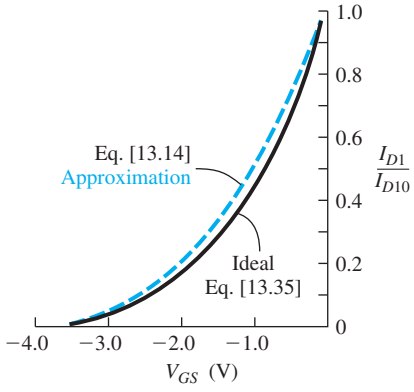


Figure 13.13 | Comparison of Equations (13.14) and (13.35) for the I_D versus V_{GS} characteristics of a JFET biased in the saturation region.

voltage. We may note that, for n-channel depletion mode JFET, both V_{GS} and V_p are negative and, for the p-channel depletion mode device, both are positive. Figure 13.13 shows the comparison between Equations (13.14) and (13.35).

13.2.3 Transconductance

The transconductance is the transistor gain of the JFET; it indicates the amount of control the gate voltage has on the drain current. The transconductance is defined as

$$g_m = \frac{\partial I_D}{\partial V_{GS}} \tag{13.37}$$

Using the expressions for the ideal drain current derived in the last section, we can write the expressions for the transconductance.

The drain current for an n-channel depletion mode device in the nonsaturation region is given by Equation (13.29). We can then determine the transconductance of the transistor in the same region as

$$g_{mL} = \frac{\partial I_{D1}}{\partial V_{GS}} = \frac{3I_{P1}}{V_{p0}}\sqrt{\frac{V_{bi} - V_{GS}}{V_{p0}}}\left[\sqrt{\left(\frac{V_{DS}}{V_{bi} - V_{GS}}\right) + 1} - 1\right] \tag{13.38}$$

Taking the limit as V_{DS} becomes small, the transconductance becomes

$$g_{mL} \approx \frac{3I_{P1}}{2V_{p0}} \cdot \frac{V_{DS}}{\sqrt{V_{p0}(V_{bi} - V_{GS})}} \tag{13.39}$$

We can also write Equation (13.39) in terms of the conductance parameter G_{01} as

$$g_{mL} = \frac{G_{01}}{2} \cdot \frac{V_{DS}}{\sqrt{V_{p0}(V_{bi} - V_{GS})}} \tag{13.40}$$

The ideal drain current in the saturation region for the JFET is given by Equation (13.35). The transconductance in the saturation region is then found to be

$$g_{ms} = \frac{\partial I_{D1}(\text{sat})}{\partial V_{GS}} = \frac{3I_{P1}}{V_{p0}}\left(1 - \sqrt{\frac{V_{bi} - V_{GS}}{V_{p0}}}\right) = G_{01}\left(1 - \sqrt{\frac{V_{bi} - V_{GS}}{V_{p0}}}\right) \tag{13.41a}$$

Using the current–voltage approximation given by Equation (13.14), we can also write the transconductance as

$$g_{ms} = \frac{-2I_{DSS}}{V_p}\left(1 - \frac{V_{GS}}{V_p}\right) \tag{13.41b}$$

Since V_p is negative for the n-channel JFET, g_{ms} is positive.

EXAMPLE 13.4

Objective: Determine the maximum transconductance of an n-channel depletion mode JFET biased in the saturation region.

Consider the silicon JFET described in Example 13.3. We had calculated $I_{P1} = 0.522$ mA, $V_{bi} = 0.814$ V, and $V_{p0} = 4.35$ V.

■ Solution

The maximum transconductance occurs when $V_{GS} = 0$. Then Equation (13.41a) can be written as

$$g_{ms}(\text{max}) = \frac{3I_{P1}}{V_{p0}}\left(1 - \sqrt{\frac{V_{bi}}{V_{p0}}}\right) = \frac{3(0.522)}{4.35}\left(1 - \sqrt{\frac{0.814}{4.35}}\right) = 0.204 \text{ mA/V}$$

■ Comment

The saturation transconductance is a function of V_{GS} and becomes zero when $V_{GS} = V_P$.

■ EXERCISE PROBLEM

Ex 13.4 Determine the maximum transconductance of the n-channel JFET described in Exercise Problem Ex 13.3.

[Ans. $g_{ms}(\text{max}) = 0.109$ mA/V]

The experimental transconductance may deviate from this ideal expression due to a source series resistance. This effect will be considered later in the discussion of the small signal model of the JFET.

13.2.4 The MESFET

So far in our discussion, we have explicitly considered the pn JFET. The MESFET is the same basic device except that the pn junction is replaced by a Schottky barrier rectifying junction. The simplified MESFET geometry is shown in Figure 13.9b. MESFETs are usually fabricated in gallium arsenide. We will neglect any depletion region that may exist between the n channel and the substrate. We have also limited our discussion to depletion mode devices, wherein a gate-to-source voltage is applied to turn the transistor off. Enhancement mode GaAs MESFETs can be fabricated—their basic operation is discussed in Section 13.1.2. We can also consider enhancement mode GaAs pn JFETs.

Since the electron mobility in GaAs is much larger than the hole mobility, we will concentrate our discussion on n-channel GaAs MESFETs or JFETs. The definition of internal pinchoff voltage, given by Equation (13.3), also applies to these devices. In considering the enhancement mode JFET, the term threshold voltage is commonly used in place of pinchoff voltage. For this reason, we shall use the term threshold voltage in our discussion of MESFETs.

For the n-channel MESFET, the threshold voltage is defined from Equation (13.4) as

$$V_{bi} - V_T = V_{p0} \quad \text{or} \quad \boxed{V_T = V_{bi} - V_{p0}} \tag{13.42}$$

For an n-channel depletion mode JFET, $V_T < 0$, and for the enhancement mode device, $V_T > 0$. We can see from Equation (13.42) that $V_{bi} > V_{p0}$ for an enhancement mode n-channel JFET.

DESIGN EXAMPLE 13.5

Objective: Determine the channel thickness of a GaAs MESFET to achieve a specified threshold voltage.

Consider an n-channel GaAs MESFET at $T = 300$ K with a gold Schottky barrier contact. Assume the barrier height is $\phi_{Bn} = 0.89$ V. The n-channel doping is $N_d = 2 \times 10^{15}$ cm^{-3}. Design the channel thickness such that $V_T = +0.25$ V.

■ Solution

We find that

$$\phi_n = V_t \ln\left(\frac{N_c}{N_d}\right) = (0.0259) \ln\left(\frac{4.7 \times 10^{17}}{2 \times 10^{15}}\right) = 0.141 \text{ V}$$

The built-in potential barrier is then

$$V_{bi} = \phi_{Bn} - \phi_n = 0.89 - 0.141 = 0.749 \text{ V}$$

The threshold voltage, from Equation (13.42), is

$$V_T = V_{bi} - V_{p0}$$

or

$$V_{p0} = V_{bi} - V_T = 0.749 - 0.25 = 0.499 \text{ V}$$

Now

$$V_{p0} = \frac{eа^2 N_d}{2\epsilon_s}$$

or

$$0.499 = \frac{a^2 (1.6 \times 10^{-19}) (2 \times 10^{15})}{2(13.1) (8.85 \times 10^{-14})}$$

The channel thickness is then

$$a = 0.601 \ \mu\text{m}$$

Comment

For this enhancement mode n-channel MESFET, the internal pinchoff voltage is less than the built-in potential barrier. A smaller channel thickness would result in a larger threshold voltage.

EXERCISE PROBLEM

Ex 13.5 Consider an n-channel GaAs MESFET with a gate barrier height of $\phi_{Bn} = 0.85$ V. The channel doping concentration is $N_d = 5 \times 10^{15}$ cm^{-3} and the channel thickness is $a = 0.40$ μm. Calculate the internal pinchoff voltage and the threshold voltage.

(Ans. $V_{po} = 0.5520$ V, $V_T = 0.1180$ V)

The design of enhancement mode JFETs implies the use of narrow channel thicknesses and low channel doping concentrations to achieve this condition. The precise control of the channel thickness and doping concentration necessary to achieve internal pinchoff voltages of a few tenths of a volt makes the fabrication of enhancement mode MESFETs difficult.

EXAMPLE 13.6

Objective: Calculate the forward-bias gate voltage required in an n-channel GaAs enhancement mode pn JFET to open up a channel.

Consider a GaAs n-channel pn JFET at $T = 300$ K with $N_a = 10^{18}$ cm^{-3}, $N_d = 3 \times 10^{15}$ cm^{-3}, and $a = 0.70$ μm. Determine the forward-bias gate voltage required to open a channel region that is 0.10 μm thick with zero drain voltage.

Solution

The built-in potential barrier is

$$V_{bi} = V_t \ln\left(\frac{N_a N_d}{n_i^2}\right) = (0.0259)\ln\left[\frac{(10^{18})(3 \times 10^{15})}{(1.8 \times 10^6)^2}\right] = 1.25 \text{ V}$$

The internal pinchoff voltage is

$$V_{p0} = \frac{ea^2 N_d}{2\epsilon_s} = \frac{(1.6 \times 10^{-19})(0.7 \times 10^{-4})^2(3 \times 10^{15})}{2(13.1)(8.85 \times 10^{-14})} = 1.01 \text{ V}$$

which gives a threshold voltage of

$$V_T = V_{bi} - V_{p0} = 0.24 \text{ V}$$

The channel depletion width is given by Equation (13.1). Setting $h = 0.60$ μm will yield an undepleted channel thickness of 0.1 μm. Solving for V_{GS}, we obtain

$$V_{GS} = V_{bi} - \frac{eh^2 N_d}{2\epsilon_s} = 1.25 - \frac{(1.6 \times 10^{-19})(0.6 \times 10^{-4})^2(3 \times 10^{15})}{2(13.1)(8.85 \times 10^{-14})}$$

$$= 1.25 - 0.745 = 0.50 \text{ V}$$

Comment

An applied gate voltage of 0.50 V is greater than the threshold voltage, so the induced depletion region will be smaller than the metallurgical channel thickness. An n-channel region is

then formed between the source and drain contacts. The forward-bias gate voltage must not be too large or an undesirable gate current will be present in the device.

EXERCISE PROBLEM

Ex 13.6 An n-channel GaAs MESFET has a gate barrier height of $\phi_{Bn} = 0.89$ V. The channel doping concentration is $N_d = 10^{16}$ cm^{-3}. What channel thickness is required to yield a threshold voltage of $V_T = 0.25$ V?

(Ans. $a = 0.280\ \mu m$)

Ideally, the *I–V* characteristics of the enhancement mode device are the same as the depletion mode device—the only real difference is the relative values of the internal pinchoff voltage. The current in the saturation region is given by Equation (13.35) as

$$I_{D1} = I_{D1}(\text{sat}) = I_{P1}\left\{1 - 3\left(\frac{V_{bi} - V_{GS}}{V_{p0}}\right)\left[1 - \frac{2}{3}\sqrt{\frac{V_{bi} - V_{GS}}{V_{p0}}}\right]\right\}$$

The threshold voltage for the n-channel device is defined in Equation (13.42) as $V_T = V_{bi} - V_{p0}$, so we can also write

$$V_{bi} = V_T + V_{p0} \tag{13.43}$$

Substituting this expression for V_{bi} into Equation (13.35), we obtain

$$I_{D1}(\text{sat}) = I_{P1}\left\{1 - 3\left[1 - \left(\frac{V_{GS} - V_T}{V_{p0}}\right)\right] + 2\left[1 - \left(\frac{V_{GS} - V_T}{V_{p0}}\right)\right]^{3/2}\right\} \tag{13.44}$$

Equation (13.44) is valid for $V_{GS} \geq V_T$.

When the transistor first turns on, we have $(V_{GS} - V_T) \ll V_{p0}$. Equation (13.44) can then be expanded into a Taylor series and we obtain

$$I_{D1}(\text{sat}) \approx I_{P1}\left[\frac{3}{4}\left(\frac{V_{GS} - V_T}{V_{p0}}\right)\right]^2 \tag{13.45}$$

Substituting the expressions for I_{P1} and V_{p0}, Equation (13.45) becomes

$$I_{D1}(\text{sat}) = \frac{\mu_n \epsilon_s W}{2aL}(V_{GS} - V_T)^2 \quad \text{for } V_{GS} \geq V_T \tag{13.46}$$

We can now write Equation (13.46) as

$$\boxed{I_{D1}(\text{sat}) = k_n(V_{GS} - V_T)^2} \tag{13.47}$$

where

$$\boxed{k_n = \frac{\mu_n \epsilon_s W}{2aL}} \tag{13.48}$$

The factor k_n is called a *conduction parameter*. The form of Equation (13.47) is the same as for a MOSFET.

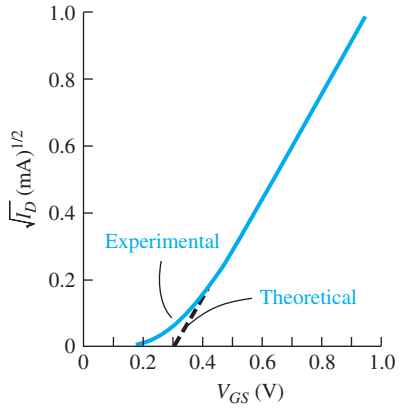


Figure 13.14 | Experimental and theoretical $\sqrt{I_D}$ versus V_{GS} characteristics of an enhancement mode JFET.

The square root of Equation (13.47), or $\sqrt{I_{D1}(\text{sat})}$ versus V_{GS}, is plotted as the ideal dotted curve shown in Figure 13.14. The ideal curve intersects the voltage axis at the threshold voltage, V_T. The solid line shows an experimental plot. Equation (13.46) does not describe the experimental results well near the threshold voltage. The ideal current–voltage relationship is derived assuming an abrupt depletion approximation for the pn junction. However, when the depletion region extends almost through the channel, a more accurate model of the space charge region must be used to more accurately predict the drain current characteristics near threshold. We consider the subthreshold conduction in Section 13.3.3.

DESIGN EXAMPLE 13.7

Objective: Design the channel width of an n-channel GaAs enhancement-mode pn JFET to produce a specified current for a given bias.

Consider the GaAs JFET described in Example 13.6. In addition, assume $\mu_n = 8000\ \text{cm}^2/\text{V-s}$ and $L = 1.2\ \mu\text{m}$. Design the width such that $I_{D1} = 75\ \mu\text{A}$ with an applied voltage of $V_{GS} = 0.5\ \text{V}$.

■ Solution

In the saturation region, the current is given by

$$I_{D1} = k_n(V_{GS} - V_T)^2$$

or

$$75 \times 10^{-6} = k_n(0.5 - 0.24)^2$$

The conduction parameter is then

$$k_n = 1.109\ \text{mA/V}^2$$

The conduction parameter, from Equation (13.48), is given by

$$k_n = \frac{\mu_n \epsilon_s W}{2aL}$$

or

$$1.109 \times 10^{-3} = \frac{(8000)(13.1)(8.85 \times 10^{-14})(W)}{2(0.70 \times 10^{-4})(1.2 \times 10^{-4})}$$

The required channel width is then

$$W = 20.1\ \mu\text{m}$$

Comment

The saturation current will obviously increase if V_{GS} is increased or if the width of the transistor is increased.

EXERCISE PROBLEM

Ex 13.7 Consider the GaAs MESFET described in Exercise Problem Ex 13.5. In addition, assume $\mu_n = 7000\ \text{cm}^2/\text{V-s}$, $L = 0.8\ \mu\text{m}$, and $W = 25\ \mu\text{m}$. Calculate the conduction parameter k_n and the current $I_{D1}(\text{sat})$ for $V_{GS} = 0.50$ V.
[Ans. $k_n = 3.17\ \text{mA/V}^2$, $I_{D1}(\text{sat}) = 0.325\ \text{mA}$]

The transconductance of the enhancement mode device operating in the saturation region can also be derived. Using Equation (13.47), we can write

$$g_{ms} = \frac{\partial I_{D1}(\text{sat})}{\partial V_{GS}} = 2k_n(V_{GS} - V_T) \tag{13.49}$$

The transconductance increases as V_{GS} increases for the enhancement mode device as it did for the depletion mode device.

TEST YOUR UNDERSTANDING

TYU 13.1 Consider a GaAs pn junction n-channel FET. The p^+ gate doping concentration is $N_a = 5 \times 10^{18}\ \text{cm}^{-3}$ and the n-channel doping concentration is $N_d = 5 \times 10^{15}\ \text{cm}^{-3}$. The zero-bias depletion width is to be $1.2a$; that is, the channel is completely depleted at zero bias. Determine the value of a and the pinchoff voltage.
(Ans. $a = 0.513\ \mu\text{m}$, $V_p = +0.397\ \text{V}$)

TYU 13.2 The pinchoff current I_{P1} given by Equation (13.28) and the pinchoff voltage given by Equation (13.26c) also apply to a p-channel JFET in which μ_n is replaced by μ_p and N_d is replaced by N_a. Assume a p-channel silicon JFET has the following parameters: $N_d = 5 \times 10^{18}\ \text{cm}^{-3}$, $N_a = 2 \times 10^{16}\ \text{cm}^{-3}$, $a = 0.50\ \mu\text{m}$, $L = 5\ \mu\text{m}$, $W = 40\ \mu\text{m}$, and $\mu_p = 400\ \text{cm}^2/\text{V-s}$. Calculate the pinchoff current I_{P1} and the maximum drain current $I_{D1}(\text{sat})$ for $V_{GS} = 0$.
[Ans. $I_{P1} = 0.659\ \text{mA}$, $I_{D1}(\text{sat}) = 0.256\ \text{mA}$]

*13.3 | NONIDEAL EFFECTS

As with any semiconductor device, there are nonideal effects that will change the ideal device characteristics. In all of the previous discussions, we have considered an ideal transistor with a constant channel length and constant mobility; we have also

neglected gate currents. However, when a JFET is biased in the saturation region, the effective electrical channel length is a function of V_{DS}. This nonideal effect is called channel length modulation. In addition, when a transistor is biased near or in the saturation region, the electric field in the channel can become large enough so that the majority carriers reach their saturation velocity. At this point, the mobility is no longer a constant. The magnitude of the gate current will affect the input impedance, which may need to be taken into account in a circuit design.

13.3.1 Channel Length Modulation

The expression for the drain current is inversely proportional to the channel length L as given, for example, by Equation (13.27). In deriving the current equations, we have implicitly assumed that the channel length was constant. However, the effective channel length can change. Figure 13.5 shows the space charge region in the channel when the transistor is biased in the saturation region. The neutral n-channel length decreases as V_{DS} increases; thus, the drain current will increase. The change in the effective channel length and the corresponding change in drain current is called channel length modulation.

The pinchoff current, Equation (13.28), is modified by the channel length modulation and can be written as

$$I'_{P1} = \frac{\mu_n (eN_d)^2 W a^3}{6\epsilon_s L'} \tag{13.50}$$

where

$$L' \approx L - \frac{1}{2}\Delta L \tag{13.51}$$

If we assume the channel depletion region shown in Figure 13.5 extends equally into the channel and drain regions, then as a first approximation, we will include the factor $\frac{1}{2}$ in the expression for L'.

The drain current can be written as

$$I'_{D1} = I_{D1} \cdot \frac{I'_{P1}}{I_{P1}} = I_{D1}\left(\frac{L}{L - \frac{1}{2}\Delta L}\right) \tag{13.52}$$

where I_{D1} is the ideal drain current predicted by Equation (13.35). Another form of the current–voltage characteristic in the saturation region is given by

$$I'_{D1}(\text{sat}) = I_{D1}(\text{sat})(1 + \lambda V_{DS}) \tag{13.53}$$

The effective channel length L' supports the $V_{DS}(\text{sat})$ voltage, and the space charge region length ΔL in the channel supports the drain voltage beyond the saturation value. Neglecting charges in the space charge region due to current flow, the depletion length ΔL is then, to a first approximation, given by

$$\Delta L = \left[\frac{2\epsilon_s (V_{DS} - V_{DS}(\text{sat}))}{eN_d}\right]^{1/2} \tag{13.54}$$

Since the effective channel length changes with V_{DS}, the drain current is now a function of V_{DS}. The small-signal output impedance at the drain terminal can be defined as

$$r_{ds} = \frac{\partial V_{DS}}{\partial I'_{D1}} \approx \frac{\Delta V_{DS}}{\Delta I'_{D1}} \tag{13.55}$$

EXAMPLE 13.8

Objective: Calculate the small-signal output resistance at the drain terminal due to channel length modulation effects.

Consider an n-channel depletion mode silicon JFET with a channel doping of $N_d = 3 \times 10^{15}$ cm^{-3}. Calculate r_{ds} for the case when V_{DS} changes from $V_{DS}(1) = V_{DS}(\text{sat}) + 2.0$ to $V_{DS}(2) = V_{DS}(\text{sat}) + 2.5$. Assume $L = 10\ \mu$m and $I_{D1} = 4.0$ mA.

■ Solution

We have that

$$r_{ds} = \frac{\Delta V_{DS}}{\Delta I'_{D1}} = \frac{V_{DS}(2) - V_{DS}(1)}{\Delta I'_{D1}(2) - I'_{D1}(1)}$$

We can calculate the change in the channel length for the two voltages as

$$\Delta L(2) = \left[\frac{2\epsilon_s(V_{DS}(2) - V_{DS}(\text{sat}))}{eN_d}\right]^{1/2} = \left[\frac{2(11.7)(8.85 \times 10^{-14})(2.5)}{(1.6 \times 10^{-19})(3 \times 10^{15})}\right]^{1/2} = 1.04\ \mu\text{m}$$

and

$$\Delta L(1) = \left[\frac{2(11.7)(8.85 \times 10^{-14})(2.0)}{(1.6 \times 10^{-19})(3 \times 10^{15})}\right]^{1/2} = 0.929\ \mu\text{m}$$

The drain currents are then

$$I'_{D1}(2) = I_{D1}\left(\frac{L}{L - \frac{1}{2}\Delta L(2)}\right) = 4.0\left(\frac{10}{9.48}\right)$$

and

$$I'_{D1}(1) = I_{D1}\left(\frac{L}{L - \frac{1}{2}\Delta L(1)}\right) = 4.0\left(\frac{10}{9.54}\right)$$

The output resistance can be calculated as

$$r_{ds} = \frac{2.5 - 2.0}{4\left(\frac{10}{9.48}\right) - 4\left(\frac{10}{9.54}\right)} = 18.9\ \text{k}\Omega$$

■ Comment

This value of output resistance is significantly less than the ideal value of infinity.

■ EXERCISE PROBLEM

Ex 13.8 Repeat Example 13.8 if the channel doping concentration increases to $N_d = 10^{16}$ cm^{-3}. All other parameters remain the same.

(Ans. $r_{ds} = 39.4\text{k}\Omega$)

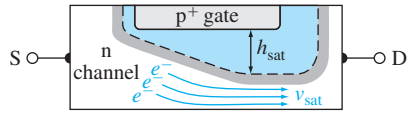


Figure 13.15 | Cross section of JFET showing carrier velocity and space charge width saturation effects.

For high-frequency MESFETs, typical channel lengths are on the order of 1 μm. Channel length modulation and other effects become very important in short-channel devices.

13.3.2 Velocity Saturation Effects

We have seen that the drift velocity of a carrier in silicon saturates with increasing electric field. This velocity saturation effect implies that the mobility is not a constant. For very short channels, the carriers can easily reach their saturation velocity, which changes the *I–V* characteristics of the JFET.

Figure 13.15 shows the channel region with an applied drain voltage. As the channel narrows at the drain terminal, the velocity of the carriers increases since the current through the channel is constant. The carriers first saturate at the drain end of the channel. The depletion region will reach a saturation thickness, so we can write

$$I_{D1}(\text{sat}) = eN_d v_{\text{sat}}(a - h_{\text{sat}})W \tag{13.56}$$

where v_{sat} is the saturation velocity and h_{sat} is the saturation depletion width. This saturation effect occurs at a drain voltage smaller than the $V_{DS}(\text{sat})$ value determined previously. Both $I_{DS}(\text{sat})$ and $V_{DS}(\text{sat})$ will be smaller than previously calculated.

Figure 13.16 shows normalized plots of I_D versus V_{DS}. Figure 13.16a is for the case of a constant mobility and Figure 13.16b is for the case of velocity saturation. Since the *I–V* characteristics change when velocity saturation occurs, the transconductance will also change—the transconductance will become smaller; hence, the effective gain of the transistor decreases when velocity saturation occurs.

13.3.3 Subthreshold and Gate Current Effects

The subthreshold current is the drain current in the JFET that exists when the gate voltage is below the pinchoff or threshold value. The subthreshold conduction is shown in Figure 13.14. When the JFET is biased in the saturation region, the drain current varies quadratically with gate-to-source voltage. When V_{GS} is below the threshold value, the drain current varies exponentially with gate-to-source voltage. Near threshold, the abrupt depletion approximation does not accurately model the channel region: A more detailed potential profile in the space charge region must be used. However, these calculations are beyond the scope of this chapter.

When the gate voltage is approximately 0.5 to 1.0 V below threshold in an n-channel MESFET, the drain current reaches a minimum value and then slowly increases as the gate voltage decreases. The drain current in this region is the gate leakage current. Figure 13.17 is a plot of the drain current versus V_{GS} for the three regions

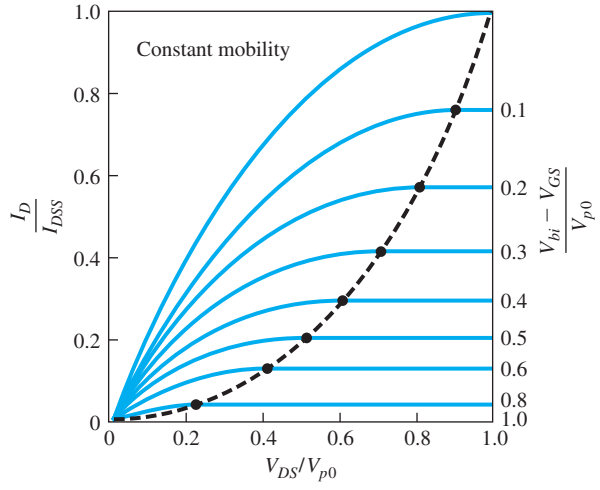


Figure 13.16 | Normalized I_D versus V_{DS} plots for a constant mobility and field-dependent mobility.
(From Sze [19].)

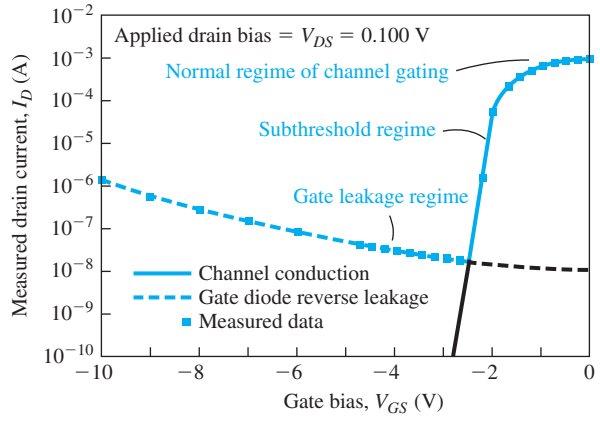


Figure 13.17 | Measured drain current versus V_{GS} for a GaAs MESFET showing the normal drain current, subthreshold current, and gate leakage current.
(From Daring [2].)

of gate voltage. The curve illustrates that the drain current becomes small below threshold, but is not zero. The minimum drain current may need to be accounted for in low-power circuit applications.

*13.4 | EQUIVALENT CIRCUIT AND FREQUENCY LIMITATIONS

In order to analyze a transistor circuit, one needs a mathematical model or equivalent circuit of the transistor. One of the most useful models is the small-signal equivalent circuit, which applies to transistors used in linear amplifier circuits. This equivalent circuit will introduce frequency effects in the transistor through the equivalent capacitor–resistor circuits. The various physical factors in the JFET affecting the frequency limitations are considered here and a transistor cutoff frequency, which is a figure of merit, is then defined.

13.4.1 Small-Signal Equivalent Circuit

The cross section of an n-channel pn JFET is shown in Figure 13.18, including source and drain series resistances. The substrate may be semi-insulating gallium arsenide or it may be a p^+ type substrate.

Figure 13.19 shows a small-signal equivalent circuit for the JFET. The voltage $V_{g's'}$ is the internal gate-to-source voltage that controls the drain current. The r_{gs} and C_{gs} parameters are the gate-to-source diffusion resistance and junction capacitance, respectively. The gate-to-source junction is reverse biased for depletion mode devices and has only a small forward-bias voltage for enhancement mode devices, so that normally r_{gs} is large. The parameters r_{gd} and C_{gd} are the gate-to-drain resistance and capacitance, respectively. The resistance r_{ds} is the finite drain resistance, which is a function of the channel length modulation effect. The C_{ds} capacitance is mainly a drain-to-source parasitic capacitance and C_s is the drain-to-substrate capacitance.

The ideal small-signal equivalent circuit is shown in Figure 13.20a. All diffusion resistances are infinite, the series resistances are zero, and at low frequency the

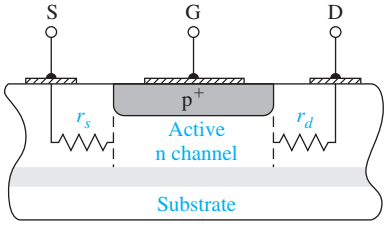


Figure 13.18 | Cross section of JFET with source and drain series resistance.

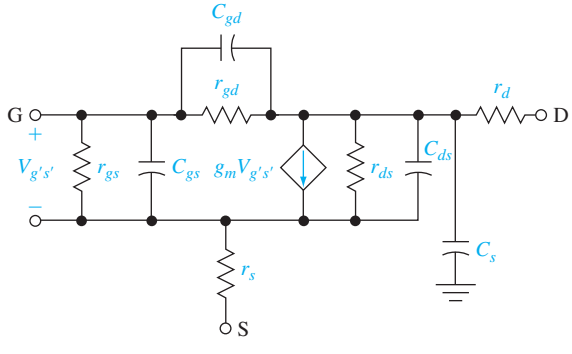


Figure 13.19 | Small-signal equivalent circuit of JFET.

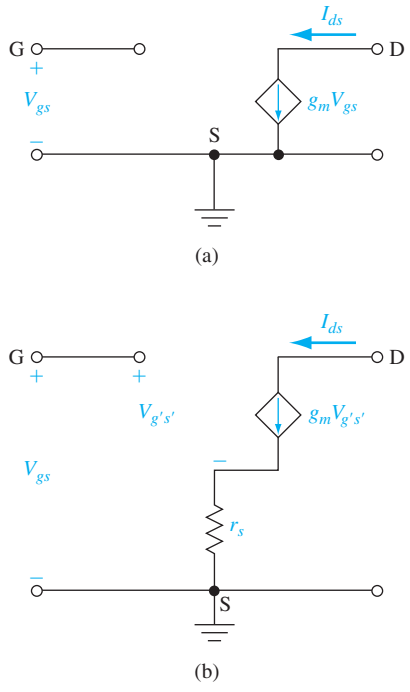


Figure 13.20 | (a) Ideal low-frequency small-signal equivalent circuit. (b) Ideal equivalent circuit including r_s.

capacitances become open circuits. The small-signal drain current is now

$$I_{ds} = g_m V_{gs} \tag{13.57}$$

which is a function only of the transconductance and the input-signal voltage.

The effect of the source series resistance can be determined using Figure 13.20b. We have

$$I_{ds} = g_m V_{g's'} \tag{13.58}$$

The relation between V_{gs} and $V_{g's'}$ can be found from

$$V_{gs} = V_{g's'} + (g_m V_{g's'}) r_s = (1 + g_m r_s) V_{g's'} \tag{13.59}$$

Equation (13.58) can then be written as

$$I_{ds} = \left(\frac{g_m}{1 + g_m r_s}\right) V_{gs} = g'_m V_{gs} \tag{13.60}$$

The effect of the source resistance is to reduce the effective transconductance or transistor gain.

Recall that g_m is a function of the dc gate-to-source voltage, so g'_m will also be a function of V_{GS}. Equation (13.41b) is the relation between g_m and V_{GS} when the

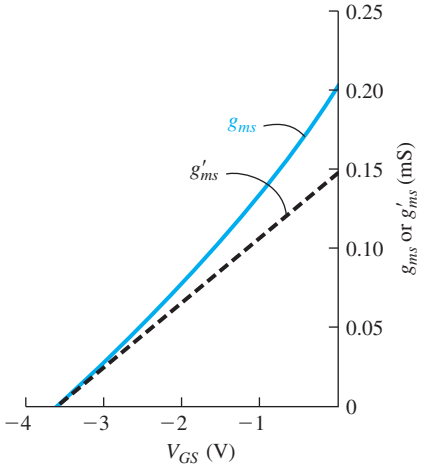


Figure 13.21 | JFET transconductance versus V_{GS} (a) without and (b) with a source series resistance.

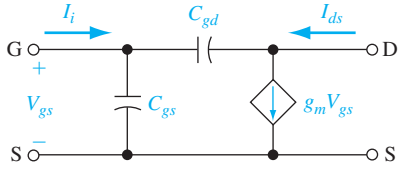


Figure 13.22 | A small-signal equivalent circuit with capacitance.

transistor is biased in the saturation region. Figure 13.21 shows a comparison between the theoretical and experimental transconductance values using the parameters from Example 13.4 and letting $r_s = 2000\ \Omega$. (A value of $r_s = 2000\ \Omega$ may seem excessive, but keep in mind that the active thickness of the semiconductor may be on the order of 1 μm or less; thus, a large series resistance may result if special care is not taken.)

13.4.2 Frequency Limitation Factors and Cutoff Frequency

There are two frequency limitation factors in a JFET. The first is the channel transit time. If we assume a channel length of 1 μm and assume carriers are traveling at their saturation velocity, then the transit time is on the order of

$$\tau_t = \frac{L}{v_s} = \frac{1 \times 10^{-4}}{1 \times 10^{+7}} = 10 \text{ ps} \tag{13.61}$$

The channel transit time is normally not the limiting factor except in very high frequency devices.

The second frequency limitation factor is the capacitance charging time. Figure 13.22 is a simplified equivalent circuit that includes the primary capacitances and ignores the diffusion resistances. The output current will be the short-circuit current. As the frequency of the input-signal voltage V_{gs} increases, the impedance of C_{gd} and C_{gs} decreases so the current through C_{gd} will increase. For a constant $g_m V_{gs}$, the I_{ds} current will then decrease. The output current then becomes a function of frequency.

If the capacitance charging time is the limiting factor, then the cutoff frequency f_T is defined as the frequency at which the magnitude of the input current I_i is equal to the magnitude of the ideal output current $g_m V_{gs}$ of the intrinsic transistor. We have,

when the output is short-circuited,

$$I_i = j\omega\,(C_{gs} + C_{gd})V_{gs} \tag{13.62}$$

If we let $C_G = C_{gs} + C_{gd}$, then at the cutoff frequency

$$|I_i| = 2\pi f_T C_G V_{gs} = g_m V_{gs} \tag{13.63}$$

or

$$f_T = \frac{g_m}{2\pi C_g} \tag{13.64}$$

From Equation (13.41b), the maximum possible transconductance is

$$g_{ms}\,(\max) = G_{01} = \frac{e\mu_n N_d W a}{L} \tag{13.65}$$

and the minimum gate capacitance is

$$C_G\,(\min) = \frac{\epsilon_s W L}{a} \tag{13.66}$$

where a is the maximum space charge width. The maximum cutoff frequency can be written as

$$\boxed{f_T = \frac{e\mu_n N_d a^2}{2\pi\epsilon_s L^2}} \tag{13.67}$$

EXAMPLE 13.9

Objective: Calculate the cutoff frequency of a silicon JFET.

Consider a silicon JFET with the following parameters:

$$\mu_n = 1000\ \text{cm}^2/\text{V-s} \qquad a = 0.60\ \mu\text{m}$$

$$N_d = 10^{16}\ \text{cm}^{-3} \qquad L = 5\ \mu\text{m}$$

Solution

Substituting the parameters into Equation (13.67), we have

$$f_T = \frac{e\mu_n N_d a^2}{2\pi\epsilon_s L^2} = \frac{(1.6\times10^{-19})(1000)(10^{16})\,(0.6\times10^{-4})^2}{2\pi(11.7)(8.85\times10^{-14})(5\times10^{-4})^2} = 3.54\ \text{GHz}$$

Comment

This example shows that even silicon JFETs can have relatively large cutoff frequencies.

EXERCISE PROBLEM

Ex 13.9 The parameters of an n-channel silicon JFET are $\mu_n = 1000\ \text{cm}^2/\text{V-s}$, $N_d = 5\times10^{15}\ \text{cm}^{-3}$, $a = 0.50\ \mu\text{m}$, and $L = 2\ \mu\text{m}$. Determine the cutoff frequency.

(Ans. $f_T = 7.69$ GHz)

For gallium arsenide JFETs or MESFETs with very small geometries, the cutoff frequency is even larger. The channel transit time may also become a factor in very high frequency devices, in which case the expression for cutoff frequency would need to be modified.

One application of GaAs FETs is in ultrafast digital integrated circuits. Conventional GaAs MESFET logic gates can achieve propagation delay times in the sub nanosecond range. These delay times are at least comparable to, if not shorter than, fast ECL, but the power dissipation is three orders of magnitude smaller than in the ECL circuits. Enhancement mode GaAs JFETs have been used as drivers in logic circuits, and depletion mode devices may be used as loads. Propagation delay times of as low as 45 ps have been observed. Special JFET structures may be used to further increase the speed. These structures include the modulation-doped field-effect transistor, which is discussed in the following section.

TEST YOUR UNDERSTANDING

TYU 13.3 Consider a p-channel silicon JFET that has parameters $a = 0.50\ \mu$m, $\mu_p = 400\ \text{cm}^2/\text{V-s}$, $N_a = 2 \times 10^{16}\ \text{cm}^{-3}$, and $L = 4\ \mu$m. Calculate the cutoff frequency.

(Ans. $f_T = 3.07\ \text{GHz}$)

TYU 13.4 An n-channel GaAs pn JFET has parameters $a = 0.50\ \mu$m, $L = 1\ \mu$m, $N_d = 3 \times 10^{15}\ \text{cm}^{-3}$, and $\mu_n = 6500\ \text{cm}^2/\text{V-s}$. Determine the cutoff frequency.

(Ans. $f_T = 107\ \text{GHz}$)

*13.5 | HIGH ELECTRON MOBILITY TRANSISTOR

As frequency needs, power capacity, and low noise performance requirements increase, the gallium arsenide MESFET is pushed to its limit of design and performance. These requirements imply a very small FET with a short channel length, large saturation current, and large transconductance. These requirements are generally achieved by increasing the channel doping under the gate. In all of the devices we have considered, the channel region is in a doped layer of bulk semiconductor with the majority carriers and doping impurities in the same region. The majority carriers experience ionized impurity scattering, which reduces carrier mobility and degrades device performance.

The degradation in mobility and peak velocity in GaAs due to increased doping can be minimized by separating the majority carriers from the ionized impurities. This separation can be achieved in a heterostructure that has an abrupt discontinuity in conduction and valence bands. We considered the basic heterojunction properties in Chapter 9. Figure 13.23 shows the conduction-band energy relative to the Fermi energy of an N-AlGaAs-intrinsic GaAs heterojunction in thermal equilibrium. Thermal equilibrium is achieved when electrons from the wide-bandgap AlGaAs flow into the GaAs and are confined to the potential well. However, the electrons are free to move parallel to the heterojunction interface. In this structure, the majority carrier

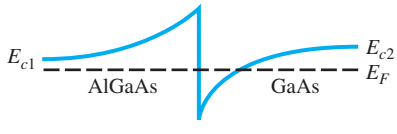


Figure 13.23 | Conduction-band edges for N-AlGaAs–intrinsic GaAs abrupt heterojunction.

electrons in the potential well are now separated from the impurity dopant atoms in the AlGaAs; thus, impurity scattering tends to be minimized.

The FETs fabricated from these heterojunctions are known by several names. The term used here is the **h**igh **e**lectron **m**obility **t**ransistor (HEMT). Other names include **mo**dulation-**d**oped **f**ield-**e**ffect **t**ransistor (MODFET), **s**electively **d**oped **h**etero junction field-effect **t**ransistor (SDHT), and **t**wo-dimensional **e**lectron **g**as **f**ield-**e**ffect **t**ransistor (TEGFET).

13.5.1 Quantum Well Structures

Figure 13.23 shows the conduction-band energy of an N-AlGaAs–intrinsic GaAs heterojunction. A two-dimensional surface channel layer of electrons is formed in the thin potential well (~80Å) in the undoped GaAs. Electron sheet carrier densities on the order of 10^{12} cm^{-2} have been obtained. An improvement in the low-field mobility of the carriers moving parallel to the heterojunction is observed since the impurity-scattering effects are reduced. At 300 K, mobilities have been reported in the range of 8500–9000 cm^2/V-s, whereas GaAs MESFETs doped to $N_d = 10^{17}$ cm^{-3} have low-field mobilities of less than 5000 cm^2/V-s. The electron mobility in the heterojunction now tends to be dominated by lattice or phonon scattering, so that as the temperature is reduced, the mobility increases rapidly.

Impurity-scattering effects can be further reduced by increasing the separation of the electrons and ionized donor impurities. The electrons in the potential well of the abrupt heterojunction shown in Figure 13.23 are separated from the donor atoms, but are still close enough to be subjected to a coulomb attraction. A thin spacer layer of undoped AlGaAs can be placed between the doped AlGaAs and the undoped GaAs. Figure 13.24 shows the energy-band diagram for this structure. Increasing the

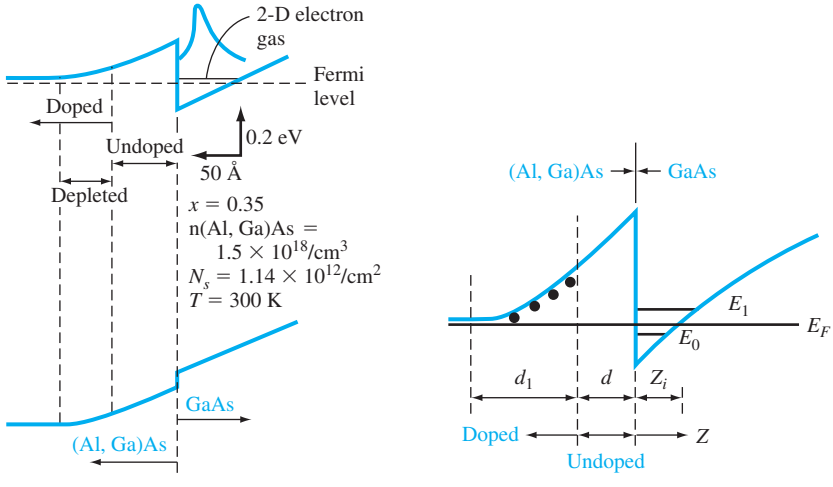


Figure 13.24 | Conduction-band edges for N-AlGaAs–undoped AlGaAs–undoped GaAs heterojunction.
(From Shur [13].)

separation between the carriers and ionized donors increases further the electron mobility, since there is even less coulomb interaction. One disadvantage of this graded heterojunction is that the electron density in the potential well tends to be smaller than in the abrupt junction.

The molecular beam epitaxial process allows the growth of very thin layers of specific semiconductor materials with specific dopings. In particular, a multilayer modulation–doped heterostructure can be formed, as shown in Figure 13.25. Several surface channel layers of electrons are formed in parallel. This structure would be equivalent to increasing the channel electron density, which would increase the current capability of the FET.

13.5.2 Transistor Performance

A typical HEMT structure is shown in Figure 13.26. The N-AlGaAs is separated from the undoped GaAs by an undoped AlGaAs spacer. A Schottky contact to the N-AlGaAs forms the gate of the transistor. This structure is a "normal" MODFET. An "inverted" structure is shown in Figure 13.27. In this case the Schottky contact is made to the undoped GaAs layer. The inverted MODFET has been investigated less than the normal structure because the normal structure has yielded superior results.

The density of electrons in the two-dimensional electron gas layer in the potential well can be controlled by the gate voltage. The electric field of the Schottky gate depletes the two-dimensional electron gas layer in the potential well when a

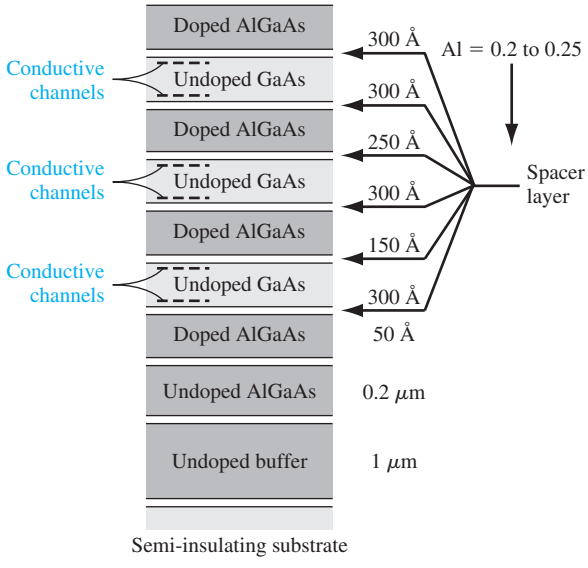


Figure 13.25 | Multilayer modulation–doped heterostructure.

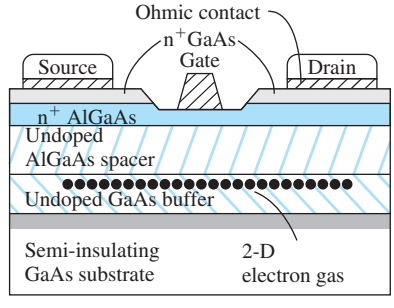


Figure 13.26 | A "normal" AlGaAs–GaAs HEMT.

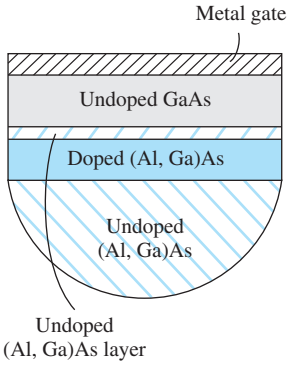


Figure 13.27 | An "inverted" GaAs–AlGaAs HEMT. (*From Shur* [*13*].)

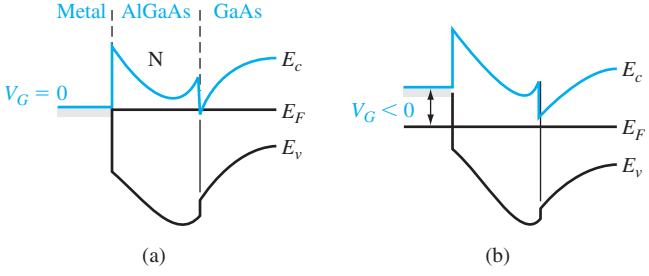


Figure 13.28 | Energy-band diagram of a normal HEMT (a) with zero gate bias and (b) with a negative gate bias.

sufficiently large negative voltage is applied to the gate. Figure 13.28 shows the energy-band diagrams of the metal–AlGaAs–GaAs structure under zero bias and with a reverse bias applied to the gate. With zero bias, the conduction-band edge in the GaAs is below the Fermi level, implying a large density of the two-dimensional electron gas. With a negative voltage applied to the gate, the conduction-band edge in the GaAs is above the Fermi level, implying that the density of the two-dimensional electron gas is very small and the current in an FET would be essentially zero.

The Schottky barrier depletes the AlGaAs layer from the surface, and the heterojunction depletes the AlGaAs layer from the heterojunction interface. Ideally the device should be designed so that the two depletion regions just overlap to prevent electron conduction through the AlGaAs layer. For depletion mode devices, the depletion layer from the Schottky gate should extend only to the heterojunction depletion layer. For enhancement mode devices, the thickness of the doped AlGaAs layer is smaller and the Schottky gate built-in potential barrier will completely

deplete the AlGaAs layer and the two-dimensional electron gas channel. A positive voltage applied to the gate of the enhancement mode device will turn on the device.

The density of the two-dimensional electron gas in a normal structure can be described using a charge control model. We may write

$$n_s = \frac{\epsilon_N}{q\,(d + \Delta d)}(V_g - V_{\text{off}}) \tag{13.68}$$

where ϵ_N is the permittivity of the N-AlGaAs, $d = d_d + d_i$ is the thickness of the doped-plus-undoped AlGaAs layer, and Δd is a correction factor given by

$$\Delta d = \frac{\epsilon_N a}{q} \approx 80 \text{ Å} \tag{13.69}$$

The threshold voltage V_{off} is given by

$$V_{\text{off}} = \phi_B - \frac{\Delta E_c}{q} - V_{p2} \tag{13.70}$$

where ϕ_B is the Schottky barrier height and V_{p2} is

$$V_{p2} = \frac{qN_d d_d^2}{2\epsilon_N} \tag{13.71}$$

A negative gate bias will reduce the two-dimensional electron gas concentration. If a positive gate voltage is applied, the density of the two-dimensional electron gas will increase. Increasing the gate voltage will increase the two-dimensional electron gas density until the conduction band of the AlGaAs crosses the Fermi level of the electron gas. Figure 13.29 shows this effect. At this point the gate loses control over the electron gas since a parallel conduction path in the AlGaAs has been formed.

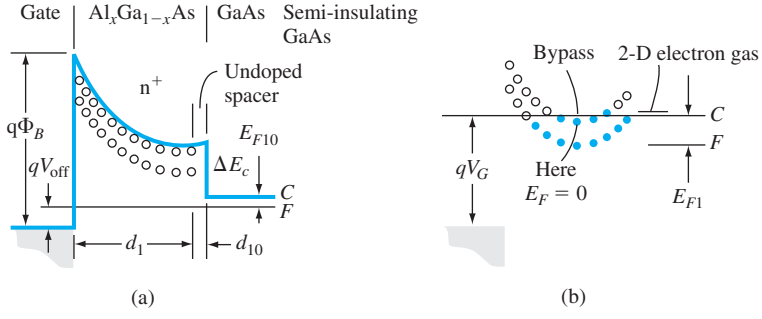


Figure 13.29 | Energy-band diagram of an enhancement mode HEMT (a) with a slight forward gate voltage, and (b) with a larger forward gate voltage that creates a conduction channel in the AlGaAs. *(From Fritzsche [5].)*

EXAMPLE 13.10

Objective: Determine the two-dimensional electron concentration for an N–AlGaAs–intrinsic GaAs heterojunction.

Consider an N-$Al_{0.3}Ga_{0.7}As$ layer doped to 10^{18} cm^{-3} and having a thickness of 500 Å. Assume an undoped spacer layer of 20 Å. Let $\phi_B = 0.85$ V and $\Delta E_c/q = 0.22$ V. The relative dielectric constant of $Al_{0.3}Ga_{0.7}As$ is $\epsilon_N = 12.2$.

■ Solution

The parameter V_{p2} is found as

$$V_{p2} = \frac{qN_d d_d^2}{2\epsilon_N} = \frac{(1.6 \times 10^{-19})(10^{18})(500 \times 10^{-8})^2}{2(12.2)(8.85 \times 10^{-14})} = 1.85 \text{ V}$$

Then the threshold voltage is

$$V_{\text{off}} = \phi_B - \frac{\Delta E_c}{q} - V_{p2} = 0.85 - 0.22 - 1.85 = -1.22 \text{ V}$$

The channel electron concentration for $V_g = 0$ is found from Equation (13.68) to be

$$n_s = \frac{(12.2)(8.85 \times 10^{-14})}{(1.6 \times 10^{-19})(500 + 20 + 80) \times 10^{-8}}[-(-1.22)] = 1.37 \times 10^{12} \text{ cm}^{-2}$$

■ Comment

The threshold voltage V_{off} is negative, making this device a depletion mode MODFET; applying a negative gate voltage will turn off the device. A value of $n_s \approx 10^{12}$ cm^{-2} is a typical channel concentration.

The current–voltage characteristics of the MODFET can be found using the charge control model and the gradual channel approximation. The channel carrier concentration can be written as

$$n_s(x) = \frac{\epsilon_N}{q(d + \Delta d)}[V_g - V_{\text{off}} - V(x)] \tag{13.72}$$

where $V(x)$ is the potential along the channel due to the drain-to-source voltage. The drain current is

$$I_D = qn_s\, v(\text{E})W \tag{13.73}$$

where v(E) is the carrier drift velocity and W is the channel width. This analysis is very similar to that for the pn JFET in Section 13.2.2.

If we assume a constant mobility, then for low V_{DS} values, we have

$$I_D = \frac{\epsilon_N \mu W}{2L(d + \Delta d)}\left[2(V_g - V_{\text{off}})\, V_{DS} - V_{DS}^2\right] \tag{13.74}$$

The form of this equation is the same as that for the pn JFET or MESFET operating in the nonsaturation region. If V_{DS} increases so that the carriers reach the saturation velocity, then

$$I_D(\text{sat}) = \frac{\epsilon_N W}{(d + \Delta d)}(V_g - V_{\text{off}} - V_0)v_{\text{sat}} \tag{13.75}$$

where v_{sat} is the saturation velocity and $V_0 = \text{E}_s L$ with E_s being the electric field in the channel that produces the saturation velocity.

Various velocity versus electric field models can be used to derive different *I–V* expressions. However, Equations (13.74) and (13.75) yield satisfactory results for most situations. Figure 13.30 shows a comparison between experimental and calculated *I–V* characteristics. As observed in the figure, the current in these heterojunction devices can be quite large. The transconductance of the MODFET is defined as it was for the pn JFET and MESFET. Typical measured values at $T = 300$ K are in the range of 250 mS/mm. Higher values have been reported. These transconductance values are significantly larger than for either the pn JFET or the MESFET.

HEMTs may also be fabricated with multiple heterojunction layers. This device type is shown in Figure 13.31. A single heterojunction for an AlGaAs–GaAs interface has a maximum two-dimensional electron sheet density on the order of 1×10^{12} cm^{-2}. This concentration can be increased by fabricating two or more AlGaAs–GaAs interfaces in the same epitaxial layer. The device current capacity is increased, and power performance is improved. The multichannel HEMT behaves as multiple single-channel HEMTs connected in parallel and modulated by the same gate but with slightly different threshold voltages. The maximum transconductance will not scale directly with the number of channels because of the change in threshold voltage with each channel. In addition, the effective channel length increases as the distance between the gate and channel increases.

HEMTs can be used in high-speed logic circuits. They have been used in flip-flop circuits operating at clock frequencies of 5.5 GHz at $T = 300$ K; the clock frequency can be increased at lower temperatures. Small-signal, high-frequency amplifiers have also been investigated. HEMTs showing low noise and reasonable gains have been operated at 35 GHz. The maximum frequency increases as the channel length decreases. Cutoff frequencies on the order of 100 GHz have been measured with channel lengths of 0.25 μm.

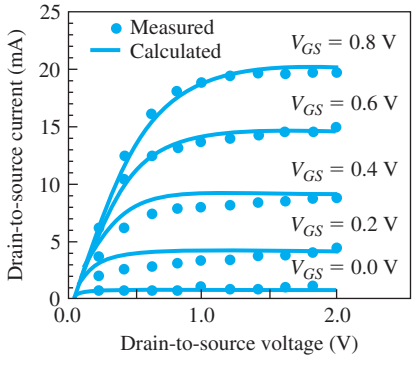


Figure 13.30 | Current–voltage characteristics of an enhancement mode HEMT, in which solid curves are numerical calculations and dots are measured points.
(From Shur [13].)

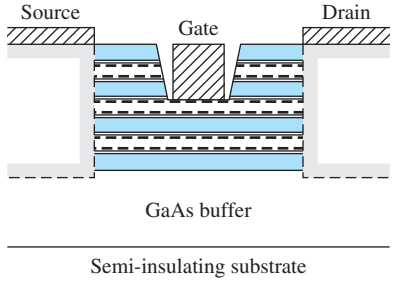


Figure 13.31 | A multilayer HEMT.

It seems clear that HEMTs are inherently superior to other FET technologies in terms of achieving higher speeds of operation, lower power dissipation, and lower noise. These advantages derive directly from the superior transport properties obtained by using undoped GaAs as the channel layer for the FET. One way to achieve an adequate carrier concentration in an undoped channel is to accumulate the carriers at a semiconductor heterojunction interface, as we have seen. The disadvantage of the HEMT is that the fabrication processes for the heterojunction are more complicated.

13.6 | SUMMARY

- The physics, characteristics, and operation of the junction field-effect transistor are considered in this chapter.
- The current in a JFET is controlled by an electric field applied perpendicular to the direction of current. The current is in the channel region between the source and drain contacts. In a pn JFET, the channel forms one side of a pn junction that is used to modulate the channel conductance.
- Two primary parameters of the JFET are the internal pinchoff voltage V_{p0} and the pinchoff voltage V_p. The internal pinchoff voltage is defined as a positive quantity and is the total gate-to-channel potential that causes the junction space charge layer to completely fill the channel region. The pinchoff voltage is defined as the gate voltage that must be applied to achieve the pinchoff condition.
- The ideal current–voltage relationship is derived. The transconductance, or transistor gain, is the rate of change of drain current with respect to the corresponding change in gate-to-source voltage.
- Three nonideal effects are considered; channel-length modulation, velocity saturation, and subthreshold current. Each of these effects changes the ideal current–voltage relationship.
- A small-signal equivalent circuit of the JFET is developed. The equivalent circuit includes capacitances that introduce frequency effects in the transistor. Two physical factors affect the frequency limitation; channel transit time and capacitance charging time. The capacitance charging time constant is normally the limiting factor in short channel devices.
- The high-electron mobility transistor (HEMT) structure utilizes a heterojunction. A two-dimensional electron gas is confined to a potential well at the heterojunction interface. However, the electrons are free to move parallel to the interface. These electrons are separated from the ionized donors so that ionized impurity scattering effects are minimized, resulting in a high mobility.

GLOSSARY OF IMPORTANT TERMS

capacitance charging time The time associated with charging or discharging the input gate capacitance with a change in the input gate signal.

channel conductance The ratio of a differential change in drain current to the corresponding differential change in drain-to-source voltage in the limit as the drain-to-source voltage approaches zero.

channel conductance modulation The process whereby the channel conductance changes with gate voltage; this is the basic field-effect transistor action.

channel length modulation The change in effective channel length with drain-to-source voltage with the JFET biased in the saturation region.

conduction parameter The multiplying factor k_n in the expression for drain current versus gate-to-source voltage for the enhancement mode MESFET.

cutoff frequency A figure of merit for the transistor defined to be the frequency at which the ratio of the small-signal input gate current to small-signal drain current is equal to unity.

depletion mode JFET A JFET in which a gate-to-source voltage must be applied to create pinchoff and turn the device off.

enhancement mode JFET A JFET in which pinchoff exists at zero gate voltage and a gate-to-source voltage must be applied to induce a channel, turning the device on.

internal pinchoff voltage The total potential drop across the gate junction at pinchoff.

output resistance The ratio of a differential change in drain-to-source voltage to the corresponding differential change in drain current at a constant gate-to-source voltage.

pinchoff The condition whereby the gate junction space charge region extends completely through the channel so that the channel is completely depleted of free carriers.

CHECKPOINT

After studying this chapter, the reader should have the ability to:

- Describe the basic operation of the pn JFET and MESFET.
- Discuss how current is contained in the channel region of a GaAs MESFET with a semi-insulating substrate.
- Sketch the *I–V* characteristics of a depletion mode JFET.
- Discuss how the internal pinchoff voltage is defined and how the pinchoff voltage is defined.
- Define transconductance for a JFET.
- Discuss the concept of an enhancement mode MESFET.
- Discuss three nonideal effects in a JFET including channel-length modulation, velocity saturation effects, and subthreshold effects.
- Sketch the small-signal equivalent circuit of a JFET.
- Discuss the frequency limitation factors and define the cutoff frequency.
- Sketch the cross section of a simple HEMT.
- Describe the advantages of a HEMT compared to a MESFET.

REVIEW QUESTIONS

1. Sketch the cross section of a p-channel pn JFET and indicate voltage polarities for device operation.
2. Sketch cross sections of a p-channel pn JFET showing the depletion regions when biased in the nonsaturation region and in the saturation region.
3. What is the mechanism of current saturation in a pn JFET?
4. Sketch the cross section of an n-channel GaAs MESFET.
5. What is the mechanism of current saturation in a MESFET?
6. Define internal pinchoff voltage and pinchoff voltage for a pn JFET.

7. Define threshold voltage for a MESFET.
8. Sketch the small-signal equivalent circuit of a JFET.
9. Define two frequency limitation factors for a JFET. Define the condition for cutoff frequency.
10. Sketch the cross section of an AlGaAs–GaAs HEMT. Sketch the conduction energy band across the heterojunction.
11. What is the principal advantage of a HEMT compared to a MESFET?

PROBLEMS

(*Note:* Assume $T = 300$ K for the following problems unless otherwise stated.)

Section 13.1 JFET Concepts

13.1 (*a*) Draw the structure of a p-channel JFET similar to the structure shown in Figure 13.2. (*b*) Qualitatively discuss the *I–V* characteristics, including current directions and voltage polarities, similar to those shown in Figures 13.3 and 13.4.

13.2 Consider the n-channel JFET in Figure P13.2. The p-type substrate is connected to the n-type source terminal. Sketch the space charge regions for various V_{GS} values when $V_{DS} = 0$ and for various V_{DS} values when $V_{GS} = 0$.

Section 13.2 The Device Characteristics

13.3 An n-channel GaAs pn JFET at $T = 300$ K has parameters $N_d = 3 \times 10^{16}$ cm^{-3}, $N_a = 2 \times 10^{18}$ cm^{-3}, and $a = 0.40$ μm. (*a*) Calculate the (*i*) internal pinchoff voltage V_{po} and (*ii*) pinchoff voltage V_p. (*b*) Determine the minimum undepleted channel thickness, $a - h$, for $V_{GS} = -0.5$ V and for (*i*) $V_{DS} = 0$, (*ii*) $V_{DS} = 0.5$ V, and (*iii*) $V_{DS} = 2.5$ V. (*c*) Find V_{DS}(sat) for (*i*) $V_{GS} = 0$ and (*ii*) $V_{GS} = -1.0$ V.

13.4 Repeat Problem 13.3 for an n-channel silicon pn JFET with the same geometrical and electrical parameters.

13.5 Consider a p-channel GaAs pn JFET at $T = 300$ K. The parameters are $N_d = 10^{18}$ cm^{-3} and $a = 0.65$ μm. (*a*) Determine the channel doping concentration such that the internal pinchoff voltage is $V_{po} = 2.75$ V. (*b*) Using the results of part (*a*), what is the pinchoff voltage V_p? (*c*) For $V_{SD} = 0$, determine the value of V_{GS} such that the minimum undepleted channel thickness is 0.15 μm. (*d*) For $V_{GS} = 0$, find the value of V_{SD} such that the channel is just pinched off at the drain terminal.

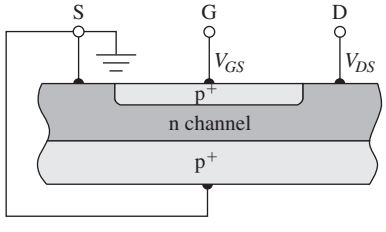


Figure P13.2 | Figure for Problem 13.2.

13.6 Repeat Problem 13.5 for a p-channel silicon pn JFET with the same geometrical and electrical parameters.

13.7 The parameters of a p-channel silicon pn JFET are $N_d = 3 \times 10^{18}$ cm^{-3} and $N_a = 2 \times 10^{16}$ cm^{-3}. (*a*) Determine the metallurgical channel thickness, *a*, such that the pinchoff voltage is $V_p = +3.0$ V. (*b*) Using the results of part (*a*), determine the internal pinchoff voltage V_{pO}. (*c*) Determine V_{SD} (sat) for (*i*) $V_{GS} = 0$ and (*ii*) $V_{GS} = 1.5$ V.

13.8 A p-channel GaAs pn JFET has the same parameters as given in Problem 13.7. Repeat the calculations for parts (*a*), (*b*), and (*c*).

13.9 The doping concentrations in a silicon n-channel pn JFET are $N_a = 4 \times 10^{18}$ cm^{-3} and $N_d = 4 \times 10^{16}$ cm^{-3}. (*a*) Design the channel metallurgical thickness, *a*, such that V_{DS} (sat) = 5.0 V for $V_{GS} = 0$. (*b*) Using the results of part (*a*), find the (*i*) internal pinchoff voltage V_{pO} and (*ii*) pinchoff voltage V_p.

13.10 Consider a p-channel GaAs pn JFET. The doping concentrations are $N_d = 10^{18}$ cm^{-3} and $N_a = 5 \times 10^{15}$ cm^{-3}. (*a*) Design the channel metallurgical thickness, a, such that V_{SD} (sat) = 3.5 V for $V_{GS} = +1.0$ V. (*b*) Using the results of part (*a*), determine the (*i*) internal pinchoff voltage V_{pO} and (*ii*) pinchoff voltage V_p.

13.11 An n-channel silicon JFET at $T = 300$ K has the following parameters:

$$N_a = 10^{19}\ \text{cm}^{-3} \qquad N_d = 10^{16}\ \text{cm}^{-3}$$
$$a = 0.50\ \mu\text{m} \qquad L = 20\ \mu\text{m}$$
$$W = 400\ \mu\text{m} \qquad \mu_n = 1000\ \text{cm}^2\text{/V-s}$$

Ignoring velocity saturation effects, calculate (*a*) I_{P1}; (*b*) V_{DS}(sat) for (*i*) $V_{GS} = 0$, (*ii*) $V_{GS} = V_p/4$, (*iii*) $V_{GS} = V_p/2$, and (*iv*) $V_{GS} = 3V_p/4$; and (*c*) I_{D1}(sat) for the same V_{GS} values in part (*b*). (*d*) Using the results from parts (*b*) and (*c*), plot the *I–V* characteristics.

13.12 Consider the JFET described in Problem 13.11. Compute and plot the channel conductance, g_d, as a function of V_{GS} for $0 < |V_{GS}| < |V_p|$.

13.13 Consider an n-channel GaAs JFET at $T = 300$ K with the following parameters:

$$N_a = 5 \times 10^{18}\ \text{cm}^{-3} \qquad N_d = 2 \times 10^{16}\ \text{cm}^{-3}$$
$$a = 0.35\ \mu\text{m} \qquad L = 10\ \mu\text{m}$$
$$W = 30\ \mu\text{m} \qquad \mu_n = 8000\ \text{cm}^2\text{/V-s}$$

Ignoring velocity saturation effects, calculate (*a*) G_{01}; (*b*) V_{DS}(sat) for $V_{GS} = 0$ and $V_{GS} = V_p/2$; and (*c*) I_{D1}(sat) for $V_{GS} = 0$ and $V_{GS} = V_p/2$. (*d*) Sketch the *I–V* characteristics using the results from parts (*b*) and (*c*).

13.14 Using the parameters from Problem 13.11, calculate the maximum transconductance in the saturation region. Normalize this transconductance to millisiemens per unit width, or mS/mm.

13.15 (*a*) Calculate the maximum transconductance for the transistor described in Problem 13.13 (*b*) Determine the maximum transconductance if the channel length is reduced to 2 μm.

13.16 The Schottky barrier height, ϕ_{Bn}, of a metal–n-GaAs MESFET is 0.90 V. The channel doping is $N_d = 1.5 \times 10^{16}$ cm^{-3}, and the channel thickness is $a = 0.5$ μm. $T = 300$ K. (*a*) Calculate the internal pinchoff voltage V_{p0} and the threshold voltage V_T. (*b*) Determine whether the MESFET is depletion type or enhancement type.

13.17 Consider an n-channel GaAs MESFET at $T = 300$ K with a gold Schottky barrier contact. Assume $\phi_{Bn} = 0.89$ V. The channel thickness is $a = 0.35$ μm. (*a*) Determine the uniform channel doping so that the threshold voltage is $V_T = 0.10$ V. (*b*) Using the results of part (*a*), determine the threshold voltage at $T = 400$ K.

13.18 The barrier height of the metal contact in an n-channel GaAs MESFET is $\phi_{Bn} = 0.87$ V. The channel doping concentration is $N_d = 2 \times 10^{16}$ cm^{-3}. (*a*) Determine the channel thickness, *a*, such that the internal pinchoff voltage is $V_{pO} = 1.5$ V. (*b*) Using the results of part (*a*), find the threshold voltage V_T. (*c*) Calculate the minimum undepleted channel width for $V_{GS} = +0.4$ V when (*i*) $V_{DS} = 0$, (*ii*) $V_{DS} = 1$ V, and (*iii*) $V_{DS} = 4$ V.

13.19 Two n-channel GaAs MESFETs have barrier heights of $\phi_{Bn} = 0.87$ V. (*a*) The channel doping concentration in device 1 is $N_d = 5 \times 10^{15}$ cm^{-3} and the channel metallurgical thickness is $a = 0.50$ μm. Determine the threshold voltage. (*b*) The channel doping concentration in device 2 is $N_d = 3 \times 10^{16}$ cm^{-3}. Find the metallurgical channel thickness, *a*, such that the threshold voltage is the same as that of device 1.

13.20 Consider an n-channel GaAs MESFET at $T = 300$ K with $\phi_{Bn} = 0.85$ V and $a = 0.25$ μm. Determine the channel doping concentration such that $V_T = 0.5$ V.

13.21 An n-channel silicon MESFET is fabricated using a gold contact. The n-channel doping is $N_d = 10^{16}$ cm^{-3} and the temperature is $T = 300$ K. When a gate voltage of $V_{GS} = 0.35$ V is applied with $V_{DS} = 0$, the undepleted channel thickness is 0.075 μm. (*a*) Determine the channel thickness dimension *a* and the threshold voltage V_T. (*b*) Determine the value of V_{DS}(sat) for $V_{GS} = 0.35$ V.

13.22 The barrier height of an n-channel GaAs MESFET is $\phi_{Bn} = 0.90$ V. The metallurgical channel thickness is $a = 0.65$ μm and the channel doping concentration is $N_d = 2 \times 10^{16}$ cm^{-3}. (*a*) Determine (*i*) V_{bi}, (*ii*) V_{pO}, and (*iii*) V_T. (*b*) Find V_{DS} (sat) for (*i*) $V_{GS} = -1.0$ V, (*ii*) $V_{GS} = -2.0$ V, and (*iii*) $V_{GS} = -3.0$ V.

13.23 The parameters of an n-channel GaAs MESFET are $V_T = 0.15$ V, $a = 0.25$ μm, $L = 1.5$ μm, $W = 12$ μm, and $\mu_n = 6500$ cm^2/V-s. (*a*) Determine the conduction parameter k_n. (*b*) Find I_{D1} (sat) for (*i*) $V_{GS} = 0.25$ V and (*ii*) $V_{GS} = 0.45$ V. (*c*) Find V_{DS} (sat) for (*i*) $V_{GS} = 0.25$ V and (*ii*) $V_{GS} = 0.45$ V.

13.24 An n-channel GaAs MESFET has the same parameters as described in Problem 13.23 except for the channel width. (*a*) The maximum transconductance is to be $g_{ms} = 1.25$ mA/V at $V_{GS} = 0.45$ V. Determine the required channel width *W*. (*b*) Using the results of part (*a*), find I_{D1} (sat) for (*i*) $V_{GS} = 0.25$ V and (*ii*) $V_{GS} = 0.45$ V.

13.25 Use Equation (13.27) to plot I_{D1} versus V_{DS} for a given value of V_{GS}. If V_{DS} is allowed to become larger than V_{DS}(sat), then I_{D1} decreases from a peak value that occurs at V_{DS}(sat). From these plots, determine V_{DS}(sat) at several values of V_{GS}. Compare these values with those determined from Equation (13.12).

13.26 This problem is to compare the JFET drain current as given by Equation (13.14) with that given by Equation (13.35). Choose device parameters such that the drain currents at $V_{GS} = 0$ are the same from the two equations.

Section 13.3 Nonideal Effects

13.27 A uniformly doped n-channel silicon pn JFET has the following parameters: $N_a = 10^{18}$ cm^{-3}, $N_d = 3 \times 10^{16}$ cm^{-3}, $a = 0.50$ μm, and $\mu_n = 850$ cm^2/V-s. The maximum drain-to-source voltage is $V_{DS} = 10$ V. (*a*) For $V_{GS} = 0$, the effective channel

length L' is to be no less than 90 percent of the original channel length. Determine the minimum value of L. (*b*) Repeat part (*a*) for $V_{GS} = -3$ V.

***13.28** If the change in the channel length, ΔL, is assumed small, derive an approximate expression in terms of channel parameters for λ given in Equation (13.53). (Note: The parameter λ may not be a constant. However, justify using Equation (13.53) by plotting the expression for λ over the range $1.5\, V_{DS}(\text{sat}) \leq V_{DS} \leq 3.0\, V_{DS}(\text{sat})$. Use typical parameter values.)

***13.29** As a first approximation, assume that the electric field in the channel of an n-channel silicon JFET is uniform through the channel. Also, assume that the drift velocity versus electric field for the electrons is given by the piecewise linear approximation given in Figure P13.29. Let:

$$N_a = 5 \times 10^{18}\ \text{cm}^{-3} \qquad N_d = 4 \times 10^{16}\ \text{cm}^{-3}$$
$$L = 2\ \mu\text{m} \qquad W = 30\ \mu\text{m}$$
$$a = 0.50\ \mu\text{m}$$

(*a*) Determine V_{DS} at which velocity saturation occurs. Let $V_{GS} = 0$. (*b*) For $V_{GS} = 0$, determine h_{sat}. (*c*) Calculate $I_{D1}(\text{sat})$ if velocity saturation occurs. (*d*) If the electron mobility is a constant and equal to $\mu_n = 1000\ \text{cm}^2/\text{V-s}$, calculate $I_{D1}(\text{sat})$ if velocity saturation did not occur.

***13.30** (*a*) Repeat Problem 13.29 if $L = 1\ \mu$m and all other parameters remain the same. (*b*) If velocity saturation occurs, does the relation $I_{D1}(\text{sat}) \propto L^{-1}$ still apply? Explain.

13.31 The channel length of an n-channel GaAs MESFET is $L = 2\ \mu$m. Assume that the average horizontal electric field in the channel is E = 5 kV/cm. Calculate the transit time of an electron through the channel assuming (*a*) a constant mobility of $\mu_n = 8000\ \text{cm}^2/\text{V-s}$ applies and (*b*) velocity saturation applies.

13.32 The channel length of an n-channel silicon MESFET is $L = 2\ \mu$m. Assume that the average horizontal electric field in the channel is E = 10 kV/cm. Calculate the transit time of an electron through the channel assuming (*a*) a constant mobility of $\mu_n = 1000\ \text{cm}^2/\text{V-s}$ applies and (*b*) velocity saturation applies.

13.33 Consider a one-sided silicon n-channel JFET at $T = 300$ K, pinched off as shown in Figure P13.33. The source-to-gate and drain-to-gate reverse-biased currents are split

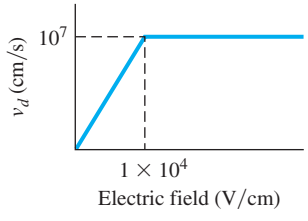


Figure P13.29 | Figure for Problem 13.29.

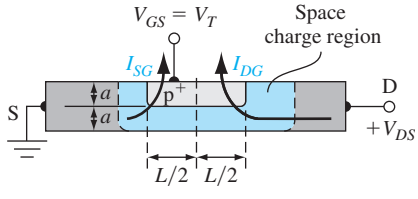


Figure P13.33 | Figure for Problem 13.33.

*Asterisks next to problems indicate problems that are more difficult.

geometrically as shown. Assume that the reverse-biased currents are dominated by the generation current. Assume the following parameters:

$$N_a = 5 \times 10^{18}\ \text{cm}^{-3} \qquad N_d = 3 \times 10^{16}\ \text{cm}^{-3}$$
$$\tau_0 = 5 \times 10^{-8}\ \text{s} \qquad a = 0.30\ \mu\text{m}$$
$$W = 30\ \mu\text{m} \qquad L = 2.4\ \mu\text{m}$$

Calculate I_{DG} for (*a*) $V_{DS} = 0$, (*b*) $V_{DS} = 1$ V, and (*c*) $V_{DS} = 5$ V. [Use Equation (8.42) and consider the volume of the depletion region.]

Section 13.4 Equivalent Circuit and Frequency Limitations

13.34 The source series resistance of a MESFET will reduce the value of transconductance, g_{ms}. Assume the doping in the source region of a GaAs MESFET is $N_d = 7 \times 10^{16}$ cm^{-3} and the dimensions are $a = 0.3\ \mu$m, $L = 1.5\ \mu$m, and $W = 5.0\ \mu$m. Let $\mu_n = 4500$ cm^2/V-s and $\phi_{Bn} = 0.89$ V. (*a*) Determine the ideal value of g_{ms} for $V_{GS} = 0$. (*b*) Determine the value of r_s for which the value of g'_{ms} is 80 percent of the ideal value. (*c*) Determine the maximum effective distance from the edge of the channel to the source terminal so that r_s is no larger than the value determined in part (*b*).

13.35 Estimate the cutoff frequency of the MESFET in Problem 13.34.

13.36 An n-channel GaAs MESFET at $T = 300$ K has the following parameters: $\phi_{Bn} = 0.90$ V, $N_d = 4 \times 10^{16}$ cm^{-3}, $\mu_n = 7500$ cm^2/V-s, $a = 0.30\ \mu$m, $W = 5\ \mu$m, and $L = 1.2\ \mu$m. Calculate the cutoff frequency using (*a*) the constant mobility model and (*b*) the saturation velocity model.

13.37 Consider a silicon n-channel pn JFET. The parameters are $a = 0.40\ \mu$m, $\mu_n = 1000$ cm^2/V-s, and $N_d = 2 \times 10^{16}$ cm^{-3}. Determine the cutoff frequency for (*a*) $L = 3\ \mu$m and (*b*) $L = 1.5\ \mu$m.

13.38 A silicon p-channel pn JFET has parameters $\mu_p = 420$ cm^2/V-s, $a = 0.40\ \mu$m, and $N_a = 2 \times 10^{16}$ cm^{-3}. Determine the maximum channel length such that the cutoff frequency is (*a*) $f_T = 5$ GHz and (*b*) $f_T = 12$ GHz.

Section 13.5 High Electron Mobility Transistor

13.39 Consider an N-$Al_{0.3}Ga_{0.7}As$–intrinsic GaAs abrupt heterojunction. Assume that the AlGaAs is doped to $N_d = 3 \times 10^{18}$ cm^{-3} and has a thickness of 350 Å. Let $\phi_{Bn} = 0.89$ V, and assume that $\Delta E_c = 0.24$ eV. (*a*) Calculate V_{off} and (*b*) calculate n_s for $V_g = 0$.

13.40 If the electrons in the channel of the JFET in Problem 13.39 are traveling at a saturation velocity of 2×10^7 cm/s, determine (*a*) the transconductance per unit width at $V_g = 0$ and (*b*) the saturation current per unit width at $V_g = 0$. (Assume $V_0 = 1$ V.)

13.41 Consider an abrupt N-$Al_{0.3}Ga_{0.7}As$–intrinsic GaAs heterojunction. The N-AlGaAs is doped to $N_d = 2 \times 10^{18}$ cm^{-3}. The Schottky barrier height is 0.85 V and the heterojunction conduction-band edge discontinuity is $\Delta E_c = 0.22$ eV. Determine the thickness of the AlGaAs layer so that $V_{\text{off}} = -0.3$ V.

Summary and Review

***13.42** Design a one-sided silicon p-channel pn JFET such that $V_p = 3.2$ V, $I_{D1}(\text{sat}) = 1.2$ mA at $V_{GS} = 0$, and $f_T = 10$ GHz. Determine the required values of L, W, and N_a.

***13.43** Design a one-sided GaAs n-channel MESFET with a barrier height of $\phi_{Bn} = 0.89$ V such that $V_T = +0.12$ V, $I_{DSS} = 2.0\ \mu$A at $V_{GS} = 0.45$ V, and $f_T = 50$ GHz. Assume $\mu_n = 7500\ \text{cm}^2/\text{V-s}$.

***13.44** Design a pair of complementary n-channel and p-channel silicon JFETs so that $I_{DSS} = 1$ mA and $|V_p| = 3.2$ V for each device at $T = 300$ K. If the devices are to operate for $0 \leq V_{DS} \leq 5$ V, comment on velocity saturation and channel length modulation effects in your design.

READING LIST

1. Chang, C. S., and D. Y. S. Day. "Analytic Theory for Current-Voltage Characteristics and Field Distribution of GaAs MESFETs." *IEEE Transactions on Electron Devices* 36, no. 2 (February 1989), pp. 269–280.
2. Daring, R. B. "Subthreshold Conduction in Uniformly Doped Epitaxial GaAs MESFETs." *IEEE Transactions on Electron Devices* 36, no. 7 (July 1989), pp. 1264–1273.
3. Dimitrijev, S. *Principles of Semiconductor Devices*. New York: Oxford University Press, 2006.
4. Drummond, T. J., W. T. Masselink, and H. Morkoc. "Modulation-Doped GaAs/(Al,Ga)As Heterojunction Field-Effect Transistors: MODFETs." *Proceedings of the IEEE* 74, no. 6 (June 1986), pp. 773–812.
5. Fritzsche, D. "Heterostructures in MODFETs." *Solid-State Electronics* 30, no. 11 (November 1987), pp. 1183–1195.
6. Ghandhi, S. K. *VLSI Fabrication Principles: Silicon and Gallium Arsenide*. New York: John Wiley & Sons, 1983.
7. Kano, K. *Semiconductor Devices*. Upper Saddle River, NJ: Prentice Hall, 1998.
8. Liao, S. Y. *Microwave Solid-State Devices*. Englewood Cliffs, NJ: Prentice Hall, 1985.
9. Ng, K. K. *Complete Guide to Semiconductor Devices*. New York: McGraw-Hill, 1995.
10. Pierret, R. F. *Field Effect Devices*. Vol. 4 of the *Modular Series on Solid State Devices*. 2nd ed. Reading, MA: Addison-Wesley, 1990.
11. _____. *Semiconductor Device Fundamentals*. Reading, MA: Addison-Wesley, 1996.
12. Roulston, D. J. *An Introduction to the Physics of Semiconductor Devices*. New York: Oxford University Press, 1999.

*13. Shur, M. *GaAs Devices and Circuits*. New York: Plenum Press, 1987.

14. _____. *Introduction to Electronic Devices*. New York: John Wiley & Sons, 1996.
15. Singh, J. *Semiconductor Devices: An Introduction*. New York: McGraw-Hill, 1994.
16. _____. *Semiconductor Devices: Basic Principles*. New York: John Wiley & Sons, 2001.

17. Streetman, B. G., and S. K. Banerjee. *Solid State Electronic Devices,* 6th ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2006.

18. Sze, S. M. *High-Speed Semiconductor Devices*. New York: John Wiley & Sons, 1990.

19. _____. *Semiconductor Devices: Physics and Technology*. New York: John Wiley & Sons, 1985.

20. Sze, S. M. and K. K. Ng. Physics of Semiconductor Devices, 3rd ed. Hoboken, NJ: John Wiley & Sons, 2007.

21. Turner, J. A., R. S. Butlin, D. Parker, R. Bennet, A. Peake, and A. Hughes. "The Noise and Gain Performance of Submicron Gate Length GaAs FETs." *GaAs FET Principles and Technology*. Edited by J. V. Di-Lorenzo and D. D. Khandelwal. Dedham, MA: Artech House, 1982.

22. Yang, E. S. *Microelectronic Devices*. New York: McGraw-Hill, 1988.

*Indicates reference that is at an advanced level compared to this text.

CHAPTER 14

Optical Devices

In previous chapters, we have considered the basic physics of transistors that are used to amplify or switch electrical signals. Semiconductor devices can be designed to convert optical energy into electrical energy, and to convert electrical signals into optical signals. These devices are used in broadband communications and data transmission over optical fibers. The general classification of these devices is called *optoelectronics*.

In this chapter, we discuss the basic principles of solar cells, several photodetectors, light emitting diodes, and laser diodes. Solar cells and photodetectors convert optical energy into electrical energy; light emitting diodes and laser diodes convert electrical signals into optical signals. ■

14.0 | PREVIEW

In this chapter, we will:

- Discuss and analyze photon absorption in a semiconductor and present absorption coefficient data for several semiconductor materials.
- Consider the basic principles of solar cells, analyze their *I–V* characteristics, and discuss the conversion efficiency.
- Present various types of solar cells, including homojunction, heterojunction, and amorphous silicon solar cells.
- Discuss the basic principles of photodetectors, including photoconductors, photodiodes, and phototransistors.
- Derive the output current characteristics of the various photodectors.
- Present and analyze the basic operation of the **L**ight **E**mitting **D**iode (LED).
- Discuss the basic principles and operation of the laser diode.

14.1 | OPTICAL ABSORPTION

In Chapter 2, we discussed the wave–particle duality principle and indicated that light waves could be treated as particles, which are referred to as photons. The energy of a photon is $E = h\nu$ where h is Plank's constant and ν is the frequency. We can also relate the wavelength and energy by

$$\lambda = \frac{c}{\nu} = \frac{hc}{E} = \frac{1.24}{E}\,\mu\text{m} \tag{14.1}$$

where E is the photon energy in eV and c is the speed of light.

There are several possible photon–semiconductor interaction mechanisms. For example, photons can interact with the semiconductor lattice whereby the photon energy is converted into heat. Photons can also interact with impurity atoms, either donors or acceptors, or they can interact with defects within the semiconductor. However, the basic photon interaction process of greatest interest is the interaction with valence electrons. When a photon collides with a valence electron, enough energy may be imparted to elevate the electron into the conduction band. Such a process generates electron–hole pairs and creates excess carrier concentrations. The behavior of excess carriers in a semiconductor was considered in Chapter 6.

14.1.1 Photon Absorption Coefficient

When a semiconductor is illuminated with light, the photons may be absorbed or they may propagate through the semiconductor, depending on the photon energy and on the bandgap energy E_g. If the photon energy is less than E_g, the photons are not readily absorbed. In this case, the light is transmitted through the material and the semiconductor appears to be transparent.

If $E = h\nu > E_g$, the photon can interact with a valence electron and elevate the electron into the conduction band. The valence band contains many electrons and the conduction band contains many empty states, so the probability of this interaction is high when $h\nu > E_g$. This interaction creates an electron in the conduction band and a hole in the valence band—an electron–hole pair. The basic absorption processes for different values of $h\nu$ are shown in Figure 14.1. When $h\nu > E_g$, an electron–hole

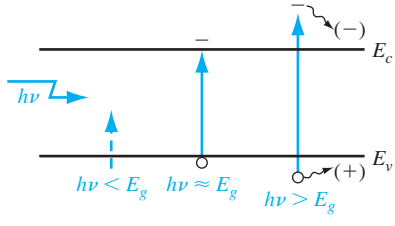


Figure 14.1 | Optically generated electron–hole pair formation in a semiconductor.

pair is created and the excess energy may give the electron or hole additional kinetic energy, which will be dissipated as heat in the semiconductor.

The intensity of the photon flux is denoted by $I_\nu(x)$ and is expressed in terms of energy/cm^2-s. Figure 14.2 shows an incident photon intensity at a position x and the photon flux emerging at a distance $x + dx$. The energy absorbed per unit time in the distance dx is given by

$$\alpha I_\nu(x)\,dx \tag{14.2}$$

where α is the absorption coefficient. The absorption coefficient is the relative number of photons absorbed per unit distance, given in units of cm^{-1}.

From Figure 14.2, we can write

$$I_\nu(x + dx) - I_\nu(x) = \frac{dI_\nu(x)}{dx} \cdot dx = -\alpha I_\nu(x)\,dx \tag{14.3}$$

or

$$\frac{dI_\nu(x)}{dx} = -\alpha I_\nu(x) \tag{14.4}$$

If the initial condition is given as $I_\nu(0) = I_{\nu 0}$, then the solution to the differential equation, Equation (14.4), is

$$I_\nu(x) = I_{\nu 0} e^{-\alpha x} \tag{14.5}$$

The intensity of the photon flux decreases exponentially with distance through the semiconductor material. The photon intensity as a function of x for two general values of absorption coefficient is shown in Figure 14.3. If the absorption coefficient is large, the photons are absorbed over a relatively short distance.

The absorption coefficient in the semiconductor is a very strong function of photon energy and bandgap energy. Figure 14.4 shows the absorption coefficient α plotted as a function of wavelength for several semiconductor materials. The absorption coefficient increases very rapidly for $h\nu > E_g$, or for $\lambda < 1.24/E_g$. The absorption

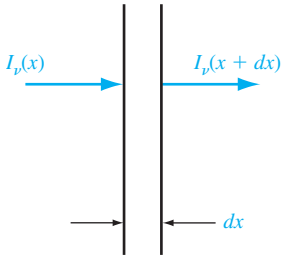


Figure 14.2 | Optical absorption in a differential length.

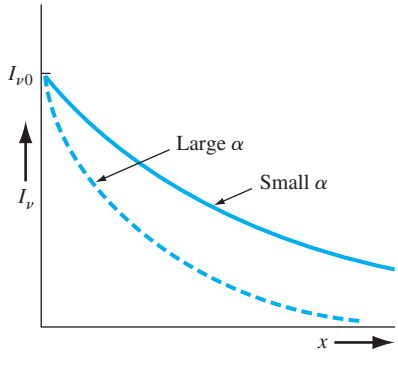


Figure 14.3 | Photon intensity versus distance for two absorption coefficients.

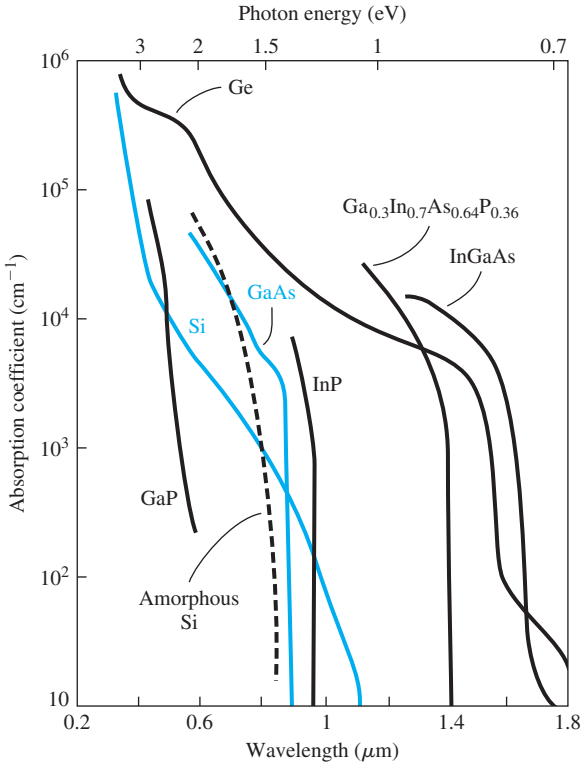


Figure 14.4 | Absorption coefficient as a function of wavelength for several semiconductors. *(From Shur [13].)*

coefficients are very small for $h\nu < E_g$, so the semiconductor appears transparent to photons in this energy range.

EXAMPLE 14.1

Objective: Calculate the thickness of a semiconductor that will absorb 90 percent of the incident photon energy.

Consider silicon and assume that in the first case the incident wavelength is $\lambda = 1.0\ \mu$m and in the second case, the incident wavelength is $\lambda = 0.5\ \mu$m.

■ Solution

From Figure 14.4, the absorption coefficient is $\alpha \approx 10^2\ \text{cm}^{-1}$ for $\lambda = 1.0\ \mu$m. If 90 percent of the incident flux is to be absorbed in a distance d, then the flux emerging at $x = d$ will be 10 percent of the incident flux. We can write

$$\frac{I_\nu(d)}{I_{\nu 0}} = 0.1 = e^{-\alpha d}$$

Solving for the distance d, we have

$$d = \frac{1}{\alpha}\ln\left(\frac{1}{0.1}\right) = \frac{1}{10^2}\ln(10) = 0.0230\ \text{cm}$$

In the second case, the absorption coefficient is $\alpha \approx 10^4$ cm^{-1} for $\lambda = 0.5$ μm. The distance d, then, in which 90 percent of the incident flux is absorbed, is

$$d = \frac{1}{10^4} \ln\left(\frac{1}{0.1}\right) = 2.30 \times 10^{-4} \text{ cm} = 2.30 \ \mu\text{m}$$

Comment

As the incident photon energy increases, the absorption coefficient increases rapidly, so that the photon energy can be totally absorbed in a very narrow region at the surface of the semiconductor.

EXERCISE PROBLEM

Ex 14.1 Consider a slab of silicon 5 μm thick. Determine the percentage of photon energy that will pass through the slab if the photon wavelength is (*a*) $\lambda = 0.8$ μm and (*b*) $\lambda = 0.6$ μm.

[Ans. (*a*) 60.7%; (*b*) 10.5%]

The relation between the bandgap energies of some of the common semiconductor materials and the light spectrum is shown in Figure 14.5. We may note that silicon and gallium arsenide will absorb all of the visible spectrum, whereas gallium phosphide, for example, will be transparent to the red spectrum.

14.1.2 Electron–Hole Pair Generation Rate

We have shown that photons with energy greater than E_g can be absorbed in a semiconductor, thereby creating electron–hole pairs. The intensity $I_\nu(x)$ is in units of

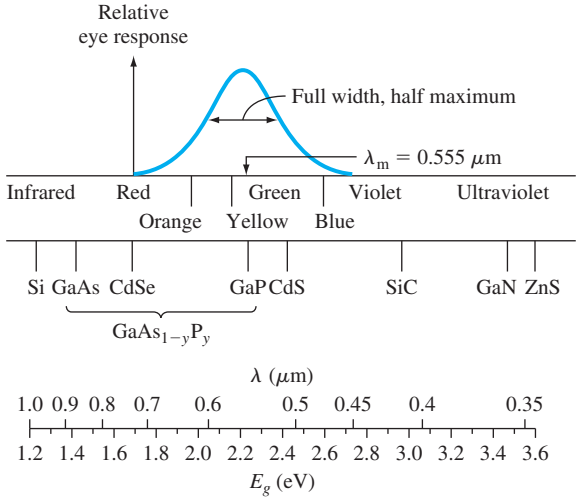


Figure 14.5 | Light spectrum versus wavelength and energy. Figure includes relative response of the human eye. *(From Sze [18].)*

energy/cm^2-s and $\alpha I_\nu(x)$ is the rate at which energy is absorbed per unit volume. If we assume that one absorbed photon at an energy $h\nu$ creates one electron–hole pair, then the generation rate of electron–hole pairs is

$$g' = \frac{\alpha I_\nu(x)}{h\nu} \tag{14.6}$$

which is in units of #/cm^3-s. We may note that the ratio $I_\nu(x)/h\nu$ is the photon flux. If, on the average, one absorbed photon produces less than one electron–hole pair, then Equation (14.6) must be multiplied by an efficiency factor.

EXAMPLE 14.2

Objective: Calculate the generation rate of electron–hole pairs given an incident intensity of photons.

Consider gallium arsenide at $T = 300$ K. Assume the photon intensity at a particular point is $I_\nu(x) = 0.05$ W/cm^2 at a wavelength of $\lambda = 0.75$ μm. This intensity is typical of sunlight, for example.

■ Solution

The absorption coefficient for gallium arsenide at this wavelength is $\alpha \approx 0.9 \times 10^4$ cm^{-1}. The photon energy, using Equation (14.1), is

$$E = h\nu = \frac{1.24}{0.75} = 1.65 \text{ eV}$$

Then, from Equation (14.6) and including the conversion factor between joules and eV, we have, for a unity efficiency factor,

$$g' = \frac{\alpha I_\nu(x)}{h\nu} = \frac{(0.9 \times 10^4)(0.05)}{(1.6 \times 10^{-19})(1.65)} = 1.70 \times 10^{21} \text{ cm}^{-3}\text{-s}^{-1}$$

If the incident photon intensity is a steady-state intensity, then, from Chapter 6, the steady-state excess carrier concentration is $\delta n = g'\tau$, where τ is the excess minority carrier lifetime. If $\tau = 10^{-7}$ s, for example, then

$$\delta n = (1.70 \times 10^{21})(10^{-7}) = 1.70 \times 10^{14} \text{ cm}^{-3}$$

■ Comment

This example gives an indication of the magnitude of the electron–hole generation rate and the magnitude of the excess carrier concentration. Obviously, as the photon intensity decreases with distance in the semiconductor, the generation rate also decreases.

■ EXERCISE PROBLEM

Ex 14.2 A photon flux with an intensity of $I_{\nu 0} = 0.10$W/cm^2 and at a wavelength of $\lambda = 1$ μm is incident on the surface of silicon. Neglecting any reflection from the surface, determine the generation rate of electron–hole pairs at a depth of (*a*) $x = 5$ μm and (*b*) $x = 20$ μm from the surface.

[Ans. (*a*) 4.79×10^{19} cm^{-3}-s^{-1}; (*b*) 1.13×10^{19} cm^{-3}-s^{-1}]

TEST YOUR UNDERSTANDING

TYU 14.1 (*a*) A photon flux with an intensity of $I_{\nu 0} = 0.10$ W/cm^2 is incident on the surface of silicon. The wavelength of the incident photon signal is $\lambda = 1$ μm. Neglecting any reflection from the surface, determine the photon flux intensity at a depth of (*i*) $x = 5$ μm and (*ii*) $x = 20$ μm from the surface. (*b*) Repeat part (*a*) for a wavelength of $\lambda = 0.60$ μm. [Ans. (*a*) (*i*) 0.0951 W/cm^2, (*ii*) 0.0819 W/cm^2; (*b*) (*i*) 0.0135 W/cm^2, (*ii*) 3.35×10^{-5} W/cm^2]

14.2 | SOLAR CELLS

A solar cell is a pn junction device with no voltage directly applied across the junction. The solar cell converts photon power into electrical power and delivers this power to a load. These devices have long been used for the power supply of satellites and space vehicles, and also as the power supply to some calculators. We will first consider the simple pn junction solar cell with uniform generation of excess carriers. We will also discuss briefly the heterojunction and amorphous silicon solar cells.

14.2.1 The pn Junction Solar Cell

Consider the pn junction shown in Figure 14.6 with a resistive load. Even with zero bias applied to the junction, an electric field exists in the space charge region as shown in the figure. Incident photon illumination can create electron–hole pairs in the space charge region that will be swept out producing the photocurrent I_L in the reverse-biased direction as shown.

The photocurrent I_L produces a voltage drop across the resistive load which forward biases the pn junction. The forward-bias voltage produces a forward-bias current I_F as indicated in the figure. The net pn junction current, in the reverse-biased direction, is

$$I = I_L - I_F = I_L - I_S\left[\exp\left(\frac{eV}{kT}\right) - 1\right] \tag{14.7}$$

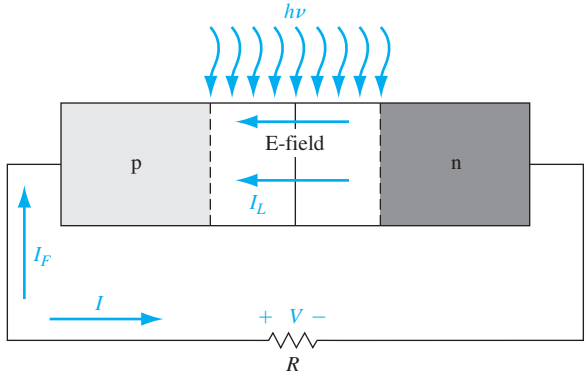


Figure 14.6 | A pn junction solar cell with resistive load.

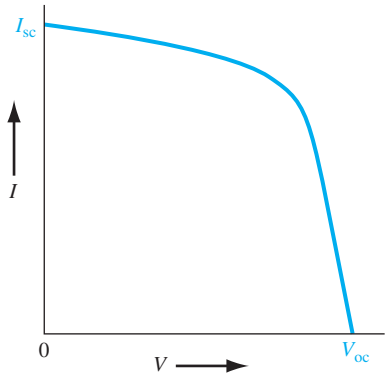


Figure 14.7 | *I–V* characteristics of a pn junction solar cell.

where the ideal diode equation has been used. As the diode becomes forward biased, the magnitude of the electric field in the space charge region decreases, but does not go to zero or change direction. The photocurrent is always in the reverse-biased direction and the net solar cell current is also always in the reverse-biased direction.

There are two limiting cases of interest. The short-circuit condition occurs when $R = 0$ so that $V = 0$. The current in this case is referred to as the *short-circuit current,* or

$$I = I_{sc} = I_L \tag{14.8}$$

The second limiting case is the open-circuit condition and occurs when $R \rightarrow \infty$. The net current is zero and the voltage produced is the *open-circuit voltage.* The photocurrent is just balanced by the forward-biased junction current, so we have

$$I = 0 = I_L - I_S\left[\exp\left(\frac{eV_{oc}}{kT}\right) - 1\right] \tag{14.9}$$

We can find the open circuit voltage V_{oc} as

$$V_{oc} = V_t \ln\left(1 + \frac{I_L}{I_S}\right) \tag{14.10}$$

A plot of the diode current I as a function of the diode voltage V from Equation (14.7) is shown in Figure 14.7. We may note the short-circuit current and open-circuit voltage points on the figure.

EXAMPLE 14.3

Objective: Calculate the open-circuit voltage of a silicon pn junction solar cell.

Consider a silicon pn junction at $T = 300$ K with the following parameters:

$$N_a = 5 \times 10^{18}\ \text{cm}^{-3} \qquad N_d = 10^{16}\ \text{cm}^{-3}$$
$$D_n = 25\ \text{cm}^2/\text{s} \qquad D_p = 10\ \text{cm}^2/\text{s}$$
$$\tau_{n0} = 5 \times 10^{-7}\ \text{s} \qquad \tau_{p0} = 10^{-7}\ \text{s}$$

Let the photocurrent density be $J_L = I_L/A = 15\ \text{mA/cm}^2$.

Solution

We have that

$$J_S = \frac{I_S}{A} = \left(\frac{eD_n n_{p0}}{L_n} + \frac{eD_p p_{n0}}{L_p}\right) = en_i^2\left(\frac{D_n}{L_n N_a} + \frac{D_p}{L_p N_d}\right)$$

We may calculate

$$L_n = \sqrt{D_n \tau_{n0}} = \sqrt{(25)(5 \times 10^{-7})} = 35.4\ \mu\text{m}$$

and

$$L_p = \sqrt{D_p \tau_{p0}} = \sqrt{(10)(10^{-7})} = 10.0\ \mu\text{m}$$

Then

$$J_S = (1.6 \times 10^{-19})(1.5 \times 10^{10})^2 \times \left[\frac{25}{(35.4 \times 10^{-4})(5 \times 10^{18})} + \frac{10}{(10 \times 10^{-4})(10^{16})}\right]$$

$$= 3.6 \times 10^{-11}\ \text{A/cm}^2$$

Then from Equation (14.10), we can find

$$V_{oc} = V_t \ln\left(1 + \frac{I_L}{I_S}\right) = V_t \ln\left(1 + \frac{J_L}{J_S}\right) = (0.0259)\ln\left(1 + \frac{15 \times 10^{-3}}{3.6 \times 10^{-11}}\right) = 0.514\ \text{V}$$

Comment

We may determine the built-in potential barrier of this junction to be $V_{bi} = 0.8556$ V. Taking the ratio of the open-circuit voltage to the built-in potential barrier, we find that $V_{oc}/V_{bi} = 0.60$. The open-circuit voltage will always be less than the built-in potential barrier.

EXERCISE PROBLEM

Ex 14.3 Consider a GaAs pn junction solar cell with the following parameters: $N_a = 10^{17}\ \text{cm}^{-3}$, $N_d = 2 \times 10^{16}\ \text{cm}^{-3}$, $D_n = 190\ \text{cm}^2/\text{s}$, $D_p = 10\ \text{cm}^2/\text{s}$, $\tau_{n0} = 10^{-7}$ s, and $\tau_{p0} = 10^{-8}$ s. Assume a photocurrent density of $J_L = 20\ \text{mA/cm}^2$ is generated in the solar cell. (*a*) Calculate the open-circuit voltage and (*b*) determine the ratio of open-circuit voltage to built-in potential barrier.

[Ans. (*a*) $V_{oc} = 0.971$ V; (*b*) $V_{oc}/V_{bi} = 0.783$]

The power delivered to the load is

$$P = I \cdot V = I_L \cdot V - I_S\left[\exp\left(\frac{eV}{kT}\right) - 1\right] \cdot V \tag{14.11}$$

We may find the current and voltage which will deliver the maximum power to the load by setting the derivative equal to zero, or $dP/dV = 0$. Using Equation (14.11), we find

$$\frac{dP}{dV} = 0 = I_L - I_S\left[\exp\left(\frac{eV_m}{kT}\right) - 1\right] - I_S V_m\left(\frac{e}{kT}\right)\exp\left(\frac{eV_m}{kT}\right) \tag{14.12}$$

where V_m is the voltage that produces the maximum power. We may rewrite Equation (14.12) in the form

$$\left(1 + \frac{V_m}{V_t}\right)\exp\left(\frac{eV_m}{kT}\right) = 1 + \frac{I_L}{I_S} \tag{14.13}$$

The value of V_m may be determined by trial and error. Figure 14.8 shows the maximum power rectangle where I_m is the current when $V = V_m$.

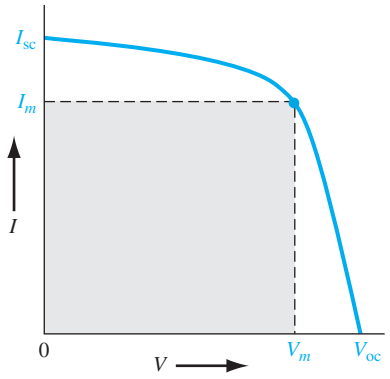


Figure 14.8 | Maximum power rectangle of the solar cell *I–V* characteristics.

14.2.2 Conversion Efficiency and Solar Concentration

The conversion efficiency of a solar cell is defined as the ratio of output electrical power to incident optical power. For the maximum power output, we can write

$$\eta = \frac{P_m}{P_{in}} \times 100\% = \frac{I_m V_m}{P_{in}} \times 100\% \tag{14.14}$$

The maximum possible current and the maximum possible voltage in the solar cell are I_{sc} and V_{oc}, respectively. The ratio $I_m V_m / I_{sc} V_{oc}$ is called the fill factor and is a measure of the realizable power from a solar cell. Typically, the fill factor is between 0.7 and 0.8.

The conventional pn junction solar cell has a single semiconductor bandgap energy. When the cell is exposed to the solar spectrum, a photon with energy less than E_g will have no effect on the electrical output power of the solar cell. A photon with energy greater than E_g will contribute to the solar cell output power, but the fraction of photon energy that is greater than E_g will eventually only be dissipated as heat. Figure 14.9 shows the solar spectral irradiance (power per unit area per unit wavelength) where air mass zero represents the solar spectrum outside the earth's atmosphere and air mass one is the solar spectrum at the earth's surface at noon. The maximum efficiency of a silicon pn junction solar cell is approximately 28 percent. Nonideal factors, such as series resistance and reflection from the semiconductor surface, will lower the conversion efficiency typically to the range of 10 to 15 percent.

A large optical lens can be used to concentrate sunlight onto a solar cell so that the light intensity can be increased up to several hundred times. The short-circuit current increases linearly with light concentration while the open-circuit voltage increases only slightly with concentration. Figure 14.10 shows the ideal solar cell efficiency at 300 K for two values of solar concentration. We can see that the conversion efficiency increases only slightly with optical concentration. The primary advantage of using concentration techniques is to reduce the overall system cost since an optical lens is less expensive than an equivalent area of solar cells.

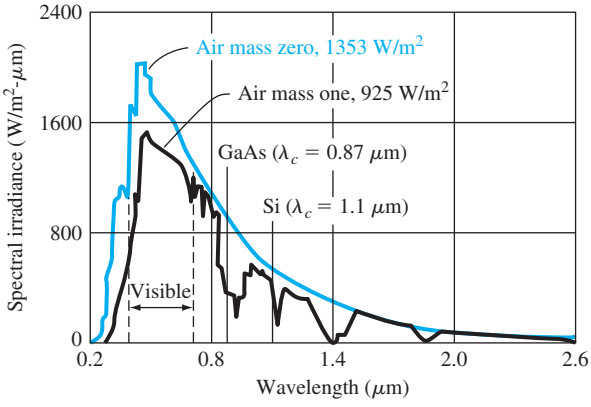


Figure 14.9 | Solar spectral irradiance.
(From Sze [18].)

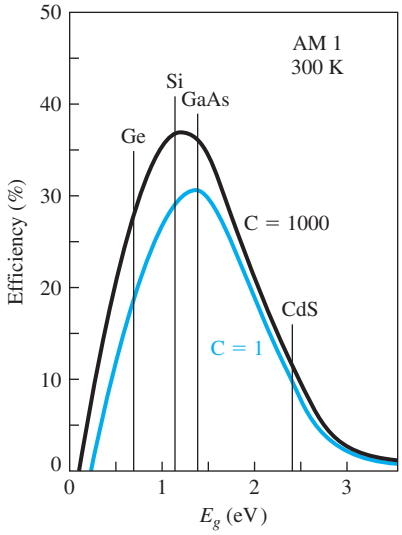


Figure 14.10 | Ideal solar cell efficiency at $T = 300$ K for C = 1 sun and for a C = 1000 sun concentrations as a function of bandgap energy.
(From Sze [18].)

14.2.3 Nonuniform Absorption Effects

We have seen from the previous section that the photon absorption coefficient in a semiconductor is a very strong function of the incident photon energy or wavelength. Figure 14.4 shows the absorption coefficient as a function of wavelength for several semiconductor materials. As the absorption coefficient increases, more photon energy will be absorbed near the surface than deeper into the semiconductor. In this case, then, we will not have uniform excess carrier generation in a solar cell.

The number of photons absorbed per cm^3 per second as a function of distance x from the surface can be written as

$$\alpha \Phi_0 e^{-\alpha x} \tag{14.15}$$

where Φ_0 is the incident photon flux ($cm^{-2}\ s^{-1}$) on the surface of the semiconductor. We can also take into account the reflection of photons from the surface. Let $R(\lambda)$ be the fraction of photons that are reflected. (For bare silicon, $R \approx 35$ percent.) If we assume that each photon absorbed creates one electron–hole pair, then the generation rate of electron–hole pairs as a function of distance x from the surface is

$$G_L = \alpha(\lambda)\Phi_0(\lambda)[1-R(\lambda)]e^{-\alpha(\lambda)x} \tag{14.16}$$

where each parameter may be a function of the incident wavelength. Figure 14.11 shows the excess minority carrier concentrations in this pn solar cell for two values of wavelength and for the case when $s = 0$ at the surface.

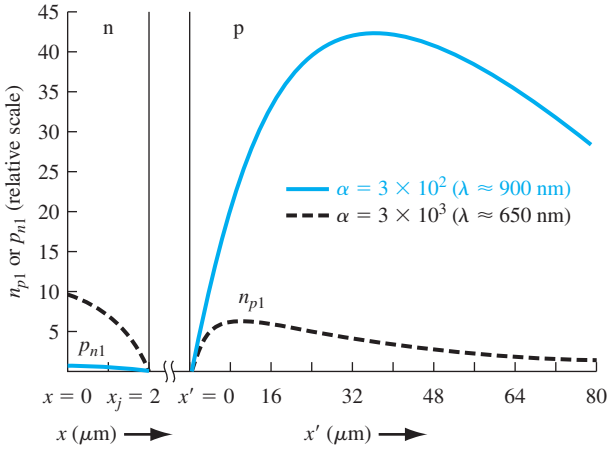


Figure 14.11 | Steady-state, photon-induced normalized minority carrier concentration in the pn junction solar cell for two values of incident photon wavelength ($x_j = 2$ μm, $W = 1$ μm, $L_p = L_n = 40$ μm).

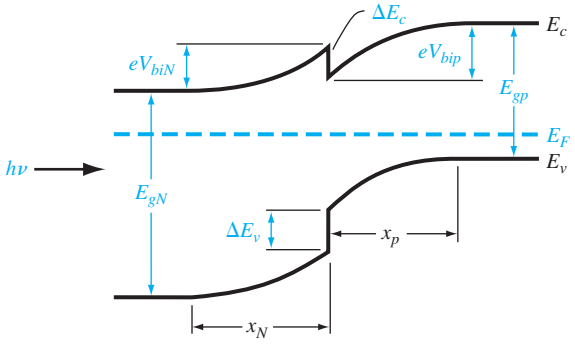


Figure 14.12 | The energy-band diagram of a pN heterojunction in thermal equilibrium.

14.2.4 The Heterojunction Solar Cell

As we have mentioned in previous chapters, a heterojunction is formed between two semiconductors with different bandgap energies. A typical pN heterojunction energy-band diagram in thermal equilibrium is shown in Figure 14.12. Assume that photons are incident on the wide-bandgap material. Photons with energy less than E_{gN} will pass through the wide-bandgap material, which acts as an optical window, and photons with energies greater than E_{gp} will be absorbed in the narrow bandgap material. On the average, excess carriers created in the depletion region and within a diffusion length of the junction will be collected and will contribute to the photocurrent. Photons with an energy greater than E_{gN} will be absorbed in the wide-bandgap

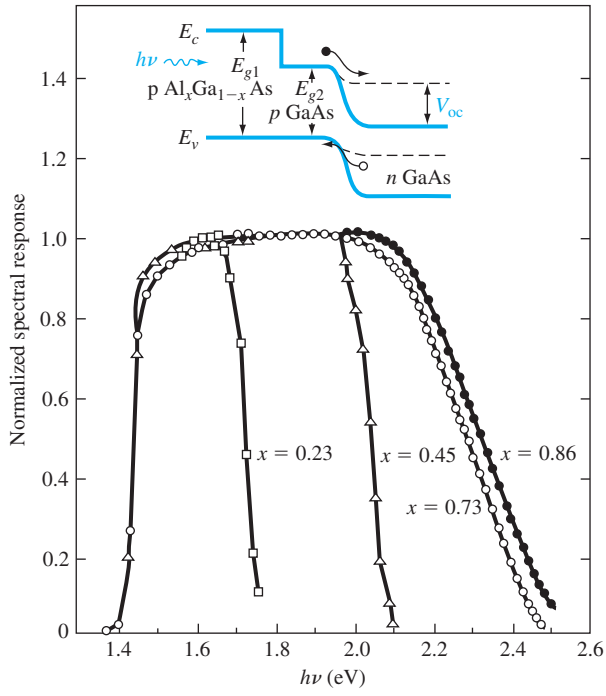


Figure 14.13 | The normalized spectral response of several AlGaAs/GaAs solar cells with different compositions. *(From Sze [17].)*

material, and excess carriers generated within one diffusion length of the junction will be collected. If E_{gN} is large enough, then the high-energy photons will be absorbed in the space charge region of the narrow-bandgap material. This heterojunction solar cell should have better characteristics than a homojunction cell, especially at the shorter wavelengths.

A variation of the heterojunction is shown in Figure 14.13. A pn homojunction is formed and then a wide-bandgap material is grown on top. Again, the wide-bandgap material acts as an optical window for photon energies $h\nu < E_{g1}$. Photons with energies $E_{g2} < h\nu < E_{g1}$ will create excess carriers in the homojunction and photons with energies $h\nu > E_{g1}$ will create excess carriers in the window type material. If the absorption coefficient in the narrow bandgap material is high, then essentially all of the excess carriers will be generated within a diffusion length of the junction, so the collection efficiency will be very high. Figure 14.13 also shows the normalized spectral response for various mole fractions x in the $Al_xGa_{1-x}As$.

14.2.5 Amorphous Silicon Solar Cells

Single-crystal silicon solar cells tend to be expensive and are limited to approximately 6 inches in diameter. A system powered by solar cells requires, in general,

a very large area solar cell array to generate the required power. Amorphous silicon solar cells provide the possibility of fabricating large area and relatively inexpensive solar cell systems.

When silicon is deposited by CVD techniques at temperatures below 600°C, an amorphous film is formed regardless of the type of substrate. In amorphous silicon, there is only very short range order, and no crystalline regions are observed. Hydrogen may be incorporated in the silicon to reduce the number of dangling bonds, creating a material called hydrogenated amorphous silicon.

The density of states versus energy for amorphous silicon is shown in Figure 14.14. Amorphous silicon contains large numbers of electronic energy states within the normal bandgap of single-crystal silicon. However, because of the short-range order, the effective mobility is quite small, typically in the range between 10^{-6} and 10^{-3} cm^2/V-s. The mobilities in the states above E_c and below E_v are between 1 and 10 cm^2/V-s. Consequently, conduction through the energy states between E_c and E_v is negligible because of the low mobility. Because of the difference in mobility values, E_c and E_v are referred to as the mobility edges and the energy between E_c and E_v is referred to as the mobility gap. The mobility gap can be modified by adding specific types of impurities. Typically, the mobility gap is on the order of 1.7 eV.

Amorphous silicon has a very high optical absorption coefficient, so most sunlight is absorbed within approximately 1 μm of the surface. Consequently, only a

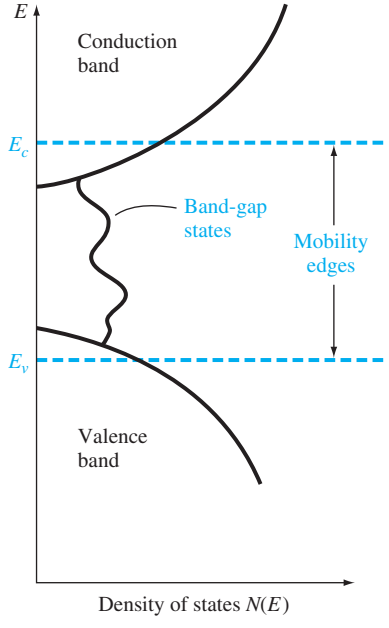


Figure 14.14 | Density of states versus energy of amorphous silicon.
(From Yang [22].)

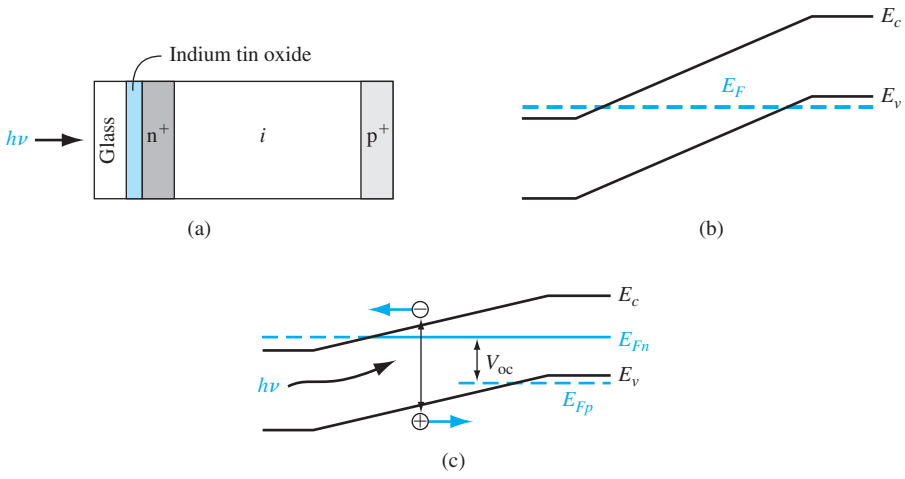


Figure 14.15 | The (a) cross section, (b) energy-band diagram at thermal equilibrium, and (c) energy-band diagram under photon illumination of an amorphous silicon PIN solar cell. *(From Yang [22].)*

very thin layer of amorphous silicon is required for a solar cell. A typical amorphous silicon solar cell is a PIN device shown in Figure 14.15. The amorphous silicon is deposited on an optically transparent indium tin oxide–coated glass substrate. If aluminum is used as the back contact, it will reflect any transmitted photons back through the PIN device. The n^+ and p^+ regions can be quite thin while the intrinsic region may be in the range of 0.5 to 1.0 μm thick. The energy-band diagram for the thermal equilibrium case is shown in the figure. Excess carriers generated in the intrinsic region are separated by the electric field and produce the photocurrent, as we have discussed. Conversion efficiencies are smaller than in single-crystal silicon, but the reduced cost makes this technology attractive. Amorphous silicon solar cells approximately 40 cm wide and many meters long have been fabricated.

TEST YOUR UNDERSTANDING

TYU 14.2 Consider a silicon pn junction solar cell with the parameters given in Example 14.3. Determine the required photocurrent density to produce an open-circuit voltage of $V_{oc} = 0.60$ V.

(Ans. $J_L = 0.414$ mA/cm^2)

TYU 14.3 Consider the silicon pn junction solar cell described in Example 14.3. Let the solar intensity increase by a factor of 10. Calculate the open-circuit voltage.

(Ans. $V_{oc} = 0.574$ V)

TYU 14.4 The silicon pn junction solar cell described in TYU 14.2 has a cross-sectional area of 1 cm^2. Determine the maximum power that can be delivered to a load.

(Ans. 0.205 W)

14.3 | PHOTODETECTORS

There are several semiconductor devices that can be used to detect the presence of photons. These devices are known as photodetectors; they convert optical signals into electrical signals. When excess electrons and holes are generated in a semiconductor, there is an increase in the conductivity of the material. This change in conductivity is the basis of the photoconductor, perhaps the simplest type of photodetector. If electrons and holes are generated within the space charge region of a pn junction, then they will be separated by the electric field and a current will be produced. The pn junction is the basis of several photodetector devices including the photodiode and the phototransistor.

14.3.1 Photoconductor

Figure 14.16 shows a bar of semiconductor material with ohmic contacts at each end and a voltage applied between the terminals. The initial thermal-equilibrium conductivity is

$$\sigma_0 = e(\mu_n n_0 + \mu_p p_0) \tag{14.17}$$

If excess carriers are generated in the semiconductor, the conductivity becomes

$$\sigma = e[\mu_n(n_0 + \delta n) + \mu_p(p_0 + \delta p)] \tag{14.18}$$

where δn and δp are the excess electron and hole concentrations, respectively. If we consider an n-type semiconductor, then, from charge neutrality, we can assume that

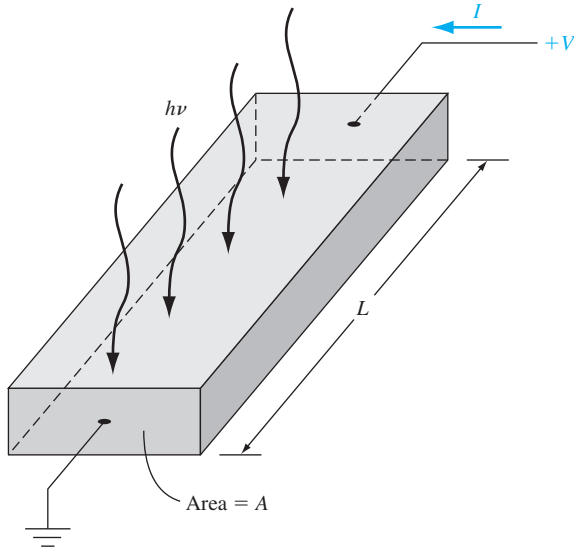


Figure 14.16 | A photoconductor.

$\delta n = \delta p \equiv \delta p$. We will use δp as the concentration of excess carriers. In steady state, the excess carrier concentration is given by $\delta p = G_L \tau_p$, where G_L is the generation rate of excess carriers (cm^{-3}-s^{-1}) and τ_p is the excess minority carrier lifetime.

The conductivity from Equation (14.18) can be rewritten as

$$\sigma = e(\mu_n n_0 + \mu_p p_0) + e(\delta p)(\mu_n + \mu_p) \tag{14.19}$$

The change in conductivity due to the optical excitation, known as the *photoconductivity,* is then

$$\Delta\sigma = e(\delta p)(\mu_n + \mu_p) \tag{14.20}$$

An electric field is induced in the semiconductor by the applied voltage, which produces a current. The current density can be written as

$$J = (J_0 + J_L) = (\sigma_0 + \Delta\sigma)\mathrm{E} \tag{14.21}$$

where J_0 is the current density in the semiconductor prior to optical excitation and J_L is the photocurrent density. The photocurrent density is $J_L = \Delta\sigma \cdot \mathrm{E}$. If the excess electrons and holes are generated uniformly throughout the semiconductor, then the photocurrent is given by

$$I_L = J_L \cdot A = \Delta\sigma \cdot A\mathrm{E} = eG_L\tau_p(\mu_n + \mu_p)A\mathrm{E} \tag{14.22}$$

where A is the cross-sectional area of the device. The photocurrent is directly proportional to the excess carrier generation rate, which in turn is proportional to the incident photon flux.

If excess electrons and holes are not generated uniformly throughout the semiconductor material, then the total photocurrent is found by integrating the photoconductivity over the cross-sectional area.

Since $\mu_n \mathrm{E}$ is the electron drift velocity, the electron transit time, that is, the time required for an electron to flow through the photoconductor, is

$$t_n = \frac{L}{\mu_n \mathrm{E}} \tag{14.23}$$

The photocurrent, from Equation (14.22), can be rewritten as

$$I_L = eG_L\left(\frac{\tau_p}{t_n}\right)\left(1 + \frac{\mu_p}{\mu_n}\right)AL \tag{14.24}$$

We may define a photoconductor gain, Γ_{ph}, as the ratio of the rate at which charge is collected by the contacts to the rate at which charge is generated within the photoconductor. We can write the gain as

$$\Gamma_{\mathrm{ph}} = \frac{I_L}{eG_L AL} \tag{14.25}$$

which, using Equation (14.24), can be written

$$\Gamma_{\mathrm{ph}} = \frac{\tau_p}{t_n}\left(1 + \frac{\mu_p}{\mu_n}\right) \tag{14.26}$$

EXAMPLE 14.4

Objective: Calculate the photoconductor gain of a silicon photoconductor.

Consider an n-type silicon photoconductor with a length $L = 100\ \mu$m, cross-sectional area $A = 10^{-7}$ cm^2, and minority carrier lifetime $\tau_p = 10^{-6}$ s. Let the applied voltage be $V = 10$ volts.

■ Solution

The electron transit time is determined as

$$t_n = \frac{L}{\mu_n \mathrm{E}} = \frac{L^2}{\mu_n V} = \frac{(100 \times 10^{-4})^2}{(1350)(10)} = 7.41 \times 10^{-9}\ \mathrm{s}$$

The photoconductor gain is then

$$\Gamma_{\mathrm{ph}} = \frac{\tau_p}{t_n}\left(1 + \frac{\mu_p}{\mu_n}\right) = \frac{10^{-6}}{7.41 \times 10^{-9}}\left(1 + \frac{480}{1350}\right) = 1.83 \times 10^2$$

■ Comment

The fact that a photoconductor—a bar of semiconductor material—has a gain may be surprising.

■ EXERCISE PROBLEM

Ex 14.4 Consider the photoconductor described in Example 14.4. Determine the photocurrent if $G_L = 10^{21}$ cm^{-3} s^{-1} and E = 10 V/cm. Also assume that $\mu_n = 1000$ cm^2/V-s and $\mu_p = 400$ cm^2/V-s.

(Ans. $I_{ph} = 0.2224$ A)

Let's consider physically what happens to a photon-generated electron, for example. After the excess electron is generated, it drifts very quickly out of the photoconductor at the anode terminal. In order to maintain charge neutrality throughout the photoconductor, another electron immediately enters the photoconductor at the cathode and drifts toward the anode. This process will continue during a time period equal to the mean carrier lifetime. At the end of this period, on the average, the photoelectron will recombine with a hole.

The electron transit time, using the parameters from Example 14.4, is $t_n = 7.41 \times 10^{-9}$ s. In a simplistic sense, the photoelectron will circulate around the photoconductor circuit 135 times during the 10^{-6} s time duration, which is the mean carrier lifetime. If we take into account the photon-generated hole, the total number of charges collected at the photoconductor contacts for every electron generated is 183.

When the optical signal ends, the photocurrent will decay exponentially with a time constant equal to the minority carrier lifetime. From the photoconductor gain expression, we would like a large minority carrier lifetime, but the switching speed is enhanced by a small minority carrier lifetime. There is obviously a trade-off between gain and speed. In general, the performance of a photodiode, which we will discuss next, is superior to that of a photoconductor.

14.3.2 Photodiode

A photodiode is a pn junction diode operated with an applied reverse-biased voltage. We will initially consider a long diode in which excess carriers are generated

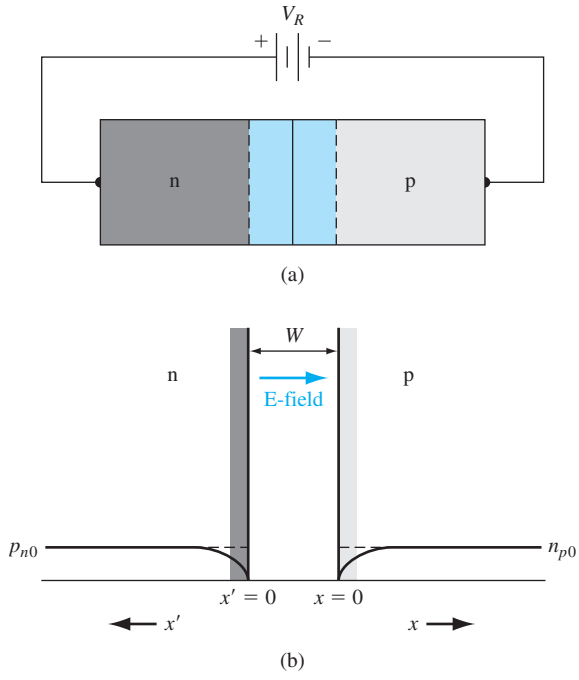


Figure 14.17 | (a) A reverse-biased pn junction. (b) Minority carrier concentration in the reverse-biased pn junction.

uniformly throughout the semiconductor device. Figure 14.17a shows the reverse-biased diode and Figure 14.17b shows the minority carrier distribution in the reverse-biased junction prior to photon illumination.

Let G_L be the generation rate of excess carriers. The excess carriers generated within the space charge region are swept out of the depletion region very quickly by the electric field; the electrons are swept into the n region and the holes into the p region. The photon-generated current density from the space charge region is given by

$$J_{L1} = e\int G_L\, dx \tag{14.27}$$

where the integral is over the space charge region width. If G_L is constant throughout the space charge volume, then

$$J_{L1} = eG_LW \tag{14.28}$$

where W is the space charge width. We may note that J_{L1} is in the reverse-biased direction through the pn junction. This component of photocurrent responds very quickly to the photon illumination and is known as the prompt photocurrent.

We may note, by comparing Equations (14.28) and (14.25), that the photodiode gain is unity. The speed of the photodiode is limited by the carrier transport through

the space charge region. If we assume that the saturation drift velocity is 10^7 cm/s and the depletion width is 2 μm, the transit time is $\tau_t = 20$ ps. The ideal modulating frequency has a period of $2\tau_t$, so the frequency is $f = 25$ GHz. This frequency response is substantially higher than that of photoconductors.

Excess carriers are also generated within the neutral n and p regions of the diode. The excess minority carrier electron distribution in the p region is found from the ambipolar transport equation, which is

$$D_n \frac{\partial^2(\delta n_p)}{\partial x^2} + G_L - \frac{\delta n_p}{\tau_{n0}} = \frac{\partial(\delta n_p)}{\partial t} \tag{14.29}$$

We will assume that the E-field is zero in the neutral regions. In steady state, $\partial(\delta n_p)/\partial t = 0$, so that Equation (14.29) can be written as

$$\frac{d^2(\delta n_p)}{dx^2} - \frac{\delta n_p}{L_n^2} = -\frac{G_L}{D_n} \tag{14.30}$$

where $L_n^2 = D_n \tau_{n0}$.

The solution to Equation (14.30) can be found as the sum of the homogeneous and particular solutions. The homogeneous solution is found from the equation

$$\frac{d^2(\delta n_{\text{ph}})}{dx^2} - \frac{\delta n_{\text{ph}}}{L_n^2} = 0 \tag{14.31}$$

where δn_{ph} is the homogeneous solution and is given by

$$\delta n_{\text{ph}} = Ae^{-x/Ln} + Be^{+x/L_n} \qquad (x \geq 0) \tag{14.32}$$

One boundary condition is that δn_{ph} must remain finite, which implies that $B \equiv 0$ for the "long" diode.

The particular solution is found from

$$-\frac{\delta n_{pp}}{L_n^2} = -\frac{G_L}{D_n} \tag{14.33}$$

which yields

$$\delta n_{pp} = \frac{G_L L_n^2}{D_n} = \frac{G_L(D_n \tau_{n0})}{D_n} = G_L \tau_{n0} \tag{14.34}$$

The total steady-state solution for the excess minority carrier electron concentration in the p region is then

$$\delta n_p = Ae^{-x/L_n} + G_L \tau_{n0} \tag{14.35}$$

The total electron concentration is zero at $x = 0$ for the reverse-biased junction. The excess electron concentration $x = 0$ is then

$$\delta n_p(x = 0) = -n_{p0} \tag{14.36}$$

Using the boundary condition from Equation (14.36), the electron concentration given by Equation (14.35) becomes

$$\delta n_p = G_L \tau_{n0} - (G_L \tau_{n0} + n_{p0})e^{-x/L_n} \tag{14.37}$$

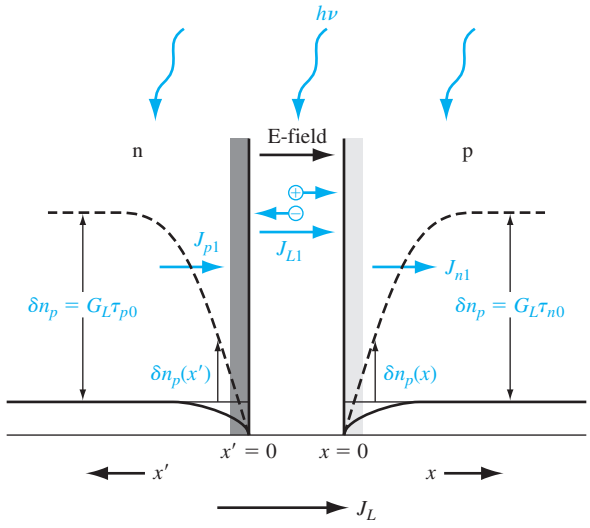


Figure 14.18 | Steady-state, photoinduced minority carrier concentrations and photocurrents in a "long" reverse-biased pn junction.

We can find the excess minority carrier hole concentration in the n region using the same type of analysis. Using the x' notation shown in Figure 14.17, we can write

$$\delta p_n = G_L\tau_{p0} - (G_L\tau_{p0} + p_{n0})e^{-x'/L_p} \tag{14.38}$$

Equations (14.37) and (14.38) are plotted in Figure 14.18. We may note that the steady-state values far from the space charge region are the same as were given previously.

The gradient in the minority carrier concentrations will produce diffusion currents in the pn junction. The diffusion current density at $x = 0$ due to minority carrier electrons is

$$\begin{aligned} J_{n1} &= eD_n \frac{d(\delta n_p)}{dx}\bigg|_{x=0} = eD_n \frac{d}{dx}[G_L\tau_{n0} - (G_L\tau_{n0} + n_{p0})e^{-x/L_n}]\bigg|_{x=0} \\ &= \frac{eD_n}{L_n}(G_L\tau_{n0} + n_{p0}) \end{aligned} \tag{14.39}$$

Equation (14.39) can be written as

$$J_{n1} = eG_L L_n + \frac{eD_n n_{p0}}{L_n} \tag{14.40}$$

The first term in Equation (14.40) is the steady-state photocurrent density while the second term is the ideal reverse saturation current density due to the minority carrier electrons.

The diffusion current density (in the x direction) at $x' = 0$ due to the minority carrier holes is

$$J_{p1} = eG_L L_p + \frac{eD_p p_{n0}}{L_p} \tag{14.41}$$

Similarly, the first term is the steady-state photocurrent density and the second term is the ideal reverse saturation current density.

The total steady-state diode photocurrent density for the long diode is now

$$J_L = eG_L W + eG_L L_n + eG_L L_p = e(W + L_n + L_p)G_L \tag{14.42}$$

Again note that the photocurrent is in the reverse-biased direction through the diode. The photocurrent given by Equation (14.42) is the result of assuming uniform generation of excess carriers throughout the structure, a long diode, and steady state.

The time response of the diffusion components of the photocurrent is relatively slow, since these currents are the results of the diffusion of minority carriers toward the depletion region. The diffusion components of photocurrent are referred to as the delayed photocurrent.

EXAMPLE 14.5

Objective: Calculate the steady-state photocurrent density in a reverse-biased, long pn diode.

Consider a silicon pn diode at $T = 300$ K with the following parameters:

$$N_a = 10^{16}\ \text{cm}^{-3} \qquad N_d = 10^{16}\ \text{cm}^{-3}$$
$$D_n = 25\ \text{cm}^2/\text{s} \qquad D_p = 10\ \text{cm}^2/\text{s}$$
$$\tau_{n0} = 5 \times 10^{-7}\ \text{s} \qquad \tau_{p0} = 10^{-7}\ \text{s}$$

Assume that a reverse-biased voltage of $V_R = 5$ volts is applied and let $G_L = 10^{21}\ \text{cm}^{-3}\text{-s}^{-1}$.

Solution

We may calculate various parameters as follows:

$$L_n = \sqrt{D_n \tau_{n0}} = \sqrt{(25)(5 \times 10^{-7})} = 35.4\ \mu\text{m}$$

$$L_p = \sqrt{D_p \tau_{p0}} = \sqrt{(10)(10^{-7})} = 10.0\ \mu\text{m}$$

$$V_{bi} = V_t \ln\left(\frac{N_a N_d}{n_i^2}\right) = (0.0259) \ln\left[\frac{(10^{16})(10^{16})}{(1.5 \times 10^{10})^2}\right] = 0.695\ \text{V}$$

$$W = \left\{\frac{2\epsilon_s}{e}\left(\frac{N_a + N_d}{N_a N_d}\right)(V_{bi} + V_R)\right\}^{1/2}$$

$$= \left\{\frac{2(11.7)(8.85 \times 10^{-14})}{1.6 \times 10^{-19}} \cdot \frac{(2 \times 10^{16})}{(10^{16})(10^{16})} \cdot (0.695 + 5)\right\}^{1/2} = 1.21\ \mu\text{m}$$

Finally, the steady-state photocurrent density is

$$J_L = e(W + L_n + L_p)G_L$$

$$= (1.6 \times 10^{-19})(1.21 + 35.4 + 10.0) \times 10^{-4}(10^{21}) = 0.75\ \text{A/cm}^2$$

Comment

Again, keep in mind that this photocurrent is in the reverse-biased direction through the diode and is many orders of magnitude larger than the reverse-biased saturation current density in the pn junction diode.

EXERCISE PROBLEM

Ex 14.5 The doping concentrations of the photodiode described in Example 14.5 are changed to $N_a = N_d = 10^{15}$ cm^{-3}. (*a*) Determine the steady-state photocurrent density. (*b*) Calculate the ratio of prompt photocurrent to steady-state photocurrent.

[Ans. (*a*) $J_L = 0.787$ A/cm^2; (*b*) $J_L/J_L = 0.0773$]

In this example calculation, $L_n \gg W$ and $L_p \gg W$. In many pn junction structures, the assumption of a long diode will not be valid, so the photocurrent expression will have to be modified. In addition, the photon energy absorption may not be uniform throughout the pn structure. The effect of nonuniform absorption will be considered in the next section.

14.3.3 PIN Photodiode

In many photodetector applications, the speed of response is important; therefore, the prompt photocurrent generated in the space charge region is the only photocurrent of interest. To increase the photodetector sensitivity, the depletion region width should be made as large as possible. This can be achieved in a PIN photodiode.

The PIN diode consists of a p region and an n region separated by an intrinsic region. A sketch of a PIN diode is shown in Figure 14.19a. The intrinsic region width W is much larger than the space charge width of a normal pn junction. If a reverse bias is applied to the PIN diode, the space charge region extends completely through the intrinsic region.

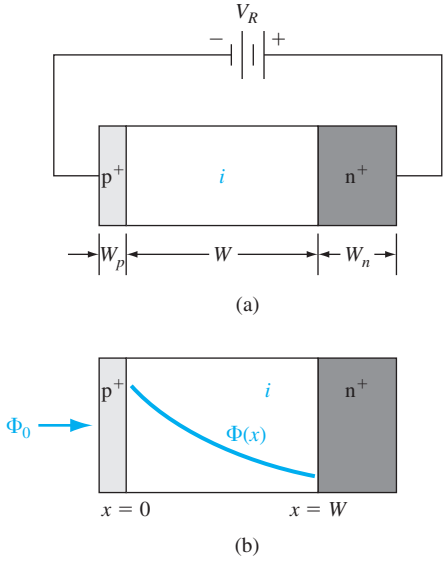


Figure 14.19 | (a) A reverse-biased PIN photodiode. (b) Geometry showing nonuniform photon absorption.

Assume that a photon flux Φ_0 is incident on the p^+ region. If we assume that the p^+ region width W_p is very thin, then the photon flux, as a function of distance, in the intrinsic region is $\Phi(x) = \Phi_0 e^{-\alpha x}$, where α is the photon absorption coefficient. This nonlinear photon absorption is shown in Figure 14.19b. The photocurrent density generated in the intrinsic region can be found as

$$J_L = e\int_0^W G_L\,dx = e\int_0^W \Phi_0\alpha e^{-\alpha x}\,dx = e\Phi_0(1-e^{-\alpha W}) \tag{14.43}$$

This equation assumes that there is no electron–hole recombination within the space charge region and also that each photon absorbed creates one electron–hole pair.

EXAMPLE 14.6

Objective: Calculate the photocurrent density in a PIN photodiode.

Consider a silicon PIN diode with an intrinsic region width of $W = 20\ \mu$m. Assume that the photon flux is 10^{17} cm^{-2}-s^{-1} and the absorption coefficient is $\alpha = 10^3$ cm^{-1}.

■ Solution

The generation rate of electron–hole pairs at the front edge of the intrinsic region is

$$G_{L1} = \alpha\Phi_0 = (10^3)(10^{17}) = 10^{20}\ \text{cm}^{-3}\text{-s}^{-1}$$

and the generation rate at the back edge of the intrinsic region is

$$G_{L2} = \alpha\Phi_0 e^{-\alpha W} = (10^3)(10^{17})\exp\left[-(10^3)(20\times 10^{-4})\right]$$
$$= 0.135\times 10^{20}\ \text{cm}^{-3}\text{-s}^{-1}$$

The generation rate is obviously not uniform throughout the intrinsic region. The photocurrent density is then

$$J_L = e\Phi_0(1-e^{-\alpha W})$$
$$= (1.6\times 10^{-19})(10^{17})\{1-\exp\left[-(10^3)(20\times 10^{-4})\right]\}$$
$$= 13.8\ \text{mA/cm}^2$$

■ Comment

The prompt photocurrent density of a PIN photodiode will be larger than that of a regular photodiode since the space charge region is larger in a PIN photodiode.

■ EXERCISE PROBLEM

Ex 14.6 Repeat Example 14.6 for photon absorption coefficients of (*a*) $\alpha = 10^2$ cm^{-1} and (*b*) $\alpha = 10^4$ cm^{-1}.

[Ans. (*a*) $J_L = 2.90$ mA/cm^2; (*b*) $J_L = 16.0$ mA/cm^2]

In most situations, we will not have a long diode; thus, the steady-state photocurrent described by Equation (14.42) will not apply for most photodiodes.

14.3.4 Avalanche Photodiode

The avalanche photodiode is similar to the pn or PIN photodiode except that the bias applied to the avalanche photodiode is sufficiently large to cause impact ionization.

Electron–hole pairs are generated in the space charge region by photon absorption, as we have discussed previously. The photon-generated electrons and holes now generate additional electron–hole pairs through impact ionization. The avalanche photodiode now has a current gain introduced by the avalanche multiplication factor.

The electron–hole pairs generated by photon absorption and by impact ionization are swept out of the space charge region very quickly. If the saturation velocity is 10^7 cm/s in a depletion region that is 10 μm wide, then the transit time is

$$\tau_t = \frac{10^7}{10 \times 10^{-4}} = 100 \text{ ps}$$

The period of a modulation signal would be $2\tau_t$, so that the frequency would be

$$f = \frac{1}{2\tau_t} = \frac{1}{200 \times 10^{-12}} = 5 \text{ GHz}$$

If the avalanche photodiode current gain is 20, then the gain-bandwidth product is 100 GHz. The avalanche photodiode could respond to light waves modulated at microwave frequencies.

14.3.5 Phototransistor

A bipolar transistor can also be used as a photodetector. The phototransistor can have high gain through the transistor action. An npn bipolar phototransistor is shown in Figure 14.20a. This device has a large base–collector junction area and is usually operated with the base open circuited. Figure 14.20b shows the block diagram of the phototransistor. Electrons and holes generated in the reverse-biased B–C junction are swept out of the space charge region, producing a photocurrent I_L. Holes are swept

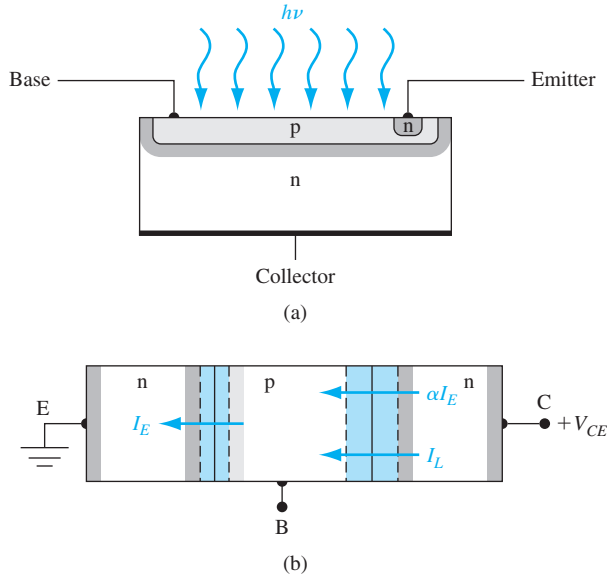


Figure 14.20 | (a) A bipolar phototransistor. (b) Block diagram of the open-base phototransistor.

into the p-type base, making the base positive with respect to the emitter. Since the B–E becomes forward-biased, electrons will be injected from the emitter back into the base, leading to the normal transistor action.

From Figure 14.20b, we see that

$$I_E = \alpha I_E + I_L \tag{14.44}$$

where I_L is the photon-generated current and α is the common base current gain. Since the base is an open circuit, we have $I_C = I_E$, so Equation (14.44) can be written as

$$I_C = \alpha I_C + I_L \tag{14.45}$$

Solving for I_C, we find

$$I_C = \frac{I_L}{1 - \alpha} \tag{14.46}$$

Relating α to β, the dc common emitter current gain, Equation (14.46) becomes

$$I_C = (1 + \beta) I_L \tag{14.47}$$

Equation (14.47) shows that the basic B–C photocurrent is multiplied by the factor $(1 + \beta)$. The phototransistor, then, amplifies the basic photocurrent.

With the relatively large B–C junction area, the frequency response of the phototransistor is limited by the B–C junction capacitance. Since the base is essentially the input to the device, the large B–C capacitance is multiplied by the Miller effect, so the frequency response of the phototransistor is further reduced. The phototransistor, however, is a lower-noise device than the avalanche photodiode.

Phototransistors can also be fabricated in heterostructures. The injection efficiency is increased as a result of the bandgap differences, as we discussed in Chapter 12. With the bandgap difference, the lightly doped base restriction no longer applies. A fairly heavily doped, narrow-base device can be fabricated with a high blocking voltage and a high gain.

TEST YOUR UNDERSTANDING

TYU 14.5 Consider a long silicon pn junction photodiode with the parameters given in Example 14.5. The cross-sectional area is $A = 10^{-3}$ cm^2. Assume the photodiode is reverse biased by a 5-volt battery in series with a 5 kΩ load resistor. An optical signal at a wavelength of $\lambda = 1$ μm is incident on the photodiode producing a uniform generation rate of excess carriers throughout the entire device. Determine the incident intensity such that the voltage across the load resistor is 0.5 V.

(Ans. $I_a = 0.266$ W/cm^2)

14.4 | PHOTOLUMINESCENCE AND ELECTROLUMINESCENCE

In the first section of this chapter, we have discussed the creation of excess electron–hole pairs by photon absorption. Eventually, excess electrons and holes recombine, and in direct bandgap materials the recombination process may result in the emission of a photon. The general property of light emission is referred to as luminescence.

When excess electrons and holes are created by photon absorption, photon emission from the recombination process is called photoluminescence.

Electroluminescence is the process of generating photon emission when the excitation of excess carriers is a result of an electric current caused by an applied electric field. We are mainly concerned here with injection electroluminescence, the result of injecting carriers across a pn junction. The light emitting diode and the pn junction laser diode are examples of this phenomenon. In these devices, electric energy, in the form of a current, is converted directly into photon energy.

14.4.1 Basic Transitions

Once electron–hole pairs are formed, there are several possible processes by which the electrons and holes can recombine. Some recombination processes may result in photon emission from direct bandgap materials, whereas other recombination processes in the same material may not.

Figure 14.21a shows the basic interband transitions. Curve (i) corresponds to an intrinsic emission very close to the bandgap energy of the material. Curves (ii) and (iii) correspond to energetic electrons or holes. If either of these recombinations result in the emission of a photon, the energy of the emitted photon will be slightly larger than the bandgap energy. There will then be an emission spectrum and a bandwidth associated with the emission.

The possible recombination processes involving impurity or defect states are shown in Figure 14.21b. Curve (i) is the conduction band to acceptor transition, curve (ii) is the donor to valence-band transition, curve (iii) is the donor to acceptor transition, and curve (iv) is the recombination due to a deep trap. Curve (iv) is a nonradiative process corresponding to the Shockley–Read–Hall recombination process discussed in Chapter 6. The other recombination processes may or may not result in the emission of a photon.

Figure 14.21c shows the Auger recombination process, which can become important in direct bandgap materials with high doping concentrations. The Auger recombination process is a nonradiative process. The Auger recombination, in one case, shown in curve (i), is a recombination between an electron and hole, accompanied by the transfer of energy to another free hole. Similarly, in the second case, the recombination between an electron and hole can result in the transfer of energy to a free electron as shown in curve (ii). The third particle involved in this process will eventually lose its energy to the lattice in the form of heat. The process involving two holes and an electron would occur predominantly in heavily doped p-type materials, and the process involving two electrons and a hole would occur primarily in a heavily doped n-type material.

The recombination processes shown in Figure 14.21a indicate that the emission of a photon is not necessarily at a single, discrete energy, but can occur over a range of energies. The spontaneous emission rate generally has the form

$$I(\nu) \propto \nu^2 (h\nu - E_g)^{1/2} \exp\left[\frac{-(h\nu - E_g)}{kT}\right] \tag{14.48}$$

where E_g is the bandgap energy. Figure 14.22 shows the emission spectra from gallium arsenide. The peak photon energy decreases with temperature because the

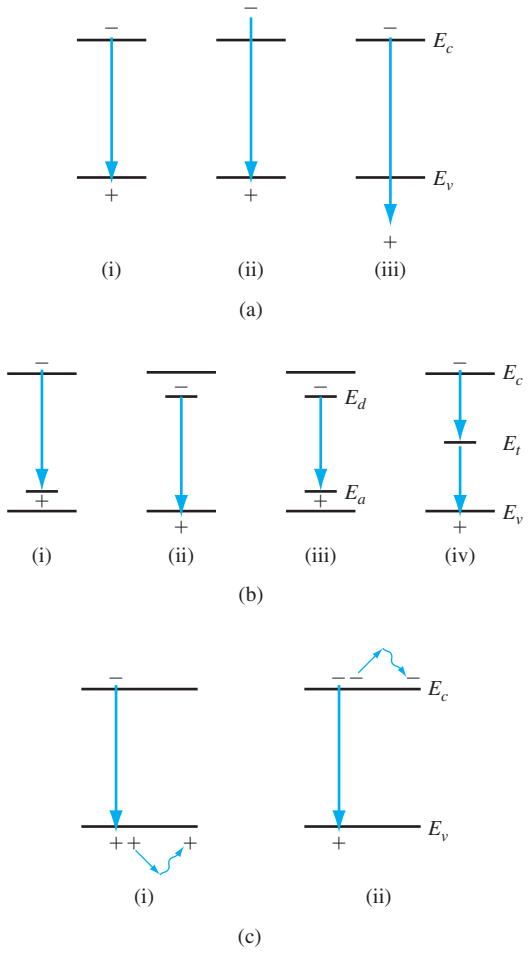


Figure 14.21 | Basic transitions in a semiconductor.

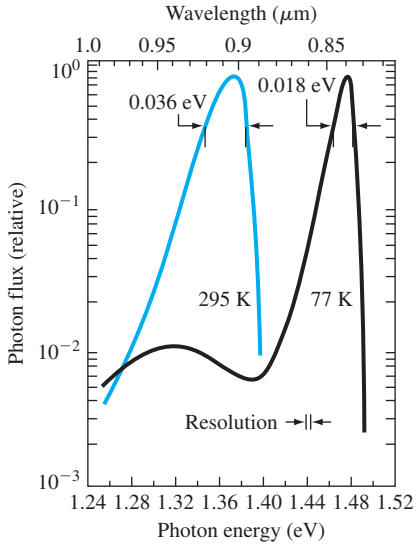


Figure 14.22 | GaAs diode emission spectra at $T = 300$ K and $T = 77$ K. *(From Sze and Ng [17].)*

bandgap energy decreases with temperature. We will show that the bandwidth of the emission spectra can be greatly reduced in a laser diode by using an optical resonator.

14.4.2 Luminescent Efficiency

We have shown that not all recombination processes are radiative. An efficient luminescent material is one in which radiative transitions predominate. The quantum efficiency is defined as the ratio of the radiative recombination rate to the total recombination rate for all processes. We can write

$$\eta_q = \frac{R_r}{R} \tag{14.49}$$

where η_q is the quantum efficiency, R_r is the radiative recombination rate, and R is the total recombination rate of the excess carriers. Since the recombination rate is

inversely proportional to lifetime, we can write the quantum efficiency in terms of lifetimes as

$$\eta_q = \frac{\tau_{nr}}{\tau_{nr} + \tau_r} \tag{14.50}$$

where τ_{nr} is the nonradiative lifetime and τ_r is the radiative lifetime. For a high luminescent efficiency, the nonradiative lifetimes must be large; thus, the probability of a nonradiative recombination is small compared to the radiative recombination.

The interband recombination rate of electrons and holes will be directly proportional to the number of electrons available and directly proportional to the number of available empty states (holes). We can write

$$R_r = Bnp \tag{14.51}$$

where R_r is the band-to-band radiative recombination rate and B is the constant of proportionality. The values of B for direct-bandgap materials are on the order of 10^6 larger than for indirect bandgap materials. The probability of a direct band-to-band radiative recombination transition in an indirect bandgap material is very unlikely.

One problem encountered with the emission of photons from a direct bandgap material is the reabsorption of the emitted photons. In general, the emitted photons will have energies $h\nu > E_g$, which means that the absorption coefficient is not zero for this energy. In order to generate a light output from a light emitting device, the process must take place near the surface. One possible solution to the reabsorption problem is to use heterojunction devices. These are discussed in later sections.

14.4.3 Materials

An important direct bandgap semiconductor material for optical devices is gallium arsenide. Another compound material that is of great interest is $Al_xGa_{1-x}As$. This material is a compound semiconductor in which the ratio of aluminum atoms to gallium atoms can be varied to achieve specific characteristics. Figure 14.23 shows the bandgap energy as a function of the mole fraction between aluminum and gallium. We can note from the figure that for $0 < x < 0.45$, the alloy material is a direct bandgap material. For $x > 0.45$, the material becomes an indirect bandgap material, not suitable for optical devices. For $0 < x < 0.35$, the bandgap energy can be expressed as

$$E_g = 1.424 + 1.247x \text{ eV} \tag{14.52}$$

Another compound semiconductor used for optical devices is the $GaAs_{1-x}P_x$ system. Figure 14.24a shows the bandgap energy as a function of the mole fraction x. For $0 \leq x \leq 0.45$, this material is also a direct bandgap material, and for $x > 0.45$, the bandgap becomes indirect. Figure 14.24b is the E versus k diagram, showing how the bandgap changes from direct to indirect as the mole fraction changes.

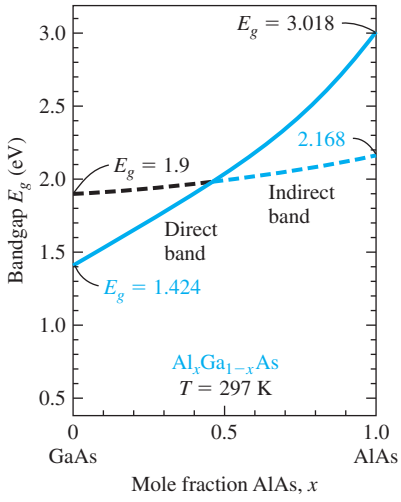


Figure 14.23 | Bandgap energy of $Al_xGa_{1-x}As$ as a function of the mole fraction x.
(From Sze [18].)

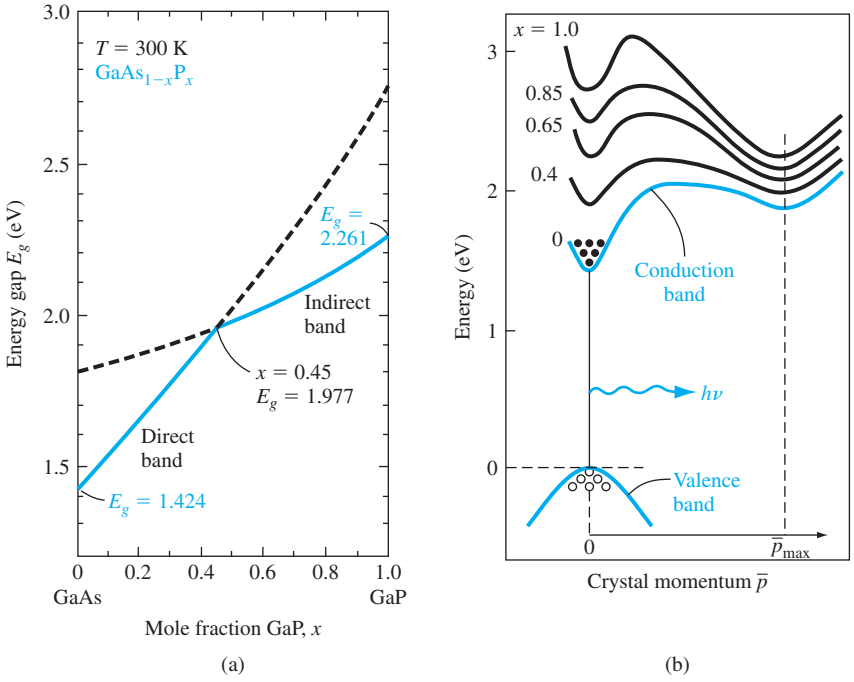


Figure 14.24 | (a) Bandgap energy of $GaAs_{1-x}P_x$ as a function of mole fraction x. (b) E versus k diagram of $GaAs_{1-x}P_x$ for various values of x.
(From Sze [18].)

EXAMPLE 14.7

Objective: Determine the output wavelength of a $GaAs_{1-x}P_x$ material for two different mole fractions.

Consider first GaAs and then $GaAs_{1-x}P_x$.

■ Solution

GaAs has a bandgap energy of $E_g = 1.42$ eV. This material would produce a photon output at a wavelength of

$$\lambda = \frac{1.24}{E} = \frac{1.24}{1.42} = 0.873 \ \mu\text{m}$$

This wavelength is in the infrared range and not in the visible range. If we desire a visible output with a wavelength of $\lambda = 0.653$ μm, for example, the bandgap energy would have to be

$$E = \frac{1.24}{\lambda} = \frac{1.24}{0.653} = 1.90 \text{ eV}$$

This bandgap energy would correspond to a mole fraction of approximately $x = 0.4$.

■ Comment

By changing the mole fraction in the $GaAs_{1-x}P_x$ system, the output can change from the infrared to the red spectrum.

■ EXERCISE PROBLEM

Ex 14.7 Determine the output wavelength of a $GaAs_{1-x}P_x$ material for mole fractions of (*a*) $x = 0.15$ and (*b*) $x = 0.30$.

[Ans. (*a*) $\lambda = 0.775$ μm; (*b*) $\lambda = 0.705$ μm]

14.5 | LIGHT EMITTING DIODES

Photodetectors and solar cells convert optical energy into electrical energy—the photons generate excess electrons and holes, which produce an electric current. We might also apply a voltage across a pn junction resulting in a diode current, which in turn can produce photons and a light output. This inverse mechanism is called injection electroluminescence. This device is known as a **L**ight **E**mitting **D**iode (LED). The spectral output of an LED may have a relatively wide wavelength bandwidth of between 30 and 40 nm. However, this emission spectrum is narrow enough so that a particular color is observed, provided the output is in the visible range.

14.5.1 Generation of Light

As we have discussed previously, photons may be emitted if an electron and hole recombine by a direct band-to-band recombination process in a direct bandgap material. The emission wavelength, from Equation (14.1), is

$$\lambda = \frac{hc}{E_g} = \frac{1.24}{E_g} \ \mu\text{m} \tag{14.53}$$

where E_g is the bandgap energy measured in electron-volts.

When a voltage is applied across a pn junction, electrons and holes are injected across the space charge region where they become excess minority carriers. These excess minority carriers diffuse into the neutral semiconductor regions where they recombine with majority carriers. If this recombination process is a direct band-to-band process, photons are emitted. The diode diffusion current is directly proportional to the recombination rate, so the output photon intensity will also be proportional to the ideal diode diffusion current. In gallium arsenide, electroluminescence originates primarily on the p side of the junction because the efficiency for electron injection is higher than that for hole injection.

14.5.2 Internal Quantum Efficiency

The *internal quantum efficiency* of an LED is the fraction of diode current that produces luminescence. The internal quantum efficiency is a function of the injection efficiency and a function of the percentage of radiative recombination events compared with the total number of recombination events.

The three current components in a forward-biased diode are the minority carrier electron diffusion current, the minority carrier hole diffusion current, and the space charge recombination current. These current densities can be written, respectively, as

$$J_n = \frac{eD_n n_{p0}}{L_n}\left[\exp\left(\frac{eV}{kT}\right) - 1\right] \tag{14.54a}$$

$$J_p = \frac{eD_p p_{n0}}{L_p}\left[\exp\left(\frac{eV}{kT}\right) - 1\right] \tag{14.54b}$$

and

$$J_R = \frac{en_i W}{2\tau_0}\left[\exp\left(\frac{eV}{2kT}\right) - 1\right] \tag{14.54c}$$

The recombination of electrons and holes within the space charge region is, in general, through traps near midgap and is a nonradiative process. Since luminescence is due primarily to the recombination of minority carrier electrons in GaAs, we can define an injection efficiency as the fraction of electron current to total current. Then

$$\gamma = \frac{J_n}{J_n + J_p + J_R} \tag{14.55}$$

where γ is the injection efficiency. We can make γ approach unity by using an n^+p diode so that J_p is a small fraction of the diode current and by forward biasing the diode sufficiently so that J_R is a small fraction of the total diode current.

Once the electrons are injected into the p region, not all electrons will recombine radiatively. We can define the radiative and nonradiative recombination rates as

$$R_r = \frac{\delta n}{\tau_r} \tag{14.56a}$$

and

$$R_{nr} = \frac{\delta n}{\tau_{nr}} \tag{14.56b}$$

where τ_r and τ_{nr} are the radiative and nonradiative recombination lifetimes, respectively, and δn is the excess carrier concentration. The total recombination rate is

$$R = R_r + R_{nr} = \frac{\delta n}{\tau} = \frac{\delta n}{\tau_r} + \frac{\delta n}{\tau_{nr}} \tag{14.57}$$

where τ is the net excess carrier lifetime.

The radiative efficiency is defined as the fraction of recombinations that are radiative. We can write

$$\eta = \frac{R_r}{R_r + R_{nr}} = \frac{\frac{1}{\tau_r}}{\frac{1}{\tau_r} + \frac{1}{\tau_{nr}}} = \frac{\tau}{\tau_r} \tag{14.58}$$

where η is the radiative efficiency. The nonradiative recombination rate is proportional to N_t, which is the density of nonradiative trapping sites within the forbidden bandgap. Obviously, the radiative efficiency increases as N_t is reduced.

The internal quantum efficiency is now written as

$$\eta_i = \gamma\eta \tag{14.59}$$

The radiative recombination rate is proportional to the p-type doping. As the p-type doping increases, the radiative recombination rate increases. However, the injection efficiency decreases as the p-type doping increases; therefore, there is an optimum doping that maximizes the internal quantum efficiency.

14.5.3 External Quantum Efficiency

One very important parameter of the LED is the *external quantum efficiency:* the fraction of generated photons that are actually emitted from the semiconductor. The external quantum efficiency is normally a much smaller number than the internal quantum efficiency. Once a photon has been produced in the semiconductor, there are three loss mechanisms the photon may encounter: photon absorption within the semiconductor, Fresnel loss, and critical angle loss.

Figure 14.25 shows a pn junction LED. Photons can be emitted in any direction. Since the emitted photon energy must be $h\nu \geq E_g$, these emitted photons can be reabsorbed within the semiconductor material. The majority of photons will actually be emitted away from the surface and reabsorbed in the semiconductor.

Photons must be emitted from the semiconductor into air; thus, the photons must be transmitted across a dielectric interface. Figure 14.26 shows the incident, reflected, and transmitted waves. The parameter $\overline{n}_2$ is the index of refraction for the semiconductor and $\overline{n}_1$ is the index of refraction for air. The reflection coefficient is

$$\Gamma = \left(\frac{\overline{n}_2 - \overline{n}_1}{\overline{n}_2 + \overline{n}_1}\right)^2 \tag{14.60}$$

This effect is called Fresnel loss. The reflection coefficient Γ is the fraction of incident photons that are reflected back into the semiconductor.

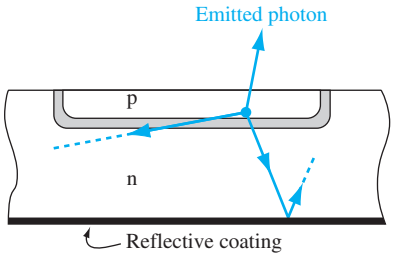


Figure 14.25 | Schematic of photon emission at the pn junction of an LED.

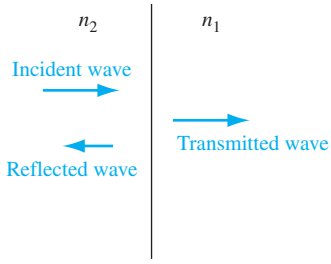


Figure 14.26 | Schematic of incident, reflected, and transmitted photons at a dielectric interface.

EXAMPLE 14.8

Objective: Calculate the reflection coefficient at a semiconductor–air interface.

Consider the interface between a GaAs semiconductor and air.

■ Solution

The index of refraction for GaAs is $\bar{n}_2 = 3.8$ at a wavelength of $\lambda = 0.70\ \mu$m and the index of refraction for air is $\bar{n}_1 = 1.0$. The reflection coefficient is

$$\Gamma = \left(\frac{\bar{n}_2 - \bar{n}_1}{\bar{n}_2 + \bar{n}_1}\right)^2 = \left(\frac{3.8 - 1.0}{3.8 + 1.0}\right)^2 = 0.34$$

■ Comment

A reflection coefficient of $\Gamma = 0.34$ means that 34 percent of the photons incident from the gallium arsenide onto the semiconductor–air interface are reflected back into the semiconductor.

■ EXERCISE PROBLEM

Ex 14.8 At a wavelength of $\lambda = 0.70\ \mu$m, the index of refraction for GaAs is $\bar{n}_2 = 3.8$ and that for GaP is $\bar{n}_2 = 3.2$. Consider a $GaAs_{1-x}P_x$ material with a mole fraction $x = 0.40$. Assuming the index of refraction is a linear function of the mole fraction, determine the reflection coefficient, Γ, at the $GaAs_{0.6}P_{0.4}$–air interface.

(Ans. $\Gamma = 0.315$)

Photons incident on the semiconductor–air interface at an angle are refracted as shown in Figure 14.27. If the photons are incident on the interface at an angle greater than the critical angle θ_c, the photons experience total internal reflection. The critical angle is determined from Snell's law and is given by

$$\theta_c = \sin^{-1}\left(\frac{\bar{n}_1}{\bar{n}_2}\right) \tag{14.61}$$

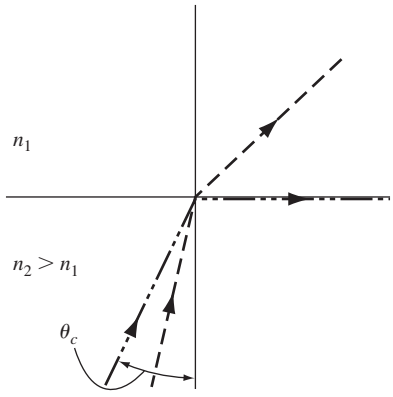


Figure 14.27 | Schematic showing refraction and total internal reflection at the critical angle at a dielectric interface.

EXAMPLE 14.9

Objective: Calculate the critical angle at a semiconductor–air interface.

Consider the interface between GaAs and air.

■ Solution

For GaAs, $\overline{n}_2 = 3.8$ at a wavelength of $\lambda = 0.70\ \mu$m and for air, $\overline{n}_1 = 1.0$. The critical angle is

$$\theta_c = \sin^{-1}\left(\frac{\overline{n}_1}{\overline{n}_2}\right) = \sin^{-1}\left(\frac{1.0}{3.8}\right) = 15.3^\circ$$

■ Comment

Any photon that is incident at an angle greater than 15.3° will be reflected back into the semiconductor.

■ EXERCISE PROBLEM

Ex 14.9 Repeat Example 14.9 for $GaAs_{0.6}P_{0.4}$. See Exercise Problem Ex 14.8 for a discussion of the dielectric constant.

(Ans. $\theta_c = 16.3^\circ$)

Figure 14.28a shows the external quantum efficiency plotted as a function of the p-type doping concentration and Figure 14.28b is a plot of the external efficiency as a function of junction depth below the surface. Both figures show that the external quantum efficiency is in the range of 1 to 3 percent.

14.5.4 LED Devices

The wavelength of the output signal of an LED is determined by the bandgap energy of the semiconductor. Gallium arsenide, a direct bandgap material, has a bandgap energy of $E_g = 1.42$ eV, which yields a wavelength of $\lambda = 0.873\ \mu$m. Comparing this wavelength to the visible spectrum, which is shown in Figure 14.5, the output

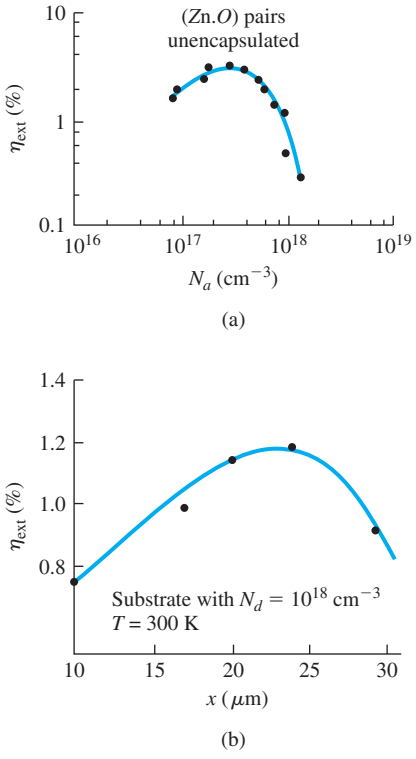


Figure 14.28 | (a) External quantum efficiency of a GaP LED versus acceptor doping. (b) External quantum efficiency of a GaAs LED versus junction depth. *(From Yang [22].)*

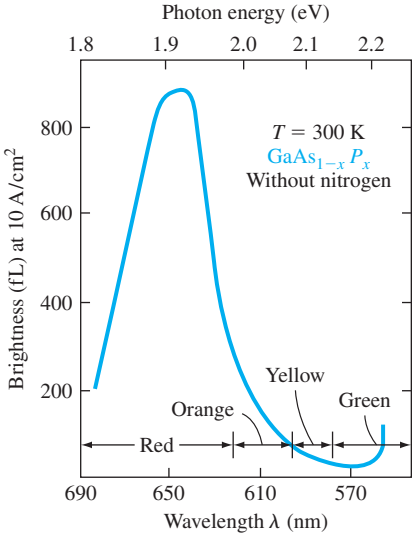


Figure 14.29 | Brightness of GaAsP diodes versus wavelength (or versus bandgap energy). *(From Yang [22].)*

of a GaAs LED is not in the visible range. For a visible output, the wavelength of the signal should be in the range of 0.4 to 0.72 μm. This range of wavelengths corresponds to bandgap energies between approximately 1.7 and 3.1 eV.

$GaAs_{1-x}P_x$ is a direct bandgap material for $0 \leq x \leq 0.45$, as shown in Figure 14.24. At $x = 0.40$, the bandgap energy is approximately $E_g = 1.9$ eV, which would produce an optical output in the red range. Figure 14.29 shows the brightness of $GaAs_{1-x}P_x$ diodes for different values of x. The peak also occurs in the red range. By using planar technology, $GaAs_{0.6}P_{0.4}$ monolithic arrays have been fabricated for numeric and alphanumeric displays. When the mole fraction x is greater than 0.45, the material changes to an indirect bandgap semiconductor so that the quantum efficiency is greatly reduced.

$GaAl_xAs_{1-x}$ can be used in a heterojunction structure to form an LED. A device structure is shown in Figure 14.30. Electrons are injected from the wide-bandgap N-$GaAl_{0.7}As_{0.3}$ into the narrow-bandgap p-$GaAl_{0.6}As_{0.4}$. The minority carrier electrons

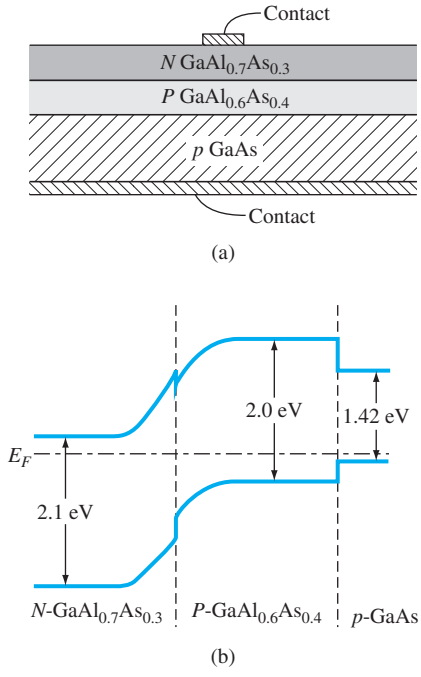


Figure 14.30 | The (a) cross section and (b) thermal equilibrium energy-band diagram of a GaAlAs heterojunction LED. *(From Yang [22].)*

in the p material can recombine radiatively. Since $E_{gp} < E_{gN}$, the photons are emitted through the wide-bandgap N material with essentially no absorption. The wide bandgap N material acts as an optical window and the external quantum efficiency increases.

14.6 | LASER DIODES

The photon output of the LED is due to an electron giving up energy as it makes a transition from the conduction band to the valence band. The LED photon emission is spontaneous in that each band-to-band transition is an independent event. The spontaneous emission process yields a spectral output of the LED with a fairly wide bandwidth. If the structure and operating condition of the LED are modified, the device can operate in a new mode, producing a coherent spectral output with a bandwidth of wavelengths less than 0.1 nm. This new device is a laser diode, where laser stands for **L**ight **A**mplification by **S**timulated **E**mission of **R**adiation. Although there are many different types of lasers, we are here concerned only with the pn junction laser diode.

14.6.1 Stimulated Emission and Population Inversion

Figure 14.31a shows the case when an incident photon is absorbed and an electron is elevated from an energy state E_1 to an energy state E_2. This process is known as induced absorption. If the electron spontaneously makes the transition back to the lower energy level with a photon being emitted, we have a spontaneous emission process as indicated in Figure 14.31b. On the other hand, if there is an incident photon at a time when an electron is in the higher energy state as shown in Figure 14.31c, the incident photon can interact with the electron, causing the electron to make a transition downward. The downward transition produces a photon. Since this process was initiated by the incident photon, the process is called *stimulated* or *induced emission*. Note that this stimulated emission process has produced two photons; thus, we can have optical gain or amplification. The two emitted photons are in phase so that the spectral output will be coherent.

In thermal equilibrium, the electron distribution in a semiconductor is determined by the Fermi–Dirac statistics. If the Boltzmann approximation applies, then we can write

$$\frac{N_2}{N_1} = \exp\left[\frac{-(E_2 - E_1)}{kT}\right] \tag{14.62}$$

where N_1 and N_2 are the electron concentrations in the energy levels E_1 and E_2, respectively, and where $E_2 > E_1$. In thermal equilibrium, $N_2 < N_1$. The probability of an induced absorption event is exactly the same as that of an induced emission event. The number of photons absorbed is proportional to N_1 and the number of additional photons emitted is proportional to N_2. In order to achieve optical amplification or for lasing action to occur, we must have $N_2 > N_1$; this is called population inversion. We cannot achieve lasing action at thermal equilibrium.

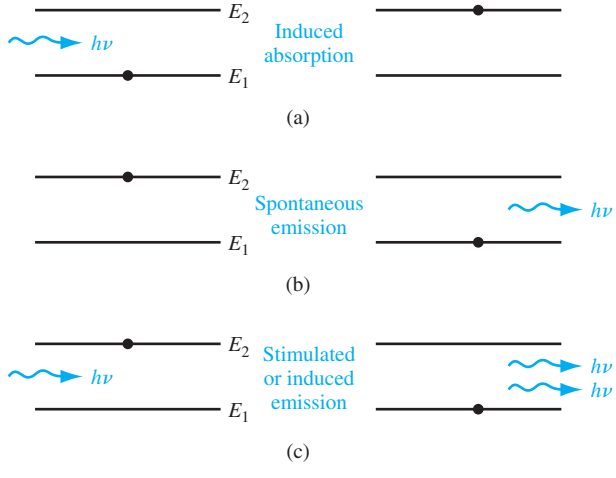


Figure 14.31 | Schematic diagram showing (a) induced absorption, (b) spontaneous emission, and (c) stimulated emission processes.

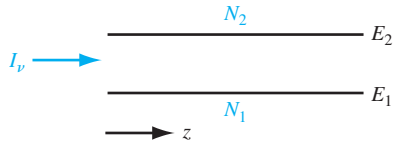


Figure 14.32 | Light propagating in z direction through a material with two energy levels.

Figure 14.32 shows the two energy levels with a light wave at an intensity I_ν propagating in the z direction. The change in intensity as a function of z can be written as

$$\frac{dI_\nu}{dz} \propto \frac{\#\text{ photons emitted}}{\text{cm}^3} - \frac{\#\text{ photons absorbed}}{\text{cm}^3}$$

or

$$\frac{dI_\nu}{dz} = N_2 W_i \cdot h\nu - N_1 W_i \cdot h\nu \tag{14.63}$$

where W_i is the induced transition probability. Equation (14.63) assumes no loss mechanisms and neglects the spontaneous transitions.

Equation (14.63) can be written as

$$\frac{dI_\nu}{dz} = \gamma(\nu) I_\nu \tag{14.64}$$

where $\gamma(\nu) \propto (N_2 - N_1)$ and is the amplification factor. From Equation (14.64), the intensity is

$$I_\nu = I_\nu(0) e^{\gamma(\nu) z} \tag{14.65}$$

Amplification occurs when $\gamma(\nu) > 0$ and absorption occurs when $\gamma(\nu) < 0$.

We can achieve population inversion and lasing in a forward-biased pn homojunction diode, if both sides of the junction are degenerately doped. Figure 14.33a shows the energy-band diagram of a degenerately doped pn junction in thermal equilibrium. The Fermi level is in the conduction band in the n-region and the Fermi level is in the valence band in the p region. Figure 14.33b shows the energy bands of the pn junction when a forward bias is applied. The gain factor in a pn homojunction diode is given by

$$\gamma(\nu) \propto \left\{1 - \exp\left[\frac{h\nu - (E_{Fn} - E_{Fp})}{kT}\right]\right\} \tag{14.66}$$

In order for $\gamma(\nu) > 1$, we must have $h\nu < (E_{Fn} - E_{Fp})$, which implies that the junction must be degenerately doped since we also have the requirement that $h\nu \geq E_g$. In the vicinity of the junction, there is a region in which population inversion occurs. There are large numbers of electrons in the conduction band directly above a large number of empty states. If band-to-band recombination occurs, photons will be emitted with energies in the range $E_g < h\nu < (E_{Fn} - E_{Fp})$.

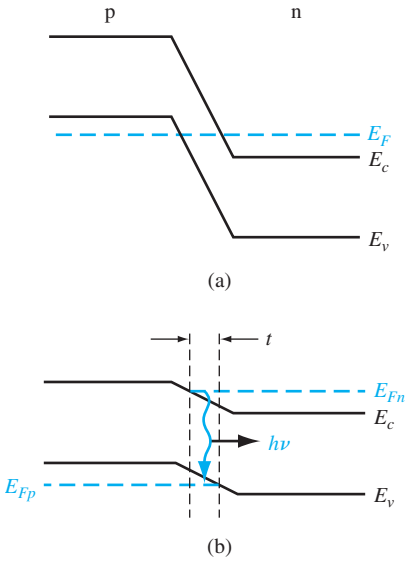


Figure 14.33 | (a) Degenerately doped pn junction at zero bias. (b) Degenerately doped pn junction under forward bias with photon emission.

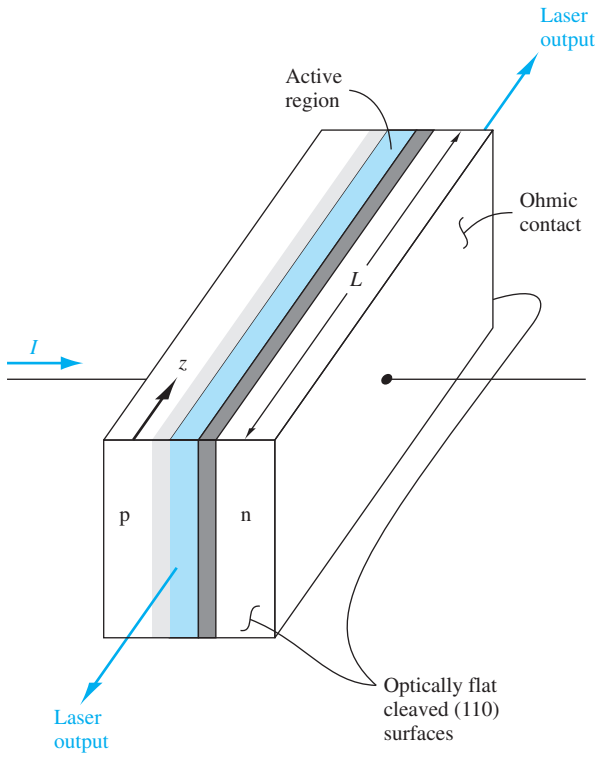


Figure 14.34 | A pn junction laser diode with cleaved (110) planes forming the Fabry-Perot cavity. *(After Yang [22].)*

14.6.2 Optical Cavity

Population inversion is one requirement for lasing action to occur. Coherent emission output is achieved by using an optical cavity. The cavity will cause a buildup of the optical intensity from positive feedback. A resonant cavity consisting of two parallel mirrors is known as a Fabry–Perot resonator. The resonant cavity can be fabricated, for example, by cleaving a gallium arsenide crystal along the (110) planes as shown in Figure 14.34. The optical wave propagates through the junction in the z direction, bouncing back and forth between the end mirrors. The mirrors are actually only partially reflecting so that a portion of the optical wave will be transmitted out of the junction.

For resonance, the length of the cavity L must be an integral number of half wavelengths, or

$$N\left(\frac{\lambda}{2}\right) = L \tag{14.67}$$

where N is an integer. Since λ is small and L is relatively large, there can be many resonant modes in the cavity. Figure 14.35a shows the resonant modes as a function of wavelength.

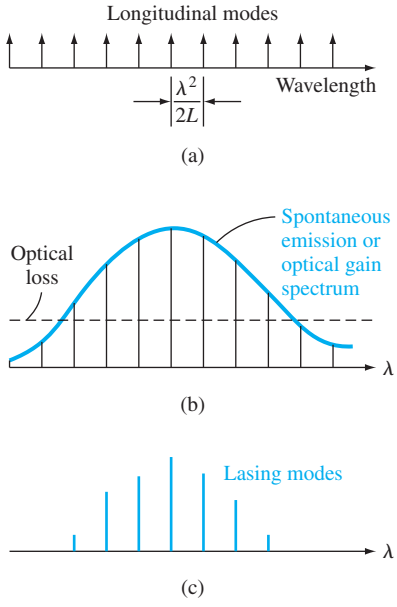


Figure 14.35 | Schematic diagram showing (a) resonant modes of a cavity with length L, (b) spontaneous emission curve, and (c) actual emission modes of a laser diode.
(After Yang [22].)

When a forward-bias current is applied to the pn junction, spontaneous emission will initially occur. The spontaneous emission spectrum is relatively broadband and is superimposed on the possible lasing modes as shown in Figure 14.35b. In order for lasing to be initiated, the spontaneous emission gain must be larger than the optical losses. By positive feedback in the cavity, lasing can occur at several specific wavelengths as indicated in Figure 14.35c.

14.6.3 Threshold Current

The optical intensity in the device can be written from Equation (14.65) as $I_\nu \propto e^{\gamma(\nu)z}$, where $\gamma(\nu)$ is the amplification factor. We have two basic loss mechanisms. The first is the photon absorption in the semiconductor material. We can write

$$I_\nu \propto e^{-\alpha(\nu)z} \tag{14.68}$$

where $\alpha(\nu)$ is the absorption coefficient. The second loss mechanism is due to the partial transmission of the optical signal through the ends, or through the partially reflecting mirrors.

At the onset of lasing, which is known as threshold, the optical loss of one round trip through the cavity is just offset by the optical gain. The threshold condition is then expressed as

$$\Gamma_1\Gamma_2 \exp[(2\gamma_t(\nu) - 2\alpha(\nu))L] = 1 \tag{14.69}$$

where Γ_1 and Γ_2 are the reflectivity coefficients of the two end mirrors. For the case when the optical mirrors are cleaved (110) surfaces of gallium arsenide, the reflectivity coefficients are given approximately by

$$\Gamma_1 = \Gamma_2 = \left(\frac{\bar{n}_2 - \bar{n}_1}{\bar{n}_2 + \bar{n}_1}\right)^2 \tag{14.70}$$

where $\bar{n}_2$ and $\bar{n}_1$ are the index of refraction parameters for the semiconductor and air, respectively. The parameter $\gamma_t(\nu)$ is the optical gain at threshold.

The optical gain at threshold, $\gamma_t(\nu)$, may be determined from Equation (14.69) as

$$\gamma_t(\nu) = \alpha + \frac{1}{2L}\ln\left(\frac{1}{\Gamma_1\Gamma_2}\right) \tag{14.71}$$

Since the optical gain is a function of the pn junction current, we can define a threshold current density as

$$J_{th} = \frac{1}{\beta}\left[\alpha + \frac{1}{2L}\ln\left(\frac{1}{\Gamma_1\Gamma_2}\right)\right] \tag{14.72}$$

where β can be determined theoretically or experimentally. Figure 14.36 shows the threshold current density as a function of the mirror losses. We may note the relatively high threshold current density for a pn junction laser diode.

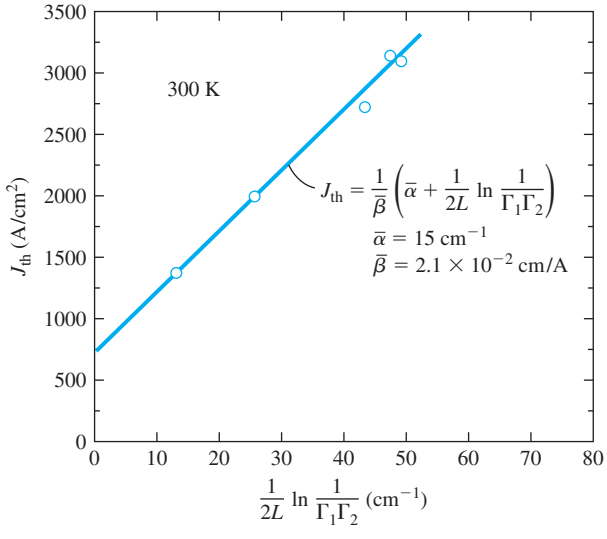


Figure 14.36 | Threshold current density of a laser diode as a function of Fabry-Perot cavity end losses.
(After Yang [22].)

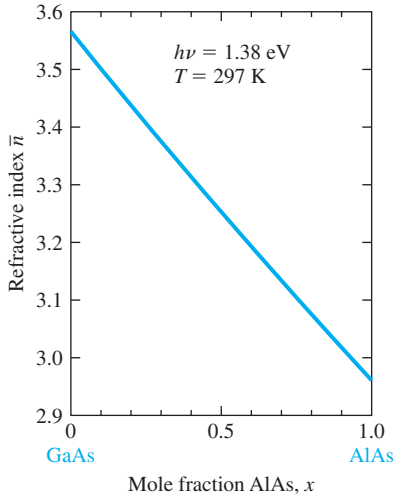


Figure 14.37 | Index of refraction of $Al_xGa_{1-x}As$ as a function of mole fraction x.
(From Sze [18].)

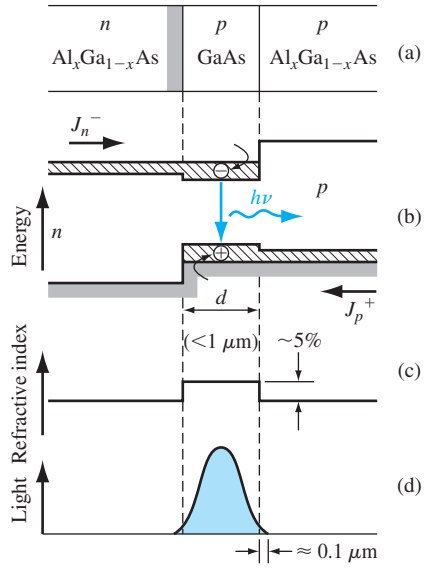


Figure 14.38 | (a) Basic double heterojunction structure. (b) Energy-band diagram under forward bias. (c) Refractive index change through the structure. (d) Confinement of light in the dielectric waveguide.
(From Yang [22].)

14.6.4 Device Structures and Characteristics

We have seen that in a homojunction LED, the photons may be emitted in any direction, which lowers the external quantum efficiency. Significant improvement in device characteristics can be made if the emitted photons are confined to a region near the junction. This confinement can be achieved by using an optical dielectric waveguide. The basic device is a three-layered, double heterojunction structure known as a double heterojunction laser. A requirement for a dielectric waveguide is that the index of refraction of the center material be larger than that of the other two dielectrics. Figure 14.37 shows the index of refraction for the AlGaAs system. We may note that GaAs has the highest index of refraction.

An example of a double heterojunction laser is shown in Figure 14.38a. A thin p-GaAs layer is between P-AlGaAs and N-AlGaAs layers. A simplified energy-band diagram is shown in Figure 14.38b for the forward-biased diode. Electrons are injected from the N-AlGaAs into the p-GaAs. Population inversion is easily obtained since the conduction band potential barrier prevents the electrons from diffusing into the P-AlGaAs region. Radiative recombination is then confined to the p-GaAs region. Since the index of refraction of GaAs is larger than that of AlGaAs, the light wave is also confined to the GaAs region. An optical cavity can be formed by cleaving the semiconductor perpendicular to the N-AlGaAs–p-GaAs junction.

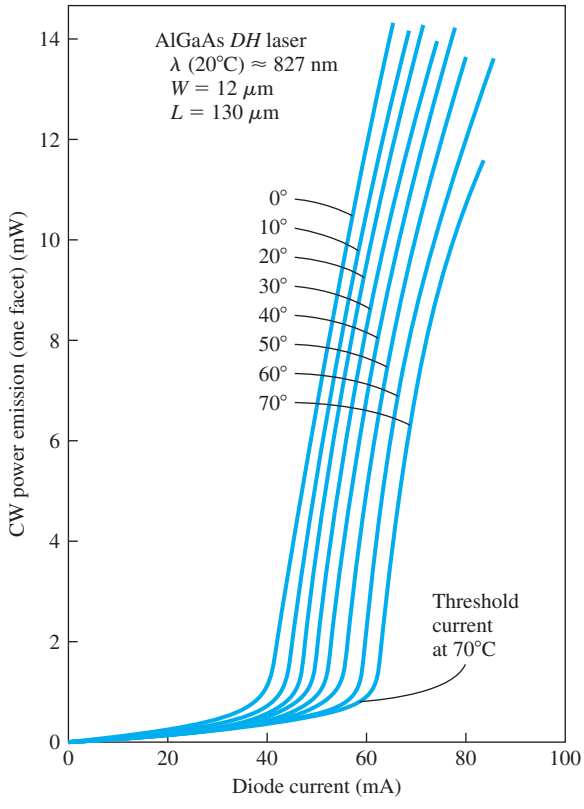


Figure 14.39 | Typical output power versus laser diode current at various temperatures.
(From Yang [22].)

Typical optical output versus diode current characteristics are shown in Figure 14.39. The threshold current is defined to be the current at the breakpoint. At low currents, the output spectrum is very wide and is the result of the spontaneous transitions. When the diode current is slightly above the threshold value, the various resonant frequencies are observed. When the diode current becomes large, a single dominant mode with a narrow bandwidth is produced.

The performance of the laser diode can be further improved if a very narrow recombination region is used with a somewhat wider optical waveguide. Very complex structures using multilayers of compound semiconductor materials have been fabricated in a continuing effort to improve semiconductor laser performance.

14.7 | SUMMARY

- The absorption or emission of light (photons) in semiconductors leads to the study of a general class of devices called optoelectronics. A few of these devices have been discussed and analyzed in this chapter.

- The photon absorption process has been discussed and the absorption coefficient data for semiconductors has been presented.
- Solar cells convert optical power into electrical power. The simple pn junction solar cell was initially considered. The short-circuit current, open-circuit voltage, and maximum power were considered.
- Heterojunction and amorphous silicon solar cells were also considered. Heterojunction cells can be fabricated that tend to increase the conversion efficiency and produce relatively large open-circuit voltages. Amorphous silicon offers the possibility of low-cost, large-area solar cell arrays.
- Photodetectors are semiconductor devices that convert optical signals into electrical signals. The photoconductor is perhaps the simplest type of photodetector. The change in conductivity of the semiconductor due to the creation of excess electrons and holes by the incident photons is the basis of this device.
- Photodiodes are diodes that have reverse-biased voltages applied. Excess carriers that are created by incident photons in the space-charge region are swept out by the electric field creating a photocurrent. The photocurrent is directly proportional to the incident photon intensity. PIN and avalanche photodiodes are variations of the basic photodiode.
- The photocurrent generated in a phototransistor is multiplied by the transistor gain. However, the time response of the phototransistor may be slower than that of a photodiode because of the Miller effect and Miller capacitance.
- The inverse mechanism of photon absorption in a pn junction is injection electroluminescence. The recombination of excess electrons and holes in a direct bandgap semiconductor can result in the emission of photons.
- The light emitting diodes (LEDs) are the class of pn junction diodes whose photon output is a result of spontaneous recombinations of excess electrons and holes. A fairly wide bandwidth in the output signal, on the order of 30 nm, is a result of the spontaneous process.
- The output of a laser diode is the result of stimulated emission. An optical cavity, or Fabry–Perot resonator, is used in conjunction with a diode so that the photon output is in phase, or coherent. Multilayered heterojunction structures can be fabricated to improve the laser diode characteristics.

GLOSSARY OF IMPORTANT TERMS

absorption coefficient The relative number of photons absorbed per unit distance in a semiconductor and denoted by the parameter α.

conversion efficiency The ratio of output electrical power to incident optical power in a solar cell.

delayed photocurrent The component of photocurrent in a semiconductor device due to diffusion currents.

external quantum efficiency The ratio of emitted photons to generated photons in a semiconductor device.

fill factor The ratio $I_m V_m$ to $I_{sc} V_{oc}$, which is a measure of the realizable power from a solar cell. The parameters I_m and V_m are the current and voltage at the maximum power point, respectively, and I_{sc} and V_{oc} are the short-circuit current and open-circuit voltage.

fresnel loss The ratio of reflected to incident photons at an interface due to a change in the index of refraction.

internal quantum efficiency The fraction of diode current that produces luminescence.

LASER diode An acronym for **L**ight **A**mplification by **S**timulated **E**mission of **R**adiation; the stimulated emission of photons produced in a forward-biased pn junction in conjunction with an optical cavity.

LED An acronym for **L**ight **E**mitting **D**iode; the spontaneous photon emission due to electron–hole recombination in a forward-biased pn junction.

luminescence The general property of light emission.

open-circuit voltage The voltage generated across the open-circuited terminals of a solar cell.

photocurrent The current generated in a semiconductor device due to the flow of excess carriers generated by the absorption of photons.

population inversion The condition whereby the concentration of electrons in one energy state is greater than that in a lower energy state; a nonequilibrium condition.

prompt photocurrent The component of photocurrent generated within the space charge region of a semiconductor device.

radiative recombination The recombination process of electrons and holes that produces a photon, such as the direct band-to-band transition in gallium arsenide.

short-circuit current The current produced in a solar cell when the two terminals are shorted together.

stimulated emission The process whereby an electron is induced by an incident photon to make a transition to a lower energy state, emitting a second photon.

CHECKPOINT

After studying this chapter, the reader should have the ability to:

- Describe the optical absorption process in semiconductors. When is optical absorption essentially zero?
- Describe the basic operation and characteristics of a solar cell, including the short-circuit current and open-circuit voltage.
- Discuss the factors that contribute to the solar cell conversion efficiency.
- Describe the advantages and disadvantages of an amorphous silicon solar cell.
- Describe the characteristics of a photoconductor, including the concept of the photoconductor gain.
- Discuss the operation and characteristics of a simple pn junction photodiode.
- Discuss the advantages of PIN and avalanche photodiodes compared to the simple pn junction photodiode.
- Discuss the operation and characteristics of a phototransistor.
- Describe the operation of an LED.
- Describe the operation of a laser diode.

REVIEW QUESTIONS

1. Sketch the general shape of the optical absorption coefficient in a semiconductor as a function of wavelength. When does the absorption coefficient become zero?
2. Sketch the *I–V* characteristic of a pn junction solar cell. Define short-circuit current and open-circuit voltage.

3. Discuss how a pn junction solar cell becomes forward biased.
4. Write an expression for the steady-state photocurrent in a simple photoconductor.
5. What is the source of prompt photocurrent in a photodiode? Does the prompt photocurrent depend on the reverse-biased voltage? Why or why not.
6. Sketch the cross section of a phototransistor and show the currents that are created by incident photons. Explain how current gain is achieved.
7. Explain the basic operation of an LED. State two factors that affect the efficiency of the device.
8. How can different colors be obtained in an LED?
9. Discuss the difference between an LED and a laser diode.
10. Discuss the concept of population inversion in a laser diode.

PROBLEMS

Section 14.1 Optical Absorption

14.1 Determine the maximum wavelength λ of a light source that can generate electron–hole pairs in (*a*) Si, (*b*) Ge, (*c*) GaAs, and (*d*) InP.

14.2 (*a*) Two sources generate light at wavelengths of $\lambda = 480$ nm and $\lambda = 725$ nm, respectively. What are the corresponding photon energies? (*b*) Three sources generate light with photon energies of $E = 0.87$ eV, $E = 1.32$ eV, and $E = 1.90$ eV, respectively. What are the corresponding wavelengths?

14.3 (*a*) A sample of GaAs is 1.2 μm thick. The sample is illuminated with a light source that generates photons with energies of $h\nu = 1.65$ eV. Determine the (*i*) absorption coefficient and (*ii*) fraction of energy that is absorbed in the material. (*b*) Repeat part (*a*) for a sample of GaAs that is 0.80 μm thick and is illuminated with photons with energies of $h\nu = 1.90$ eV.

14.4 A light source with $h\nu = 1.3$ eV and at a power density of 10^{-2} W/cm^2 is incident on a thin slab of silicon. The excess minority carrier lifetime is 10^{-6} s. Determine the electron–hole generation rate and the steady-state excess carrier concentration. Neglect surface effects.

14.5 An n-type GaAs sample has a minority carrier lifetime of $\tau_p = 2 \times 10^{-7}$ s. Incident photons with energies $h\nu = 1.65$ eV generate an excess carrier concentration of $\delta p = 5 \times 10^{15}$ cm^{-3} at the surface of the semiconductor. (*a*) Determine the incident power required. (*b*) At what distance in the semiconductor does the generation rate drop to 10 percent of that at the surface?

14.6 Consider a silicon semiconductor that is illuminated with photons with energies $h\nu = 1.40$ eV. (*a*) Determine the thickness of the material such that 90 percent of the energy is absorbed. (*b*) Determine the thickness of the material such that 30 percent of the energy is transmitted through the material.

14.7 If the thickness of a GaAs semiconductor is 1 μm and 50 percent of the incident monochromic photon energy is absorbed, determine the incident photon energy and wavelength.

***14.8** Consider monochromatic light at an intensity $I_{\nu 0}$ incident on the surface at $x = 0$ of an n-type semiconductor that extends to $x = \infty$. Assume the electric field is zero in the semiconductor and assume a surface recombination velocity, s. Taking into

*Asterisks next to problems indicate problems that are more difficult.

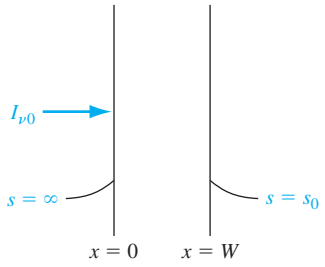


Figure P14.9 | Figure for Problem 14.9.

account the absorption coefficient, determine the steady-state excess hole concentration as a function of x.

***14.9** Monochromatic light with intensity I_{v0} is incident on a p-type semiconductor as shown in Figure P14.9. Assume the surface recombination velocity at $x = 0$ is $s = \infty$ and assume the surface recombination velocity at $x = W$ is $s = s_0$. Derive the expression for the steady-state excess electron concentration as a function of x.

Section 14.2 Solar Cells

14.10 A long silicon pn junction solar cell at $T = 300$ K has the following parameters: $N_a = 10^{16}$ cm^{-3}, $N_d = 10^{15}$ cm^{-3}, $D_n = 25$ cm^2/s, $D_p = 10$ cm^2/s, $\tau_{n0} = 10^{-6}$ s, and $\tau_{p0} = 5 \times 10^{-7}$ s. The cross-sectional area of the solar cell is 5 cm^2. The entire junction is uniformly illuminated such that the generation rate of electron–hole pairs is $G_L = 5 \times 10^{21}$ cm^{-3} s^{-1}. (*a*) Calculate the short circuit photocurrent generated in the space charge region. (*b*) Using the results of part (*a*), calculate the open-circuit voltage. (*c*) Determine the ratio of V_{oc} to V_{bi}.

14.11 A long silicon pn junction solar cell has the same parameters as described in Problem 14.10. The generated photocurrent in the cell is $I_L = 120$ mA. Determine the (*a*) open-circuit voltage, (*b*) the voltage across the junction that will produce a total solar cell current of $I = 100$ mA, (*c*) the maximum power output of the solar cell, and (*d*) the external load resistance that will produce the maximum power.

14.12 Consider the solar cell described in Problem 14.10. (*a*) The generated photocurrent is $I_L = 10$ mA. Determine (*i*) the open-circuit voltage and (*ii*) the maximum power output. (*b*) The solar cell now uses a solar concentrator such that the photocurrent increases by a factor of 10. Determine the new values of (*i*) open circuit voltage and (*ii*) maximum power output. (*c*) Determine the ratio of maximum power from part (*b*) to that from part (*a*).

14.13 Consider an ideal long n^+p junction GaAs solar cell at $T = 300$ K in which excess carriers are uniformly generated. The parameters of the diode are as follows:

$$N_d = 10^{19} \text{ cm}^{-3} \qquad D_n = 225 \text{ cm}^2/\text{s}$$
$$\tau_{n0} = \tau_{p0} = 5 \times 10^{-8} \text{ s} \qquad D_p = 7 \text{ cm}^2/\text{s}.$$

The generated photocurrent density is $J_L = 30$ mA/cm^2. Plot the open-circuit voltage as a function of the acceptor doping concentration for $10^{15} \leq N_a \leq 10^{18}$ cm^{-3}.

14.14 A long silicon pn junction solar cell with an area of 2 cm^2 has the following parameters:

$$N_d = 10^{19}\ cm^{-3} \qquad N_a = 3 \times 10^{16}\ cm^{-3}$$
$$D_p = 6\ cm^2/s \qquad D_n = 18\ cm^2/s$$
$$\tau_{p0} = 5 \times 10^{-7}\ s \qquad \tau_{n0} = 5 \times 10^{-6}\ s$$

Assume that excess carriers are uniformly generated in the solar cell and that $J_L = 25\ mA/cm^2$. Let $T = 300$ K. (*a*) Plot the *I–V* characteristics of the diode, (*b*) determine the maximum power output of the solar cell, and (*c*) calculate the external load resistance that will produce the maximum power.

14.15 A silicon solar cell at $T = 300$ K has a cross-sectional area of 6 cm^2 and a reverse saturation current of $I_S = 2 \times 10^{-9}$ A. The induced short-circuit photocurrent is $I_L = 180$ mA. Determine the (*a*) open-circuit voltage, (*b*) maximum power output, and (*c*) load resistance that will produce the maximum output power. (*d*) If the load resistance determined in part (*c*) is increased by 50 percent, what is the new value of the maximum output power?

14.16 Consider a silicon solar cell at $T = 300$ K with a reverse saturation current of $I_S = 10^{-10}$ A and an induced short-circuit photocurrent of $I_L = 100$ mA. (*a*) Determine V_{oc}. (*b*) Find V_m, I_m, and P_m. (*c*) How many cells, operating at the maximum output power, must be connected in series to produce an output voltage of at least 10 V? (*d*) How many of the 10 V cells in part (*c*) must be connected in parallel to produce an output power of at least 5.2 W? (*e*) Considering the results of part (*d*), what must be the load resistance connected across the solar cell system to produce the maximum output power?

***14.17** Consider the pn junction solar cell with nonuniform absorption. Derive the expression for the excess minority carrier electron concentration for the short-circuit condition and for the case when the p region is very long and the n region is short.

14.18 The absorption coefficient in amorphous silicon is approximately $10^4\ cm^{-1}$ at $h\nu = 1.7$ eV and $10^5\ cm^{-1}$ at $h\nu = 2.0$ eV. Determine the amorphous silicon thickness for each case so that 90 percent of the photons are absorbed.

Section 14.3 Photodetectors

14.19 Consider an n-type silicon photoconductor at $T = 300$ K doped at $N_d = 5 \times 10^{15}\ cm^{-3}$. The cross-sectional area is $A = 5 \times 10^{-4}\ cm^2$ and the length is $L = 120\ \mu m$. The carrier parameters are $\mu_n = 1200\ cm^2/V\text{-}s$, $\mu_p = 400\ cm^2/V\text{-}s$, $\tau_{n0} = 5 \times 10^{-7}$ s, and $\tau_{p0} = 10^{-7}$ s. The photoconductor is uniformly illuminated such that the generation rate of electron–hole pairs is $G_L = 10^{21}\ cm^{-3}\ s^{-1}$. For 3 volts applied to the photoconductor, determine (*a*) the thermal equilibrium current, (*b*) the steady-state excess carrier concentration, (*c*) the photoconductivity, (*d*) the steady-state photocurrent, and (*e*) the photocurrent gain.

14.20 Excess carriers are uniformly generated in a GaAs photoconductor at a rate of $G_L = 10^{21}\ cm^{-3}\text{-}s^{-1}$. The area is $A = 10^{-4}\ cm^2$ and the length is $L = 100\ \mu m$. The other parameters are:

$$N_d = 5 \times 10^{16}\ cm^{-3} \qquad N_a = 0$$
$$\mu_n = 8000\ cm^2/V\text{-}s \qquad \mu_p = 250\ cm^2/V\text{-}s$$
$$\tau_{n0} = 10^{-7}\ s \qquad \tau_{p0} = 10^{-8}\ s.$$

If a voltage of 5 volts is applied, calculate (*a*) the steady-state excess carrier concentration, (*b*) the photoconductivity, (*c*) the steady-state photocurrent, and (*d*) the photoconductor gain.

***14.21** Consider an n-type silicon photoconductor that is 1 μm thick, 50 μm wide, and has an applied electric field in the longitudinal dimension of 50 V/cm. If the incident photon flux is $\Phi_0 = 10^{16}$ cm^{-2}-s^{-1} and the absorption coefficient is $\alpha = 5 \times 10^4$ cm^{-1}, calculate the steady-state photocurrent if $\mu_n = 1200$ cm^2/V-s, $\mu_p = 450$ cm^2/V-s, and $\tau_{p0} = 2 \times 10^{-7}$ s.

14.22 A long silicon pn junction photodiode has the following parameters at $T = 300$ K: $N_a = 10^{16}$ cm^{-3}, $N_d = 2 \times 10^{15}$ cm^{-3}, $D_p = 10$ cm^2/s, $D_n = 25$ cm^2/s, $\tau_{p0} = 10^{-7}$ s, and $\tau_{n0} = 5 \times 10^{-7}$ s. The cross-sectional area of the diode is $A = 10^{-3}$ cm^2. Assume that a reverse-biased voltage of 5 volts is applied and that a uniform generation rate for electron–hole pairs of $G_L = 10^{21}$ cm^{-3} s^{-1} exists throughout the entire photodiode. (*a*) Determine the prompt component of photocurrent. (*b*) Find the steady-state excess carrier concentrations in the p and n regions far from the junction. (*c*) Determine the total steady-state photocurrent.

***14.23** Starting with the ambipolar transport equation for minority carrier holes, derive Equation (14.41) using the geometry shown in Figure 14.17.

14.24 Three silicon PIN photodiodes A, B, and C, at $T = 300$ K have intrinsic region widths of 2, 10, and 80 μm, respectively. A photon flux of $\Phi_0 = 5 \times 10^{17}$ cm^{-2} s^{-1} is incident on the surface of each diode as shown in Figure 14.19. (*a*) For an absorption coefficient of $\alpha = 10^4$ cm^{-1}, calculate the prompt photocurrent density in each diode. (*b*) Repeat part (*a*) for an absorption coefficient of $\alpha = 5 \times 10^2$ cm^{-1}.

14.25 Consider a silicon PIN photodiode at $T = 300$ K with the geometry shown in Figure 14.19. The intrinsic region width is 100 μm. Assume that a reverse-biased voltage is applied such that the intrinsic region is completely depleted. The incident photon power is $I_{v0} = 0.080$ W/cm^2, the absorption coefficient is $\alpha = 10^3$ cm^{-1}, and the photon energy is 1.5 eV. Neglect any absorption in the p$^+$ top layer of the photodiode. (*a*) Determine the steady-state electron–hole generation rate, G_L, versus distance in the intrinsic region. (*b*) Determine the steady-state photocurrent density.

14.26 A silicon PIN photodiode at $T = 300$ K has the geometry shown in Figure 14.19. The intrinsic region width is 20 μm and is fully depleted. (*a*) The electron–hole pair generation rate in the intrinsic region is $G_L = 10^{21}$ cm^{-3} s^{-1} and is uniform throughout the intrinsic region. Calculate the steady-state photocurrent density for this condition. (*b*) The generation rate of electron–hole pairs is $G_L = 10^{21}$ cm^{-3} s^{-1} at $x = 0$ and the absorption coefficient is $\alpha = 10^3$ cm^{-1}. Determine the steady-state photocurrent density for this situation.

14.27 Consider a silicon PIN photodiode exposed to sunlight. Calculate the intrinsic region width so that at least 90 percent of all photons with wavelengths $\lambda \leq 1$ μm are absorbed in the intrinsic region. Neglect any absorption in the p$^+$ or n$^+$ regions.

Section 14.4 Photoluminescence and Electroluminescence

14.28 Consider the $Al_xGa_{1-x}As$ system. Determine the range of the direct bandgap energies possible and the corresponding range of wavelengths.

14.29 Consider the $GaAs_{1-x}P_x$ system. (*a*) For a mole fraction $x = 0.2$, determine the (*i*) bandgap energy and (*ii*) corresponding photon wavelength. (*b*) Repeat part (*a*) for a mole fraction $x = 0.32$.

14.30 Using Figure 14.23, determine the mole fraction x in $Al_xGa_{1-x}As$ such that the material would emit light at a wavelength of $\lambda = 0.670\ \mu m$. What is the corresponding bandgap energy?

14.31 Repeat Problem 14.30 for the $GaAs_{1-x}P_x$ system.

Section 14.5 Light Emitting Diodes

14.32 Consider a pn junction GaAs LED. Assume that photons are generated uniformly in all directions in a plane perpendicular to the junction at a distance of 0.50 μm from the surface. (*a*) Taking into account total internal reflection, calculate the fraction of photons that have the potential of being emitted from the semiconductor. (*b*) Using the results of part (*a*) and including Fresnel loss, determine the fraction of generated photons that will be emitted from the semiconductor into air (neglect absorption losses).

***14.33** In a pn junction LED, consider a point source in the semiconductor at the junction and assume that photons are emitted uniformly in all directions. Show that (neglecting photon absorption) the external quantum efficiency of the LED is given by

$$\eta_{ext} = \frac{2\bar{n}_1\bar{n}_2}{(\bar{n}_1 + \bar{n}_2)^2}(1 - \cos\theta_c)$$

where $\bar{n}_1$ and $\bar{n}_2$ are the index of refraction parameters for the air and semiconductor, respectively, and θ_c is the critical angle.

Section 14.6 Laser Diodes

14.34 Consider an optical cavity. If $N \gg 1$, show that the wavelength separation between two adjacent resonant modes is $\Delta\lambda = \lambda^2/2L$.

14.35 If the photon output of a laser diode is equal to the bandgap energy, find the wavelength separation between adjacent resonant modes in a GaAs laser with $L = 75\ \mu m$.

READING LIST

1. Bhattacharya, P. *Semiconductor Optoelectronic Devices,* 2nd ed. Upper Saddle River, NJ: Prentice Hall, 1997.
2. Carlson, D. E. "Amorphous Silicon Solar Cells." *IEEE Transactions on Electron Devices* ED-24 (April 1977), pp. 449–53.
3. Fonash, S. J. *Solar Cell Device Physics.* New York: Academic Press, 1981.
4. Kano, K. *Semiconductor Devices.* Upper Saddle River, NJ: Prentice Hall, 1998.
5. Kressel, H. *Semiconductor Devices for Optical Communications: Topics in Applied Physics.* Vol. 39. New York: Springer-Verlag, 1987.
6. MacMillan, H. F., H. C. Hamaker, G. F. Virshup, and J. G. Werthen. "Multijunction III-V Solar Cells: Recent and Projected Results." *Twentieth IEEE Photovoltaic Specialists Conference* (1988), pp. 48–54.
7. Madan, A. "Amorphous Silicon: From Promise to Practice." *IEEE Spectrum* 23 (September 1986), pp. 38–43.
8. Pankove, J. I. *Optical Processes in Semiconductors.* New York: Dover Publications, 1971.

9. Pierret, R. F. *Semiconductor Device Fundamentals.* Reading, MA: Addison-Wesley, 1996.

10. Roulston, D. J. *An Introduction to the Physics of Semiconductor Devices.* New York: Oxford University Press, 1999.

11. Schroder, D. K. *Semiconductor Material and Devices Characterization,* 3rd ed. Hoboken, NJ: John Wiley and Sons, 2006.

12. Shur, M. *Introduction to Electronic Devices.* New York: John Wiley and Sons, 1996.

***13.** _____. *Physics of Semiconductor Devices.* Englewood Cliffs, NJ: Prentice Hall, 1990.

14. Singh, J. Optoelectronics: An Introduction to Materials and Devices. New York: McGraw-Hill, 1996.

15. _____. *Semiconductor Devices: Basic Principles.* New York: John Wiley and Sons, 2001.

16. Streetman, B. G., and S. K. Banerjee. *Solid State Electronic Devices,* 6th ed. Upper Saddle River, NJ: Pearson Prentice-Hall, 2006.

17. Sze, S. M. *Semiconductor Devices: Physics and Technology.* New York: John Wiley and Sons, 1985.

18. Sze, S. M. and K. K. Ng. Physics of Semiconductor Devices, 3rd ed. Hoboken, NJ: John Wiley and Sons, 2007.

***19.** Wang, S. *Fundamentals of Semiconductor Theory and Device Physics.* Englewood Cliffs, NJ: Prentice Hall, 1989.

20. Wilson, J., and J. F. B. Hawkes. *Optoelectronics: An Introduction.* Englewood Cliffs, NJ: Prentice Hall, 1983.

***21.** Wolfe, C. M, N. Holonyak, Jr., and G. E. Stillman. *Physical Properties of Semiconductors.* Englewood Cliffs, NJ: Prentice Hall, 1989.

22. Yang, E. S. *Microelectronic Devices.* New York: McGraw-Hill, 1988.

*Indicates reference that is at an advanced level compared to this text.

CHAPTER 15

Semiconductor Microwave and Power Devices

In previous chapters, we have discussed the basic physics, operation, and characteristics of diodes and transistors. We have analyzed the frequency response as well as the current–voltage characteristics of these semiconductor devices. However, we have not specifically considered the generation of microwave signals using semiconductor devices or the power capabilities of semiconductor transistors.

In this chapter, we first consider three semiconductor devices that are used to generate microwave signals. These devices include the tunnel diode, GUNN diode, and IMPATT diode. A basic principle of oscillators is that a region of negative differential resistance must exist. We consider the process by which a region of negative differential resistance is created in each device and discuss the basic operation of these devices.

Second, we discuss three specialized semiconductor power devices, including power bipolar transistors and power MOSFETs. We have considered the basic physics of these devices in previous chapters, and analyzed the current–voltage characteristics without specifically considering the current or voltage limitations or the power dissipation within the devices. In this chapter, we discuss the limitations in current and voltage, and the power capabilities of the devices. Finally, we discuss the operation and characteristics of a four-layered structure called a thyristor. ■

15.0 | PREVIEW

In this chapter, we will:

- Discuss the concept of negative differential resistance in a tunnel diode and derive an expression for the maximum resistance cutoff frequency.
- Discuss the concept of negative differential mobility in GaAs and discuss the process by which this characteristic can lead to microwave oscillations in a GUNN diode.
- Discuss the operation of an IMPATT diode oscillator and determine the process by which a dynamic negative resistance is created.

- Present the basic geometry and electrical characteristics of a power bipolar transistor. The limiting current and voltage factors will be analyzed, and the safe operating area of the BJT will be considered.
- Present the basic geometry and electrical characteristics of a power MOSFET. The limiting current and voltage factors will be analyzed, and the safe operating area of the MOSFET will be considered.
- Discuss the operation of a four-layer switching device that is generally referred to as a Thyristor. The operation of several structures will be analyzed.

15.1 | TUNNEL DIODE

The tunnel diode, also known as the Esaki diode, has been briefly discussed in Section 8.5 of the book. Recall that the device is a pn junction in which both the n and p regions are degenerately doped. With the very high doping concentrations, the space charge region width is very narrow ($W \approx 0.5 \times 10^{-6}$ cm $= 50$ Å).

The forward-bias current–voltage characteristics are again shown in Figure 15.1a. For small forward-bias voltages ($V < V_p$), electrons in the conduction band on the n side are directly opposite empty states in the valence band of the p region (see Figure 8.29). Electrons tunnel through the potential barrier into the empty states producing a tunneling current. For forward-bias voltages in the range $V_p < V < V_v$, the number of electrons on the n side directly opposite empty states on the p side decreases so that the tunneling current decreases. For $V > V_v$, the normal diode diffusion currents dominate.

A decrease in current with an increase in voltage produces a region of negative differential resistance in the range $V_p < V < V_v$. A negative differential resistance phenomenon is necessary for oscillators.

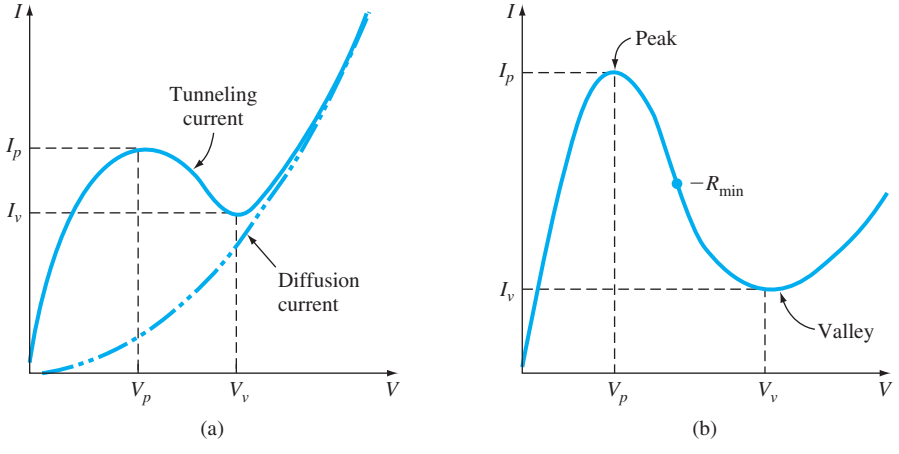


Figure 15.1 | (a) Forward-bias current–voltage characteristics of a tunnel diode. (b) Expanded plot of *I–V* characteristics.

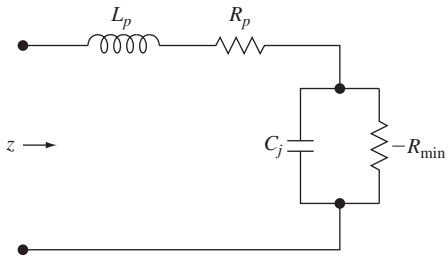


Figure 15.2 | Equivalent circuit of the tunnel diode.

Figure 15.1b shows an expanded plot of the *I–V* characteristics in the tunneling range. A point is shown on the curve where the minimum value of negative resistance occurs. (Note that $R_{\min}$ is a positive quantity.) The equivalent circuit of the tunnel diode for the case when the diode is biased at the $-R_{\min}$ point is shown in Figure 15.2. The parameter C_j is the junction capacitance, and the parameters L_p and R_p are the parasitic or interconnect line inductance and resistance, respectively.

The small signal input impedance can be written as

$$Z = \left[R_p - \frac{R_{\min}}{1 + \omega^2 R_{\min}^2 C_j^2}\right] + j\omega\left[L_p - \frac{\omega R_{\min}^2 C_j}{1 + \omega^2 R_{\min}^2 C_j^2}\right] \tag{15.1}$$

The resistive part of the impedance goes to zero at a frequency of

$$f_r = \frac{1}{2\pi R_{\min} C_j}\sqrt{\frac{R_{\min}}{R_p} - 1} \tag{15.2}$$

For frequencies $f > f_r$, the resistive part of the impedance becomes positive so that the diode loses its negative differential resistance characteristic. The operating frequency must then occur at $f_o < f_r$. The frequency f_r is referred to as the *maximum resistive cutoff frequency*.

The tunneling process is a majority carrier effect so the diode does not exhibit time delays due to minority carrier diffusion, which means that the diode is capable of operating at microwave frequencies. However, due to the relatively small voltage range in which the diode exhibits the negative resistance characteristic, the tunnel diode is not used extensively.

15.2 | GUNN DIODE

Another negative differential resistance device is the GUNN diode, or **T**ransferred-**E**lectron **D**evice (TED). The transferred-electron phenomenon is demonstrated in a few semiconductors in which conduction electrons in a high-mobility band are scattered to a low-mobility band by a high electric field. In Chapter 5, we discussed the drift velocity of electrons in GaAs versus electric field. Figure 15.3 again shows a plot of this characteristic. InP also shows this same characteristic.

Figure 15.4 shows an expanded plot of the energy-band structure in GaAs that is given in Figure 5.8. For small electric fields, essentially all of the electrons in the

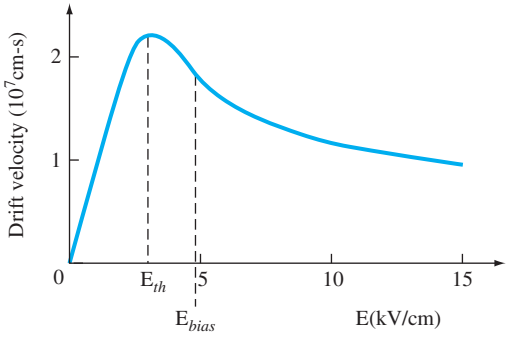


Figure 15.3 | Electron drift velocity versus electric field for GaAs.

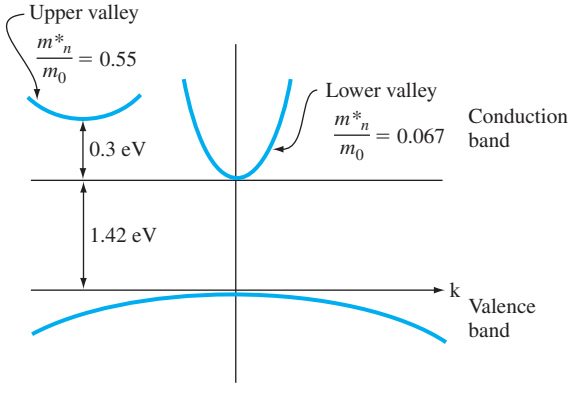


Figure 15.4 | Energy-band structure of GaAs showing the lower valley and upper valley in the conduction band.

conduction band exist in the lower valley of the E versus k diagram, where the density of states electron effective mass is small. A small effective mass leads to a large mobility value.

As the electric field increases above a threshold or critical value, E_{th}, the electrons gain more than the 0.3 eV energy separating the two valleys so that electrons can be scattered into the upper valley, where the density of states electron effective mass is much larger. The larger effective mass yields a smaller mobility. The intervalley transfer mechanism with a change in mobility results in a decreasing average drift velocity of electrons with electric field, or a negative differential electron mobility. The maximum negative differential electron mobility in GaAs is approximately -2400 cm^2/V-s.

Consider a two-terminal n-type GaAs device with ohmic contacts at the ends that is biased in the negative mobility region ($E_{bias} > E_{th}$) as shown in Figure 15.5a. A small space charge region may develop in the material near the cathode as shown in Figure 15.5b. As a result, the electric field increases in this region as shown in Figure 15.5c. (Special device structures can be fabricated to ensure that the space charge fluctuations are generated near the cathode.)

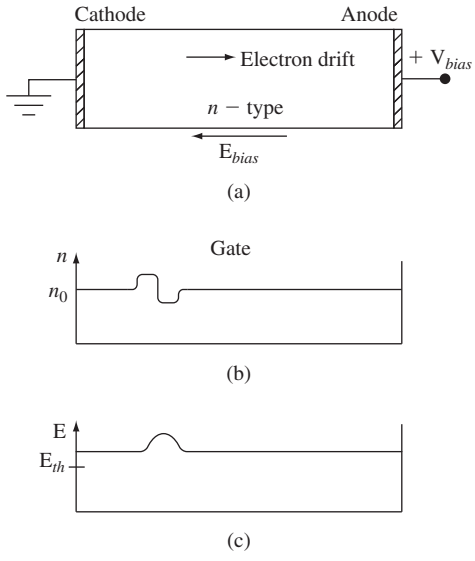


Figure 15.5 | (a) A simplified two-terminal GaAs device. (b) Electron concentration versus distance showing a space charge formation. (c) Electric field versus distance.

In discussing excess carrier behavior in Chapter 6, we found the time behavior of a net charge density in a semiconductor to be given by

$$\delta Q(t) = \delta Q(0)e^{-t/\tau_d} \tag{15.3}$$

where τ_d is the dielectric relaxation time constant and is on the order of a picosecond. Normally, a small space charge region would be quickly neutralized. The dielectric relaxation time constant is given by $\tau_d = \epsilon/\sigma$, where σ is the semiconductor conductivity. If the GaAs is biased in the negative mobility region, then the conductivity is negative and the exponent in Equation (15.3) becomes positive, so the space charge region, now called a *domain,* can actually build up as it drifts toward the anode. As the domain grows (Figure 15.6a), the electric field in this region increases which means that the electric field in the remaining material decreases. The E field in the material outside of the domain can drop below the critical value, as indicated in Figure 15.6b, while the E field within the domain remains above the critical value. For this reason, only one domain will normally be established in the material at any given time.

As the domain reaches the anode, a current pulse is induced in the external circuit. After the domain reaches the anode, another domain may form near the cathode and the process repeats itself. Thus, a series of current pulses may be generated as shown in Figure 15.7. The time between current pulses is the time for the domain to drift through the device. The oscillation frequency is given by

$$f = 1/\tau = v_d/L \tag{15.4}$$

where v_d is the average drift velocity and L is the length of the drift region.

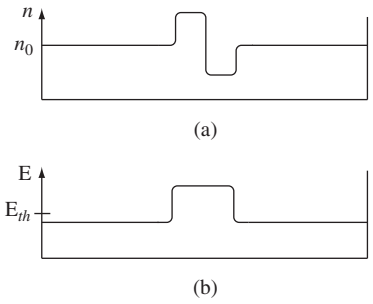


Figure 15.6 | (a) Electron concentration versus distance showing a domain. (b) Electric field versus distance.

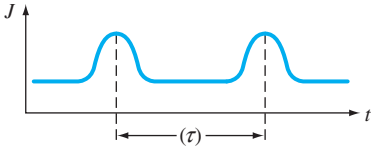


Figure 15.7 | Current pulses versus time in the GaAs device.

The oscillation mechanism just described is called the transit-time mode. More complex modes of operation are possible. Studies have shown that the efficiency of the transit-time device is largest when the product $n_0 L$ is a few times 10^{12} cm^{-2}. For this case, the domain fills about one-half of the drift region length and produces a current output that is nearly sinusoidal. The maximum dc-to-rf conversion efficiency is approximately 10 percent.

Oscillations in the frequency range of 1 to 100 GHz or higher can be obtained. If the device is operated in a pulsed mode, a peak output power in the range of hundreds of watts can be produced. Transferred-electron devices are now used as the microwave source in many radar systems.

15.3 | IMPATT DIODE

The term IMPATT stands for **IMP**act ionization **A**valanche **T**ransit-**T**ime. The IMPATT diode consists of a high-field avalanche region and a drift region that produces a dynamic negative resistance at microwave frequencies. The negative resistance characteristic produced in this device is a result of a time delay so that the ac current and voltage components are out of phase, and is a different phenomenon compared to the tunnel diode, for example. The tunnel diode has a negative dI/dV region in the I–V characteristic.

One example of an IMPATT diode is a p^+-n-i-n^+ structure as shown in Figure 15.8a. Typical doping concentrations (magnitudes) are shown in Figure 15.8b. The device is reverse biased so that the n and intrinsic regions are completely depleted. The electric field in the device is shown in Figure 15.8c. We may note that $\int E dx = V_B$ where V_B is the applied reverse-biased voltage. The value of V_B is very close to the breakdown voltage. The avalanche region is localized near the pn junction. The electric field in the intrinsic region is nearly constant and the intrinsic layer provides the drift region.

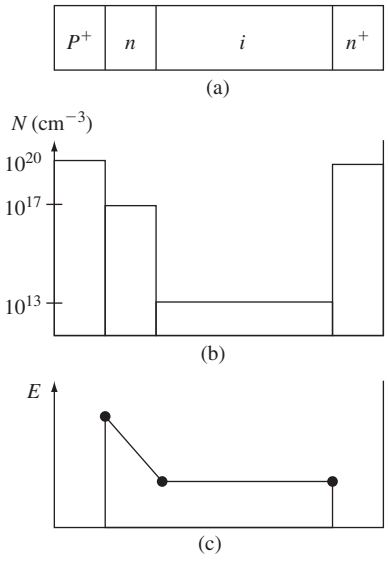


Figure 15.8 | (a) An IMPATT diode structure. (b) Typical doping concentrations in the IMPATT diode. (c) Electric field versus distance through the IMPATT diode.

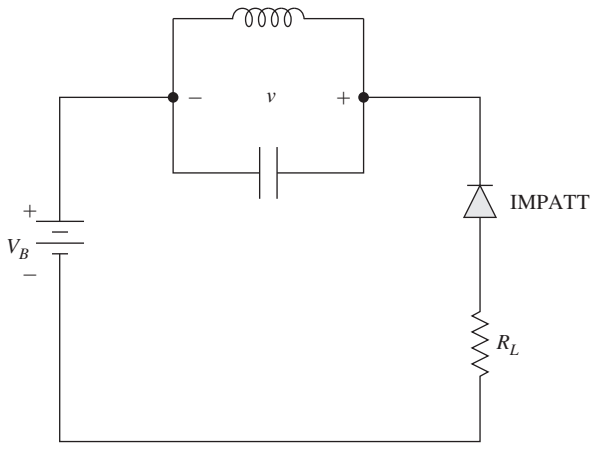


Figure 15.9 | Circuit for an IMPATT diode oscillator.

Figure 15.9 shows the circuit for an IMPATT diode oscillator. An LC resonant circuit is required for the oscillator operation. During the positive ac voltage across the LC circuit as shown in the figure, the diode goes into breakdown and electron–hole pairs are generated at the p^+n junction. The generated electrons flow back into the p^+ region, while the holes start drifting through the depleted intrinsic region. In general, the holes will travel at their saturation velocity. During the negative ac voltage, the device operates below the breakdown voltage so electron–hole pairs are no longer produced.

There is an inherent $\pi/2$ phase shift between the peak value of the avalanche voltage at the p^+n junction and the injection of the holes into the intrinsic drift region due to the finite buildup time of the avalanche generated electron–hole pairs. A further delay of $\pi/2$ is then required during the drift process to provide the total 180 degrees of phase shift between the current and voltage at the output terminal. The transit time of the holes is $\tau = L/v_s$, where L is the length of the drift region and v_s is the saturation velocity of the holes. The LC circuit resonant frequency must be designed to be equal to the device resonant frequency, which is given by

$$f = \frac{1}{2\tau} = \frac{v_s}{2L} \tag{15.5}$$

When the holes reach the n^+ cathode, the current is at a maximum value and the voltage is at its minimum value. The ac current and ac voltage are 180 degrees out of phase with respect to each other producing the dynamic negative resistance.

Devices can be designed to operate in the 100 GHz or higher frequency range and produce power outputs of a few watts. The efficiency of these devices is in the range of 10 to 15 percent, and these devices provide the highest continuous output power of all the semiconductor microwave devices. As with most semiconductor device designs, other structures can be fabricated to provide specialized output characteristics.

15.4 | POWER BIPOLAR TRANSISTORS

In our previous discussions, we have ignored any physical transistor limitations in terms of maximum current, voltage, and power. We implicitly assumed that the transistors are capable of handling the current and voltage, and could handle the power dissipated within the device without suffering any damage.

However, with power transistors, we must be concerned with various transistor limitations. The limitations involve maximum rated current (on the order of amperes), maximum rated voltage (on the order of 100 V), and maximum rated power (on the order of watts or tens of watts).[1]

15.4.1 Vertical Power Transistor Structure

Figure 15.10 shows the structure of a vertical npn power transistor. We have considered vertical npn bipolar transistors previously. However, with small switching devices, the collector terminal is still formed at the surface. In the vertical configuration for the power bipolar transistor, the collector terminal is at the "bottom" of the device. This configuration is preferred since it maximizes the cross-sectional area through which current is flowing in the device. In addition, the doping concentrations

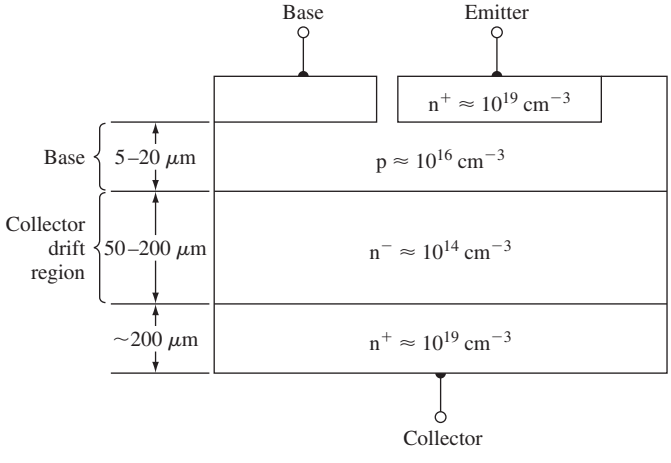


Figure 15.10 | Cross section of typical vertical npn power BJT.

[1]We must note that, in general, the maximum rated current and maximum rated voltage cannot occur at the same time.

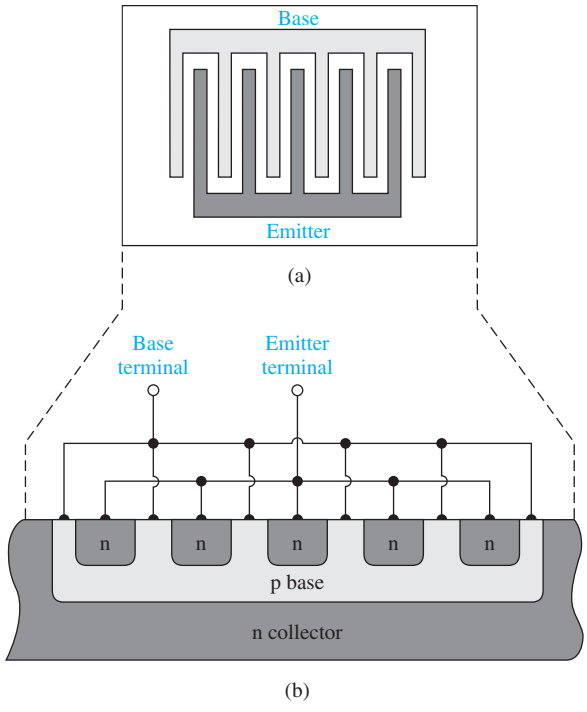


Figure 15.11 | An interdigitated bipolar transistor structure showing the top view and cross-sectional view.

and dimensions are not the same as we have encountered in small switching transistors. The primary collector region has a low-doped impurity concentration so that a large base–collector voltage can be applied without initiating breakdown. Another n region, with a higher doping concentration, reduces collector resistance and makes contact with the external collector terminal. The base region is also much wider than normally encountered in small devices. A large base–collector voltage implies a relatively large space charge width being induced in both the collector and base regions. A relatively large base width is required to prevent punch-through breakdown.

Power transistors must also be large-area devices in order to handle large currents. We have previously considered the interdigitated structure that is repeated in Figure 15.11. Relatively small emitter widths are required to prevent the emitter current crowding effects that were discussed in Section 12.4.4.

15.4.2 Power Transistor Characteristics

The relatively wide base width implies a much smaller current gain β for power transistors compared to small switching transistors, and large area device implies a larger junction capacitance and hence lower cutoff frequency for a power transistor compared to a small switching transistor. Table 15.1 compares the parameters of a

Table 15.1 | Comparison of the characteristics and maximum ratings of small-signal and power BJTs

Parameter	Small-signal BJT (2N2222A)	Power BJT (2N3055)	Power BJT (2N6078)
V_{CE} (max) (V)	40	60	250
T_C (max) (A)	0.8	15	7
P_D (max) (W) (at $T = 25°C$)	1.2	115	45
β	35–100	5–20	12–70
f_T (MHz)	300	0.8	1

general-purpose small-signal BJT to those of two power BJTs. The current gain is generally smaller in the power transistors, typically in the range of 20 to 100, and may be a strong function of collector current and temperature. Figure 15.12 shows typical current gain versus collector current characteristics for the 2N3055 power BJT at various temperatures.

The *maximum rated collector current* $I_{C,\max}$ may be related to the maximum current that the wires connecting the semiconductor to the external terminals can handle, the collector current at which the current gain falls below a minimum specified value, or the current that leads to the maximum power dissipation when the transistor is biased in saturation.

The *maximum rated voltage* in a BJT is generally associated with avalanche breakdown in the reverse-biased base–collector junction. In the common-emitter configuration, the breakdown voltage mechanism also involves the transistor gain, as well as the breakdown phenomenon in the pn junction. This is discussed in Section 12.4.6. Typical I_C versus V_{CE} characteristics are shown in Figure 15.13.

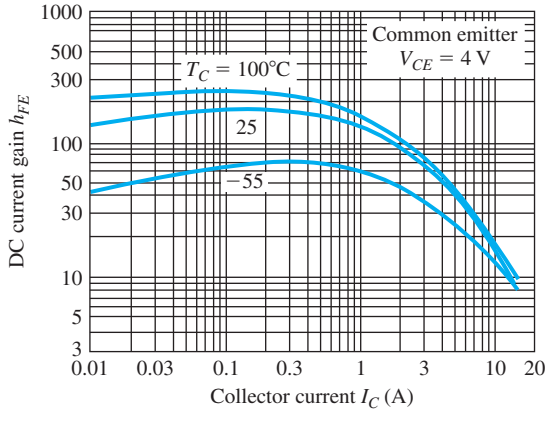


Figure 15.12 | Typical dc beta characteristics (h_{FE} versus I_C) for 2N3055.

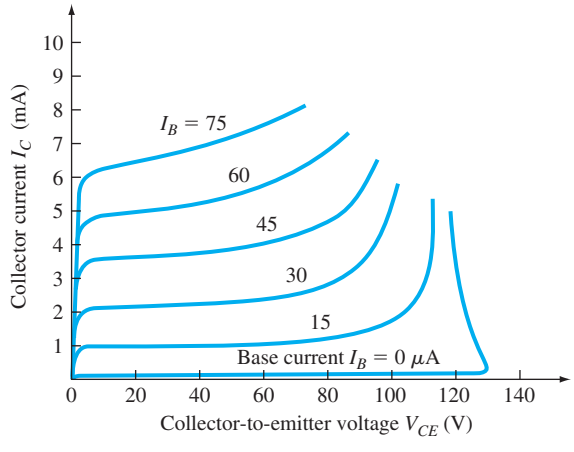


Figure 15.13 | Typical collector current versus collector–emitter voltage characteristics of a bipolar transistor, showing breakdown effects.

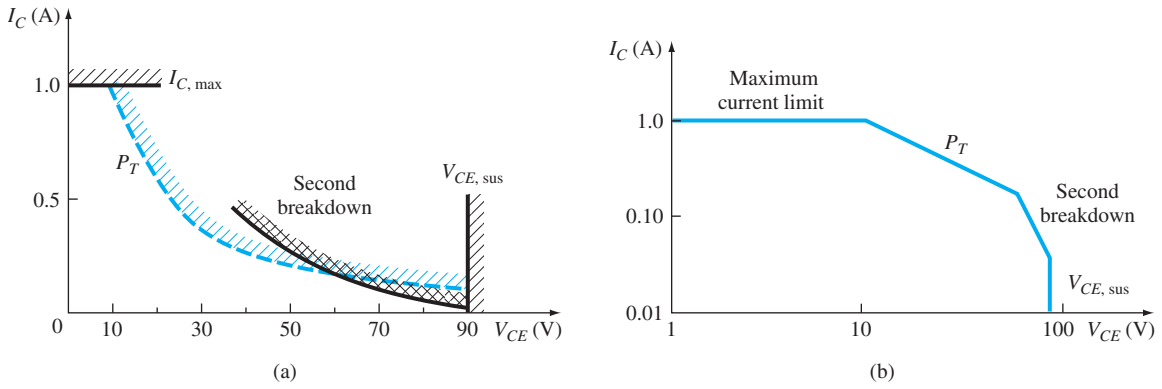


Figure 15.14 | The safe operating area (SOA) of a bipolar transistor plotted on (a) linear scales and (b) logarithmic scales.

When the transistor is biased in the forward-active mode, the collector current begins to increase significantly before the actual breakdown voltage is reached. All the curves tend to merge to the same collector–emitter voltage once breakdown has occurred. This voltage, $V_{CE,\,sus}$, is the minimum voltage necessary to sustain the transistor in breakdown.

Another breakdown effect is called *second breakdown,* which occurs in a BJT operating at high voltage and high current. Slight nonuniformities in current density produce local regions of increased heating that increases the minority carrier concentrations in the semiconductor material, which in turn increases the current in these regions. This effect results in positive feedback, and the current continues to increase, producing a further increase in temperature, until the semiconductor material may actually melt, creating a short circuit between the collector and emitter.

The average power dissipated in a BJT must be kept below a specified maximum value, to ensure that the temperature of the device remains below a maximum value. If we assume the collector current and collector–emitter voltage are dc values, then at the *maximum rated power* P_T for the transistor, we can write

$$P_T = V_{CE} I_C \tag{15.6}$$

Equation (15.6) neglects the $V_{BE} I_B$ component of power dissipation in the transistor.

The maximum current, voltage, and power limitations can be illustrated on the I_C versus V_{CE} characteristics as shown in Figure 15.14. The average power limitation, P_T, is a hyperbola described by Equation (15.6). The region where the transistor can be operated safely is known as the safe operating area (SOA) and is bounded by $I_{C,\max}$, $V_{CE,\,sus}$, P_T, and the transistor's second breakdown characteristic curve. Figure 15.14a shows the safe operating area using linear current and voltage scales. Figure 15.14b shows the same characteristics using log scales.

EXAMPLE 15.1

Objective: Determine the required current, voltage, and power rating of a power BJT.

Consider the common-emitter circuit in Figure 15.15. The parameters are $R_L = 10\ \Omega$ and $V_{CC} = 35$ V.

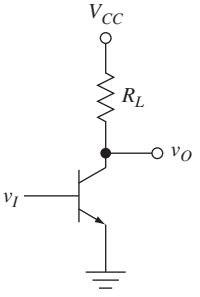


Figure 15.15 | Bipolar common-emitter circuit.

■ Solution

For $V_{CE} \approx 0$, the maximum collector current is

$$I_C\,(\text{max}) = \frac{V_{CC}}{R_L} = \frac{35}{10} = 3.5\ \text{A}$$

For $I_C = 0$, the maximum collector–emitter voltage is

$$V_{CE}\,(\text{max}) = V_{CC} = 35\ \text{V}$$

The load line is given by

$$V_{CE} = V_{CC} - I_C R_L$$

and must remain within the SOA, as shown in Figure 15.16.

The transistor power dissipation is

$$P_T = V_{CE} I_C = (V_{CC} - I_C R_L) I_C = V_{CC} I_C - I_C^2 R_L$$

The current at which the maximum power occurs is found by setting the derivative of this equation equal to zero as follows:

$$\frac{dP_T}{dI_C} = 0 = V_{CC} - 2 I_C R_L$$

which yields

$$I_C = \frac{V_{CC}}{2R_L} = \frac{35}{2(10)} = 1.75\ \text{A}$$

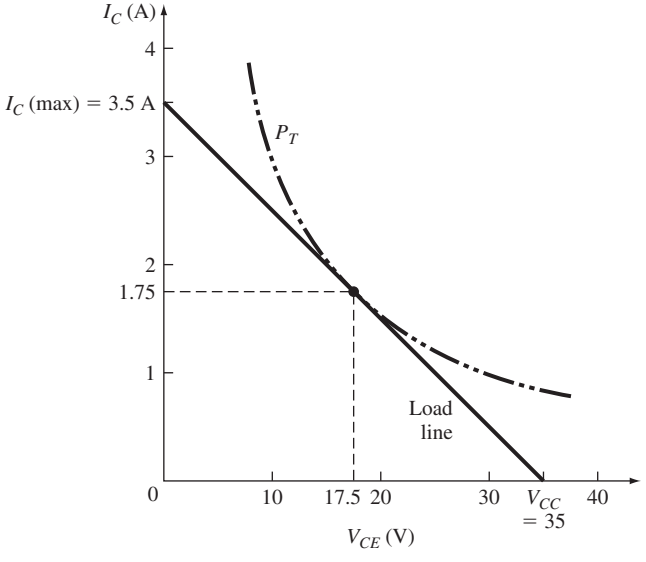


Figure 15.16 | Load line and maximum power curve for Example 15.1.

The collector–emitter voltage at this maximum power point is

$$V_{CE} = V_{CC} - I_C R_L = 35 - (1.75)(10) = 17.5 \text{ V}$$

The maximum power dissipated in the transistor occurs at the center of the load line. The maximum transistor power dissipation is therefore

$$P_T = V_{CE} I_C = (17.5)(1.75) = 30.6 \text{ W}$$

Comment

To find a transistor for a given application, safety factors are normally used. For this example, a transistor with a current rating greater than 3.5 A, a voltage rating greater than 35 V, and a power rating greater than 30.6 W would be required for the application just described.

EXERCISE PROBLEM

Ex 15.1 Assume the BJT in the common-emitter circuit shown in Figure 15.15 has limiting factors of $I_{C,\max} = 5\text{A}$, $V_{CE,\text{sus}} = 75$ V, and $P_T = 30$ *W*. Neglecting second breakdown effects, determine the minimum value of R_L such that the Q-point of the transistor always stays within the safe operating area for (*a*) $V_{CC} = 60$ V, (*b*) $V_{CC} = 40$ V, and (*c*) $V_{CC} = 20$ V. In each case, determine the maximum collector current and maximum transistor power dissipation.

[Ans. (*a*) $R_L = 30\ \Omega$, $I_{C(\max)} = 2$ A, $P_{(\max)} = 30$ W; (*b*) $R_L = 13.3\ \Omega$, $I_{C(\max)} = 3$ A, $P_{(\max)} = 30$ W; (*c*) $R_L = 4\ \Omega$, $I_{C(\max)} = 5$ A, $P_{(\max)} = 25$ W]

15.4.3 Darlington Pair Configuration

As mentioned, the base width of a power BJT is relatively wide so that the current gain is then relatively small. One method that is used to increase the effective current gain is to use a Darlington pair such as shown in Figure 15.17. Considering the currents, we see that

$$i_C = i_{CA} + i_{CB} = \beta_A i_B + \beta_B i_{EA} = \beta_A i_B + \beta_B (1 + \beta_A) i_B \tag{15.7}$$

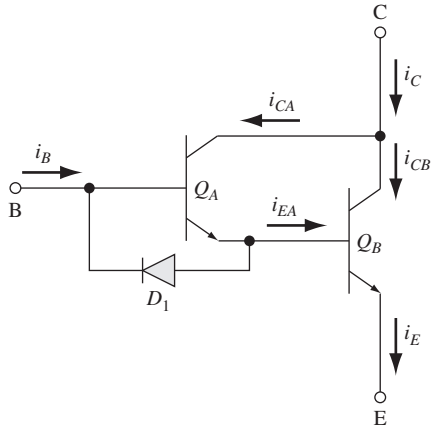


Figure 15.17 | An npn Darlington pair configuration.

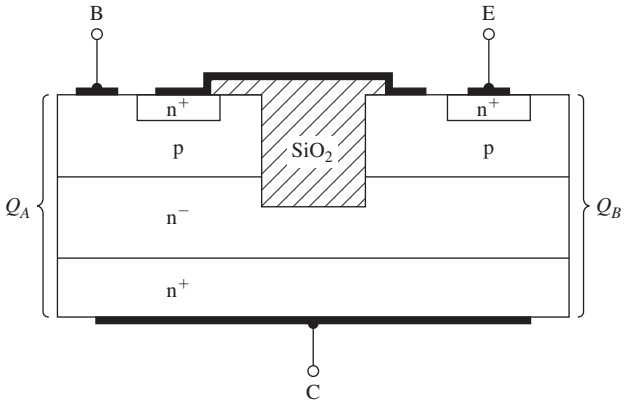


Figure 15.18 | An integrated circuit implementation of the npn Darlington pair configuration.

The overall common-emitter current gain is then

$$\frac{i_C}{i_B} = \beta_A\beta_B + \beta_A + \beta_B \tag{15.8}$$

Thus, if the gain of each individual transistor is $\beta_A = \beta_B = 15$, then the overall gain of the Darlington pair is $i_C/i_B = 255$. This overall gain is then substantially larger than that of the individual device. A diode may be incorporated as shown in Figure 15.17 to aid in turning off the transistor Q_B. A reverse current out of the base of Q_B through the diode will pull charge out of the base of this transistor and turn the device off faster than when no diode is used.

The Darlington pair shown in Figure 15.17 is typically used in the output stage of a power amplifier when an npn bipolar transistor is required. A pnp Darlington pair may also be used to increase the effective current gain of a power pnp device.

The integrated circuit configuration of the npn Darlington pair may be as shown in Figure 15.18. The silicon dioxide that is shown completely penetrates through the p-type base region so that the base regions of the two transistors are isolated.

TEST YOUR UNDERSTANDING

TYU 15.1 Consider the vertical power silicon BJT shown in Figure 15.10. Assume that a reverse-biased voltage of 200 V is applied to the base–collector junction. Calculate the space charge width that extends into the (*a*) collector region and (*b*) base region.

[Ans. (*a*) $x_d(n) = 0.506\ \mu m$; (*b*) $x_d(p) = 0.950\ \mu m$]

TYU 15.2 For the emitter–follower circuit in Figure 15.19, the parameters are $V_{CC} = 10$ V and $R_E = 200\ \Omega$. The transistor current gain is $\beta = 150$, and the current and voltage limitations are $I_{C,\max} = 200$ mA and $V_{CE,\text{sus}} = 50$ V. Determine the minimum transistor power rating such that the transistor Q-point is always inside the safe operating area.

(Ans. $P_{\max} = 0.5$ W)

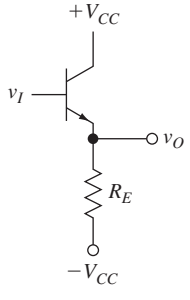


Figure 15.19 | Figure for Exercise TYU 15.2.

15.5 | POWER MOSFETs

The basic operation of the power MOSFET is the same as that of any MOSFET. However, the current handling capability of these devices is usually in the ampere range, and the drain-to-source blocking voltage may be in the range of 50 to 100 volts or even higher. One big advantage that a power MOSFET has over a bipolar power device is that the control signal is applied to the gate whose input impedance is extremely large. Even during switching between on and off states, the gate current is small, so that relatively large currents can be switched with very small control currents.

15.5.1 Power Transistor Structures

Large currents can be obtained in a MOSFET with a very large channel width. To achieve a large channel width device with good characteristics, power MOSFETs are fabricated with a repetitive pattern of small cells operating in parallel. To achieve a large blocking voltage, a vertical structure is used. There are two basic power MOSFET structures. The first is called a *DMOS* device and is shown in Figure 15.20. The DMOS device uses a double diffusion process: The p-base or the p-substrate region and the n^+ source contact are diffused through a common window defined by the edge of the gate. The p-base region is diffused deeper than the n^+ source, and the difference in the lateral diffusion distance between the p-base and the n^+ source defines the surface channel length.

Electrons enter the source terminal and flow laterally through the inversion layer under the gate to the n-drift region. The electrons then flow vertically through the n-drift region to the drain terminal. The conventional current direction is from the drain to the source. The n-drift region must be moderately doped so that the drain breakdown voltage is sufficiently large. However, the thickness of the n-drift region should also be as thin as possible to minimize drain resistance.

The second power MOSFET structure, shown in Figure 15.21, is a *VMOS* structure. The vertical channel or VMOS power device is a nonplanar structure that requires a different type of fabrication process. In this case, a p-base or p-"substrate" diffusion

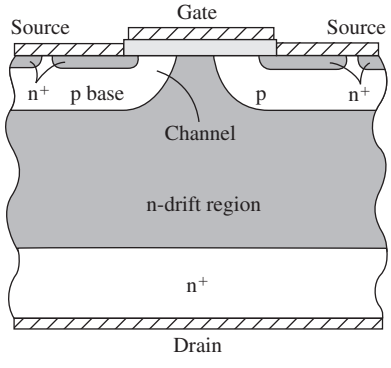


Figure 15.20 | Cross section of a double-diffused MOS (DMOS) transistor.

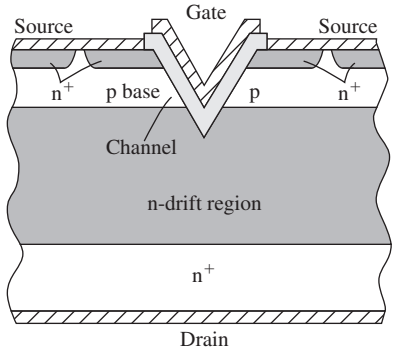


Figure 15.21 | Cross section of a vertical channel MOS (VMOS) transistor.

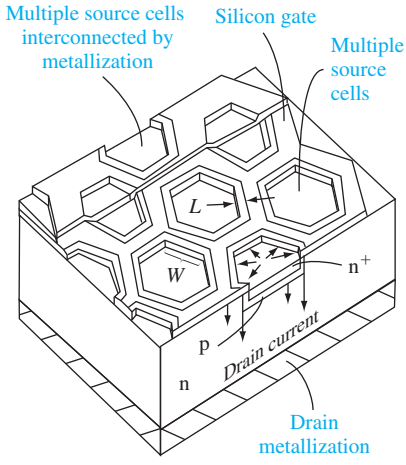


Figure 15.22 | A HEXFET structure.

is performed over the entire surface followed by the n^+ source diffusion. A V-shaped groove is then formed, extending through the n-drift region. It has been found that certain chemical solutions etch the (111) planes in silicon at a much slower rate than the other planes. If (100) oriented silicon is etched through a window at the surface, these chemical etches will create a V-shaped groove. A gate oxide is then grown in the V-shaped groove and the metal gate material is deposited. An electron inversion layer is formed in the base or substrate so that current is again essentially a vertical current between the source and the drain. The relatively low-doped n-drift region supports the drain voltage since the depletion region extends mainly into this low-doped region.

We mentioned that many individual MOSFET cells are connected in parallel to fabricate a power MOSFET with the proper width-to-length ratio. Figure 15.22 shows a HEXFET structure. Each cell is a DMOS device with an n^+ polysilicon gate. The HEXFET has a very high packing density—it may be on the order of 10^5 cells per cm^2. In the VMOS structure, the anisotropic etching of the grooves must be along the [110] direction on the (100) surface. This constraint limits the design options available for this type of device.

15.5.2 Power MOSFET Characteristics

Table 15.2 lists the basic parameters of two n-channel power MOSFETs. The drain currents are in the ampere range and the breakdown voltages are in the hundreds of volts range.

An important parameter of a power MOSFET is the *on resistance,* which can be written as

$$R_{on} = R_S + R_{CH} + R_D \tag{15.9}$$

where R_S is the resistance associated with the source contact, R_{CH} is the channel resistance, and R_D is the resistance associated with the drain contact. The R_S and R_D

Table 15.2 | Characteristics of two power MOSFETs

Parameter	2N6757	2N6792
V_{DS}(max) (V)	150	400
I_D(max) (at $T = 25°C$)	8	2
P_D(W)	75	20

resistance values are not necessarily negligible in power MOSFETs since small resistances and high currents can produce considerable power dissipation.

In the linear region of operation, we may write the channel resistance as

$$R_{CH} = \frac{L}{W\mu_n C_{ox}(V_{GS} - V_T)} \tag{15.10}$$

We have noted in previous chapters that mobility decreases with increasing temperature. The threshold voltage varies only slightly with temperature so that, as current in a device increases and produces additional power dissipation, the temperature of the device increases, the carrier mobility decreases, and R_{CH} increases, which inherently limits the channel current. The resistances R_S and R_D are proportional to semiconductor resistivity and so are also inversely proportional to mobility and have the same temperature characteristics as R_{CH}. Figure 15.23 shows a typical "on-resistance" characteristic as a function of drain current.

The increase in resistance with temperature provides stability for the power MOSFET. If the current in any particular cell begins to increase, the resulting temperature rise will increase the resistance, thus limiting the current. With this particular characteristic, the total current in a power MOSFET tends to be evenly distributed among the parallel cells, not concentrated in any single cell, a condition that can cause burnout.

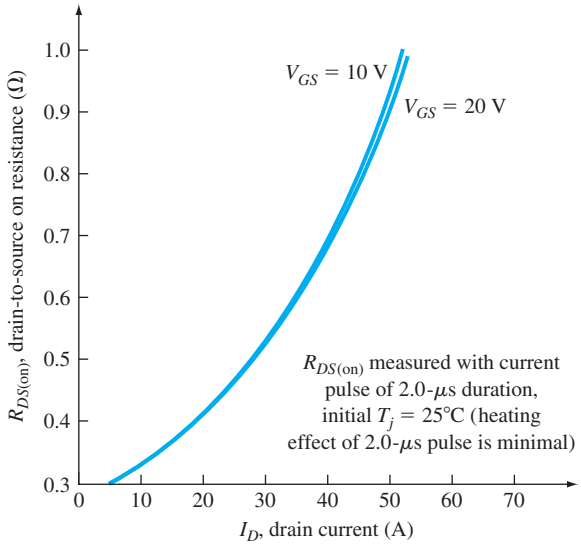


Figure 15.23 | Typical drain-to-source resistance versus drain current characteristics of a MOSFET.

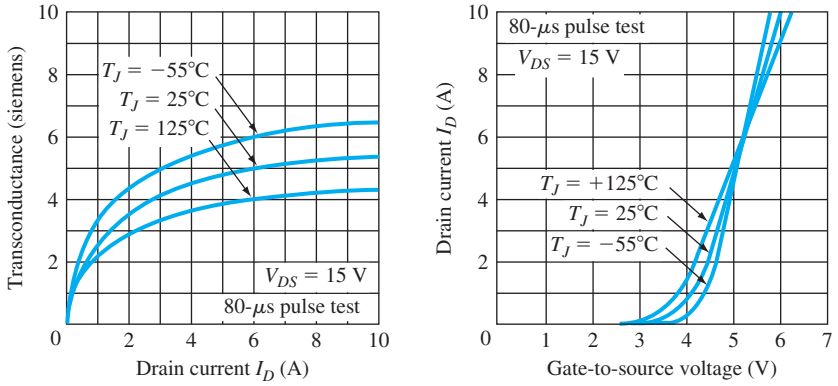


Figure 15.24 | Typical characteristics for high-power MOSFETs at various temperatures: (a) transconductance versus drain current; (b) drain current versus gate-to-source voltage.

Power MOSFETs differ from bipolar power transistors in both operating principles and performance. The superior performance characteristics of power MOSFETs are faster switching times, no second breakdown, and stable gain and response time over a wide temperature range. Figure 15.24a shows the transconductance of the 2N6757 versus temperature. The variation with temperature of the MOSFET transconductance is less than the variation in the BJT current gain that is shown in Figure 15.12. Figure 15.24b is a plot of drain current versus gate-to-source voltage at three different temperatures. We may note that at high current, the current decreases with temperature at a constant gate-to-source voltage, providing the stability that has been discussed.

Power MOSFETs must operate in a SOA. As with power BJTs, the SOA is defined by three factors: the maximum drain current, $I_{D,\max}$, rated breakdown voltage, BV_{DSS}, and the maximum power dissipation given by $P_T = V_{DS}I_D$. The SOA is shown in Figure 15.25a in which the current and voltage are plotted on linear scales. The

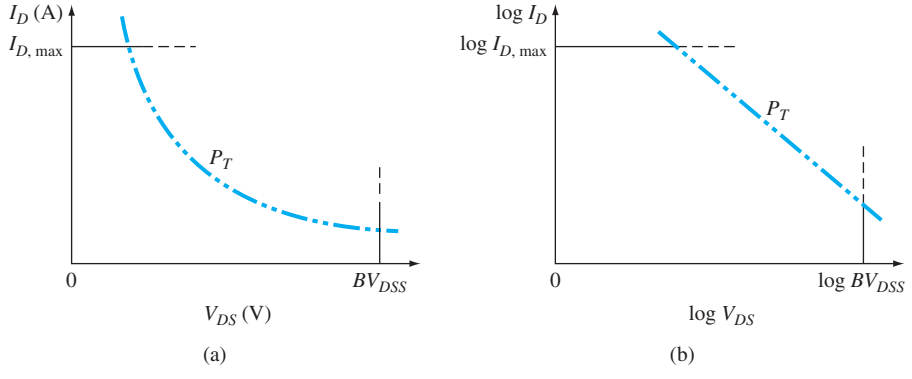


Figure 15.25 | The safe operating area (SOA) of a MOSFET plotted on (a) linear scales and (b) logarithmic scales.

same SOA curve is shown in Figure 15.25b in which the current and voltage are plotted on log scales.

EXAMPLE 15.2

Objective: Find the optimum drain resistor in a MOSFET inverter circuit.

A MOSFET inverter circuit is shown in Figure 15.26. Two different MOSFETs are being considered for use in the circuit. The parameters for devices *A* and *B* are given.

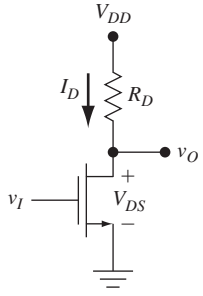


Figure 15.26 | A MOSFET inverter circuit.

Device A	Device B
$BV_{DSS} = 35$ V	$BV_{DSS} = 35$ V
$P_T = 30$ W	$P_T = 30$ W
$I_{D,\max} = 6$ A	$I_{D,\max} = 4$ A

■ Solution

The SOA curves for the two devices are shown in Figure 15.27.

The load line for the inverter circuit using device A is shown as curve A. The load line intersects the voltage axis at $V_{DD} = 24$ V. This curve is tangent to the maximum power curve and intersects the current axis at $I_D = 5$ A. Note that, if we had wanted the load line to intersect the maximum rated current of $I_{D,\max} = 6$ A, the load line would have gone outside of the SOA.

For the load line *A*, the drain resistance is

$$R_D = \frac{V_{DD}}{I_D} = \frac{24}{5} = 4.8\ \Omega$$

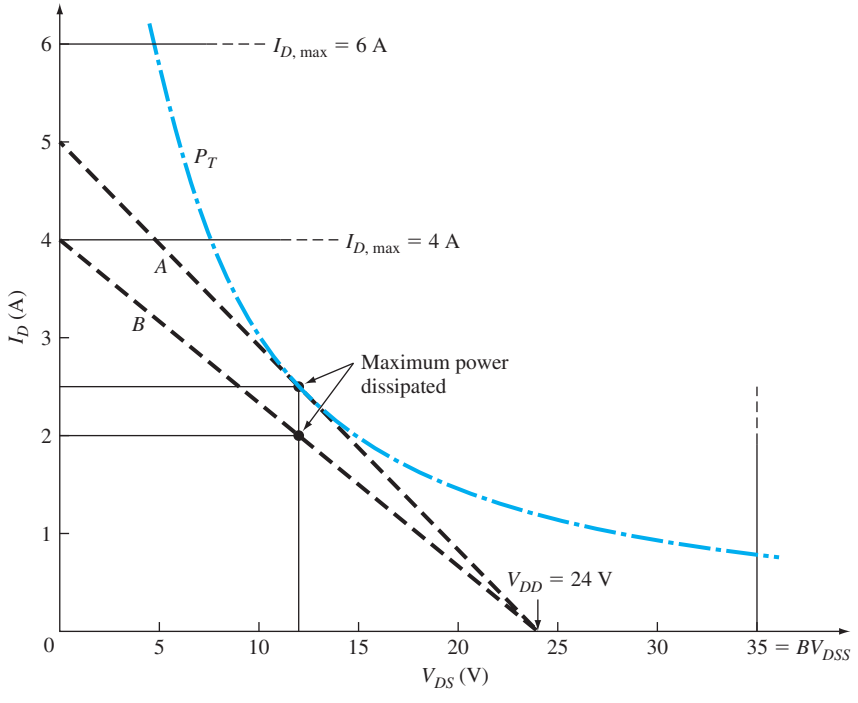


Figure 15.27 | Safe operating area and load lines for devices in Example 15.2.

The current at the maximum power point (using the results from Example 15.1) is

$$I_D = \frac{V_{DD}}{2R_D} = \frac{24}{2(4.8)} = 2.5 \text{ A}$$

and the corresponding drain-to-source voltage is

$$V_{DS} = V_{DD} - I_D R_D = 24 - (2.5)(4.8) = 12 \text{ V}$$

The maximum power that may be dissipated in the transistor is $P = V_{DS} I_D = (12)(2.5) = 30 \text{ W} = P_T$, which corresponds to the maximum rated power. This point is shown on the curve.

The load line for the inverter circuit using device B is shown as curve B. The load line intersects the voltage axis at $V_{DD} = 24$ V as before. This curve can now intersect the current axis at the maximum rated drain current of $I_{D,\max} = 4$ A. We see that the load line falls within the SOA of the transistor.

For load line B, the drain resistance is

$$R_D = \frac{V_{DD}}{I_D} = \frac{24}{4} = 6 \ \Omega$$

The current at the maximum power point is

$$I_D = \frac{V_{DD}}{2R_D} = \frac{24}{2(6)} = 2 \text{ A}$$

and the corresponding drain-to-source voltage is

$$V_{DS} = V_{DD} - I_D R_D = 24 - (2)(6) = 12 \text{ V}$$

The maximum power that may be dissipated in the transistor is $P = V_{DD} I_D = (12)(2) = 24$ W, which is less than the maximum rated power. This point is also shown on the curve.

Conclusion

We see that if device A is used, the drain resistor is determined by the maximum power. However, if device B is used, the drain resistor is determined by the maximum rated current of the device.

EXERCISE PROBLEM

Ex 15.2 Consider the common-source circuit shown in Figure 15.26. Determine the required current, voltage, and power ratings of the MOSFET for (*a*) $R_D = 12 \ \Omega$, $V_{DD} = 24$ V and (*b*) $R_D = 8 \ \Omega$, $V_{DD} = 40$ V.

[Ans. (*a*) $BV_{DSS} = 24$ V, $I_{D,\max} = 2$ A, $P_T = 12$ W; (*b*) $BV_{DSS} = 40$ V, $I_{D,\max} = 5$ A, $P_T = 50$ W]

15.5.3 Parasitic BJT

The MOSFET has a parasitic BJT as an inherent part of its structure. The parasitic BJT may be seen in both the DMOS and VMOS structures shown in Figures 15.20 and 15.21. The source terminal corresponds to the n-type emitter, the p-type base or substrate region corresponds to the p-type base, and the n-type drain corresponds to the n-type collector. This is also shown schematically in Figure 15.28. The channel length of the MOSFET corresponds to the base width of the parasitic BJT. Since this length is normally quite small, the current gain β of the BJT can be larger than unity.

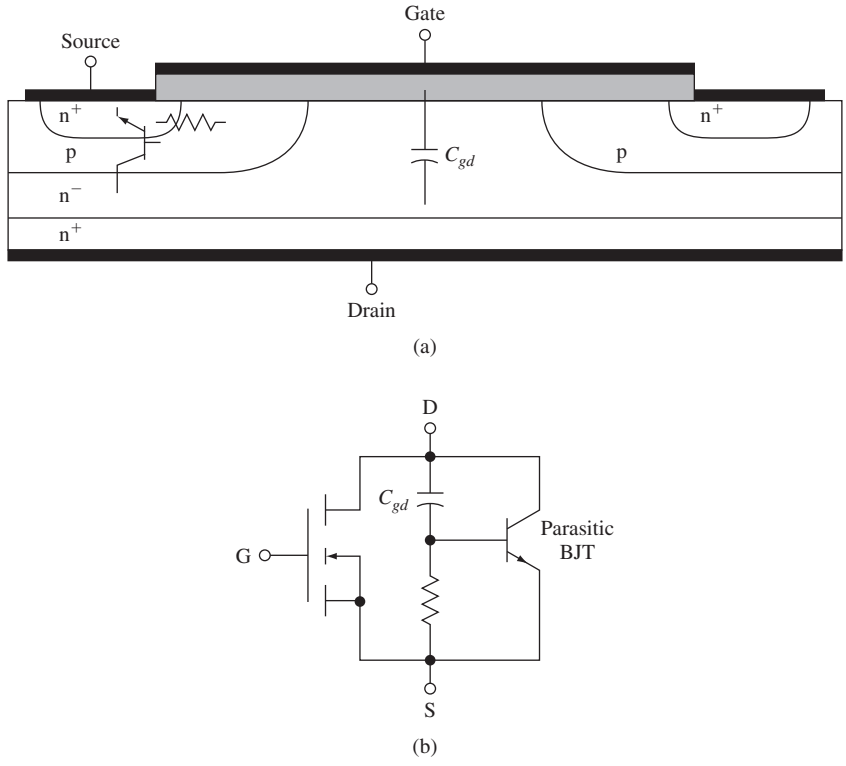


Figure 15.28 | (a) Cross section of vertical MOSFET showing parasitic BJT and distributed resistance; (b) equivalent circuit of MOSFET and parasitic BJT with distributed parameters.

The BJT should be cutoff at all times, which means the source-to-body voltage (emitter-to-base voltage) should be as close to zero as possible. We see from the geometries of Figures 15.20 and 15.21 that the source ohmic contact also goes across the p-type body region so that this junction voltage is zero during steady-state operation of the transistor. However, the BJT may be turned on during high-speed switching of the MOSFET.

Figure 15.28b shows that the base and the collector of the parasitic BJT are connected by the gate-to-drain capacitance. A parasitic or distributed resistance also connects the base to the emitter of the BJT. When the MOSFET is being turned off, the drain-to-source voltage increases and induces a current in the gate-to-drain capacitance in the direction from the parasitic collector terminal to the parasitic base terminal. This induced current may be large enough to induce a voltage in the parasitic resistance that is sufficient to forward bias the base–emitter junction and therefore turn the BJT on. The turned on BJT may then induce a large drain current that can cause burnout of the MOSFET. This breakdown mechanism is known as *snapback breakdown* and has been discussed briefly in Section 11.4.1.

The current–voltage characteristics are shown in Figure 11.22. Devices can be designed to minimize the parasitic or distributed base–emitter resistance to minimize this problem.

15.6 | THE THYRISTOR

One of the important applications of electronic devices is in switching between an off or blocking state to an on or low-impedance state. Thyristor is the name given to a general class of semiconductor pnpn switching devices that exhibit bistable regenerative switching characteristics. We have considered the transistor, which may be switched on with the application of a base drive or a gate voltage. The base drive or gate voltage must be applied as long as the transistor is to remain on. There are a number of applications in which it is useful to have a device remain in a blocking state until switched to the low-impedance state by a control signal, which then does not necessarily have to remain on. These devices are efficient in switching large currents at low frequencies, such as industrial control circuits operating at 60 Hz.

A **S**emiconductor **C**ontrolled **R**ectifier (SCR) is the common name given to a three-terminal thyristor. The SCR (sometimes referred to as a silicon controlled rectifier) is a four-layer pnpn structure with a gate control terminal. As with most semiconductor devices, there are several variations of the device structure. We consider the basic SCR operation and limitations, and then discuss some variations of the basic four-layer device.

15.6.1 The Basic Characteristics

The four-layer pnpn structure is shown in Figure 15.29a. The upper p region is called the anode and the lower n region is called the cathode. If a positive voltage is applied to the anode, the device is said to be forward biased. However, the junction J_2 is reverse biased so that only a very small current exists. If a negative voltage is applied to the anode, then junctions J_1 and J_3 are reverse biased—again only a very small current will exist. Figure 15.29b shows the I–V characteristics for these conditions. The voltage V_p is the breakdown voltage of the J_2 junction. For properly designed devices, the blocking voltage can be several thousand volts.

To consider the characteristics of the device as it goes into its conducting state, we can model the structure as coupled npn and pnp bipolar transistors. Figure 15.30a shows how we can split the four-layer structure and Figure 15.30b shows the two-transistor equivalent circuit with the associated currents. Since the base of the pnp device is the same as the collector of the npn transistor, the base current I_{B1} must in fact be the same as the collector current I_{C2}. Similarly, since the collector of the pnp transistor is the same as the base of the npn device, the collector current I_{C1} must be the same as the base current I_{B2}. In this bias configuration, the B–C of the pnp and the B–C of the npn devices are reverse biased, while the B–E junctions are both forward biased. The parameters α_1 and α_2 are the common base current gains of the pnp and npn transistors, respectively.

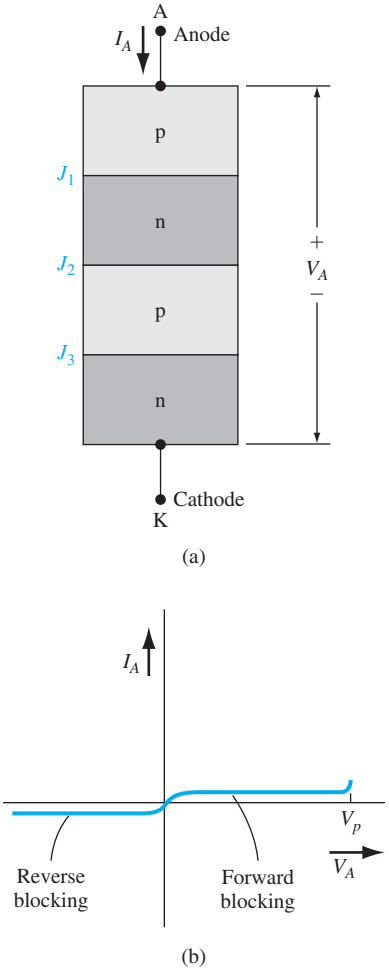


Figure 15.29 | (a) The basic four-layer pnpn structure. (b) The initial current–voltage characteristic of the pnpn device.

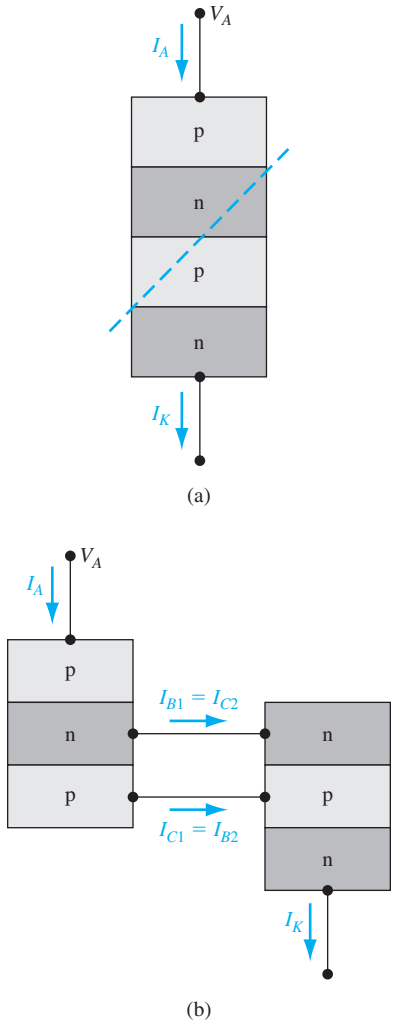


Figure 15.30 | (a) The splitting of the basic pnpn structure. (b) Two two-transistor equivalent circuit of the four-layer pnpn device.

We can write

$$I_{C1} = \alpha_1 I_A + I_{C01} = I_{B2} \tag{15.11a}$$

and

$$I_{C2} = \alpha_2 I_K + I_{C02} = I_{B1} \tag{15.11b}$$

where I_{C01} and I_{C02} are the reverse B–C junction saturation currents in the two devices. In this particular configuration, $I_A = I_K$ and $I_{C1} + I_{C2} = I_A$. If we add Equations (15.11a) and (15.11b), we obtain

$$I_{C1} + I_{C2} = I_A = (\alpha_1 + \alpha_2)\, I_A + I_{C01} + I_{C02} \tag{15.12}$$

The anode current I_A, from Equation (15.12), can be found as

$$I_A = \frac{I_{C01} + I_{C02}}{1 - (\alpha_1 + \alpha_2)} \tag{15.13}$$

As long as $(\alpha_1 + \alpha_2)$ is much smaller than unity, the anode current is small, as we have indicated in Figure 15.29b.

The common base current gains, α_1 and α_2, are very strong functions of collector current as we discussed in Chapter 12. For small values of V_A, the collector current in each device is just the reverse saturation current, which is very small. The small collector current implies that both α_1 and α_2 are much smaller than unity. The four-layer structure maintains this blocking condition until the junction J_2 starts into breakdown or until a current is induced in the J_2 junction by some external means.

Consider, initially, the condition when the applied anode voltage is sufficiently large to cause the J_2 junction to start into avalanche breakdown. This effect is shown in Figure 15.31a. The electrons generated by impact ionization are swept into the n_1 region, making the n_1 region more negative, and the holes generated by impact ionization are swept into the p_2 region, making the p_2 region more positive. The more negative voltage of the n_1 region and the more positive voltage of the p_2 region means that the forward-bias junction voltages V_1 and V_3 both increase. The increase in the respective B–E junction voltages causes an increase in current, which results in an increase in the common-base current gains α_1 and α_2, causing a further increase in

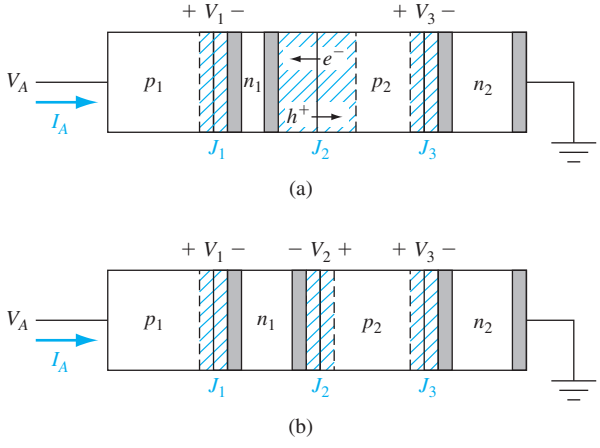


Figure 15.31 | (a) The pnpn device when the J_2 junction starts into avalanche breakdown. (b) The junction voltages in the pnpn structure when the device is in the high-current, low-impedance state.

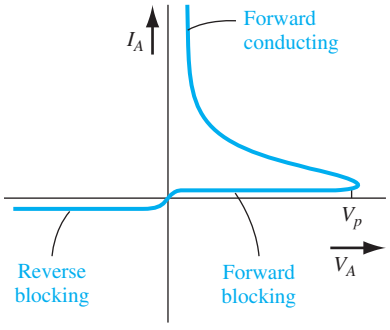


Figure 15.32 | The current–voltage characteristics of the pnpn device.

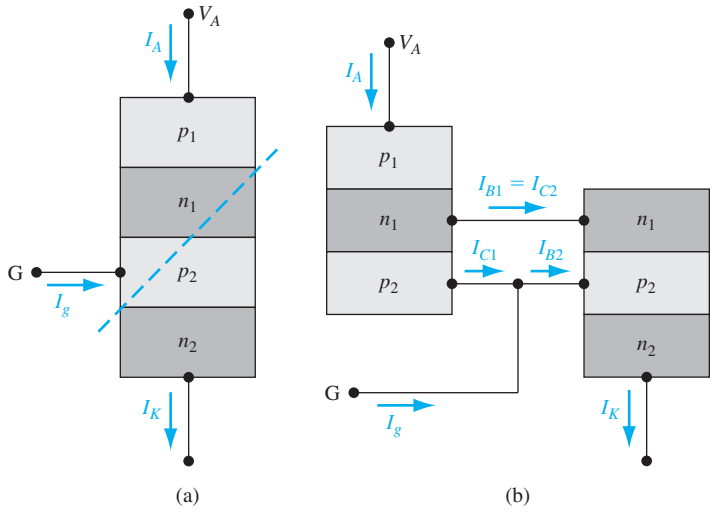


Figure 15.33 | (a) The three-terminal SCR. (b) The two-transistor equivalent circuit of the three-terminal SCR.

I_A as seen in Equation (15.13). We now have a regenerative positive feedback situation, so the current I_A will increase very rapidly.

As the anode current I_A increases and $\alpha_1 + \alpha_2$ increases, the two equivalent bipolar transistors are driven into saturation and the junction J_2 becomes forward biased. The total voltage across the device decreases and is approximately equal to one diode drop as shown in Figure 15.31b. The current in the device is limited by the external circuit. If the current is allowed to increase, ohmic losses may become important so that the voltage drop across the device may increase slightly with current. The I_A versus V_A characteristic is shown in Figure 15.32.

15.6.2 Triggering the SCR

In the last section, we considered the case when the four-layer pnpn device is turned on by the avalanche breakdown process in the center junction. The turn-on condition can also be initiated by other means. Figure 15.33a shows three-terminal SCR in which the third terminal is the gate control. We can determine the effect of the gate current by reconsidering Equations (15.11a) and (15.11b).

Figure 15.33b again shows the two-transistor equivalent circuit including the gate current. We can write

$$I_{C1} = \alpha_1 I_A + I_{C01} \tag{15.14a}$$

and

$$I_{C2} = \alpha_2 I_K + I_{C02} \tag{15.14b}$$

We now have $I_K = I_A + I_g$ and we can still write $I_{C1} + I_{C2} = I_A$. Adding Equations (15.14a) and (15.14b), we find that

$$I_{C1} + I_{C2} = I_A = (\alpha_1 + \alpha_2)I_A + \alpha_2 I_g + I_{C01} + I_{C02} \tag{15.15}$$

Solving for I_A, we find

$$I_A = \frac{\alpha_2 I_g + (I_{C01} + I_{C02})}{1 - (\alpha_1 + \alpha_2)} \tag{15.16}$$

We can think of the gate current as the flow of holes into the p_2 region. The additional holes increase the potential of this region, which increases the forward-biased B–E voltage of the npn bipolar transistor, and the transistor action. The transistor action of the npn increases the collector current I_{C2}, which starts the transistor action of the pnp bipolar transistor, and the entire pnpn device can be turned on into its low-impedance state. The gate current required to switch the SCR into its on condition is typically in the milliamp range. SCR can be turned on with a small gate current, which can control hundreds of amperes of anode current. The gate current can be turned off and the SCR will remain in its conducting state. The gate loses control of the device once the SCR is triggered into its conducting state. The current–voltage characteristics of the SCR as a function of gate current is shown in Figure 15.34.

A simple application of an SCR in a half-wave control circuit is shown in Figure 15.35a. The input signal is an ac voltage and a trigger pulse will control the turn-on of the SCR. We assume that the trigger pulse occurs at time t_1 during the ac voltage cycle. Prior to t_1, the SCR is off so that the current in the load is zero; thus, there is a zero output voltage. At $t = t_1$, the SCR is triggered on and the input voltage appears across the load (neglecting the voltage drop across the SCR). The SCR turns off when the anode-to-cathode voltage becomes zero even though the trigger pulse has been turned off prior to this time. The time at which the SCR is triggered during the voltage cycle can be varied, changing the amount of power delivered to the load. Full-wave control circuits can be designed to increase efficiency and degree of control.

The gate allows control of the turn-on of the SCR. However, the four-layer pnpn structure can also be triggered on by other means. In many integrated circuits, parasitic pnpn structures exist. One such example is the CMOS structure that we considered in Chapter 10. A transient ionizing radiation pulse can trigger the parasitic

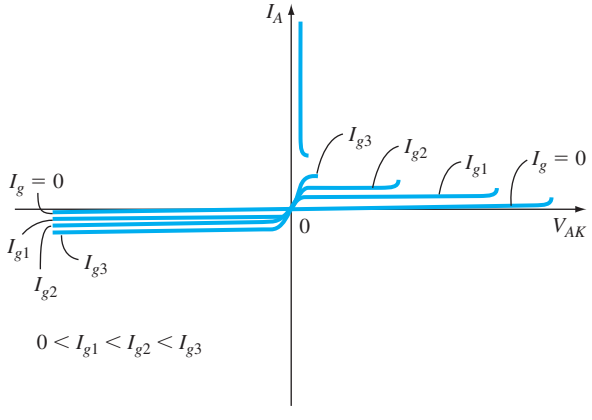


Figure 15.34 | Current–voltage characteristics of an SCR.

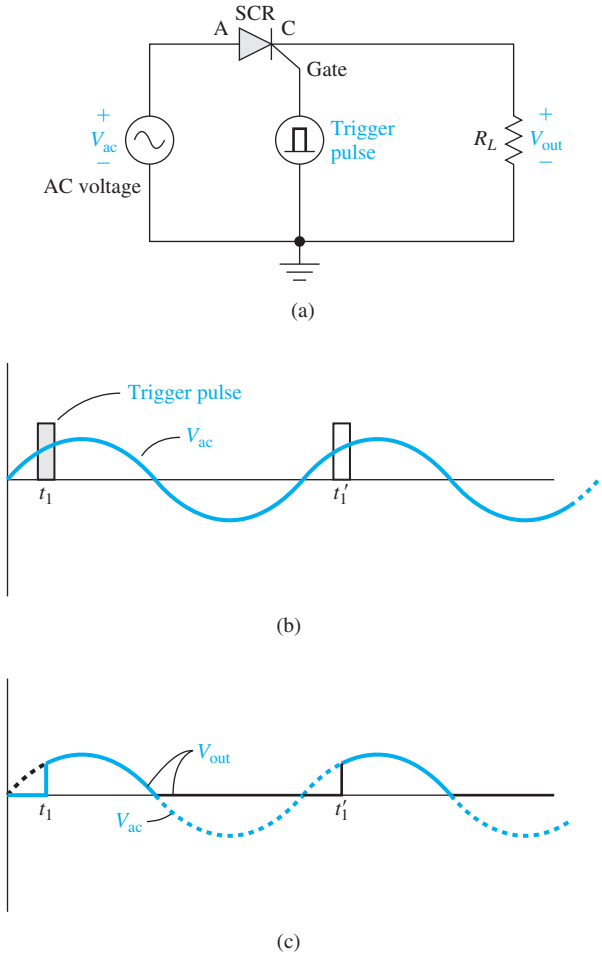


Figure 15.35 | (a) Simple SCR circuit. (b) Input ac voltage signal and trigger pulse. (c) Output voltage versus time.

four-layer device by generating electron–hole pairs, particularly in the J_2 junction, producing a photocurrent. The photocurrent is equivalent to a gate current in an SCR so the parasitic device can be switched into its conducting state. Again, once the device is switched on, it will remain in its conducting state even when the radiation ceases. An optical signal can also trigger the device in the same manner by generating electron–hole pairs.

Another triggering mechanism in the pnpn device is dV/dt triggering. If the forward-bias anode voltage is applied rapidly, the voltage across the J_2 junction will also change quickly. This changing reverse-biased J_2 junction voltage means that the space charge region width is increasing; thus, electrons are being removed from the n_1 side of the junction and holes are being removed from the p_2 side of the junction.

If dV/dt is large, the rate of removal of these carriers is rapid, which leads to a large transient current that is equivalent to a gate current and can trigger the device into a low-impedance conducting state. In SCR devices, a dV/dt rating is usually specified. However, in parasitic pnpn structures, the dV/dt triggering mechanism is a potential problem.

15.6.3 SCR Turn-Off

Switching the four-layer pnpn structure from its conducting state to its blocking state can be accomplished if the current I_A is reduced below the value creating the $\alpha_1 + \alpha_2 = 1$ condition. This critical I_A current is called the holding current. If a parasitic four-layer structure is triggered into the conducting state, the effective anode current in the device must be reduced below the corresponding holding current in order to turn off the device. This requirement essentially implies that all power supplies must be turned off in order to bring the parasitic device back into its blocking state.

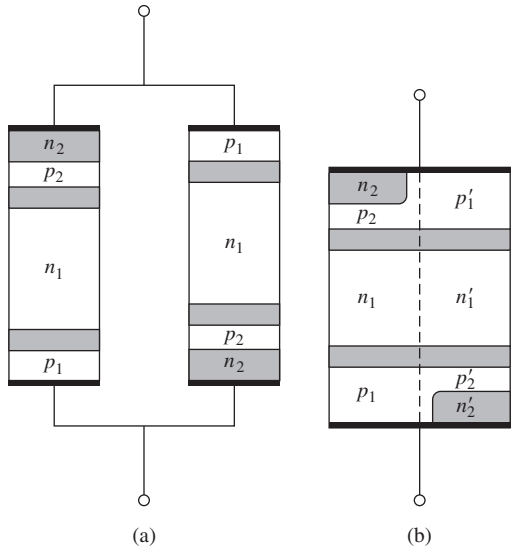
The SCR can be triggered on by supplying holes to the p_2 region of the device. The SCR can perhaps be turned off by removing holes from this same region. If the reverse gate current is large enough to bring the npn bipolar transistor out of saturation, then the SCR can be switched from the conducting state into the blocking state. However, the lateral dimensions of the device may be large enough so that nonuniform biasing in the J_2 and J_3 junctions occurs during a negative gate current and the device will remain in the low-impedance conducting state. The four-layer pnpn device must be specifically designed for a turn-off capability.

15.6.4 Device Structures

Many thyristor structures have been fabricated with specific characteristics for specific applications. We consider a few of these types of device to gain an appreciation for the variety of structures.

Basic SCR There are many variations of diffusion, implantation, and epitaxial growth that can be used in the fabrication of the SCR device. The basic structure is shown in Figure 15.36. The p_1 and p_2 regions are diffused into a fairly high resistivity n_1 material. The n^+ cathode is formed and the p^+ gate contact is made. High thermal conductivity materials can be used for the anode and cathode ohmic contacts to aid in heat dissipation for high-power devices. The n_1 region width may be on the order of 250 μm in order to support very large reverse-biased voltages across the J_2 junction. The p_1 and p_2 regions may be on the order of 75 μm wide, while the n^+ and p^+ regions are normally quite thin.

Bilateral Thyristor Since thyristors are often used in ac power applications, it may be useful to have a device that switches symmetrically in the positive and negative cycles of the ac voltage. There are a number of such devices, but the basic concept is to connect two conventional thyristors in antiparallel as shown in Figure 15.37a. The integration of this concept into a single device is shown in Figure 15.37b. Symmetrical n regions can be diffused into a pnp structure. Figure 15.37c shows the current–voltage characteristics in which the triggering into the conduction mode



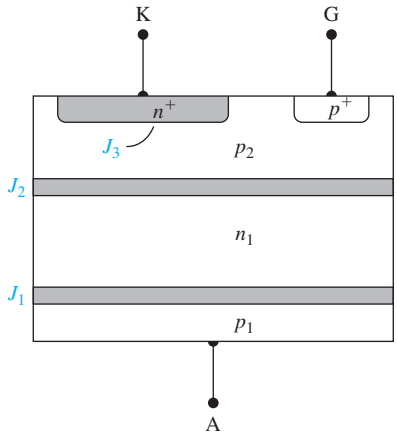


Figure 15.36 | The basic SCR device structure.

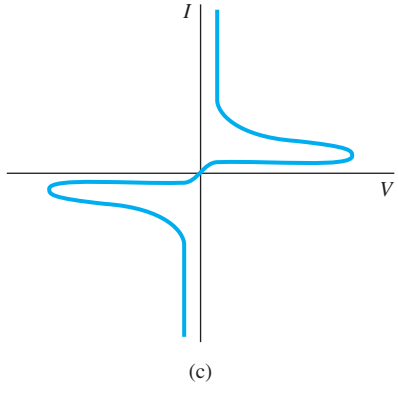


Figure 15.37 | (a) The antiparallel connection of two thyristors to form a bilateral device. (b) The bilateral thyristor as an integrated device. (c) The current–voltage characteristics of the bilateral thyristor. *(From Ghandhi [7].)*

would be due to breakdown triggering. The two terminals alternately share the role of anode and cathode during successive half cycles of the ac voltage.

Triggering by a gate control is more complex for this device since a single gate region must serve for both of the antiparallel thyristors. One such device is known as a *triac*. Figure 15.38a shows the cross section of such a device. This device can be triggered into conduction by gate signals of either polarity and with anode-to-cathode voltages of either polarity.

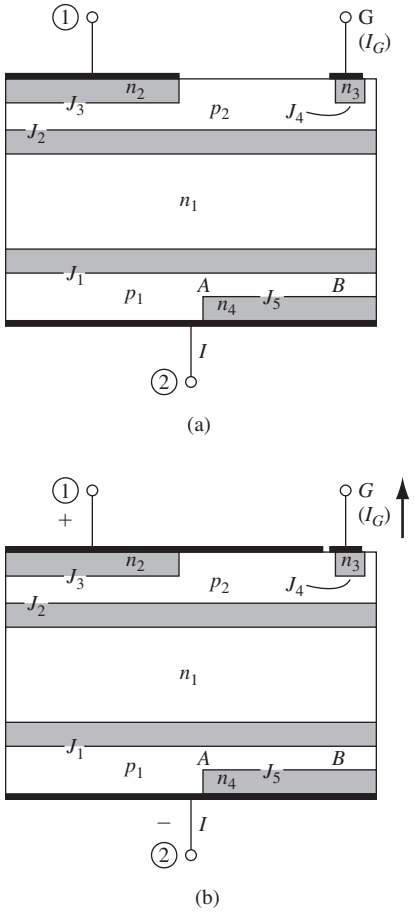


Figure 15.38 | (a) The triac device. (b) The triac with a specific bias configuration. *(From Ghandhi [7].)*

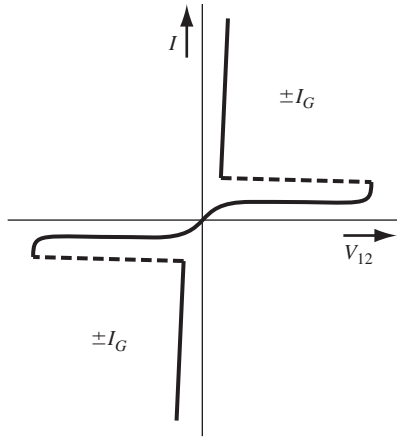


Figure 15.39 | The current–voltage characteristics of the triac.

One particular gate control situation is shown in Figure 15.38b. Terminal 1 is positive with respect to terminal 2, and a negative gate voltage is applied with respect to terminal 1, so the gate current is negative. This polarity arrangement induces the current I_1 and the junction J_4 becomes forward biased. Electrons are injected from n_3, diffuse across p_2, and are collected in the n_1 region. In this case $n_3p_2n_1$ behaves like a saturated transistor. The collected electrons in n_1 lower the potential of n_1 with respect to p_2. The current across the p_2n_1 junction increases, which can trigger the $p_2n_1p_1n_4$ thyristor into its conducting mode.

We can show that the other combinations of gate, anode, and cathode voltages will also trigger the triac into conduction. Figure 15.39 shows the terminal characteristics.

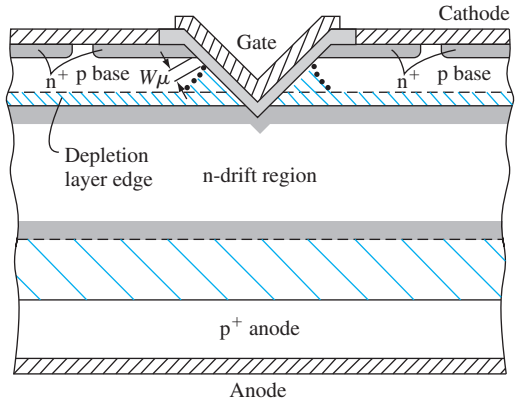


Figure 15.40 | The V groove MOS gated thyristor.
(From Baliga [1].)

MOS Gated Thyristor The operation of a MOS gated thyristor is based upon controlling the gain of the npn bipolar transistor. Figure 15.40 shows a V-groove MOS gated thyristor. The MOS gate structure must extend into the n-drift region. If the gate voltage is zero, the depletion edge in the p-base remains essentially flat and parallel to the junction J_2; the gain of the npn transistor is low. This effect is shown in the figure by the dashed line. When a positive gate voltage is applied, the surface of the p base becomes depleted—the depletion region in the p base adjacent to the gate is shown by the dotted line. The undepleted base width W_μ of the npn bipolar device narrows and the gain of the device increases.

At a gate voltage approximately equal to the threshold voltage, electrons from the n^+ emitter are injected through the depletion region into the n-drift region. The potential of the n-drift region is lowered, which further forward biases the p^+ anode to n-drift junction voltage, and the regenerative process is initiated. The gate voltage required to initiate turn-on is approximately the threshold voltage of the MOS device. One advantage of this device is that the input impedance to the control terminal is very high; relatively large currents can be switched with very small capacity coupled gate currents.

MOS Turn-Off Thyristor The MOS turn-off thyristor can both turn on and turn off the anode current by applying a signal to a MOS gate terminal. The basic device structure is shown in Figure 15.41. By applying a positive gate voltage, the n^+pn bipolar transistor can be turned on as just discussed. Once the thyristor is turned on, the device can be turned off by applying a negative gate voltage: the negative gate voltage turns on the p-channel MOS transistor that effectively short circuits the B–E junction of the n^+pn bipolar transistor. Holes that now enter the p-base have an alternative path to the cathode. If the resistance of the p-channel MOSFET becomes low enough, all current will be diverted away from the n^+p emitter and the n^+pn device will effectively be turned off.

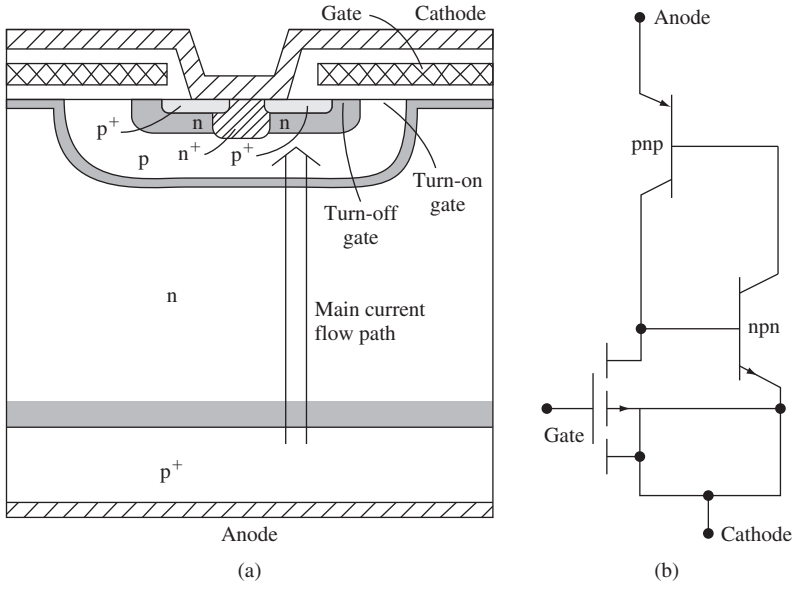


Figure 15.41 | (a) The MOS turn-off thyristor. (b) Equivalent circuit for the MOS turn-off thyristor.
(From Baliga [1].)

15.7 | SUMMARY

- The concept of a negative differential resistance in the *I–V* characteristic of the tunnel diode is used in the design of a microwave tunnel diode oscillator. The expression for the maximum resistance cutoff frequency is derived.
- The operation of a microwave GUNN diode oscillator is based on the concept of negative differential mobility.
- The IMPATT diode oscillator uses injection and drift time delays to create a region of differential negative resistance.
- The power BJT has a vertical configuration and an interdigitated base–emitter surface structure. The collector drift region (doping and width) determines the rated blocking voltage of the BJT, while the base width must be sufficiently wide to avoid punch-through breakdown at the rated blocking voltage.
- A power BJT is characterized by the maximum rated collector current, maximum rated voltage, and maximum rated power dissipation. These three parameters define the SOA of the transistor.
- A power MOSFET has a vertical configuration and an interdigitated gate–source surface structure. Two specific devices considered are the DMOS and VMOS structures. The drain-drift region (doping and width) determines the rated blocking voltage of the MOSFET, while the channel length of the base (body) must be sufficiently wide to avoid punch-through breakdown at the rated blocking voltage.
- A power MOSFET is characterized by the maximum rated drain current, maximum rated voltage, and maximum rated power dissipation. These three parameters define the SOA of the transistor.

- The "on resistance" of a MOSFET has a positive temperature coefficient so that the power MOSFET is more stable versus temperature than a power BJT. This characteristic allows MOSFETs to be fabricated in parallel to increase the current capability of the device.
- The thyristor refers to a general class of pnpn switching devices that can be switched between a high-impedance, low-current state and a low-impedance, high-current state. These devices exhibit a bistable regenerative positive feedback switching characteristic.
- The basic pnpn device can be modeled as coupled npn and pnp bipolar transistors. In the "on" state, both bipolar transistors are driven into saturation, creating the high-current, low-voltage condition. In the "off" or blocking state, large voltages can be applied to the device and the current is essentially zero.
- The turn-on characteristics of the thyristor can be controlled through a gate control terminal. The three-terminal thyristors are referred to as semiconductor controlled rectifiers (SCRs).

GLOSSARY OF IMPORTANT TERMS

double-diffused MOSFET (DMOS) A power MOSFET in which the source and channel regions are formed using a double diffusion process.

HEXFET The structure of a power MOSFET in which many individual MOSFETs are placed in parallel in a hexagonal configuration.

maximum rated current The maximum allowed current in a power transistor such that proper operation is maintained.

maximum rated power The maximum allowed power dissipation in a power transistor such that no permanent damage is done to the transistor.

maximum rated voltage The maximum allowed applied voltage to a power transistor such that breakdown is not initiated.

negative differential mobility A region in the drift velocity versus electric field characteristic of a semiconductor material in which the drift velocity decreases with an increase in the electric field.

negative differential resistance A region in the *I*–*V* characteristic of a device in which the current decreases while the voltage increases.

on resistance The effective resistance between source and drain of a power MOSFET.

safe operating area The allowed current–voltage regions of operation for a power transistor bounded by the maximum rated current, maximum rated voltage, and maximum power.

second breakdown A breakdown effect in a power BJT in which high temperature causes a thermal runaway process.

SCR (semiconductor controlled rectifier) The common name given to a three-terminal thyristor.

thyristor The name given to a general class of semiconductor pnpn switching devices exhibiting bistable regenerative switching characteristics.

transferred-electron effect The phenomenon in which conduction electrons are scattered from a lower energy, high-mobility band into a higher energy, low-mobility band.

triac The name of a bilateral three-terminal thyristor.

V-groove MOSFET (VMOS) A power MOSFET in which the channel region is formed along a V-shaped groove formed in the surface of the semiconductor.

CHECKPOINT

After studying this chapter, the reader should have the ability to:

- Explain how a region of negative differential resistance is developed in the *I–V* characteristic of the tunnel diode.
- Discuss the concept of negative differential mobility in GaAs and discuss how this phenomenon leads to the generation of domains in a GUNN diode.
- Discuss the operation of an IMPATT diode oscillator.
- Sketch the cross section of a power BJT and discuss the voltage and current limitations of the device.
- Discuss the reason the current gain of a power BJT is generally smaller than that of a small switching BJT.
- Sketch the safe operating area of a power BJT.
- Describe the reason for and the operation of a Darlington configuration.
- Sketch the cross section of the DMOS and VMOS power MOSFET structures.
- Sketch the safe operating area of a power MOSFET.
- Describe why the "on resistance" of a power MOSFET has a positive temperature coefficient.
- Describe the switching characteristics of a pnpn device.
- Describe the switching characteristics of a semiconductor controlled rectifier.

REVIEW QUESTIONS

1. Describe how a negative differential resistance region in the *I–V* characteristic of the tunnel diode is generated.
2. Describe how a negative differential mobility region in the drift velocity versus electric field characteristic in GaAs is developed.
3. Describe how a negative differential resistance characteristic is produced in the IMPATT diode.
4. Why is the doping concentration in the collector drift region low and why is the drift region width large in a power BJT?
5. Why does a power BJT have an interdigitated base–emitter structure?
6. Sketch the safe operating area of a power BJT.
7. Discuss how a DMOS structure of a power MOSFET is formed.
8. Discuss the voltage limitation of a power MOSFET.
9. Define the "on resistance" of a power MOSFET and show that the on resistance has a positive temperature coefficient.
10. Discuss how the gate terminal of a semiconductor controlled rectifier can control the switching characteristics.

PROBLEMS

Section 15.1 Tunnel Diode

15.1 Sketch the energy band diagrams of a tunnel diode in which both the n and p regions are degenerately doped for the case of (*a*) zero bias, (*b*) $0 < V < V_p$, (*c*) $V_p < V < V_v$, and (*d*) $V > V_v$.

15.2 The parameters in Figure 15.1b are $I_p = 20$ mA, $I_v = 2$ mA, $V_p = 0.15$ V, and $V_v = 0.60$ V. Assuming a straight-line approximation to the *I–V* characteristics between these two points, calculate the value of differential negative resistance.

15.3 For values of $R_{\min} = 10\ \Omega$, $R_p = 1\ \Omega$, and $C_j = 2$ nF, determine the maximum resistance cutoff frequency of a tunnel diode.

Section 15.2 GUNN Diode

15.4 (*a*) A GaAs transferred-electron device has a doping concentration of $N_d = 10^{15}$ cm^{-3}. Determine (*i*) the minimum device length, (*ii*) the time between current pulses, and (*iii*) the oscillation frequency (assume $v_d = 1.5 \times 10^7$ cm/s). (*b*) Repeat part (*a*) for a doping concentration of $N_d = 10^{16}$ cm^{-3}.

15.5 The drift region length of a GUNN diode is $L = 15\ \mu$m. The voltage across the diode oscillates between 8 and 10 V. (*a*) Determine the average electric field in the device. (*b*) Using Figure 15.3, find the average electron drift velocity. (*c*) Using the results of part (*b*), find the frequency of oscillation.

Section 15.3 IMPATT Diode

15.6 Find the frequency of oscillation of a silicon IMPATT diode with a drift region length of $L = 10\ \mu$m.

Section 15.4 Power Bipolar Transistors

15.7 Consider the vertical npn power bipolar transistor shown in Figure 15.10. The doping concentrations are $N_E = 10^{18}$ cm^{-3}, $N_B = 8 \times 10^{15}$ cm^{-3}, and $N_C = 6 \times 10^{14}$ cm^{-3}. The neutral base width is 2 μm, the electron diffusion coefficient in the base is $D_B = 20$ cm^2/s, and the B–E cross-sectional area is 0.4 cm^2. (*a*) The excess electron concentration in the base at the edge of the B–E junction is $\delta n_p(0) = 10^{14}$ cm^{-3}. Determine (*i*) the B–E voltage and (*ii*) the approximate collector current. (*b*) Determine the (*i*) B–E voltage at the edge of high injection and (*ii*) the approximate resulting collector current.

15.8 Consider the npn power bipolar transistor described in Problem 15.7. (*a*) Determine the expected B–C avalanche breakdown voltage. (*b*) Find the punch-through voltage. (*c*) What is the expected B–E avalanche breakdown voltage?

15.9 A silicon pnp power BJT is to be designed. The base doping concentration is $N_B = 5 \times 10^{15}$ cm^{-3}. The base–collector junction breakdown voltage is to be $BV_{CBO} = 1000$ V. Determine the maximum collector doping concentration and the minimum base and collector region widths.

15.10 (*a*) Assume that $BV_{CBO} = 300$ V for a power BJT. Determine BV_{CEO} for (*i*) $\beta = 10$ and (*ii*) $\beta = 50$. Assume $n = 3$ (see Equation (12.63)). (*b*) Repeat part (*a*) for $BV_{CBO} = 125$ V.

15.11 The effective β of a Darlington pair is found to be $\beta_{eff} = 180$. The driver BJT, Q_A, has a current gain $\beta_A = 25$. (*a*) What is β of the output transistor Q_B? (*b*) If the rated collector current of Q_B is $I_{CB,\max} = 20$ A, what must be the rated collector current of Q_A?

15.12 The maximum current, voltage, and power rating of an npn power BJT are 2 A, 120 V, and 30 W, respectively. (*a*) Sketch and label the safe operating area for this transistor using linear current and voltage scales. (*b*) Determine R_L such that the

maximum power is delivered to the load if the quiescent collector–emitter voltage is 60 V. For this value of R_L, what is the maximum current and maximum voltage? (*c*) Determine the value of R_L such that the maximum current and maximum power can be obtained. (*d*) Determine the value of R_L such that the maximum voltage and maximum power can be obtained.

15.13 The common-emitter circuit in Figure 15.15 is biased at $V_{CC} = 12$ V. The power rating of the transistor is $P_T = 10$ W. (*a*) Determine R_L such that the maximum power is delivered to the load. (*b*) What must be the current rating of the transistor, $I_{C,\max}$?

15.14 The transistor in the common-emitter circuit in Figure 15.15 has parameters $P_T = 2.5$ W, $V_{CE,\text{sus}} = 25$ V, and $I_{C,\max} = 500$ mA. Let $R_L = 100\ \Omega$. What is the value of V_{CC} such that the maximum power is delivered to the load?

Section 15.5 Power MOSFETs

15.15 A power MOSFET is used in the inverter circuit shown in Figure 15.26 in which $V_{DD} = 200$ V and $R_D = 100\ \Omega$. The on resistance of the transistor is $R_{on} = 2\ \Omega$ at a junction temperature of 25°C. The on resistance increases linearly with temperature and is 3 Ω at a junction temperature of 100°C. Plot the power dissipated in the transistor as a function of junction temperature.

15.16 Three MOSFETs are to be used in parallel to sink 5 A of load current when they are on. (*a*) The on resistances of the three devices are $R_{\text{on1}} = 1.8\ \Omega$, $R_{\text{on2}} = 2\ \Omega$, and $R_{\text{on3}} = 2.2\ \Omega$. Calculate the current in each device and the power dissipated in each device. (*b*) For some unknown reason, the on resistance of the second device increases to $R_{\text{on2}} = 3.6\ \Omega$. Recalculate the current in each device and the power dissipated in each device.

15.17 Consider a silicon DMOS power MOSFET shown in Figure 15.20. The source doping concentration is 5×10^{17} cm^{-3} and the base doping concentration is 10^{15} cm^{-3}. (*a*) Design the drain doping concentration, channel length, and drain drift region width to support a blocking voltage of 200 V. (*b*) Repeat part (*a*) such that the blocking voltage is 80 V.

15.18 A power MOSFET is connected in a common-source configuration as shown in Figure 15.26. The transistor parameters are $K_n = 0.20$ A/V^2, $V_T = 2$ V, $I_{D,\max} = 8$ A, $BV_{DSS} = 80$ V, and $P_T = 45$ W. The circuit parameters are $V_{DD} = 60$ V and $R_L = 10\ \Omega$. (*a*) Sketch and label the safe operating area for the transistor using linear current and voltage scales. Sketch the load line on the same curve. (*b*) Calculate the power dissipated in the transistor for $V_{GS} = 4$, 6, and 8 V. Is there a possibility of damaging the transistor? Explain.

15.19 Consider the power MOSFET described in Problem 15.18. (*a*) For $V_{DD} = 60$ V, determine the value of R_L such that the maximum power is delivered to the load and the transistor remains biased in the safe operating area. For this case, what is the maximum allowed drain current? (*b*) For $R_L = 10\ \Omega$, determine the maximum value V_{DD} such that the maximum power is delivered to the load and the transistor remains biased in the safe operating area.

Section 15.6 The Thyristor

15.20 One condition for switching a thyristor is that $\alpha_1 + \alpha_2 = 1$. Show that this condition corresponds to $\beta_1\beta_2 = 1$, where β_1 and β_2 are the common-emitter current gains of the pnp and npn bipolar transistors in the equivalent circuit of the thyristor.

15.21 Explain how a pulse of ionizing radiation could trigger a basic CMOS structure into a high-current, low-impedance state.

15.22 Show that the triac can be triggered into its ON state by gate signals of either polarity and with anode-to-cathode voltages of either polarity. Consider each voltage polarity combination.

READING LIST

1. Baliga, B. J. *Modern Power Devices*. New York: John Wiley and Sons, 1987.

2. _____. *Power Semiconductor Devices*. Boston: PWS Publishing, 1996.

3. Dimitrijev, S. *Principles of Semiconductor Devices*. New York: Oxford University Press, 2006.

4. Esaki, L. "Discovery of the Tunnel Diode." *IEEE Trans. Elec. Dev.,* ED-23 (1976).

5. Fisher, M. J. *Power Electronics*. Boston: PWS-Kent Publishing, 1991.

6. Gentry, F. E., F. W. Gutzwiller, N. Holonyak, Jr., and E. E. Von Zastrow. *Semiconductor Controlled Rectifiers: Principles and Applications of pnpn Devices*. Englewood Cliffs, NJ: Prentice Hall, 1964.

7. Ghandhi, S. K. *Semiconductor Power Devices: Physics of Operation and Fabrication Technology*. New York: John Wiley and Sons, 1977.

8. Gunn, J. B. "Microwave Oscillations of Current in III-V Semiconductors." *Solid State Comm*., 1 (1963).

9. Oxner, E. S. *Power FETs and Their Applications*. Englewood Cliffs, NJ: Prentice Hall, 1982.

10. Read, W. T. "A Proposed High Frequency, Negative Resistance Diode." *Bell Syst. Tech. J.*, 37 (1958).

11. Ridley, B. K., and T. B. Watkins. "The Possibility of Negative Resistance Effects in Semiconductors." *Proc. Phys. Soc. Lond.*, 78 (1961).

12. Roulston, D. J. *Bipolar Semiconductor Devices*. New York: McGraw-Hill, 1990.

13. Schroder, D. K. *Advanced MOS Devices: Modular Series on Solid State Devices*. Reading, MA: Addison-Wesley, 1987.

14. Shur, M. *Introduction to Electronic Devices*. New York: John Wiley and Sons, 1996.

15. Streetman, B. G., and S. K. Banerjee. *Solid State Electronic Devices*, 6th ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2006.

16. Sze, S. M. *Semiconductor Devices: Physics and Technology*. New York: John Wiley and Sons, 1985.

17. Sze, S. M. and K. K. Ng. *Physics of Semiconductor Devices*, 3rd ed. Hoboken, NJ: John Wiley and Sons, 2007.

***18.** Wang, S. *Fundamentals of Semiconductor Theory and Device Physics*. Englewood Cliffs, NJ: Prentice Hall, 1989.

19. Yang, E. S. *Microelectronic Devices*. New York: McGraw-Hill, 1988.

*Indicates reference that is at an advanced level compared to this text.

APPENDIX A

Selected List of Symbols

This list does not include some symbols that are defined and used specifically in only one section. Some symbols have more than one meaning; however, the context in which the symbol is used should make the meaning unambiguous. The usual unit associated with each symbol is given.

a	Unit cell dimension (Å), potential well width, acceleration, gradient of impurity concentration, channel thickness of a one-sided JFET (cm)
a_0	Bohr radius (Å)
c	Speed of light (cm/s)
d	Distance (cm)
e	Electronic charge (magnitude) (C), Napierian base
f	Frequency (Hz)
$f_F(E)$	Fermi–Dirac probability function
f_T	Cutoff frequency (Hz)
g	Generation rate ($cm^{-3}\ s^{-1}$)
g'	Generation rate of excess carriers ($cm^{-3}\ s^{-1}$)
$g(E)$	Density of states function ($cm^{-3}\ eV^{-1}$)
g_c, g_v	Density of states function in the conduction band and valence band ($cm^{-3}\ eV^{-1}$)
g_d	Channel conductance (S), small-signal diffusion conductance (S)
g_m	Transconductance (A/V)
g_n, g_p	Generation rate for electrons and holes ($cm^{-3}\ s^{-1}$)
h	Planck's constant (J-s), induced space charge width in a JFET (cm)
$\hbar$	Modified Planck's constant ($h/2\pi$)
h_f	Small-signal common-emitter current gain

j	Imaginary constant, $\sqrt{-1}$
k	Boltzmann's constant (J/K), wavenumber (cm^{-1})
k_n	Conduction parameter (A/V^2)
m	Mass (kg)
m_0	Rest mass of the electron (kg)
m^*	Effective mass (kg)
m_{cn}^*, m_{cp}^*	Conductivity effective mass of electron and hole (kg)
m_{dn}^*, m_{dp}^*	Density of states effective mass of electron and hole (kg)
m_n^*, m_p^*	Effective mass of electron and hole (kg)
n	Integer
n, l, m, s	Quantum numbers
n, p	Electron and hole concentration (cm^{-3})
$\overline{n}$	Index of refraction
n', p'	Constants related to the trap energy (cm^{-3})
n_{B0}, p_{E0}, p_{C0}	Thermal-equilibrium minority carrier electron concentration in the base and minority carrier hole concentration in the emitter and collector (cm^{-3})
n_d	Density of electrons in the donor energy level (cm^{-3})
n_i	Intrinsic concentration of electrons (cm^{-3})
n_0, p_0	Thermal-equilibrium concentration of electrons and holes (cm^{-3})
n_p, p_n	Minority carrier electron and minority carrier hole concentration (cm^{-3})
n_{p0}, p_{n0}	Thermal-equilibrium minority carrier electron and minority carrier hole concentration (cm^{-3})
n_s	Density of a two-dimensional electron gas (cm^{-2})
p	Momentum
p_a	Density of holes in the acceptor energy level (cm^{-3})
p_i	Intrinsic hole concentration ($= n_i$)(cm^{-3})
q	Charge (C)
r, θ, ϕ	Spherical coordinates
r_d, r_π	Small-signal diffusion resistance (Ω)
r_{ds}	Small-signal drain-to-source resistance (Ω)
r_o	Output resistance (Ω)
s	Surface recombination velocity (cm/s)
t	Time (s)
t_d	Delay time (s)
t_{ox}	Gate oxide thickness (cm or Å)
t_s	Storage time (s)

$u(x)$	Periodic wave function
v	Velocity (cm/s)
v_d	Carrier drift velocity (cm/s)
v_{ds}, v_s, v_{sat}	Carrier saturation drift velocity (cm/s)
x, y, z	Cartesian coordinates
x	Mole fraction in compound semiconductors
x_B, x_E, x_C	Neutral base, emitter, and collector region widths (cm)
x_d	Induced space charge width (cm)
x_{dB}, x_{dC}	Space charge width in base and collector (cm)
x_{BO}	Metallurgical base width (cm)
x_{dT}	Maximum space charge width (cm)
x_n, x_p	Depletion width from the metallurgical junction into n-type and p-type semiconductor regions (cm)
A	Area (cm^2)
A^*	Effective Richardson constant ($A/K^2/cm^2$)
B	Magnetic flux density (Wb/m^2)
B, E, C	Base, emitter, and collector
BV_{CBO}	Breakdown voltage of collector–base junction with emitter open (V)
BV_{CEO}	Breakdown voltage of collector–emitter with base open (V)
C	Capacitance (F)
C'	Capacitance per unit area (F/cm^2)
C_d, C_π	Diffusion capacitance (F)
C_{FB}	Flat-band capacitance (F)
C_{gs}, C_{gd}, C_{ds}	Gate-source, gate-drain, and drain-source capacitance (F)
C'_j	Junction capacitance per unit area (F/cm^2)
C_M	Miller capacitance (F)
C_n, C_p	Constants related to capture rate of electrons and holes
C_{ox}	Gate oxide capacitance per unit area (F/cm^2)
C_μ	Reverse-biased B–C junction capacitance (F)
D, S, G	Drain, source, and gate of an FET
D'	Ambipolar diffusion coefficient (cm^2/s)
D_B, D_E, D_C	Base, emitter, and collector minority carrier diffusion coefficients (cm^2/s)
D_{it}	Density of interface states ($\#/eV\text{-}cm^3$)
D_n, D_p	Minority carrier electron and minority carrier hole diffusion coefficient (cm^2/s)
E	Energy (J or eV)
E_a	Acceptor energy level (eV)

E_c, E_v	Energy at the bottom edge of the conduction band and top edge of the valence band (eV)
ΔE_c, ΔE_v	Difference in conduction band energies and valence band energies at a heterojunction (eV)
E_d	Donor energy level (eV)
E_F	Fermi energy (eV)
E_{Fi}	Intrinsic Fermi energy (eV)
E_{Fn}, E_{Fp}	Quasi-Fermi energy levels for electrons and holes (eV)
E_g	Bandgap energy (eV)
ΔE_g	Bandgap narrowing factor (eV), difference in bandgap energies at a heterojunction (eV)
E_t	Trap energy level (eV)
F	Force (N)
F_n^-, F_p^+	Electron and hole particle flux ($cm^{-2}\ s^{-1}$)
$F_{1/2}(\eta)$	Fermi–Dirac integral function
G	Generation rate of electron–hole pairs ($cm^{-3}\ s^{-1}$)
G_L	Excess carrier generation rate ($cm^{-3}\ s^{-1}$)
G_{n0}, G_{p0}	Thermal-equilibrium generation rate for electrons and holes ($cm^{-3}\ s^{-1}$)
G_{01}	Conductance (S)
I	Current (A)
I_b, I_e, I_c	Small-signal base, emitter, and collector currents (A)
I_A	Anode current (A)
I_B, I_E, I_C	Base, emitter, and collector current (A)
I_{CBO}	Reverse-biased collector–base junction current with emitter open (A)
I_{CEO}	Reverse-biased collector–emitter current with base open (A)
I_D	Diode current (A), drain current (A)
I_D (sat)	Saturation drain current (A)
I_L	Photocurrent (A)
I_{P1}	Pinchoff current (A)
I_S	Ideal reverse-biased saturation current (A)
I_{SC}	Short-circuit current (A)
I_v	Photon intensity (energy/cm^2/s)
J	Electric current density (A/cm^2)
J_{gen}	Generation current density (A/cm^2)
J_L	Photocurrent density (A/cm^2)
J_n, J_p	Electron and hole electric current density (A/cm^2)
J_n^-, J_p^+	Electron and hole particle current density ($cm^{-2}\ s^{-1}$)

J_{rec}	Recombination current density (A/cm^2)
J_{r0}	Zero-bias recombination current density (A/cm^2)
J_R	Reverse-biased current density (A/cm^2)
J_S	Ideal reverse-biased saturation current density (A/cm^2)
J_{sT}	Ideal reverse-saturation current density in a Schottky diode (A/cm^2)
L	Length (cm), inductance (H), channel length (cm)
ΔL	Channel length modulation factor (cm)
L_B, L_E, L_C	Minority carrier diffusion length in the base, emitter, and collector (cm)
L_D	Debye length (cm)
L_n, L_p	Minority carrier electron and hole diffusion length (cm)
M, M_n	Multiplication constant
N	Number density (cm^{-3})
N_a	Density of acceptor impurity atoms (cm^{-3})
N_B, N_E, N_C	Base, emitter, and collector doping concentrations (cm^{-3})
N_c, N_v	Effective density of states function in the conduction band and valence band (cm^{-3})
N_d	Density of donor impurity atoms (cm^{-3})
N_{it}	Interface state density (cm^{-2})
N_t	Trap density (cm^{-3})
P	Power (W)
$P(r)$	Probability density function
Q	Charge (C)
Q'	Charge per unit area (C/cm^2)
Q_B	Gate-controlled bulk charge (C)
Q'_n	Inversion channel charge density per unit area (C/cm^2)
Q'_{sig}	Signal charge density per unit area (C/cm^2)
Q'_{SD} (max)	Maximum space charge density per unit area (C/cm^2)
Q'_{SS}	Equivalent trapped oxide charge per unit area (C/cm^2)
R	Reflection coefficient, recombination rate (cm^{-3} s^{-1}), resistance (Ω)
$R(r)$	Radial wave function
R_c	Specific contact resistance (Ω-cm^2)
R_{cn}, R_{cp}	Capture rate for electrons and holes (cm^{-3} s^{-1})
R_{en}, R_{ep}	Emission rate for electrons and holes (cm^{-3} s^{-1})
R_n, R_p	Recombination rate for electrons and holes (cm^{-3} s^{-1})
R_{n0}, R_{p0}	Thermal-equilibrium recombination rate of electrons and holes (cm^{-3} s^{-1})

T	Temperature (K), kinetic energy (J or eV), transmission coefficient
V	Potential (V), potential energy (J or eV)
V_a	Applied forward-bias voltage (V)
V_A	Early voltage (V), anode voltage (V)
V_{bi}	Built-in potential barrier (V)
V_B	Breakdown voltage (V)
V_{BD}	Breakdown voltage at the drain (V)
V_{BE}, V_{CB}, V_{CE}	Base–emitter, collector–base, and collector–emitter voltage (V)
V_{DS}, V_{GS}	Drain-source and gate-source voltage (V)
V_{DS}(sat)	Drain-source saturation voltage (V)
V_{FB}	Flat-band voltage (V)
V_G	Gate voltage (V)
V_H	Hall voltage (V)
V_{oc}	Open-circuit voltage (V)
V_{ox}	Potential difference across an oxide (V)
V_{p0}	Pinchoff voltage (V)
V_{pt}	Punch-through voltage (V)
V_R	Applied reverse-biased voltage (V)
V_{SB}	Source-body voltage (V)
V_t	Thermal voltage (kT/e)
V_T	Threshold voltage (V)
ΔV_T	Threshold voltage shift (V)
W	Total space charge width (cm), channel width (cm)
W_B	Metallurgical base width (cm)
Y	Admittance
α	Photon absorption coefficient (cm^{-1}), ac common-base current gain
α_n, α_p	Electron and hole ionization rates (cm^{-1})
α_0	dc common-base current gain
α_T	Base transport factor
β	Common-emitter current gain
γ	Emitter injection efficiency factor
δ	Recombination factor
δn, δp	Excess electron and hole concentration (cm^{-3})
δn_p, δp_n	Excess minority carrier electron and excess minority carrier hole concentration (cm^{-3})
ϵ	Permittivity (F/cm^2)

ϵ_0	Permittivity of free space (F/cm^2)
ϵ_{ox}	Permittivity of an oxide (F/cm^2)
ϵ_r	Relative permittivity or dielectric constant
ϵ_s	Permittivity of a semiconductor (F/cm^2)
λ	Wavelength (cm or μm)
μ	Permeability (H/cm)
μ'	Ambipolar mobility (cm^2/V-s)
μ_n, μ_p	Electron and hole mobility (cm^2/V-s)
μ_0	Permeability of free space (H/cm)
ν	Frequency (Hz)
ρ	Resistivity (Ω-cm), volume charge density (C/cm^3)
σ	Conductivity (Ω^{-1} cm^{-1})
$\Delta\sigma$	Photoconductivity (Ω^{-1} cm^{-1})
σ_i	Intrinsic conductivity (Ω^{-1} cm^{-1})
σ_n, σ_p	Conductivity of n-type and p-type semiconductors (Ω^{-1} cm^{-1})
τ	Lifetime (s)
τ_n, τ_p	Electron and hole lifetime (s)
τ_{n0}, τ_{p0}	Excess minority carrier electron and hole lifetime (s)
τ_0	Lifetime in space charge region (s)
ϕ	Potential (V)
$\phi(t)$	Time-dependent wave function
$\Delta\phi$	Schottky barrier lowering potential (V)
ϕ_{Bn}	Schottky barrier height (V)
ϕ_{B0}	Ideal Schottky barrier height (V)
ϕ_{fn}, ϕ_{fp}	Potential difference (magnitude) between E_{Fi} and E_F in n-type and p-type semiconductors (V)
ϕ_{Fn}, ϕ_{Fp}	Potential difference (with sign) between E_{Fi} and E_F in n-type and p-type semiconductors (V)
ϕ_m	Metal work function (V)
ϕ'_m	Modified metal work function (V)
ϕ_{ms}	Metal–semiconductor work function difference (V)
ϕ_n, ϕ_p	Potential difference (magnitude) between E_c and E_F in n-type and between E_v and E_F in p-type semiconductor (V)
ϕ_s	Semiconductor work function (V), surface potential (V)
χ	Electron affinity (V)
χ'	Modified electron affinity (V)
$\psi(x)$	Time-independent wave function

ω	Radian frequency (s^{-1})
Γ	Reflection coefficient
E	Electric field (V/cm)
E_H	Hall electric field (V/cm)
E_{crit}	Critical electric field at breakdown (V/cm)
$\Theta(\theta)$	Angular wave function
Φ	Photon flux ($cm^{-2}\ s^{-1}$)
$\Phi(\phi)$	Angular wave function
$\Psi(x, t)$	Total wave function

APPENDIX B

System of Units, Conversion Factors, and General Constants

Table B.1 | International system of units*

Quantity	Unit	Symbol	Dimension
Length	meter	m	
Mass	kilogram	kg	
Time	second	s or sec	
Temperature	kelvin	K	
Current	ampere	A	
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 flux density	tesla	T	Wb/m^2
Inductance	henry	H	Wb/A

*The centimeter is the common unit of length and the electron-volt is the common unit of energy (see Appendix D) used in the study of semiconductors. However, the joule and in some cases the meter should be used in most formulas.

Table B.2 | Conversion factors

	Prefixes		
1 Å (angstrom) $= 10^{-8}$ cm $= 10^{-10}$ m	10^{-15}	femto-	= f
1 μm (micrometer) $= 10^{-4}$ cm	10^{-12}	pico-	= p
1 mil $= 10^{-3}$ in. $= 25.4$ μm	10^{-9}	nano-	= n
2.54 cm $=$ 1 in.	10^{-6}	micro-	= μ
1 eV $= 1.6 \times 10^{-19}$ J	10^{-3}	milli-	= m
1 J $= 10^{7}$ erg	10^{+3}	kilo-	= k
	10^{+6}	mega-	= M
	10^{+9}	giga-	= G
	10^{+12}	tera	= T

Table B.3 | Physical constants

Avogadro's number	$N_A = 6.02 \times 10^{+23}$ atoms per gram molecular weight
Boltzmann's constant	$k = 1.38 \times 10^{-23}$ J/K $= 8.62 \times 10^{-5}$ eV/K
Electronic charge (magnitude)	$e = 1.60 \times 10^{-19}$ C
Free electron rest mass	$m_0 = 9.11 \times 10^{-31}$ kg
Permeability of free space	$\mu_0 = 4\pi \times 10^{-7}$ H/m
Permittivity of free space	$\epsilon_0 = 8.85 \times 10^{-14}$ F/cm $= 8.85 \times 10^{-12}$ F/m
Planck's constant	$h = 6.625 \times 10^{-34}$ J-s $= 4.135 \times 10^{-15}$ eV-s
	$\frac{h}{2\pi} = \hbar = 1.054 \times 10^{-34}$ J-s
Proton rest mass	$M = 1.67 \times 10^{-27}$ kg
Speed of light in vacuum	$c = 2.998 \times 10^{10}$ cm/s
Thermal voltage ($T = 300$ K)	$V_t = \frac{kT}{e} = 0.0259$ V
	$kT = 0.0259$ eV

Table B.4 | Silicon, gallium arsenide, and germanium properties ($T = 300$ K)

Property	Si	GaAs	Ge
Atoms (cm^{-3})	5.0×10^{22}	4.42×10^{22}	4.42×10^{22}
Atomic weight	28.09	144.63	72.60
Crystal structure	Diamond	Zincblende	Diamond
Density (g/cm^3)	2.33	5.32	5.33
Lattice constant (Å)	5.43	5.65	5.65
Melting point (°C)	1415	1238	937
Dielectric constant	11.7	13.1	16.0
Bandgap energy (eV)	1.12	1.42	0.66
Electron affinity, χ (V)	4.01	4.07	4.13
Effective density of states in conduction band, N_c (cm^{-3})	2.8×10^{19}	4.7×10^{17}	1.04×10^{19}
Effective density of states in valence band, N_v (cm^{-3})	1.04×10^{19}	7.0×10^{18}	6.0×10^{18}
Intrinsic carrier concentration (cm^{-3})	1.5×10^{10}	1.8×10^{6}	2.4×10^{13}
Mobility (cm^2/V-s)			
Electron, μ_n	1350	8500	3900
Hole, μ_p	480	400	1900
Effective mass $\left(\frac{m^*}{m_0}\right)$			
Electrons	$m_l^* = 0.98$	0.067	1.64
	$m_t^* = 0.19$		0.082
Holes	$m_{lh}^* = 0.16$	0.082	0.044
	$m_{hh}^* = 0.49$	0.45	0.28
Density of states effective mass			
Electrons $\left(\frac{m_{dn}^*}{m_o}\right)$	1.08	0.067	0.55
Holes $\left(\frac{m_{dp}^*}{m_o}\right)$	0.56	0.48	0.37
Conductivity effective mass			
Electrons $\left(\frac{m_{cn}^*}{m_o}\right)$	0.26	0.067	0.12
Holes $\left(\frac{m_{cp}^*}{m_o}\right)$	0.37	0.34	0.21

Table B.5 | Other semiconductor parameters

Material	E_g (eV)	a (Å)	ϵ_r	χ	$\overline{n}$
Aluminum arsenide	2.16	5.66	12.0	3.5	2.97
Gallium phosphide	2.26	5.45	10	4.3	3.37
Aluminum phosphide	2.43	5.46	9.8		3.0
Indium phosphide	1.35	5.87	12.1	4.35	3.37

Table B.6 | Properties of SiO_2 and Si_3N_4 ($T = 300$ K)

Property	SiO_2	Si_3N_4
Crystal structure	[Amorphous for most integrated circuit applications]	
Atomic or molecular density (cm^{-3})	2.2×10^{22}	1.48×10^{22}
Density (g/cm^3)	2.2	3.4
Energy gap	≈ 9 eV	4.7 eV
Dielectric constant	3.9	7.5
Melting point (°C)	≈ 1700	≈ 1900

APPENDIX C

The Periodic Table

Period	Group I a	Group I b	Group II a	Group II b	Group III a	Group III b	Group IV a	Group IV b	Group V a	Group V b	Group VI a	Group VI b	Group VII a	Group VII b	Group VIII a	Group VIII	Group VIII	Group VIII b
I	1 H 1.0079																	2 He 4.003
II	3 Li 6.94		4 Be 9.02		5 B 10.82		6 C 12.01		7 N 14.01		8 O 16.00		9 F 19.00					10 Ne 20.18
III	11 Na 22.99		12 Mg 24.32		13 Al 26.97		14 Si 28.06		15 P 30.98		16 S 32.06		17 Cl 35.45					18 Ar 39.94
IV	19 K 39.09		20 Ca 40.08		21 Sc 44.96		22 Ti 47.90		23 V 50.95		24 Cr 52.01		25 Mn 54.93		26 Fe 55.85	27 Co 58.94	28 Ni 58.69	
		29 Cu 63.54		30 Zn 65.38		31 Ga 69.72		32 Ge 72.60		33 As 74.91		34 Se 78.96		35 Br 79.91				36 Kr 83.7
V	37 Rb 85.48		38 Sr 87.63		39 Y 88.92		40 Zr 91.22		41 Nb 92.91		42 Mo 95.95		43 Tc 99		44 Ru 101.7	45 Rh 102.91	46 Pd 106.4	
		47 Ag 107.88		48 Cd 112.41		49 In 114.76		50 Sn 118.70		51 Sb 121.76		52 Te 127.61		53 I 126.92				54 Xe 131.3
VI	55 Cs 132.91		56 Ba 137.36		57–71 *Rare earths*		72 Hf 178.6		73 Ta 180.88		74 W 183.92		75 Re 186.31		76 Os 190.2	77 Ir 193.1	28 Pt 195.2	
		79 Au 197.2		80 Hg 200.61		81 Tl 204.39		82 Pb 207.21		83 Bi 209.00		84 Po 210		85 At 211				86 Rn 222
VII	87 Fr 223		88 Ra 226.05		89 Ac 227		90 Th 232.12		91 Pa 231		92 U 238.07	93 Np 237	94 Pu 239	95 Am 241	96 Cm 242	97 Bk 246	98 Ct 249	99 Es 254	100 Fm 256	101 Md 256

Rare earths

VI	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu
57-71	138.92	140.13	140.92	144.27	147	150.43	152.0	156.9	159.2	162.46	164.90	167.2	169.4	173.04	174.99

The numbers in front of the symbols of the elements denote the atomic numbers; the numbers underneath are the atomic weights.

APPENDIX D

Unit of Energy—The Electron Volt

The electron volt (eV) is a unit of energy that is used constantly in the study of semiconductor physics and devices. This short discussion may help in "getting a feel" for the electron-volt.

Consider a parallel-plate capacitor with an applied voltage as shown in Figure D.1. Assume that an electron is released at $x = 0$ at time $t = 0$. We may write

$$F = m_0 a = m_0 \frac{d^2x}{dt^2} = e\mathrm{E} \tag{D.1}$$

where e is the magnitude of the electronic charge and E is the magnitude of the electric field as shown. Upon integrating, the velocity and distance versus time are given by

$$v = \frac{e\mathrm{E}t}{m_0} \tag{D.2}$$

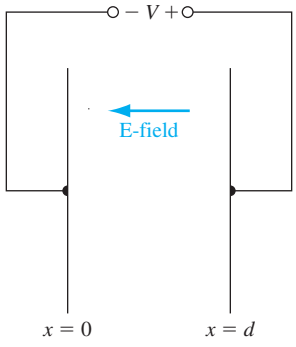


Figure D.1 | Parallel-plate capacitor.

and

$$x = \frac{e\mathrm{E}t^2}{2m_0} \tag{D.3}$$

where we have assumed that $v = 0$ at $t = 0$.

Assume that at $t = t_0$ the electron reaches the positive plate of the capacitor so that $x = d$. Then

$$d = \frac{e\mathrm{E}t_0^2}{2m_0} \tag{D.4a}$$

or

$$t_0 = \sqrt{\frac{2m_0 d}{e\mathrm{E}}} \tag{D.4b}$$

The velocity of the electron when it reaches the positive plate of the capacitor is

$$v(t_0) = \frac{e\mathrm{E}t_0}{m_0} = \sqrt{\frac{2e\mathrm{E}d}{m_0}} \tag{D.5}$$

The kinetic energy of the electron at this time is

$$T = \frac{1}{2} m_0 v(t_0)^2 = \frac{1}{2} m_0 \left(\frac{2e\mathrm{E}d}{m_0}\right) = e\mathrm{E}d \tag{D.6}$$

The electric field is

$$\mathrm{E} = \frac{\mathrm{V}}{d} \tag{D.7}$$

so that the energy is

$$T = e \cdot \mathrm{V} \tag{D.8}$$

If an electron is accelerated through a potential of 1 V, then the energy is

$$T = e \cdot \mathrm{V} = (1.6 \times 10^{-19})(1) = 1.6 \times 10^{-19} \text{ joule (J)} \tag{D.9}$$

The electron-volt (eV) unit of energy is defined as

$$\text{Electron-volt} = \frac{\text{joule}}{e} \tag{D.10}$$

Then, the electron that is accelerated through a potential of 1 V will have an energy of

$$T = 1.6 \times 10^{-19}\,\mathrm{J} = \frac{1.6 \times 10^{-19}}{1.6 \times 10^{-19}}(\mathrm{eV}) \tag{D.11}$$

or 1 eV.

We may note that the magnitude of the potential (1 V) and the magnitude of the electron energy (1 eV) are the same. However, it is important to keep in mind that the unit associated with each number is different.

APPENDIX E

"Derivation" of Schrodinger's Wave Equation

Schrodinger's wave equation is stated in Equation (2.6). The time-independent form of Schrodinger's wave equation is then developed and given by Equation (2.13). The time-independent Schrodinger's wave equation can also be developed from the classical wave equation. We may think of this development more in terms of a justification of the Schrodinger's time-independent wave equation rather than a strict derivation.

The time-independent classical wave equation, in terms of voltage, is given as

$$\frac{\partial^2 V(x)}{\partial x^2} + \left(\frac{\omega^2}{v_p^2}\right) V(x) = 0 \tag{E.1}$$

where ω is the radian frequency and v_p is the phase velocity.

If we make a change of variable and let $\psi(x) = V(x)$, then we have

$$\frac{\partial^2 \psi(x)}{\partial x^2} + \left(\frac{\omega^2}{v_p^2}\right) \psi(x) = 0 \tag{E.2}$$

We can write that

$$\frac{\omega^2}{v_p^2} = \left(\frac{2\pi\nu}{v_p}\right)^2 = \left(\frac{2\pi}{\lambda}\right)^2 \tag{E.3}$$

where ν and λ are the wave frequency and wavelength, respectively.

From the wave–particle duality principle, we can relate the wavelength and momentum as

$$\lambda = \frac{h}{p} \tag{E.4}$$

Then

$$\left(\frac{2\pi}{\lambda}\right)^2 = \left(\frac{2\pi}{h} \cdot p\right)^2 \tag{E.5}$$

and since $\hbar = \frac{h}{2\pi}$, we can write

$$\left(\frac{2\pi}{\lambda}\right)^2 = \left(\frac{p}{\hbar}\right)^2 = \frac{2m}{\hbar^2}\left(\frac{p^2}{2m}\right) \tag{E.6}$$

Now

$$\frac{p^2}{2m} = T = E - V \tag{E.7}$$

where T, E, and V are the kinetic energy, total energy, and potential energy terms, respectively.

We can then write

$$\frac{\omega^2}{v_p^2} = \left(\frac{2\pi}{\lambda}\right)^2 = \frac{2m}{\hbar^2}\left(\frac{p^2}{2m}\right) = \frac{2m}{\hbar^2}(E - V) \tag{E.8}$$

Substituting Equation (E.8) into Equation (E.2), we have

$$\frac{\partial^2\psi(x)}{\partial x^2} + \frac{2m}{\hbar^2}(E - V)\psi(x) = 0 \tag{E.9}$$

which is the one-dimensional, time-independent Schrodinger's wave equation.

APPENDIX F

Effective Mass Concepts

In Chapter 3, we have discussed the relationship between the effective masses of electrons and holes and the E versus k diagrams. In that discussion, we have limited ourselves to a one-dimensional analysis in k space.

F.1 | ENERGY-BAND STRUCTURES

GaAs Energy Bands: The E versus k diagram for GaAs is given in Figure 3.25a. The minimum conduction band energy and the maximum valence band energy occur at $k = 0$. In a three-dimensional $k_x - k_y - k_z$ coordinate system, the constant energy surface near the minimum conduction band energy is essentially spherical as shown in Figure F.1. The electron effective mass can be determined as previously discussed and is found to be $m_n^* = 0.067m_o$, where m_o is the rest mass of an electron.

Silicon Conduction Energy Band: The E versus k diagram for silicon is given in Figure 3.25b. The minimum conduction band energy is in the [100] direction. The constant energy surface near the minimum conduction band energy in the three-dimensional $k_x - k_y - k_z$ coordinate system is approximately an ellipsoid. There are actually six ellipsoid energy surfaces corresponding to the six equivalent [100] directions in the crystal as shown in Figure F.2a. The effective mass in both the k_x and k_y directions is called a transverse effective mass m_t, and the effective mass in the k_z direction is called a longitudinal effective mass m_l. These effective masses are indicated in a single ellipsoid as shown in Figure F.2b. The values of these effective masses in silicon are found to be $m_t = 0.19m_o$ and $m_l = 0.98m_o$.

Electrons are continually undergoing random scattering effects (see Chapter 5) so that, at any given time, one-third of the electrons are moving in the k_x direction with an effective mass m_t, one-third of the electrons are moving in the k_y direction with an effective mass m_t, and one-third of the electrons are moving in the k_z direction with an effective mass m_l. The effective mass parameter in the density of states function and the effective mass parameter in conductivity calculations must therefore involve some type of averaging of the transverse and longitudinal effective masses.

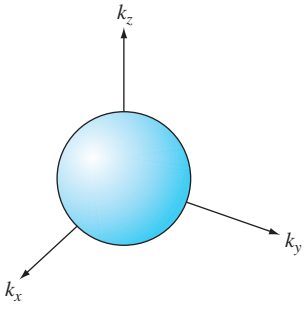


Figure F.1 | Spherical constant energy surface in the conduction band of GaAs.

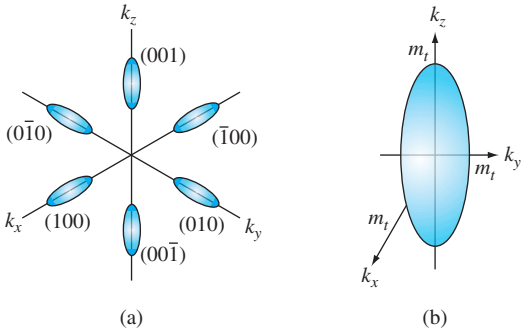


Figure F.2 | (a) Six equivalent ellipsoidal constant energy surfaces in the conduction band of silicon. (b) A single ellipsoidal energy surface showing the effective masses.

Silicon Valence Energy Band: The maximum energy of the valence band in silicon occurs at $k = 0$. The valence band actually has two branches (not shown in Figure 3.25b) with approximately parabolic shapes. The sharper parabola (larger d^2E/dk^2) corresponds to light holes and the wider parabola (smaller d^2E/dk^2) corresponds to heavy holes. The effective masses of the light and heavy holes in silicon are $m_{lh} = 0.16m_o$ and $m_{hh} = 0.49m_o$, respectively.

F.2 | DENSITY OF STATES EFFECTIVE MASS

Density of States Effective Mass—Electrons: The kinetic energy of an electron, corresponding to a constant ellipsoidal energy surface in silicon (see Figure F.2b) can be written as

$$E = \frac{p_x^2}{2m_t} + \frac{p_y^2}{2m_t} + \frac{p_z^2}{2m_l}$$

or

$$1 = \frac{p_x^2}{2m_t E} + \frac{p_y^2}{2m_t E} + \frac{p_z^2}{2m_l E}$$

The general equation of an ellipsoid in momentum space can be written as

$$1 = \frac{p_x^2}{a^2} + \frac{p_y^2}{b^2} + \frac{p_z^2}{c^2}$$

where a, b, and c are the axes of the ellipsoid. For the energy ellipsoid in Figure F.2b, we can write

$$a^2 = 2m_t E,\ b^2 = 2m_t E,\ c^2 = 2m_l E$$

The volume of an ellipsoid is proportional to the product $a \cdot b \cdot c$, so we have

$$\text{Volume} \propto \sqrt{(m_t)^2 m_l}$$

There are six energy ellipsoids, so the total volume is proportional to

$$\text{Total volume} \propto 6\sqrt{(m_t)^2 m_l}$$

In the derivation of the density of states function in the conduction band, the volume in k space (momentum space) is included. So, from Equation (3.72), the density of states function is proportional to

$$g_c(E) \propto \text{Volume} \propto (m_{dn}^*)^{3/2} = 6\sqrt{(m_t)^2 m_l}$$

The density of states function electron effective mass can then be written as

$$m_{dn}^* = 6^{2/3}[(m_t)^2 m_l]^{1/3}$$

For silicon, we have $m_t = 0.19m_o$ and $m_l = 0.98m_o$. Then, we find

$$m_{dn}^* = 6^{2/3}[(0.19m_o)^2(0.98m_o)]^{1/3} = 1.08m_o$$

where m_{dn}^* is the density of states electron effective mass.

Density of States Effective Mass—Holes: In the three-dimensional $k_x - k_y - k_z$ coordinate system, the constant energy is essentially spherical for both the heavy and light holes. The volume of a sphere in momentum space is

$$\text{Volume} \propto p^3$$

where, for the heavy and light holes, respectively, we have

$$p_{hh}^2 = 2m_{hh}E \quad \text{and} \quad p_{lh}^2 = 2m_{lh}E$$

The total volume is the sum of the two spherical volumes, so that

$$\text{Total volume} \propto (m_{hh})^{3/2} + (m_{lh})^{3/2}$$

In the derivation of the density of states function in the valence band, the volume in k space (momentum space) is included. So, from Equation (3.75), the density of states function for holes is proportional to

$$g_v(E) \propto \text{Volume} \propto (m_{dp}^*)^{3/2} = (m_{hh})^{3/2} + (m_{lh})^{3/2}$$

The density of states function effective hole mass is then

$$m_{dp}^* = [(m_{hh})^{3/2} + (m_{lh})^{3.2}]^{2/3}$$

For silicon, we have $m_{hh} = 0.49m_o$ and $m_{lh} = 0.16m_o$, so that

$$m_{dp}^* = [(0.49m_o)^{3.2} + (0.16m_o)^{3/2}]^{2/3} = 0.55m_o$$

where m_{dp}^* is the density of states hole effective mass.

F.3 | CONDUCTIVITY EFFECTIVE MASS

Conductivity Effective Mass—Electrons: From Chapter 5, the average drift velocity of a carrier due to an applied electric field is given by

$$\langle v_d \rangle = \frac{1}{2}\left(\frac{e\tau_c}{m_c^*}\right) \cdot \mathrm{E}$$

where τ_c is the mean time between collisions, E is the electric field, and m_c^* is now the conductivity effective mass.

For a simple electron gas, the electron kinetic energy can be written as

$$E = \frac{1}{2}m^*(\langle v_d \rangle)^2 = \frac{p_x^2}{2m_{cn}^*} + \frac{p_y^2}{2m_{cn}^*} + \frac{p_z^2}{2m_{cn}^*}$$

For the case of silicon and the ellipsoid energy surface, we have

$$E = \frac{p_x^2}{2m_t} + \frac{p_y^2}{2m_t} + \frac{p_z^2}{2m_l}$$

The two expressions for kinetic energy are equal if

$$3\left(\frac{1}{2m_{cn}^*}\right) = \frac{2}{2m_t} + \frac{1}{2m_l}$$

or

$$\frac{3}{m_{cn}^*} = \frac{2}{m_t} + \frac{1}{m_l}$$

Again, for electrons in silicon, we have $m_t = 0.19m_o$ and $m_l = 0.98m_o$. Then

$$\frac{3}{m_{cn}^*} = \frac{2}{0.19m_o} + \frac{1}{0.98m_o}$$

which gives $m_{cn}^* = 0.26m_o$, where m_{cn}^* is the conductivity effective mass for electrons.

Conductivity Effective Mass—Holes: From Chapter 5, the drift current density due to holes is given by

$$J = e\mu_p p\mathrm{E} = e\left(\frac{e\tau_c}{m^*}\right)p\mathrm{E}$$

Assuming the mean time between collisions is the same for heavy holes and light holes, we can write

$$J_{Total} = J_{hh} + J_{lh}$$

which can be written as

$$J_{Total} \propto e\left(\frac{e\tau_c}{m_{cp}^*}\right)(m_{dp}^*)^{3/2}$$

where p is the total hole concentration and is proportional to $(m_{dp}^*)^{3/2}$. The parameter m_{cp}^* is the conductivity effective mass for holes and m_{dp}^* is the density of states effective mass for holes.

The individual currents for heavy holes and light holes are proportional to

$$J_{hh} \propto e\left(\frac{e\tau_c}{m_{hh}}\right)(m_{hh})^{3/2} = e(e\tau_c)(m_{hh})^{1/2}$$

and

$$J_{lh} \propto e\left(\frac{e\tau_c}{m_{lh}}\right)(m_{lh})^{3/2} = e(e\tau_c)(m_{lh})^{1/2}$$

We then have

$$\frac{(m_{dp}^*)^{3/2}}{m_{cp}^*} = (m_{hh})^{1/2} + (m_{lh})^{1/2}$$

or

$$m_{cp}^* = \frac{(m_{dp}^*)^{3/2}}{(m_{hh})^{1/2} + (m_{lh})^{1/2}} = \frac{(m_{hh})^{3/2} + (m_{lh})^{3/2}}{(m_{hh})^{1/2} + (m_{lh})^{1/2}}$$

For silicon, we again have $m_{hh} = 0.49m_o$ and $m_{lh} = 0.16m_o$, so that

$$m_{cp}^* = \frac{(0.49m_o)^{3/2} + (0.16m_o)^{3/2}}{(0.49m_o)^{1/2} + (0.16m_o)^{1/2}} = 0.37m_o$$

and m_{cp}^* is the conductivity effective mass for holes.

F.4 | SUMMARY

The energy-band structure of germanium is essentially the same as silicon with four ellipsoidal energy surfaces in the conduction band and two spherical energy surfaces in the valence band corresponding to heavy and light holes. The calculations for the density of states effective masses and conductivity effective masses are then identical to those for silicon. Gallium arsenide also has two spherical energy surfaces in the valence band corresponding to heavy and light holes. So the calculations for the density of states effective mass for holes and conductivity effective mass for holes are also identical to those for silicon.

The density of states effective masses for electrons and holes are denoted as m_{dn}^* and m_{dp}^*, respectively. The conductivity effective masses for electrons and holes are denoted as m_{cn}^* and m_{cp}^*, respectively. In analyses and calculations in the text, the effective masses for electrons and holes are usually denoted simply as m_n^* and m_p^*, respectively. Whether the density of states effective mass or the conductivity effective mass is to be used should be clear from the context of the problem.

APPENDIX G

The Error Function

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt$$

$$\text{erf}(0) = 0 \qquad \text{erf}(\infty) = 1$$

$$\text{erfc}(z) = 1 - \text{erf}(z)$$

z	erf(z)	z	erf(z)
0.00	0.00000	1.00	0.84270
0.05	0.05637	1.05	0.86244
0.10	0.11246	1.10	0.88021
0.15	0.16800	1.15	0.89612
0.20	0.22270	1.20	0.91031
0.25	0.27633	1.25	0.92290
0.30	0.32863	1.30	0.93401
0.35	0.37938	1.35	0.94376
0.40	0.42839	1.40	0.95229
0.45	0.47548	1.45	0.95970
0.50	0.52050	1.50	0.96611
0.55	0.56332	1.55	0.97162
0.60	0.60386	1.60	0.97635
0.65	0.64203	1.65	0.98038
0.70	0.67780	1.70	0.98379
0.75	0.71116	1.75	0.98667
0.80	0.74210	1.80	0.98909
0.85	0.77067	1.85	0.99111
0.90	0.79691	1.90	0.99279
0.95	0.82089	1.95	0.99418
1.00	0.84270	2.00	0.99532

APPENDIX H

ANSWERS TO SELECTED PROBLEMS

Chapter 1

1.1 (*a*) 4 atoms, (*b*) 2 atoms, (*c*) 8 atoms

1.3 (*a*) 2.35 Å, (*b*) 5×10^{22} atoms/cm^3
(*c*) 2.33 gm/cm^3

1.5 (*a*) 2.447 Å, (*b*) 3.995 Å

1.7 (*a*) 3.9 Å, (*b*) 5.515 Å,
(*c*) 4.503 Å, (*d*) 9.007 Å

1.9 (*a*) 0.228 gm/cm^3, (*b*) 0.296 gm/cm^3

1.11 (*b*) $a = 2.8$ Å, (*c*) 2.28×10^{22} cm^{-3} for both Na and Cl, (*d*) 2.21 gm/cm^3

1.13 (*a*) For A and B atoms, 4.687×10^{14} cm^{-2},
(*b*) For A and B atoms, 3.315×10^{14} cm^{-2}

1.15 (*a*) (*i*) See Figure 1.10b,
(*ii*) See Figure 1.10c,
(*iii*) Same as (110) plane,
(*iv*) Intercepts at $p = 2$, $q = 3$, $s = 6$;
(*b*) Directions perpendicular to planes

1.17 (634) plane

1.19 (*a*) (*i*) 4.47×10^{14} cm^{-2}, (*ii*) 3.16×10^{14} cm^{-2},
(*iii*) 2.58×10^{14} cm^{-2};
(*b*) (*i*) 4.47×10^{14} cm^{-2}, (*ii*) 6.32×10^{14} cm^{-2},
(*iii*) 2.58×10^{14} cm^{-2};
(*c*) (*i*) 8.94×10^{14} cm^{-2}, (*ii*) 6.32×10^{14} cm^{-2},
(*iii*) 1.03×10^{15} cm^{-2}

1.21 (*a*) 1.328×10^{22} cm^{-3},
(*b*) 3.148×10^{14} cm^{-2},
(*c*) 4.74 Å, (*d*) 5.14×10^{14} cm^{-2}, 3.87 Å

1.23 1.77×10^{23} cm^{-3}

1.25 (*a*) 1.542×10^{-7}, (*b*) 2.208×10^{-5}

1.27 $d/a_o = 116$

Chapter 2

2.5 $\lambda = 0.254$ μm (gold), $\lambda = 0.654$ μm (cesium)

2.7 (*a*) (*i*) 11.2 Å, (*ii*) 3.54 Å, (*iii*) 1.12 Å; (*b*) 0.262 Å

2.9 10.3 keV

2.11 (*a*) 12.4 kV, (*b*) 0.11 Å

2.13 (*a*) (*i*) $\Delta p = 8.783 \times 10^{-26}$ kg-m/s,
(*ii*) $\Delta E = 1.31$ eV;
(*b*) (*i*) $\Delta p = 8.783 \times 10^{-26}$ kg-m/s,
(*ii*) $\Delta E = 5.55 \times 10^{-2}$ eV

2.15 (*a*) $\Delta t = 8.23 \times 10^{-16}$ s, (*b*) $\Delta p = 7.03 \times 10^{-25}$ kg-m/s

2.17 $|A| = \dfrac{1}{\sqrt{2}}$

2.19 (*a*) $P = 0.393$, (*b*) $P = 0.239$, (*c*) $P = 0.865$

2.21 (*a*) $P = 0.25$, (*b*) $P = 0.25$, (*c*) $P = 1$

2.23 (*a*) $\psi(x, t) = A \exp[-j(kx + \omega t)]$,
(*b*) $k = 8.097 \times 10^8$ m^{-1}, $\lambda = 7.76 \times 10^{-9}$ m,
$\omega = 7.586 \times 10^{13}$ rad/s

2.25 $E_1 = 6.69 \times 10^{-3}$ eV, $E_2 = 2.67 \times 10^{-2}$ eV,
$E_3 = 6.02 \times 10^{-2}$ eV

2.27 (*a*) $n = 7.688 \times 10^{29}$, (*b*) $E_{n+1} \cong 15$ mJ, (*c*) No

2.29 $\psi_1 = A \cos\left(\dfrac{\pi x}{a}\right)$, $\psi_2 = B \sin\left(\dfrac{2\pi x}{a}\right)$,
$\psi_3 = C \cos\left(\dfrac{3\pi x}{a}\right)$, $\psi_4 = D \sin\left(\dfrac{4\pi x}{a}\right)$

2.31 (*a*) $E_{n_x n_y} = \dfrac{\hbar^2}{2m}\left(\dfrac{n_x^2\pi^2}{a^2} + \dfrac{n_y^2\pi^2}{b^2}\right)$

2.33 (*a*) $\psi_1(x) = B_1 \exp(-jk_1 x)$, $k_1 = \sqrt{\dfrac{2mE}{\hbar^2}}$;
$\psi_2(x) = A_2 \exp(jk_2 x) + B_2 \exp(-jk_2 x)$,
$k_2 = \sqrt{\dfrac{2m}{\hbar^2}(E - V_O)}$
(*b*) $R = \left(\dfrac{k_2 - k_1}{k_2 + k_1}\right)^2$, $T = \dfrac{4k_1 k_2}{(k_1 + k_2)^2}$

2.35 (*a*) $T = 0.0295$, (*b*) $T = 1.24 \times 10^{-5}$, (*c*) $N = 1.357 \times 10^{10}\ \text{cm}^{-3}$

2.37 (*a*) $T = 5.875 \times 10^{-7}$, (*b*) $a = 0.842 \times 10^{-14}$ m

2.39 $T = \dfrac{4k_1 k_3}{(k_1 + k_3)^2}$

2.41 $E_1 = -13.58$ eV, $E_2 = -3.395$ eV, $E_3 = -1.51$ eV, $E_4 = -0.849$ eV

Chapter 3

3.5 (*b*) (*i*) $\alpha a = \pi$, $\alpha a = 1.729\pi$; (*ii*) $\alpha a = 2\pi$, $\alpha a = 2.617\pi$

3.9 (*a*) $\Delta E = 0.559$ eV, (*b*) $\Delta E = 2.15$ eV

3.11 (*a*) $\Delta E = 1.005$ eV, (*b*) $\Delta E = 3.635$ eV

3.13 $m^*(A) < m^*(B)$

3.15 A,B: velocity $= -x$ direction; C,D: velocity $= +x$ direction; B,C: positive mass; A,D: negative mass

3.17 A: $m^* = -0.976m_o$; B: $m^* = -0.0813m_o$

3.21 (*a*) $m_{dn}^* = 0.56m_o$, (*b*) $m_{cn}^* = 0.12m_o$

3.25 $g(E) = \dfrac{1}{\hbar\pi}\sqrt{\dfrac{2m_n^*}{E}} = \dfrac{1.055 \times 10^{18}}{\sqrt{E}}\ \text{m}^{-3}\ \text{J}^{-1}$

3.27 (*a*) (*i*) $g_v = 4.12 \times 10^{19}\ \text{cm}^{-3}$, (*ii*) $g_v = 6.34 \times 10^{19}\ \text{cm}^{-3}$; (*b*) (*i*) $g_v = 3.27 \times 10^{19}\ \text{cm}^{-3}$, (*ii*) $g_v = 5.03 \times 10^{19}\ \text{cm}^{-3}$

3.29 (*a*) 2.68, (*b*) 0.0521

3.31 (*a*) 120; (*b*) (*i*) 66, (*ii*) 495

3.33 (*a*) 0.269, (*b*) 6.69×10^{-3}, (*c*) 4.54×10^{-5}

3.35 $E_F = \dfrac{E_c + E_v}{2} = E_{midgap}$

3.37 (*a*) $E_F = 2.35$ eV, (*b*) $E_F = 5.746$ eV

3.39 (*a*) $E_1 = E_F + 4.6kT$, (*b*) $f(E_1) \cong 0.01$

3.41 (*a*) 0.00304, (*b*) 0.1496, (*c*) 0.997, (*d*) 0.50

3.43 (*a*) At $E = E_1$, $f(E) = 9.3 \times 10^{-6}$; At $E = E_2$, $1 - f(E) = 1.66 \times 10^{-19}$ (*b*) At $E = E_1$, $f(E) = 7.88 \times 10^{-18}$; At $E = E_2$, $1 - f(E) = 1.96 \times 10^{-7}$

3.45 (*a*) Si: $f(E) = 4.07 \times 10^{-10}$; Ge: $f(E) = 2.93 \times 10^{-6}$; GaAs: $f(E) = 1.24 \times 10^{-12}$; (*b*) Same values as part (a)

3.47 (*a*) $\Delta E = 0.1017$ eV, (*b*) $\Delta E = 0.2034$ eV

Chapter 4

4.1 (*a*) $n_i = 7.68 \times 10^4\ \text{cm}^{-3}$; $2.38 \times 10^{12}\ \text{cm}^{-3}$; $9.74 \times 10^{14}\ \text{cm}^{-3}$, (*b*) $n_i = 2.16 \times 10^{10}\ \text{cm}^{-3}$; $8.60 \times 10^{14}\ \text{cm}^{-3}$; $3.82 \times 10^{16}\ \text{cm}^{-3}$, (*c*) $n_i = 1.38\ \text{cm}^{-3}$; $3.28 \times 10^9\ \text{cm}^{-3}$; $5.72 \times 10^{12}\ \text{cm}^{-3}$

4.3 (*a*) $T \cong 367.5$ K, (*b*) $T \cong 417.5$ K

4.5 (*a*) 9.325×10^{-6}, (*b*) 4.43×10^{-4}, (*c*) 3.05×10^{-3}

4.7 0.0854

4.11 For $T = 200$ K, $E_{Fi} - E_{midgap} = -0.0086$ eV; For $T = 400$ K, $E_{Fi} - E_{midgap} = -0.0171$ eV; For $T = 600$ K, $E_{Fi} - E_{midgap} = -0.0257$ eV

4.13 $n_o = K \cdot kT \exp\left[\dfrac{-(E_c - E_F)}{kT}\right]$

4.15 $r_1 = 15.4$ Å, $E = 0.029$ eV

4.17 (*a*) 0.2148 eV, (*b*) 0.9052 eV, (*c*) $6.90 \times 10^3\ \text{cm}^{-3}$, (*d*) Holes, (*e*) 0.338 eV

4.19 (*a*) 0.2764 eV, (*b*) $2.414 \times 10^{14}\ \text{cm}^{-3}$, (*c*) p type

4.21 (*a*) $n_o = 6.86 \times 10^{15}\ \text{cm}^{-3}$, $p_o = 7.84 \times 10^7\ \text{cm}^{-3}$; (*b*) $E_c - E_F = 0.2153$ eV, $p_o = 7.04 \times 10^3\ \text{cm}^{-3}$

4.23 (*a*) $n_o = 7.33 \times 10^{13}\ \text{cm}^{-3}$, $p_o = 3.07 \times 10^6\ \text{cm}^{-3}$; (*b*) $n_o = 8.80 \times 10^9\ \text{cm}^{-3}$, $p_o = 3.68 \times 10^2\ \text{cm}^{-3}$

4.25 (*a*) 0.2787 eV, (*b*) 0.8413 eV, (*c*) $1.134 \times 10^9\ \text{cm}^{-3}$, (*d*) Holes, (*e*) 0.2642 eV

4.27 (*a*) $p_o = 6.68 \times 10^{14}\ \text{cm}^{-3}$, $n_o = 7.23 \times 10^4\ \text{cm}^{-3}$; (*b*) $E_F - E_v = 0.3482$ eV, $n_o = 8.49 \times 10^9\ \text{cm}^{-3}$

4.29 0.0777 eV

4.31 $E = E_c + \dfrac{1}{2}kT$, $\text{E} = E_v - \dfrac{1}{2}kT$

4.35 (*a*) $p_o = 3 \times 10^{15}\ \text{cm}^{-3}$, $n_o = 1.08 \times 10^{-3}\ \text{cm}^{-3}$; (*b*) $n_o = 3 \times 10^{16}\ \text{cm}^{-3}$, $p_o = 1.08 \times 10^{-4}\ \text{cm}^{-3}$; (*c*) $n_o = p_o = 1.8 \times 10^6\ \text{cm}^{-3}$; (*d*) $p_o = 4 \times 10^{15}\ \text{cm}^{-3}$, $n_o = 1.44 \times 10^2\ \text{cm}^{-3}$; (*e*) $n_o = 10^{14}\ \text{cm}^{-3}$, $p_o = 1.48 \times 10^7\ \text{cm}^{-3}$

4.37 (*a*) $\dfrac{n_d}{N_d} = 8.85 \times 10^{-4}$, (*b*) $f_F(E) = 2.87 \times 10^{-5}$

4.39 (*a*) n type; (*b*) $n_o = 8 \times 10^{14}\ \text{cm}^{-3}$, $p_o = 2.81 \times 10^5\ \text{cm}^{-3}$; (*c*) $N_a' = 4.8 \times 10^{15}\ \text{cm}^{-3}$, $n_o = 5.625 \times 10^4\ \text{cm}^{-3}$

4.41 $n_o = 6.88 \times 10^{11}\ \text{cm}^{-3}$, $p_o = 2.75 \times 10^{12}\ \text{cm}^{-3}$, $N_a = 2.064 \times 10^{12}\ \text{cm}^{-3}$

4.45 $n_i = 5.74 \times 10^{13}\ \text{cm}^{-3}$, $p_o = 3 \times 10^{13}\ \text{cm}^{-3}$

4.47 (*a*) n type; (*b*) $n_o = 1.125 \times 10^{16}\ \text{cm}^{-3}$, $p_o = 2 \times 10^4\ \text{cm}^{-3}$; (*c*) $N_d = 1.825 \times 10^{16}\ \text{cm}^{-3}$

4.49 For $10^{14}\ \text{cm}^{-3}$, $E_c - E_F = 0.3249$ eV, $E_F - E_{Fi} = 0.2280$ eV; $10^{15}\ \text{cm}^{-3}$, $E_c - E_F = 0.2652$ eV, $E_F - E_{Fi} = 0.2877$ eV; $10^{16}\ \text{cm}^{-3}$, $E_c - E_F = 0.2056$ eV, $E_F - E_{Fi} = 0.3473$ eV,

10^{17} cm^{-3}, $E_c - E_F = 0.1459$ eV,
$E_F - E_{Fi} = 0.4070$ eV

4.51 $T = 200$ K, $E_{Fi} - E_F = 0.4212$ eV,
$T = 400$ K, $E_{Fi} - E_F = 0.2465$ eV,
$T = 600$ K, $E_{Fi} - E_F = 0.0630$ eV

4.53 (*a*) $E_{Fi} - E_{midgap} = +0.0447$ eV;
(*b*) (*i*) Acceptors, (*ii*) $N_a = 1.97 \times 10^{13}$ cm^{-3}

4.55 (*a*) (*i*) $E_c - E_F = 0.2188$ eV,
(*ii*) $N'_d = 1.031 \times 10^{16}$ cm^{-3};
(*b*) (*i*) $E_c - E_F = 0.1594$ eV,
(*ii*) $N'_d = 1.718 \times 10^{15}$ cm^{-3}

4.57 Add acceptors, $N_a = 4 \times 10^{15}$ cm^{-3}

4.59 (*a*) 0.2009 eV, (*b*) 1.360 eV, (*c*) 0.7508 eV,
(*d*) 0.2526 eV, (*e*) 1.068 eV

Chapter 5

5.1 (*a*) $\rho = 4.808$ Ω-cm, (*b*) $\sigma = 0.208(\Omega\text{-cm})^{-1}$

5.3 (*a*) $N_d \cong 6 \times 10^{16}$ cm^{-3}, $\mu_n \cong 1050$ cm^2/V-s;
(*b*) $N_a \cong 10^{17}$ cm^{-3}, $\mu_p \cong 320$ cm^2/V-s

5.5 $\mu_n = 1116$ cm^2/V-s

5.7 (*a*) $R = 100$ Ω, (*b*) $\sigma = 0.01(\Omega\text{-cm})^{-1}$,
(*c*) $N_d = 4.63 \times 10^{15}$ cm^{-3},
(*d*) $N_a = 1.13 \times 10^{15}$ cm^{-3}

5.9 (*a*) $L = 0.0256$ cm, (*b*) $v_d = 1.56 \times 10^6$ cm/s,
(*c*) $I = 80$ mA

5.11 (*a*) Si: $t_t = 8.33 \times 10^{-11}$ s, GaAs: $t_t = 1.33 \times 10^{-11}$ s;
(*b*) Si: $t_t = 1.05 \times 10^{-11}$ s, GaAs: $t_t = 1.43 \times 10^{-11}$ s

5.13 (*a*) $p_o \cong 1.3 \times 10^{17}$ cm^{-3}, $n_o \cong 2.49 \times 10^{-5}$ cm^{-3};
(*b*) $n_o \cong 5.79 \times 10^{14}$ cm^{-3}, $p_o \cong 3.89 \times 10^5$ cm^{-3}

5.15 (*a*) (*i*) 4.39×10^{-6} $(\Omega\text{-cm})^{-1}$,
(*ii*) 2.23×10^{-2} $(\Omega\text{-cm})^{-1}$,
(*iii*) 2.56×10^{-9} $(\Omega\text{-cm})^{-1}$;
(*b*) (*i*) 5.36×10^9 Ω, (*ii*) 1.06×10^6 Ω,
(*iii*) 9.19×10^{12} Ω

5.17 $\sigma_{avg} = 3.97$ $(\Omega\text{-cm})^{-1}$

5.21 (*a*) $J = 1.60$ A/cm^2, (*b*) $T \cong 456$ K

5.23 (*a*) n type: $n_o = 5 \times 10^{16}$ cm^{-3}, $p_o = 4.5 \times 10^3$ cm^{-3};
p type: $p_o = 2 \times 10^{16}$ cm^{-3},
$n_o = 1.125 \times 10^4$ cm^{-3}; compensated:
$n_o = 3 \times 10^{16}$ cm^{-3}, $p_o = 7.5 \times 10^3$ cm^{-3};
(*b*) n type: $\mu_n \cong 1100$ cm^2/V-s;
p type: $\mu_p \cong 400$ cm^2/V-s;
compensated: $\mu_n \cong 1000$ cm^2/V-s;
(*c*) n type: $\sigma = 8.8$ $(\Omega\text{-cm})^{-1}$;
p type: $\sigma = 1.28$ $(\Omega\text{-cm})^{-1}$;
compensated: $\sigma = 4.8$ $(\Omega\text{-cm})^{-1}$;
(*d*) n type: E = 13.6 V/cm;
p type: E = 93.75 V/cm;
compensated: E = 25 V/cm

5.25 (*a*) 2388 cm^2/V-s, (*b*) 844 cm^2/V-s

5.29 $n(0) = 0.25 \times 10^{14}$ cm^{-3}

5.31 (*a*) $n(x_1) = 1.67 \times 10^{14}$ cm^{-3},
(*b*) $n(x_1) = 8.91 \times 10^{14}$ cm^{-3}

5.33 $J_{Total} = -18$ A/cm^2

5.35 $\mathrm{E} = 14.5 - 26\exp\left(\frac{x}{18}\right)$ V/cm

5.37 (*a*) $n(x) = 6.51 \times 10^{15} - (3.255 \times 10^{15})\exp\left(\frac{-x}{d}\right)$cm^{-3};
(*b*) $n(0) = 3.26 \times 10^{15}$ cm^{-3},
$n(50) = 6.19 \times 10^{15}$ cm^{-3};
(*c*) $J_{drf} = 95.08$ A/cm^2, $J_{diff} = 4.92$ A/cm^2

5.39 (*a*) $\mathrm{E} = \dfrac{24.1}{\left(\frac{x}{L} - 1\right)}$, (*b*) $\mathrm{E} = \dfrac{13.4}{\left(1 - \frac{x}{L}\right)}$

5.41 $V = -2.73$ mV

5.43 (*a*) $J_{diff} = -(1.24 \times 10^5)\exp\left(-\frac{x}{L}\right)$ A/cm^2,
(*b*) $\mathrm{E} = 2.59 \times 10^3$ V/cm

5.45 (*a*) (*i*) 29.8 cm^2/s, (*ii*) 160.6 cm^2/s;
(*b*) (*i*) 308.9 cm^2/V-s, (*ii*) 1351 cm^2/V-s

5.47 (*a*) $V_H = -0.3125$ mV, (*b*) $E_H = -1.56 \times 10^{-2}$ V/cm,
(*c*) $\mu_n = 3125$ cm^2/V-s

5.49 (*a*) $V_H = -0.825$ mV, (*b*) n type,
(*c*) $n = 4.92 \times 10^{15}$ cm^{-3}, (*d*) $\mu_n = 1015$ cm^2/V-s

Chapter 6

6.1 (*a*) $n_o = 5 \times 10^{15}$ cm^{-3}, $p_o = 4.5 \times 10^4$ cm^{-3};
(*b*) $R' = 5 \times 10^{20}$ cm^{-3} s^{-1}

6.3 (*a*) $\tau_{n0} = 8.89 \times 10^{+6}$ s,
(*b*) $G = 1.125 \times 10^9$ cm^{-3} s^{-1},
(*c*) $G = R = 1.125 \times 10^9$ cm^{-3} s^{-1}

6.7 $\dfrac{\partial F_p^+}{\partial x} = -2 \times 10^{19}$ cm^{-3} s^{-1}

6.9 (*a*) $\mu' = \mu_n \cong 1300$ cm^2/V-s;
(*b*) $D' = D_n = 33.67$ cm^2/s;
(*c*) $\tau_{nt} = \tau_{nO} = 10^{-7}$ s, $\tau_{pt} = 2.18 \times 10^4$ s

6.13 (*a*) For $0 \le t \le 10^{-6}$ s:
$\delta n = \delta p = (2 \times 10^{14})\left[1 - \exp\left(\frac{-t}{\tau_{pO}}\right)\right]$ cm^{-3},
For $t \ge 10^{-6}$ s:
$\delta n = \delta p = (2 \times 10^{14})\exp\left[\dfrac{-(t - 10^{-6})}{\tau_{pO}}\right]$ cm^{-3};
(*b*) For $0 \le t \le 10^{-6}$ s:
$\sigma = 6.0 + 0.250\left[1 - \exp\left(\frac{-t}{\tau_{pO}}\right)\right]$ $(\Omega\text{-cm})^{-1}$,
For $t \ge 10^{-6}$ s:
$\sigma = 6.0 + 0.250\exp\left[\dfrac{-(t - 10^{-6})}{\tau_{pO}}\right]$ $(\Omega\text{-cm})^{-1}$

6.15 (*a*) $\tau_{nO} = 2.5 \times 10^{-7}$ s;
(*b*) $\delta n = \delta p = (5 \times 10^{14})\left[1 - \exp\left(\frac{-t}{\tau_{nO}}\right)\right]$ cm^{-3},
$R' = (2 \times 10^{21})\left[1 - \exp\left(\frac{-t}{\tau_{nO}}\right)\right]$ cm^{-3} s^{-1};
(*c*) (*i*) 7.19×10^{-8} s, (*ii*) 1.73×10^{-7} s,
(*iii*) 3.47×10^{-7} s, (*iv*) 7.49×10^{-7} s

6.17 (*a*) (*i*) For $0 \le t \le 5 \times 10^{-7}$ s:
$\delta p = (2.5 \times 10^{14})\left[1 - \exp\left(\frac{-t}{\tau_{pO}}\right)\right]$ cm^{-3},
For $t \ge 5 \times 10^{-7}$ s:
$\delta p = (1.58 \times 10^{14}) \exp\left[\frac{-(t - 5 \times 10^{-7})}{\tau_{pO}}\right]$ cm^{-3};
(*ii*) At $t = 5 \times 10^{-7}$ s: $\delta p = 1.58 \times 10^{14}$ cm^{-3};
(*b*) (*i*) For $0 \le t \le 2 \times 10^{-6}$ s:
$\delta p = (2.5 \times 10^{14})\left[1 - \exp\left(\frac{-t}{\tau_{pO}}\right)\right]$ cm^{-3},
For $t \ge 2 \times 10^{-6}$ s:
$\delta p = (2.454 \times 10^{14}) \exp\left[\frac{-(t - 2 \times 10^{-6})}{\tau_{pO}}\right]$ cm^{-3};
(*ii*) At $t = 2 \times 10^{-6}$ s: $\delta p = 2.454 \times 10^{14}$ cm^{-3}

6.19 (*a*) $\delta n = \delta p = (2 \times 10^{14}) \exp\left(\frac{-x}{L_n}\right)$ cm^{-3},
$L_n = 5.575 \times 10^{-3}$ cm;
(*b*) $J_n = -0.1784 \exp\left(\frac{-x}{L_n}\right)$ A/cm^2,
$J_p = +0.1784 \exp\left(\frac{-x}{L_n}\right)$ A/cm^2

6.21 $\delta n(x) = (5 \times 10^{14}) \exp\left(\frac{-x}{L_n}\right)$ cm^{-3}, $L_n = 5 \times 10^{-3}$ cm;
$J_n(x) = -0.4 \exp\left(\frac{-x}{L_n}\right)$ A/cm^2,
$J_p(x) = +0.4 \exp\left(\frac{-x}{L_n}\right)$ A/cm^2

6.25 For $0 \le t \le T$: $\delta n = G'_o t$,
For $t \ge T$: $\delta n = G'_o T$

6.27 $\mu_p = 390.6$ cm^2/V-s, $D_p = 10.42$ cm^2/s

6.31 (*a*) $E_{Fi} - E_F = 0.3294$ eV;
(*b*) $E_{Fn} - E_{Fi} = 0.2697$ eV, $E_{Fi} - E_{Fp} = 0.3318$ eV

6.33 (*a*) $\delta n = \delta p = 5.05 \times 10^{14}$ cm^{-3};
(*b*) $E_{Fi} - E_{Fp} = 0.3362$ eV;
(*c*) (*i*) $E_F - E_{Fp} = kT \ln\left(\frac{p_o + \delta p}{p_o}\right)$,
(*ii*) $E_F - E_{Fp} = 2.093$ meV

6.39 (*a*) $R = \frac{-n_i}{\tau_{pO} + \tau_{nO}}$

6.41 (*a*) (*i*) $\delta p = 10^{14}$ cm^{-3},
(*ii*) $\delta p = 10^{14}\left[1 - 0.167 \exp\left(\frac{-x}{L_p}\right)\right]$ cm^{-3},
(*iii*) $\delta p = 10^{14}\left[1 - \exp\left(\frac{-x}{L_p}\right)\right]$ cm^{-3}, $L_p = 10^{-3}$ cm;
(*b*) (*i*) $\delta p(0) = 10^{14}$ cm^{-3},
(*ii*) $\delta p(0) = 0.833 \times 10^{14}$ cm^{-3},
(*iii*) $\delta p(0) = 0$

6.43 (*a*) $\delta p(x) = 10^{18}(20 \times 10^{-4} - x)$ cm^{-3},
(*b*) $\delta p(x) = 10^{18}(70 \times 10^{-4} - x)$ cm^{-3}

Chapter 7

7.1 (*a*) (*i*) 0.611 V, (*ii*) 0.671 V, (*iii*) 0.731 V;
(*b*) (*i*) 0.731 V, (*ii*) 0.790 V, (*iii*) 0.850 V

7.3 (*a*) For $N_a = N_d = 10^{14}$ cm^{-3}, $V_{bi} = 0.4561$ V
10^{15} cm^{-3}, 0.5754 V
10^{16} cm^{-3}, 0.6946 V
10^{17} cm^{-3}, 0.8139 V
(*b*) For $N_a = N_d = 10^{14}$ cm^{-3}, $V_{bi} = 0.9237$ V
10^{15} cm^{-3}, 1.043 V
10^{16} cm^{-3}, 1.162 V
10^{17} cm^{-3}, 1.282 V
(*c*) Silicon:
For $N_a = N_d = 10^{14}$ cm^{-3}, $V_{bi} = 0.2582$ V
10^{15} cm^{-3}, 0.4172 V
10^{16} cm^{-3}, 0.5762 V
10^{17} cm^{-3}, 0.7353 V
GaAs:
For $N_a = N_d = 10^{14}$ cm^{-3}, $V_{bi} = 0.7129$ V
10^{15} cm^{-3}, 0.8719 V
10^{16} cm^{-3}, 1.031 V
10^{17} cm^{-3}, 1.190 V

7.5 (*a*) n side: $E_F - E_{Fi} = 0.3653$ eV,
p side: $E_{Fi} - E_F = 0.3653$ eV;
(*b*) $V_{bi} = 0.7306$ V;
(*c*) $V_{bi} = 0.7305$ V;
(*d*) $x_n = 0.154$ μm, $x_p = 0.154$ μm,
$|E_{max}| = 4.75 \times 10^4$ V/cm

7.7 For $T = 200$ K, $V_{bi} = 1.257$ V
300 K, 1.157 V
400 K, 1.023 V

7.9 (*a*) $V_{bi} = 0.635$ V;
(*b*) $x_n = 0.864$ μm,
$x_p = 0.0864$ μm;
(*d*) $|E_{max}| = 1.34 \times 10^4$ V/cm

7.11 $T \cong 380$ K

7.13 (*a*) $V_{bi} = 0.456$ V,
(*b*) $x_n = 2.43 \times 10^{-7}$ cm,
(*c*) $x_p = 2.43 \times 10^{-3}$ cm,
(*d*) $|E_{max}| = 3.75 \times 10^2$ V/cm

7.17 (*a*) $V_{bi} = 0.8081$ V;
(*b*) $x_n = 0.2987$ μm,
$x_p = 0.0597$ μm, $W = 0.3584$ μm;
(*c*) $|E_{max}| = 1.85 \times 10^5$ V/cm;
(*d*) $C = 5.78$ pF

7.19 (*a*) $\Delta V_{bi} = 0.02845$ V, (*b*) 1.732

7.21 (*a*) 3.13, (*b*) 0.316, (*c*) 0.319

7.23 $V_{R2} = 2.58$ V

7.25 (*a*) $L = 3.306$ mH;
(*b*) (*i*) $f = 0.794$ MHz, (*ii*) $f = 1.069$ MHz

7.27 (*a*) $N_a = 6.016 \times 10^{15}$ cm^{-3}, $N_d = 1.504 \times 10^{15}$ cm^{-3};
(*b*) $N_a = 1.19 \times 10^{16}$ cm^{-3}, $N_d = 2.976 \times 10^{15}$ cm^{-3}

7.29 (*a*) $V_R = 193$ V, (*b*) $x_n = 0.5$ μm,
(*c*) $|\mathrm{E}_{\max}| = 7.65 \times 10^4$ V/cm

7.31 (*a*) $N = 5.36 \times 10^{15}$ cm^{-3}, (*b*) $A = 7.56 \times 10^{-5}$ cm^2,
(*c*) $V_R = 2.96$ V

7.33 (*a*) $V_{bi} = V_t \ln\left[\dfrac{N_{aO} N_{dO}}{n_i^2}\right]$,
(*c*) p region: $\mathrm{E} = \dfrac{-eN_{aO}}{\epsilon}(x + x_p)$;
n region: $0 < x < x_o$, $\mathrm{E} = \dfrac{eN_{dO}x}{2\epsilon} - \dfrac{eN_{dO}}{\epsilon} \times \left(x_n - \dfrac{x_o}{2}\right)$;
$x_o < x < x_n$, $\mathrm{E} = \dfrac{-eN_{dO}}{\epsilon}(x_n - x)$

7.35 (*a*) $N_a = 1.29 \times 10^{16}$ cm^{-3},
(*b*) $N_a = 2.59 \times 10^{16}$ cm^{-3}

7.37 (*a*) $V_B \cong 75$ V, (*b*) $V_B \cong 450$ V

7.39 x_n (min) $= 5.09$ μm

7.41 (*a*) $V_R = 4.35 \times 10^3$ V, (*b*) $V_R = 1.74 \times 10^4$ V
(Note that breakdown is reached first in each case.)

Chapter 8

8.1 (*a*) 60 mV, (*b*) 120 mV

8.3 (*a*) $p_n(x_n) = 4.0 \times 10^{11}$ cm^{-3},
$n_p(-x_p) = 1.0 \times 10^{11}$ cm^{-3};
(*b*) $p_n(x_n) = 9.03 \times 10^{14}$ cm^{-3},
$n_p(-x_p) = 2.26 \times 10^{14}$ cm^{-3};
(*c*) $p_n(x_n) \cong 0$, $n_p(-x_p) \cong 0$

8.5 (*a*) $I_n = 1.85$ mA, (*b*) $I_p = 4.52$ mA, (*c*) $I = 6.37$ mA

8.7 (*a*) $I = 0.244$ mA, (*b*) $I = -1.568 \times 10^{-8}$ A

8.9 $V = -59.6$ mV

8.11 (*a*) $\dfrac{N_d}{N_a} = 12.73$, (*b*) $\dfrac{N_d}{N_a} = 0.354$

8.15 (*a*) p side: $E_{Fi} - E_F = 0.329$ eV, n side:
$E_F - E_{Fi} = 0.407$ eV;
(*b*) $I_S = 4.426 \times 10^{-15}$ A, $I = 1.07$ μA;
(*c*) $\dfrac{I_p}{I} = 0.0741$

8.17 (*a*) $\delta p_n(x) = (3.81 \times 10^{14}) \exp\left(\dfrac{-x}{2.83 \times 10^{-4}}\right)$ cm^{-3},
(*b*) $J_p = 0.597$ A/cm^2, (*c*) $J_n = 1.39$ A/cm^2

8.19 (*a*) $N_p = 1.51 \times 10^4$, $N_n = 2.41 \times 10^3$;
(*b*) $N_p = 7.17 \times 10^5$, $N_n = 1.15 \times 10^5$;
(*c*) $N_p = 3.40 \times 10^7$, $N_n = 5.45 \times 10^6$

8.21 (*b*) (*i*) $\dfrac{I_S(400)}{I_S(300)} = 1383$, (*ii*) $\dfrac{I_S(400)}{I_S(300)} = 1.17 \times 10^5$

8.23 $T \cong 502$ K, reverse-biased current

8.29 (*a*) $T \cong 567$ K, $I_S = I_{\text{gen}} = 2.314$ μA;
(*b*) $V_a = 0.5366$ V

8.31 $V_a = 0.4$ V: $I_d = 7.64 \times 10^{-16}$ A, $I_{rec} = 1.35 \times 10^{-10}$ A;

0.6 V:	1.73×10^{-12} A,	6.44×10^{-9} A;
0.8 V:	3.90×10^{-9} A,	3.06×10^{-7} A;
1.0 V:	8.80×10^{-6} A,	1.45×10^{-5} A;
1.2 V:	1.99×10^{-2} A,	6.90×10^{-4} A

8.35 $J_{gen} = 1.5 \times 10^{-3}$ A/cm^2

8.37 (*a*) $r_d = 21.6$ Ω, $C_d = 11.6$ nF;
(*b*) $r_d = 216$ Ω, $C_d = 1.16$ nF

8.39 For 10 kHz, $Z = 25.9 - j0.0814$;
For 100 kHz, $Z = 25.9 - j0.814$;
For 1 MHz, $Z = 23.6 - j7.41$;
For 10 MHz, $Z = 2.38 - j7.49$

8.41 $\tau_{p0} = 1.3 \times 10^{-7}$ s; $C_d = 2.5 \times 10^{-9}$ F

8.43 (*a*) $R = 72.3$ Ω, $I = 1.38$ mA

8.45 (*a*) $V_a = 0.4896$ V, (*b*) $V_a = 0.4733$ V

8.47 (*a*) $\dfrac{t_s}{\tau_{pO}} = 0.956$, (*b*) $\dfrac{t_s}{\tau_{pO}} = 0.228$

8.49 2.21×10^{-7} s

Chapter 9

9.1 (*c*) $\phi_n = 0.206$ V, $\phi_{B0} = 0.27$ V,
$V_{bi} = 0.064$ V, $|\mathrm{E}_{\max}| = 1.41 \times 10^4$ V/cm,
(*d*) $\phi_{Bn} = 0.55$ V, $|\mathrm{E}_{\max}| = 3.26 \times 10^4$ V/cm

9.3 (*a*) ϕ_{BO} 1.09 V;
(*b*) $V_{bi} = 0.8844$ V;
(*c*) (*i*) $x_n = 0.4939$ μm, $|\mathrm{E}_{\max}| = 7.63 \times 10^4$ V/cm;
(*ii*) $x_n = 0.8728$ μm, $|\mathrm{E}_{\max}| = 1.35 \times 10^5$ V/cm

9.5 (*b*) $\phi_n = 0.1177$ V;
(*c*) $V_{bi} = 0.7623$ V;
(*d*) (*i*) $x_n = 0.7147$ μm, $|\mathrm{E}_{\max}| = 4.93 \times 10^4$ V/cm,
(*ii*) $x_n = 1.292$ μm, $|\mathrm{E}_{\max}| = 8.92 \times 10^4$ V/cm

9.7 (*a*) $V_{bi} = 0.90$ V, (*b*) $N_d = 1.05 \times 10^{16}$ cm^{-3},
(*c*) $\phi_n = 0.0985$ V, (*d*) $\phi_{Bn} = 0.9985$ V

9.13 $D'_{it} = 4.97 \times 10^{11}$ cm^{-2} eV^{-1}

9.15 (*a*) $\phi_{BO} \cong 0.63$ V; (*i*) 0.151 V, (*ii*) 0.211 V,
(*iii*) 0.270 V;
(*b*) (*i*) 0.0654 V, (*ii*) 0.1317 V, (*iii*) 0.201 V

9.21 pn junction: (*a*) 0.678 V, (*b*) 0.718 V, (*c*) 0.732 V;
Schottky junction: (*a*) 0.447 V, (*b*) 0.487 V,
(*c*) 0.501 V

9.23 pn junction: (*a*) 0.691 V, (*b*) $I = 120$ mA;
Schottky junction: (*a*) 0.445 V, (*b*) $I = 53.3$ mA

9.25 (*a*) $R = 0.1$ Ω, (*b*) $R = 1$ Ω, (*c*) $R = 10$ Ω

9.27 (*a*) $\phi_{Bn} = 0.258$ V, (*b*) $\phi_{Bn} = 0.198$ V

9.29 $N_d = 3.5 \times 10^{18}$ cm^{-3}
9.33 $|\Delta E_c| = 0.17$ eV

Chapter 10

10.1 (*a*) p type, inversion;
(*b*) p type, depletion;
(*c*) p type, accumulation;
(*d*) n type, inversion
10.3 (*a*) $N_d = 8.38 \times 10^{14}$ cm^{-3}, (*b*) $\phi_s = 0.566$ V
10.5 $\phi_{ms} = -0.9932$ V
10.7 (*a*) $V_{FB} = -1.04$ V, (*b*) $V_{FB} = -1.012$ V
10.9 $Q'_{ss}/e = 1.2 \times 10^{10}$ cm^{-2}
10.11 (*a*) $V_{TP} = -1.20$ V, (*b*) $V_{TP} = +0.210$ V,
(*c*) $V_{TP} = -1.08$ V
10.13 $N_a \cong 4 \times 10^{16}$ cm^{-3}
10.15 $N_d \cong 5 \times 10^{14}$ cm^{-3}
10.17 (*a*) $t_C = 0.863$ μm, (*b*) $V_T = -1.07$ V
10.23 (*a*) $C_{ox} = 2.876 \times 10^{-7}$ F/cm^2,
$C'_{FB} = 1.346 \times 10^{-7}$ F/cm^2,
$C'_{\min} = 3.083 \times 10^{-8}$ F/cm^2,
$C'(inv) = 2.876 \times 10^{-7}$ F/cm^2;
(*b*) C_{ox}, C'_{FB}, and $C'_{\min}$ unchanged from part (*a*),
$C'(inv) = 3.083 \times 10^{-8}$ F/cm^2;
(*c*) $V_{FB} \cong -1.10$ V, $V_T = -0.2385$ V
10.29 (*b*) $V_{FB} = -0.695$ V;
(*c*) (*i*) For $V_G = +3$ V, $V_{ox} = 0.359$ V
10.31 Point 1: Inversion, 2: Threshold, 3: Depletion,
4: Flat band, 5: Accumulation
10.33 (*a*) 0.0864 mA, (*b*) 0.1152 mA, (*c*) 0.1152 mA,
(*d*) 0.4608 mA
10.35 (*a*) $\frac{W}{L} = 9.26$, (*b*) $I_D = 3.06$ mA, (*c*) $I_D = 0.271$ mA
10.37 (*a*) $V_{GS} = 0.6$ V, I_D(sat) = 0.025 mA
1.2 V, 0.625 mA
1.8 V, 2.025 mA
2.4 V, 4.225 mA
(*c*) $V_{GS} = 0.6$ V, $I_D = 0.0222$ mA
1.2 V, 0.156 mA
1.8 V, 0.289 mA
2.4 V, 0.422 mA
10.39 (*a*) $V_{GS} = 0$ V, I_D(sat) = 0.711 mA
0.8 V, 2.84 mA
1.6 V, 6.40 mA
10.43 For $V_{SG} < 0.35$ V, $g_d = 0$;
For $V_{SG} > 0.35$ V, $g_d = 2(0.961)(V_{SG} - 0.35)$
10.45 $V_T \approx 0.2$ V, $\mu_n = 342$ cm^2/V-s
10.47 (*a*) (*i*) $k'_n = 86.29$ μA/V^2, (*ii*) $\frac{W}{L} = 7.24$;
(*b*) (*i*) $k'_p = 40.27$ μA/V^2, (*ii*) $\frac{W}{L} = 15.5$
10.49 (*a*) $g_{mL} = 0.192$ mA/V, (*b*) $g_{ms} = 2.21$ mA/V
10.51 $V_{TO} = 0.386$ V
10.53 (*a*) $V_T = -0.357$ V, (*b*) $V_{SB} = 5.43$ V
10.55 (*a*) $r_s = 198$ Ω, (*b*) 12% reduction
10.57 (*a*) $f_T = 3.18$ GHz, (*b*) $f_T = 0.83$ GHz

Chapter 11

11.1 $I_D = 10^{-15}\exp\left(\frac{V_{GS}}{(2.1)V_t}\right)$, $I_T = (10^6)I_D$,
$P = I_T \cdot V_{DD}$; for $V_{GS} = 0.5$ V,
$I_D = 9.83$ pA, $I_T = 9.83$ μA,
$P = 49.2$ μW; for $V_{GS} = 0.7$ V,
$I_D = 0.388$ nA, $I_T = 0.388$ mA,
$P = 1.94$ mW; for $V_{GS} = 0.9$ V,
$I_D = 15.4$ nA, $I_T = 15.4$ mA, $P = 77$ mW
11.3 (*a*) $\Delta L = 0.1413$ μm, (*b*) $\Delta L = 0.2816$ μm,
(*c*) $\Delta L = 0.0346$ μm, (*d*) $\Delta L = 0.1749$ μm
11.5 (*a*) (*i*) $\Delta L = 0.0735$ μm, (*ii*) $\Delta L = 0.1303$ μm,
(*iii*) $\Delta L = 0.2205$ μm;
(*b*) $L = 1.84$ μm
11.7 (*a*) (*i*) $I_D = 75.94$ μA, (*ii*) $I'_D = 78.22$ μA,
(*iii*) $r_o = 658$ kΩ;
(*b*) (*i*) $I_D = 0.30375$ mA, (*ii*) $I'_D = 0.3129$ mA,
(*iii*) $r_o = 165$ kΩ
11.9 (*a*) Assume V_{DS}(sat) = 1 V; then
$L = 3$ μm $\Rightarrow$ $E_{sat} = 3.33 \times 10^3$ V/cm
$L = 1$ μm $\Rightarrow$ $E_{sat} = 10^4$ V/cm
$L = 0.5$ μm $\Rightarrow$ $E_{sat} = 2 \times 10^4$ V/cm
(*b*) Assume $\mu_n = 500$ cm^2/V-s, $v = \mu_n E_{sat}$,
$L = 3$ μm $\Rightarrow$ $v = 1.67 \times 10^6$ cm/s
$L = 1$ μm $\Rightarrow$ $v = 5 \times 10^6$ cm/s
$L \leq 0.5$ μm $\Rightarrow$ $v \approx 10^7$ cm/s
11.13 (*a*) (*i*) $I_D = 0.7175$ mA, (*ii*) $I_D = 1.23$ mA,
(*iii*) $I_D = 1.409$ mA, (*iv*) $I_D = 1.64$ mA;
(*b*) (*i*) $I_D = 0.552$ mA, (*ii*) $I_D = 1.10$ mA,
(*iii*) $I_D = 1.38$ mA, (*iv*) $I_D = 1.38$ mA;
(*c*) For (*a*), V_{DS}(sat) = 2 V; for (*b*), V_{DS}(sat) = 1.25 V
11.15 (*a*) Both bias conditions, $I_D \approx kI_D$,
(*b*) $P \approx k^2 P$
11.17 (*a*) (*i*) I_D(max) = 2.438 mA, (*ii*) I_D(max) = 1.298 mA;
(*b*) (*i*) P(max) = 7.314 mW, (*ii*) P(max) = 2.531 mW
11.19 $V_{TO} = 0.389$ V
11.25 $\Delta V_T \rightarrow k\Delta V_T$

11.27 $W = 1.11\ \mu m$
11.29 $\Delta V_T \rightarrow k\Delta V_T$
11.31 (*a*) $t_{ox} = 400$ Å, (*b*) $t_{ox} = 600$ Å
11.33 Near punch-through, $V_{pt} = 2.08$ V;
Ideal punch-through, $V_{pt} = 4.9$ V
11.35 $L = 1.08\ \mu m$
11.37 Donor ions, $D_I = 7.19 \times 10^{11}\ cm^{-2}$
11.39 (*a*) $V_{TO} = -0.0969$ V;
(*b*) Donor ions, $D_I = 3.63 \times 10^{11}\ cm^{-2}$
11.41 For $V_{SB} = 1$ V: $\Delta V_T = 0.0443$ V
3 V: 0.0987 V
5 V: 0.138 V
11.43 $\Delta V_T = -2.09$ V

Chapter 12

12.3 (*a*) $I_S = 7.2 \times 10^{-15}$ A;
(*b*) (*i*) $I_C = 38.27\ \mu A$, (*ii*) $I_C = 0.571$ mA,
(*iii*) $I_C = 8.519$ mA
12.5 (*a*) $\beta = 65.7$;
(*b*) (*i*) $I_B = 0.5828\ \mu A$, $I_E = 38.85\ \mu A$;
(*ii*) $I_B = 8.695\ \mu A$, $I_E = 0.5797$ mA;
(*iii*) $I_B = 0.1297$ mA, $I_E = 8.649$ mA;
(*c*) $\beta = 165.7$;
(*i*) $I_B = 0.2310\ \mu A$, $I_E = 38.50\ \mu A$;
(*ii*) $I_B = 3.446\ \mu A$, $I_E = 0.5744$ mA;
(*iii*) $I_B = 51.41\ \mu A$, $I_E = 8.570$ mA
12.9 (*a*) $p_{EO} = 2.8125 \times 10^2\ cm^{-3}$,
$n_{BO} = 1.125 \times 10^4\ cm^{-3}$,
$p_{CO} = 2.25 \times 10^5\ cm^{-3}$;
(*b*) $n_B(0) = 6.064 \times 10^{14}\ cm^{-3}$,
$p_E(0) = 1.516 \times 10^{13}\ cm^{-3}$
12.11 (*a*) $V_{BE} = 0.6709$ V, (*b*) $p_E(0) = 5.0 \times 10^{13}\ cm^{-3}$
12.15 (*a*) 0.126%, (*b*) 11.32%
12.19 (*b*) $n_B(x_B) = 6.7 \times 10^{12}\ cm^{-3}$,
$p_C(0) = 9.56 \times 10^{13}\ cm^{-3}$;
(*c*) $x_B = 0.994\ \mu m$
12.21 (*a*) (*i*) $\gamma = 0.99305$, (*ii*) $\alpha_T = 0.990$,
(*iii*) $\delta = 0.990167$, (*iv*) $\alpha = 0.97345$,
(*v*) $\beta = 36.7$;
(*b*) $I_{nC} = 0.4986$ mA, $I_{pE} = 1.38\ \mu A$, $I_R = 1.39\ \mu A$
12.23 (*a*) $J_{nE} = 1.779\ A/cm^2$, $J_{pE} = 0.0425\ A/cm^2$,
$J_{nC} = 1.773\ A/cm^2$, $J_R = 3.22 \times 10^{-3}\ A/cm^2$;
(*b*) $\gamma = 0.9767$, $\alpha_T = 0.9966$, $\delta = 0.9982$,
$\alpha = 0.9716$, $\beta = 34.2$
12.25 (*b*) (*i*) $\dfrac{\alpha_T(B)}{\alpha_T(A)} = 1$, (*ii*) $\dfrac{\alpha_T(C)}{\alpha_T(A)} = 1$

12.27 (*a*) $x_B/L_B = 0.01$: $\alpha_T = 0.99995$, $\beta = 19{,}999$
0.10: 0.995 199
1.0: 0.648 1.84
10.0: $\cong 0$ $\cong 0$
(*b*) $N_B/N_E = 0.01$: $\gamma = 0.990$, $\beta = 99$
0.10: 0.909 9.99
1.0: 0.50 1.0
10.0: 0.0909 0.10
12.29 (*a*) Let $x_B = 0.80\ \mu m$, then $N_E = 4.61 \times 10^{18}\ cm^{-3}$;
(*b*) $\alpha_T = 0.99930$, $\gamma = 0.99656$
12.35 (*a*) (*i*) $r_o = 101.7\ k\Omega$, (*ii*) $g_o = 9.84 \times 10^{-6}\ (\Omega)^{-1}$,
(*iii*) $I_C = 1.22$ mA;
(*b*) (*i*) $r_o = 648\ k\Omega$, (*ii*) $g_o = 1.54 \times 10^{-6}\ (\Omega)^{-1}$,
(*iii*) $I_C = 0.253$ mA
12.37 (*a*) (*i*) $J_C = 52.16\ A/cm^2$, (*ii*) $J_C = 57.18\ A/cm^2$,
(*iii*) $J_C = 61.85\ A/cm^2$
(*b*) $V_A = 38.4$ V
12.39 (*a*) $\Delta x_{dB} = 0.1188\ \mu m$, (*b*) $\Delta I_C = 0.519$ mA,
(*c*) $V_A = 13.3$ V, (*d*) $r_o = 7.705\ k\Omega$
12.41 (*a*) $N_B = 1.83 \times 10^{15}\ cm^{-3}$,
(*b*) $N_B = 4.02 \times 10^{16}\ cm^{-3}$
12.43 $S = 1.42\ \mu m$
12.45 (*a*) $BV_{BCO} \cong 180$ V, (*b*) $BV_{ECO} = 34.5$ V,
(*c*) $BV_{EB} \cong 19$ V
12.47 (*a*) $BV_{CBO} \cong 64$ V, (*b*) $V_{pt} = 70.0$ V
12.49 $x_{BO} = 0.1483\ \mu m$
12.55 (*a*) (*i*) $\tau_e = 36.26$ ps, (*ii*) $\tau_b = 84.5$ ps,
(*iii*) $\tau_d = 22$ ps, (*iv*) $\tau_c = 0.72$ ps;
(*b*) $\tau_{ec} = 143.48$ ps;
(*c*) $f_T = 1.109$ GHz;
(*d*) $f_\beta = 8.87$ MHz

Chapter 13

13.3 (*a*) (*i*) $V_{pO} = 3.312$ V, (*ii*) $V_p = -1.984$ V;
(*b*) (*i*) $a - h = 0.103\ \mu m$, (*ii*) $a - h = 0.065\ \mu m$,
(*iii*) $a - h = 0$;
(*c*) (*i*) $V_{DS}(\text{sat}) = 1.984$ V, (*ii*) $V_{DS}(\text{sat}) = 0.984$ V
13.5 (*a*) $N_a = 9.433 \times 10^{15}\ cm^{-3}$, (*b*) $V_p = 1.47$ V,
(*c*) $V_{GS} = 0.347$ V, (*d*) $V_{SD} = 1.47$ V
13.7 (*a*) $a = 0.50\ \mu m$;
(*b*) $V_{pO} = 3.86$ V;
(*c*) (*i*) $V_{SD}(\text{sat}) = 3.0$ V, (*ii*) $V_{SD}(\text{sat}) = 1.5$ V
13.9 (*a*) $a = 0.436\ \mu m$;
(*b*) (*i*) $V_{pO} = 5.886$ V, (*ii*) $V_p = -5.0$ V

13.11 (*a*) $I_{p1} = 1.03$ mA;
(*b*) (*i*) $V_{DS}(\text{sat}) = 1.056$ V, (*ii*) $V_{DS}(\text{sat}) = 0.792$ V,
(*ii*) $V_{DS}(\text{sat}) = 0.528$ V, (*iv*) $V_{DS}(\text{sat}) = 0.264$ V;
(*c*) (*i*) $I_{D1} = 0.258$ mA, (*ii*) $I_{D1} = 0.141$ mA,
(*iii*) $I_{D1} = 0.061$ mA, (*iv*) $I_{D1} = 0.0148$ mA
13.13 (*a*) $G_{O1} = 2.69 \times 10^{-3}$ S;
(*b*) (*i*) $V_{DS}(\text{sat}) = 0.35$ V, (*ii*) $V_{DS}(\text{sat}) = 0.175$ V;
(*c*) (*i*) $I_{D1}(\text{sat}) = 50.6\ \mu$A, (*ii*) $I_{D1}(\text{sat}) = 12.4\ \mu$A
13.15 (*a*) $g_{ms}(\text{max}) = 0.295$ mS, (*b*) $g_{ms}(\text{max}) = 1.48$ mS
13.17 (*a*) $N_d = 8.1 \times 10^{15}$ cm^{-3}, (*b*) $V_T = 0.051$ V
13.19 (*a*) $V_T = -0.1103$ V, (*b*) $a = 0.2095\ \mu$m
13.21 (*a*) $a = 0.26\ \mu$m, $V_T = 0.092$ V;
(*b*) $V_{DS}(\text{sat}) = 0.258$ V
13.23 (*a*) $k_n = 1.206$ mA/V^2;
(*b*) (*i*) $I_{D1}(\text{sat}) = 12.06\ \mu$A, (*ii*) $I_{D1}(\text{sat}) = 0.1085$ mA;
(*c*) (*i*) $V_{DS}(\text{sat}) = 0.10$ V, (*ii*) $V_{DS}(\text{sat}) = 0.30$ V
13.27 (*a*) $L = 2.333\ \mu$m, (*b*) $L = 2.946\ \mu$m
13.29 (*a*) $V_{DS} = 2$ V, (*b*) $h_{sat} = 0.306\ \mu$m,
(*c*) $I_{D1}(\text{sat}) = 3.72$ mA, (*d*) $I_{D1}(\text{sat}) = 9.05$ mA
13.31 (*a*) $t_d = 5$ ps, (*b*) $t_d = 20$ ps
13.33 (*a*) $I_{DG} = 0.39$ pA, (*b*) $I_{DG} = 0.42$ pA,
(*c*) $I_{DG} = 0.50$ pA
13.35 $f_T = 9.76$ GHz
13.37 (*a*) $f_T = 8.74$ GHz, (*b*) $f_T = 35.0$ GHz
13.39 (*a*) $V_{off} = -2.07$ V, (*b*) $n_s = 3.25 \times 10^{12}$ cm^{-2}
13.41 $d = 251$ Å

Chapter 14

14.1 (*a*) 1.11 μm, (*b*) 1.88 μm, (*c*) 0.873 μm,
(*d*) 0.919 μm
14.3 (*a*) (*i*) $\alpha \cong 9 \times 10^3$ cm^{-1}, (*ii*) 0.66;
(*b*) (*i*) $\alpha \cong 2.6 \times 10^4$ cm^{-1}, (*ii*) 0.875
14.5 (*a*) $I_{vO} = 0.733$ W/cm^2, (*b*) $d = 2.56\ \mu$m
14.7 $E = 1.65$ eV, $\lambda = 0.75\ \mu$m
14.11 (*a*) $V_{oc} = 0.4847$ V, (*b*) $V = 0.4383$ V,
(*c*) $P_m = 46.5$ mW, (*d*) $R_L = 3.65\ \Omega$
14.15 (*a*) $V_{oc} = 0.474$ V, (*b*) $P_m = 67.9$ mW,
(*c*) $R_L = 2.379\ \Omega$, (*d*) $P = 55.2$ mW
14.17 $\delta n_p = \dfrac{\alpha\Phi_o\tau_n}{\alpha^2 L_n^2 - 1}\left[\exp\left(\dfrac{-x}{L_n}\right) - \exp(-\alpha x)\right]$
14.19 (*a*) $I = 120$ mA, (*b*) $\delta p = 10^{14}$ cm^{-3},
(*c*) $\Delta\sigma = 2.56 \times 10^{-2}\ (\Omega\text{-cm})^{-1}$,
(*d*) $I_L = 3.2$ mA, (*e*) $\gamma_{ph} = 3.33$
14.21 $I_L = 0.131\ \mu$A
14.25 (*a*) $G_L(x) = (3.33 \times 10^{20})\exp[-(10^3)x]$ cm^{-3} s^{-1},
(*b*) $J_L = 53.3$ mA/cm^2
14.27 $d = 230\ \mu$m
14.29 (*a*) (*i*) $E_g \cong 1.64$ eV, (*ii*) $\lambda = 0.756\ \mu$m;
(*b*) (*i*) $E_g \cong 1.78$ eV, (*ii*) $\lambda = 0.697\ \mu$m
14.31 $x \cong 0.38$, $E_g = 1.85$ eV
14.35 $\Delta\lambda = 5.08 \times 10^{-3}\ \mu$m

Chapter 15

15.1 See Figure 8.29
15.3 $f_r = 23.9$ MHz
15.5 (*a*) $E = 6 \times 10^3$ V/cm, (*b*) $v_d \cong 1.5 \times 10^7$ cm/s,
(*c*) $f = 10$ GHz
15.7 (*a*) (*i*) $V_{BE} = 0.5696$ V, (*ii*) $I_C = 0.640$ A;
(*b*) (*i*) $V_{BE} = 0.6234$ V, (*ii*) $I_C = 5.12$ A
15.9 $N_C \cong 2 \times 10^{14}$ cm^{-3}, base width $= 3.16\ \mu$m,
collector width $= 78.9\ \mu$m
15.11 (*a*) $\beta_B = 5.96$, (*b*) $I_{CA} = 3.23$ A
15.13 (*a*) $R_L = 3.60\ \Omega$, (*b*) $I_{C,\text{max}} = 3.33$ A
15.17 (*a*) Let $N_d = 10^{14}$ cm^{-3}, channel length $= 4.86\ \mu$m,
drift region $= 48.6\ \mu$m;
(*b*) Let $N_d = 10^{14}$ cm^{-3}, channel length $= 3.08\ \mu$m,
drift region $= 30.8\ \mu$m
15.19 (*a*) $R_L = 20\ \Omega$, $I_{D,\text{max}} = 3$ A;
(*b*) $V_{DD} = 42.4$ V

INDEX

C

D

E

F

I

M

N

Q

R

S