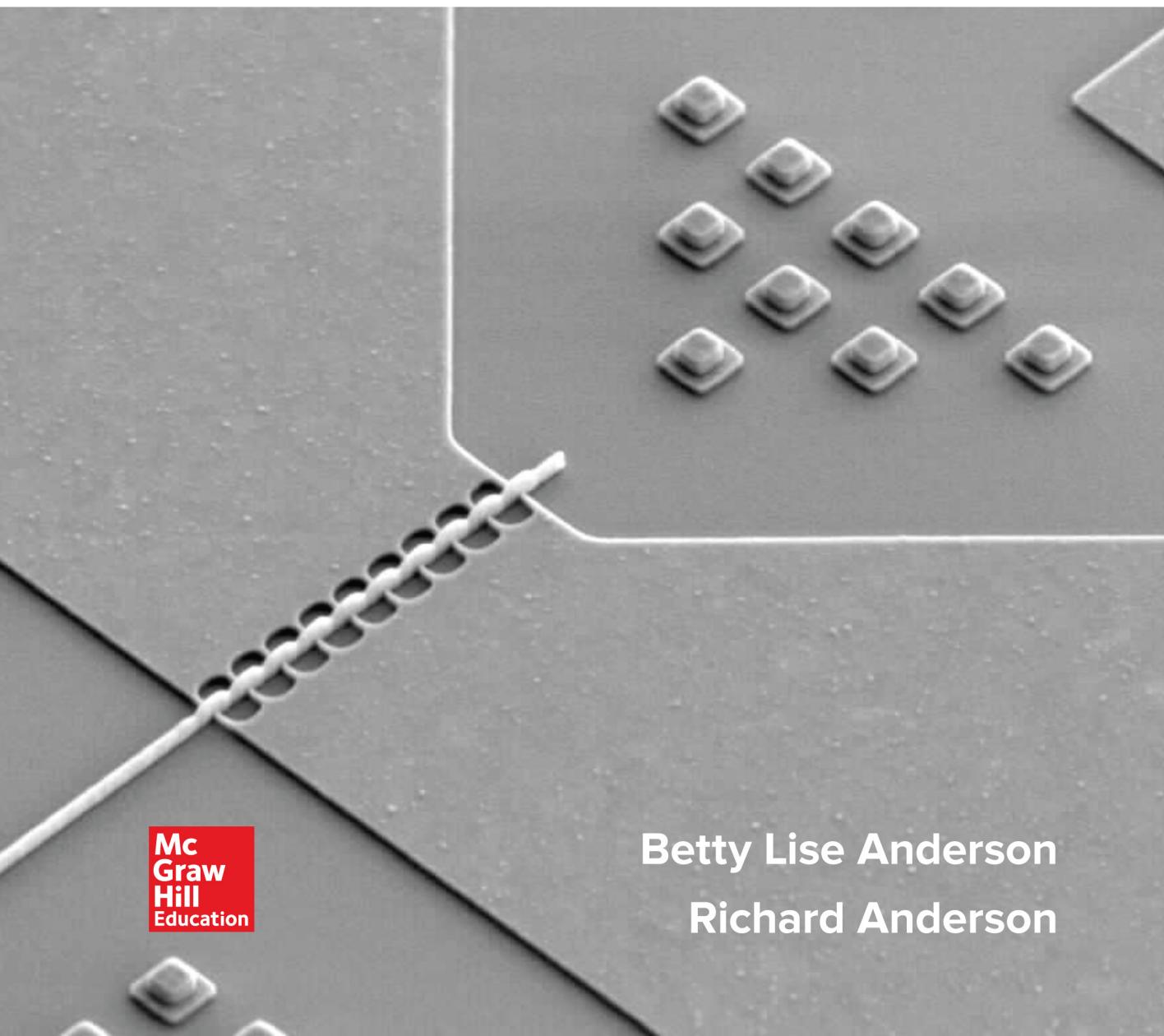


Fundamentals of **SEMICONDUCTOR DEVICES**

Second Edition



**Mc
Graw
Hill
Education**

**Betty Lise Anderson
Richard Anderson**

Fundamentals of Semiconductor Devices

Second Edition

Betty Lise Anderson
The Ohio State University

Richard L. Anderson





FUNDAMENTALS OF SEMICONDUCTOR DEVICES, SECOND EDITION

Published by McGraw-Hill Education, 2 Penn Plaza, New York, NY 10121. Copyright © 2018 by McGraw-Hill Education. All rights reserved. Printed in the United States of America. Previous edition © 2005. No part of this publication may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written consent of McGraw-Hill Education, including, but not limited to, in any network or other electronic storage or transmission, or broadcast for distance learning.

Some ancillaries, including electronic and print components, may not be available to customers outside the United States.

This book is printed on acid-free paper.

1 2 3 4 5 6 7 8 9 LCR 21 20 19 18 17

ISBN 978-0-07-352956-1

MHID 0-07-352956-7

Senior Vice President, Products & Markets: *G. Scott Virkler*

Vice President, General Manager, Products & Markets: *Marty Lange*

Vice President, Content Design & Delivery: *Betsy Whalen*

Managing Director: *Thomas Timp*

Brand Manager: *Raghathan Srinivasan/Thomas M. Scaife, Ph.D.*

Director, Product Development: *Rose Koos*

Product Developer: *Tina Bower*

Marketing Manager: *Shannon O'Donnell*

Director, Content Design & Delivery: *Linda Avenarius*

Program Manager: *Lora Neyens*

Content Project Manager: *Sherry Kane*

Buyer: *Laura Fuller*

Design: *Egzon Shaqiri*

Content Licensing Specialists: *Carrie Burger/Lorraine Buczak*

Cover Image: Courtesy IMEC

Compositor: *SPi Global*

Printer: *LSC Communications*

All credits appearing on page or at the end of the book are considered to be an extension of the copyright page.

Library of Congress Cataloging-in-Publication Data

Names: Anderson, Betty Lise, author. | Anderson, Richard L., author.

Title: Fundamentals of semiconductor devices / Betty Lise Anderson, The Ohio

State University; Richard L. Anderson.

Description: Second edition. | Dubuque : McGraw-Hill Education, 2017.

Identifiers: LCCN 2016036979 | ISBN 9780073529561 (alk. paper)

Subjects: LCSH: Semiconductors. | Transistors.

Classification: LCC TK7871.85 A495 2017 | DDC 621.3815/2—dc23 LC record

available at <https://lccn.loc.gov/2016036979>

The Internet addresses listed in the text were accurate at the time of publication. The inclusion of a website does not indicate an endorsement by the authors or McGraw-Hill Education, and McGraw-Hill Education does not guarantee the accuracy of the information presented at these sites.

BRIEF CONTENTS

Preface xii

PART 1

Materials 1

- 1** Electron Energy and States in Semiconductors 3
- 2** Homogeneous Semiconductors 48
- 3** Current Flow in Homogeneous Semiconductors 113
- 4** Nonhomogeneous Semiconductors 162

Supplement to Part 1: Introduction to Quantum Mechanics 180

PART 2

Diodes 223

- 5** Prototype pn Homojunctions 227
- 6** Additional Considerations for Diodes 302

Supplement to Part 2: Diodes 338

PART 3

Field-Effect Transistors 357

- 7** The MOSFET 367

8 Other Field-Effect Transistors 439

Supplement to Part 3: Additional Consideration for MOSFETs 493

PART 4

Bipolar Junction Transistors 539

9 Bipolar Junction Transistors: Statics 544

10 Time-Dependent Analysis of BJTs 590

Supplement to Part 4: Bipolar Devices 623

PART 5

Optoelectronic and Power Semiconductor Devices 643

11 Optoelectronic Devices 644

12 Power Semiconductor Devices 699

Appendix A Constants 750

Appendix B List of Symbols 754

Appendix C Fabrication 769

Appendix D Some Useful Integrals 792

Appendix E Useful Equations 793

Index 803

Preface xii

PART 1

Materials 1

Chapter 1

**Electron Energy and States
in Semiconductors 3**

- 1.1** Introduction and Preview 3
- 1.2** A Brief History 4
- 1.3** Application to the Hydrogen Atom 5
 - 1.3.1 The Bohr Model for The Hydrogen Atom* 5
 - 1.3.2 Application to Molecules: Covalent Bonding* 11
 - 1.3.3 Quantum Numbers and the Pauli Exclusion Principle* 13
 - 1.3.4 Covalent Bonding in Crystalline Solids* 14
- 1.4** Wave-Particle Duality 21
- 1.5** The Wave Function 22
 - 1.5.1 Probability and the Wave Function* 22
- 1.6** The Electron Wave Function 23
 - 1.6.1 The Free Electron in One Dimension* 23
 - 1.6.2 The De Broglie Relationship* 26
 - 1.6.3 The Free Electron in Three Dimensions* 27
 - 1.6.4 The Quasi-Free Electron Model* 28
 - 1.6.5 Reflection and Tunneling* 32
- 1.7** A First Look at Optical Emission and Absorption 33
- 1.8** Crystal Structures, Planes, and Directions 39
- 1.9** Summary 41

- 1.10** References 42
- 1.11** Review Questions 42
- 1.12** Problems 43

Chapter 2

Homogeneous Semiconductors 48

- 2.1** Introduction and Preview 48
- 2.2** Pseudo-Classical Mechanics for Electrons in Crystals 49
 - 2.2.1 One-Dimensional Crystals* 49
 - 2.2.2 Three-Dimensional Crystals* 55
- 2.3** Conduction Band Structure 57
- 2.4** Valence Band Structure 58
- 2.5** Intrinsic Semiconductors 60
- 2.6** Extrinsic Semiconductors 62
 - 2.6.1 Donors* 62
 - 2.6.2 Acceptors* 66
- 2.7** The Concept of Holes 68
 - 2.7.1 Hole Charge* 68
- 2.8** Effective Mass of Electrons and Holes 70
- 2.9** Density-of-States Functions for Electrons in Bands 72
 - 2.9.1 Density of States and Density-of-States Effective Mass* 72
- 2.10** Fermi-Dirac Statistics 73
 - 2.10.1 Fermi-Dirac Statistics for Electrons and Holes in Bands* 75
- 2.11** Electron and Hole Distributions with Energy 78
- 2.12** Temperature Dependence of Carrier Concentrations in Nondegenerate Semiconductors 90
 - 2.12.1 Carrier Concentrations at High Temperatures* 91

2.12.2	<i>Carrier Concentrations at Low Temperatures (Carrier Freeze-Out)</i>	95
2.13	Degenerate Semiconductors	95
2.13.1	<i>Impurity-Induced Band-Gap Narrowing</i>	96
2.13.2	<i>Apparent Band-Gap Narrowing</i>	98
2.14	Summary	101
2.14.1	<i>Nondegenerate Semiconductors</i>	102
2.14.2	<i>Degenerate Semiconductors</i>	103
2.15	References	103
2.16	Review Questions	104
2.17	Problems	104

Chapter 3**Current Flow in Homogeneous Semiconductors** 113

3.1	Introduction	113
3.2	Drift Current	113
3.3	Carrier Mobility	117
3.3.1	<i>Carrier Scattering</i>	121
3.3.2	<i>Scattering Mobility</i>	123
3.3.3	<i>Impurity Band Mobility</i>	124
3.3.4	<i>Temperature Dependence of Mobility</i>	126
3.3.5	<i>High-Field Effects</i>	126
3.4	Diffusion Current	130
3.5	Carrier Generation and Recombination	133
3.5.1	<i>Band-to-Band Generation and Recombination</i>	135
3.5.2	<i>Two-Step Processes</i>	135
3.6	Optical Processes in Semiconductors	135
3.6.1	<i>Absorption</i>	136
3.6.2	<i>Emission</i>	139
3.7	Continuity Equations	141
3.8	Minority Carrier Lifetime	144
3.8.1	<i>Rise Time</i>	146
3.8.2	<i>Fall Time</i>	146
3.9	Minority Carrier Diffusion Lengths	149
3.10	Quasi Fermi Levels	152
3.11	Summary	154

3.12	References	156
3.13	Review Questions	156
3.14	Problems	157

Chapter 4**Nonhomogeneous Semiconductors** 162

4.1	Constancy of The Fermi Level at Equilibrium	162
4.2	Graded Doping	164
4.3	Nonuniform Composition	170
4.4	Graded Doping and Graded Composition Combined	173
4.5	Summary	175
4.6	References	175
4.7	Review Questions	175
4.8	Problems	176

Supplement to Part 1**Introduction to Quantum Mechanics** 180

S1.1	Introduction	180
S1.2	The Wave Function	180
S1.3	Probability and the Wave Function	182
S1.3.1	<i>Particle in a One-Dimensional Potential Well</i>	182
S1.4	Schrödinger's Equation	184
S1.5	Applying Schrödinger's Equation to Electrons	185
S1.6	Some Results From Quantum Mechanics	186
S1.6.1	<i>The Free Electron</i>	187
S1.6.2	<i>The Quasi-Free Electron</i>	188
S1.6.3	<i>The Potential Energy Well</i>	189
S1.6.4	<i>The Infinite Potential Well in One Dimension</i>	190
S1.6.5	<i>Reflection and Transmission at a Finite Potential Barrier</i>	193
S1.6.6	<i>Tunneling</i>	195
S1.6.7	<i>The Finite Potential Well</i>	203
S1.6.8	<i>The Hydrogen Atom Revisited</i>	205
S1.6.9	<i>The Uncertainty Principle</i>	205

S1.7	Phonons	207
S1.7.1	<i>Carrier Scattering by Phonons</i>	211
S1.7.2	<i>Indirect Electron Transitions</i>	213
S1.8	Summary	217
S1.9	References	217
S1.10	Review Questions	217
S1.11	Problems	218

PART 2
Diodes 223

Chapter 5**Prototype pn Homojunctions** 227

5.1	Introduction	227
5.2	Prototype pn Junctions (Qualitative)	229
5.2.1	<i>Energy Band Diagrams of Prototype pn Junctions</i>	229
5.2.2	<i>Description of Current Flow in a pn Prototype Homojunction</i>	236
5.2.3	<i>Tunnel Diodes</i>	241
5.3	Prototype pn Homojunctions (Quantitative)	245
5.3.1	<i>Energy Band Diagram at Equilibrium (Step Junction)</i>	245
5.3.2	<i>Energy Band Diagram with Applied Voltage</i>	248
5.3.3	<i>Current-Voltage Characteristics of pn Homojunctions</i>	254
5.3.4	<i>Reverse-Bias Breakdown</i>	275
5.4	Small-Signal Impedance of Prototype Homojunctions	277
5.4.1	<i>Junction (Differential) Resistance</i>	278
5.4.2	<i>Junction (Differential) Capacitance</i>	279
5.4.3	<i>Stored-Charge Capacitance</i>	281
5.5	Transient Effects	285
5.5.1	<i>Turn-Off Transient</i>	285
5.5.2	<i>Turn-On Transient</i>	287
5.6	Effects of Temperature	291

5.7	Summary	292
5.8	Review Questions	296
5.9	Problems	296

Chapter 6**Additional Considerations for Diodes** 302

6.1	Introduction	302
6.2	Nonstep Homojunctions	302
6.2.1	<i>Linearly Graded Junctions</i>	306
6.2.2	<i>Hyperabrupt Junctions</i>	309
6.3	Semiconductor Heterojunctions	310
6.3.1	<i>The Energy Band Diagrams of Semiconductor–Semiconductor Heterojunctions</i>	310
6.3.2	<i>Tunneling-Induced Dipoles</i>	314
6.3.3	<i>Effects of Interface States</i>	318
6.3.4	<i>Effects of Lattice Mismatch on Heterojunctions</i>	322
6.4	Metal–Semiconductor Junctions	323
6.4.1	<i>Ideal Metal–Semiconductor Junctions (Electron Affinity Model)</i>	323
6.4.2	<i>Influence of Interface-Induced Dipoles</i>	325
6.4.3	<i>The Current–Voltage Characteristics of Metal–Semiconductor Junctions</i>	326
6.4.4	<i>Ohmic (Low-Resistance) Contacts</i>	330
6.4.5	<i>I–V_a Characteristics of Heterojunction Diodes</i>	331
6.5	Capacitance in Nonideal Junctions and Heterojunctions	332
6.6	Summary	332
6.7	References	333
6.8	Review Questions	333
6.9	Problems	334

Supplement to Part 2**Diodes** 338

S2.1	Introduction	338
S2.2	Dielectric Relaxation Time	338

S2.2.1	<i>Case 1: Dielectric Relaxation Time for Majority Carriers</i>	338	7.3	Drift Model for MOSFETs (Quantitative)	389
S2.2.2	<i>Case 2: Dielectric Relaxation Time for Minority Carriers</i>	341	7.3.1	<i>Long-Channel Drift MOSFET Model with Constant Channel Mobility</i>	390
S2.3	Junction Capacitance	342	7.3.2	<i>More Realistic Long-Channel Models: Effect of Fields on the Mobility</i>	404
S2.3.1	<i>Junction Capacitance in a Prototype (Step) Junction</i>	342	7.3.3	<i>Series Resistance</i>	420
S2.3.2	<i>Junction Capacitance in a Nonuniformly Doped Junction</i>	344	7.4	Comparison of Models with Experiment	421
S2.3.3	Varactors	345	7.5	Ballistic Model for MOSFETs	423
S2.3.4	<i>Stored-Charge Capacitance of Short-Base Diodes</i>	346	7.6	Some Short-Channel Effects	426
S2.4	Second-Order Effects in Schottky Diodes	348	7.6.1	<i>Dependence of Effective Channel Length on V_{DS}</i>	426
S2.4.1	<i>Tunneling Through Schottky Barriers</i>	349	7.6.2	<i>Dependence of Threshold Voltage on the Drain Voltage</i>	428
S2.4.2	<i>Barrier Lowering in Schottky Diodes Due to The Image Effect</i>	351	7.7	Subthreshold Leakage Current	429
S2.5	Summary	353	7.8	Summary	432
S2.6	Review Questions	354	7.9	References	435
S2.7	References	354	7.10	Review Questions	435
S2.8	Problems	354	7.11	Problems	436

PART 3

Field-Effect Transistors 357

The Generic FET	358
Transistors in Circuits	362
The Basis for Deriving the I_D - V_{DS} Characteristics of a FET	362

Chapter 7

The MOSFET 367

7.1	Introduction	367
7.2	MOSFETs (Qualitative)	367
7.2.1	<i>Introduction to MOS Capacitors</i>	367
7.2.2	<i>MOS Capacitor Hybrid Diagrams</i>	373
7.2.3	<i>MOSFETs at Equilibrium (Qualitative)</i>	376
7.2.4	<i>MOSFETs Not at Equilibrium (Qualitative)</i>	378

Chapter 8

Other Field-Effect Transistors 439

8.1	Introduction	439
8.2	Measurement of Threshold Voltage and Low-Field Mobility	440
8.3	Complementary MOSFETs (CMOS)	444
8.3.1	<i>Operation of The CMOS Inverter</i>	444
8.3.2	<i>Matching of CMOS Devices</i>	447
8.4	Switching in CMOS Inverter Circuits	449
8.4.1	<i>Effect of Load Capacitance</i>	449
8.4.2	<i>Propagation (Gate) Delay in CMOS Switching Circuits</i>	451
8.4.3	<i>Pass-Through Current in CMOS Switching</i>	454
8.5	Other MOSFETs	454
8.5.1	<i>Silicon on Insulator (SOI) MOSFETs</i>	454
8.5.2	<i>FinFETs</i>	463
8.5.3	<i>Nonvolatile MOSFETs</i>	465
8.6	Other FETs	468
8.6.1	<i>Heterojunction Field-Effect Transistors (HFETs)</i>	468

8.6.2	<i>Metal-Semiconductor Field-Effect Transistors (MESFETs)</i>	475
8.6.3	<i>Junction Field-Effect Transistors (JFETs)</i>	479
8.6.4	<i>Tunnel Field-Effect Transistors (TFETs)</i>	480
8.7	Bulk Channel FETs: Quantitative	484
8.8	Summary	487
8.9	References	488
8.10	Review Questions	489
8.11	Problems	489

Supplement to Part 3

Additional Consideration for MOSFETs 493

S3.1	Introduction	493
S3.2	Dependence of the Channel Charge Q_{ch} on the Longitudinal Field \mathcal{E}_L	493
S3.3	Threshold Voltage for MOSFETs	495
S3.3.1	<i>Fixed Charge</i>	496
S3.3.2	<i>Interface Trapped Charge</i>	497
S3.3.3	<i>Bulk Charge</i>	497
S3.3.4	<i>Effect of Charges on the Threshold Voltage</i>	498
S3.3.5	<i>Flat Band Voltage</i>	499
S3.3.6	<i>Threshold Voltage Control</i>	502
S3.3.7	<i>Channel Quantum Effects</i>	504
S3.4	MOSFET Analog Equivalent Circuit	506
S3.4.1	<i>Small-Signal Equivalent Circuit</i>	507
S3.4.2	<i>CMOS Amplifiers</i>	511
S3.5	Unity Current Gain Cutoff Frequency f_T	511
S3.6	MOS Capacitors	514
S3.6.1	<i>Ideal MOS Capacitance</i>	514
S3.6.2	<i>The $C-V_G$ Characteristics of Real MOS Capacitors</i>	519
S3.6.3	<i>MOSFET Parameter Analyses from $C-V_G$ Measurements</i>	520
S3.7	Dynamic Random-Access Memories (DRAMs)	521
S3.8	MOSFET Scaling [6]	523

S3.9	Device and Interconnect Degradation	526
S3.9.1	<i>MOSFET Integrated Circuit Reliability</i>	531
S3.10	Summary	532
S3.11	References	533
S3.12	Review Questions	534
S3.13	Problems	534

PART 4

Bipolar Junction Transistors 539

Chapter 9

Bipolar Junction Transistors: Statics 544		
9.1	Introduction	544
9.2	Output Characteristics (Qualitative)	548
9.3	Current Gain	550
9.4	Model of a Prototype BJT	551
9.4.1	<i>Collection Efficiency M</i>	554
9.4.2	<i>Injection Efficiency γ</i>	555
9.4.3	<i>Base Transport Efficiency α_T</i>	557
9.5	Doping Gradients in BJTs	563
9.5.1	<i>The Graded-Base Transistor</i>	565
9.5.2	<i>Effect of Base Field on β</i>	570
9.6	Heterojunction Bipolar Transistors (HBTs)	570
9.6.1	<i>Uniformly Doped HBT</i>	571
9.6.2	<i>Graded-Composition HBT: (Si: SiGe-Base: Si HBTs)</i>	575
9.6.3	<i>Double Heterojunction Bipolar Transistor, (DHBT)</i>	577
9.7	Comparison of Si-Base, SiGe-Base, and GaAs-Base HBTs	579
9.8	The Basic Ebers-Moll dc Model	579
9.9	Summary	583
9.10	References	584
9.11	Review Questions	585
9.12	Problems	586

Chapter 10**Time-Dependent Analysis of BJTs** 590

- 10.1** Introduction 590
- 10.2** Ebers-Moll ac Model 590
- 10.3** Small-Signal Equivalent Circuits 592
 - 10.3.1 Hybrid-Pi Models* 594
- 10.4** Stored-Charge Capacitance in BJTs 598
- 10.5** Frequency Response 603
 - 10.5.1 Unity Current Gain Frequency f_T* 604
 - 10.5.2 Base Transit Time t_T* 606
 - 10.5.3 Base-Collector Transit Time t_{BC}* 607
 - 10.5.4 Maximum Oscillation Frequency f_{max}* 608
- 10.6** High-Frequency Transistors 608
 - 10.6.1 Double Poly Si Self-Aligned Transistor* 608
- 10.7** BJT Switching Transistor 611
 - 10.7.1 Output Low-To-High Transition Time* 612
 - 10.7.2 Schottky-Clamped Transistor* 614
 - 10.7.3 Double Heterojunction Bipolar Transistor (DHBT)* 615
- 10.8** BJTs, MOSFETs, and BiMOS 616
 - 10.8.1 Comparison of BJTs and MOSFETs* 616
 - 10.8.2 BiMOS* 618
- 10.9** Summary 620
- 10.10** References 620
- 10.11** Review Questions 621
- 10.12** Problems 621

Supplement to Part 4**Bipolar Devices** 623

- S4.1** Introduction 623
- S4.2** Current Crowding and Base Resistance in BJTs 623
- S4.3** Base Width Modulation (Early Effect) 627
- S4.4** Avalanche Breakdown 632
- S4.5** High Injection 632
- S4.6** Base Push-Out (Kirk) Effect 633

- S4.7** Recombination in the Emitter-Base Junction 635
- S4.8** Offset Voltage in BJTs 636
- S4.9** Lateral Bipolar Transistors 637
- S4.10** Summary 638
- S4.11** References 638
- S4.12** Review Questions 639
- S4.13** Problems 639

PART 5**Optoelectronic and Power Semiconductor Devices** 643Chapter 11**Optoelectronic Devices** 644

- 11.1** Introduction and Preview 644
- 11.2** Photodetectors 644
 - 11.2.1 Generic Photodetector* 644
 - 11.2.2 Solar Cells* 652
 - 11.2.3 The pin (PIN) Photodetector* 658
 - 11.2.4 Avalanche Photodiodes* 660
- 11.3** Light-Emitting Diodes 661
 - 11.3.1 Spontaneous Emission in a Forward-Biased Junction* 661
 - 11.3.2 Blue, Ultraviolet, and White LEDs* 664
 - 11.3.3 Infrared LEDs* 664
 - 11.3.4 White LEDs and Solid-State Lighting* 671
- 11.4** Laser Diodes 674
 - 11.4.1 Optical Gain* 675
 - 11.4.2 Feedback* 677
 - 11.4.3 Gain + Feedback = Laser* 680
 - 11.4.4 Laser Structures* 682
 - 11.4.5 Other Semiconductor Laser Materials* 686
- 11.5** Image Sensors (Imagers) 686
 - 11.5.1 Charge-Coupled Devices (CCDs)* 686
 - 11.5.2 Linear Image Sensors* 688
 - 11.5.3 Area Image Sensors* 691
- 11.6** Summary 692

- 11.7 References** 693
- 11.8 Review Questions** 694
- 11.9 Problems** 694

Chapter 12

Power Semiconductor Devices 699

- 12.1 Introduction and Preview** 699
- 12.2 Rectifying Diodes** 700
 - 12.2.1 Junction Breakdown* 700
 - 12.2.2 Specific On-Resistance* 710
 - 12.2.3 Transient Losses* 718
 - 12.2.4 Merged Pin-Schottky (MPS) Diodes* 723
- 12.3 Thyristors (npnp Switching Devices)** 725
 - 12.3.1 The Four-Layer Diode Switch* 725
 - 12.3.2 Two-Transistor Model of an npnp Switch* 729
 - 12.3.3 Silicon-Controlled Rectifiers (SCRs)* 730
 - 12.3.4 TRIAC* 733
 - 12.3.5 Gate Turn-Off Thyristors (GTOs)* 735
- 12.4 The Power MOSFET** 736
- 12.5 The Insulated-Gate Bipolar Transistor** 740
- 12.6 Power MOSFET versus IGBT** 745
- 12.7 Summary** 746
- 12.8 References** 747
- 12.9 Review Questions** 748
- 12.10 Problems** 748

Appendices

Appendix A Constants 750

Appendix B List of Symbols 754

Appendix C Fabrication 769

- C.1 Introduction** 769
- C.2 Substrate Preparation** 769
 - C.2.1 The Raw Material* 770
 - C.2.2 Crystal Growth* 770
 - C.2.3 Defects* 773
 - C.2.4 Epitaxy* 774
- C.3 Doping** 777
 - C.3.1 Diffusion* 777
 - C.3.2 Ion Implantation* 778
- C.4 Lithography** 780
- C.5 Conductors and Insulators** 782
 - C.5.1 Metallization* 782
 - C.5.2 Poly Si* 783
 - C.5.3 Oxidation* 783
 - C.5.4 Silicon Nitride* 784
- C.6 Silicon Oxynitride (SiO_xN_y or SiON)** 785
- C.7 Clean Rooms** 787
- C.8 Packaging** 787
 - C.8.1 Wire Bonding* 788
 - C.8.2 Lead Frame* 789
 - C.8.3 Surface-Mount Packages* 790
- C.9 Summary** 791

Appendix D Some Useful Integrals 792

Appendix E Useful Equations 793

- General Physics 793
- Semiconductor Materials 793
- Junctions 794
- Field-Effect Transistors 796
- Bipolar Junction Transistors 798
- Optoelectronic Devices 801
- Power Semiconductor Devices 802

Index 803

PREFACE

This is a textbook on the operating principles of semiconductor devices. It is appropriate for undergraduate (junior or senior) or beginning graduate students in electrical engineering, as well as students of computer engineering, physics, and materials science. It is also useful as a reference for practicing engineers and scientists who are involved with modern semiconductor devices.

Prerequisites are courses in chemistry and physics and in basic electric circuits, which are normally taken in the freshman and sophomore years.

The text is appropriate for a two- or three-semester course on semiconductor devices. However, it can be used for a one-semester course by eliminating some of the more advanced material and assigning some of the sections as read-only. The authors have attempted to organize the material so that some of the detail derivation sections can be skipped without affecting the comprehension of other sections.

This book is divided into five parts:

1. Materials
2. Diodes
3. Field-effect transistors
4. Bipolar transistors
5. Optoelectronic and power semiconductor devices

The first four parts are followed by “Supplements” that, while not required for an understanding of the basic principle of device operation, contain related material that may be assigned at the discretion of the instructor.

Part 1, “Materials,” contains four chapters and a Supplement. The first two chapters contain considerable review material from the prerequisite courses. This material is included because it is used extensively in later chapters to explain the principles of device operation. Depending on the detailed content on the prerequisite courses, much of the material in these chapters can be relegated to reading assignments.

The level of quantum mechanics to be covered in a course like this varies widely. In this book some basic concepts are included in the main chapters of Part 1. Those wishing to cover quantum mechanics in more detail will find more extensive material in the Supplement to Part 1.

The basic operating principles of large and small devices of a particular type (e.g., diodes, field-effect transistors, bipolar junction transistors, and

photodetectors) are the same. However the relative importance of many of the parameters involved in device operation depends on the device dimensions. In this book the general behavior of devices of large dimensions is treated first. In each case, we treat “prototype” devices (such as step junctions and long channel field-effect transistors) from which the fundamental physics can be learned, and then we develop more realistic models considering second-order effects. These second-order effects can have significant influence on the electrical characteristics of modern small-geometry devices. The instructor can go into as much depth as desired or as time permits.

Topics treated that are typically omitted in undergraduate texts are:

- The differences between the electron and hole effective masses as used in density-of-states calculations and conductivity calculations.
- The differences in electron and hole mobilities (and thus diffusion coefficients) depending on whether they are majority carriers or minority carriers.
- The effects of doping gradients in the base of bipolar junction transistors (and/or the composition and heterojunction BJTs) on the current gain in switching speed.
- Band gap reduction in degenerate semiconductors. While this has little effect on the electrical characteristics of diodes or field-effect transistors, its effect in the emitter of bipolar junction transistors can reduce the current gain by orders of magnitude.
- The use of wide band-gap semiconductors (e.g., GaN and 4H-SiC) for use in high-power semiconductor devices.

While the major emphasis is on silicon and silicon-based devices, the operation of compound semiconductor devices, alloy devices (e.g., Si:Ge, AlGaAs) and heterojunction devices (junctions between semiconductors of different composition) are also considered because of the increased performance that is possible with *band-gap engineering*.

Fabrication, while an important part of semiconductor engineering, is often skipped in the interest of time. This material is introduced in Appendix C and can be assigned as read-only material if desired.

Supplemental topics are presented in a series of *Online Modules*. These modules, whose content is beyond that normally taught in a first course on semiconductor devices, contain material which supplements that of the book proper. For example, Online Module 7 describes some basic representative circuits utilizing CMOS devices. These Online Modules are available on the web for downloading.

ACKNOWLEDGMENTS

We would like to thank, first and foremost, our spouses Bill and Claire for their love, support, patience, and help. We are also grateful to the anonymous manuscript reviewers for their comments and suggestions, as well as staff at McGraw-Hill for all their help. We thank our students and their professors for valuable feedback on the first edition of the text. In particular, we thank Professor Gary Bernstein of Notre Dame University for his extensive comments and suggestions. We also wish to thank Cor Claeys and Rita Rooyackers of IMEC for scanning electron microscopy photographs of devices.

Materials

INTRODUCTION

Semiconductors form the basis of most modern electronic systems (e.g., computers, communication networks, control systems). While there are applications for other materials in electronics (e.g., magnetic materials in hard drives), this book concentrates on electronic devices that are based on semiconductors.

Understanding the operation and design of semiconductor devices begins with an understanding of the materials involved. In Part 1 of this book, we investigate the behavior of electrons in materials, starting with the atoms themselves. Then we progress to electrons in crystalline semiconductors.

We will see that classical mechanics does not provide a complete picture of electron activity in solids. In principle, one should instead use quantum mechanics to predict the electrons' behavior, but the application of quantum mechanics is not as simple as the more familiar classical or Newtonian mechanics. We will therefore introduce pseudo-classical mechanics, which modifies familiar classical equations to account for some quantum mechanical effects.

Some basic quantum mechanical concepts important for the understanding of device operation are covered in Chapter 1. (A more detailed discussion is contained in Supplement to Part 1, found after Chapter 4.) In Chapter 2, we cover pseudo-classical mechanics, which allows us to predict the reaction of electrons to complicated fields, while using simple and intuitive pseudo-classical equations.

The use of pseudo-classical mechanics will also allow us to draw and use energy band diagrams. These diagrams are indispensable for understanding and predicting the motion of the electrons and holes, and thus the current in semiconductors.

In Chapter 3, we will see that conductivity of semiconductors is controlled by the number of charge carriers available to carry current. The charge carriers in semiconductors are electrons and holes. Their numbers are controlled by the

concentrations of impurity elements that are intentionally added to the material. The carrier concentrations also depend on temperature and on whether light is shining on the sample.

It will emerge that there are two major forms of current in semiconductors, drift current and diffusion current. Drift current is caused by the presence of an electric field, whereas diffusion current arises when the carrier concentrations vary with position.

Chapter 4 covers nonhomogeneous semiconductor materials, in which the doping or the material composition itself may vary with position. These variations can lead to internal electric fields that can enhance device performance. Most modern semiconductor devices have regions of such nonhomogeneous materials.

The Supplement to Part 1 contains additional topics relevant to semiconductor materials, including a more detailed discussion of quantum mechanics and phonons.

We will start with electrons in atoms. ■

Electron Energy and States in Semiconductors

1.1 INTRODUCTION AND PREVIEW

We begin our study of semiconductors with some fundamental physics of how electrons behave in matter. The ability to control the movement of electrons in solids is the basis of semiconductor device engineering. To understand the electronic properties of these devices, it is necessary to understand the electronic properties of the materials from which they are made and how those properties are affected by impurities (intentional and unintentional), temperature, applied voltages, device structures, and optical radiation.

Since solids are composed of atoms, we start by examining the electronic properties of atoms, and then extending those results to simple molecules and solids. In particular, the results for silicon (Si) and gallium arsenide (GaAs) are emphasized, two commonly used semiconductors in integrated circuits and semiconductor devices. Several other semiconductors and semiconductor alloys important in modern devices are also discussed.

As we investigate the atom, we'll be using quantum mechanics, a branch of science that is needed to accurately describe the behavior of very small objects such as atoms and electrons. We will see as we go along that quantum mechanics is based on the idea that energy can exist only in discrete packets, or quanta. The size of a quantum is so small that it doesn't affect one's results when one is computing the momentum or velocity of large objects such as automobiles or dust particles, but the quantum description is extremely important for electrons and atoms.

An understanding of quantum mechanics is not simple to obtain, and its use to calculate properties of more than a few systems in closed form is difficult. Fortunately, however, in semiconductors the behavior of electrons of interest can be determined by *pseudo-classical mechanics*, in which classical formulas such as Newton's laws and the Lorentz equation can be used, with the true electron mass replaced by an *effective mass*. As a result, in this section, a minimal discussion

of quantum mechanics is presented. A somewhat greater discussion of quantum mechanics appropriate to some of the electronic processes in semiconductor devices is presented in the Supplement to Part 1, after Chapter 4.

The key to understanding semiconductors is to appreciate the physical interpretation of the mathematical results. Physical understanding is emphasized in this book.

1.2 A BRIEF HISTORY

In the early twentieth century, scientists were trying to develop models that would explain the results observed from such experiments as the scattering of X-rays, the photoelectric effect, and the emission and absorption spectra of atoms. In 1910, J. J. Thomson proposed a model of the atom in which a sphere of continuous positive charge is embedded with electrons, as shown in Figure 1.1a. Ernest Rutherford, in 1911, offered an improvement to the Thomson model: In the Rutherford model of the atomic structure, all of the positive charge and virtually all of the atom's mass were assumed to be concentrated in a small region in the center of the atom. This nucleus is often treated as a sphere with a radius on the order of 10^{-14} meters. The negatively charged electrons were assumed to orbit about the positively charged nucleus, much as planets orbit the sun or satellites orbit the earth.

In 1913, Neils Bohr assumed that the electrons in the Rutherford model of the atom orbited the nucleus in circles, as shown in Figure 1.1b. From this, he predicted that for the atom to be stable, the electrons could have only certain energies, or that the energies would be *quantized*. Energy and many other observables (properties that can be directly measured) are expressed in terms of Planck's constant. Planck's constant, h , has the value 6.63×10^{-34} joule–seconds.

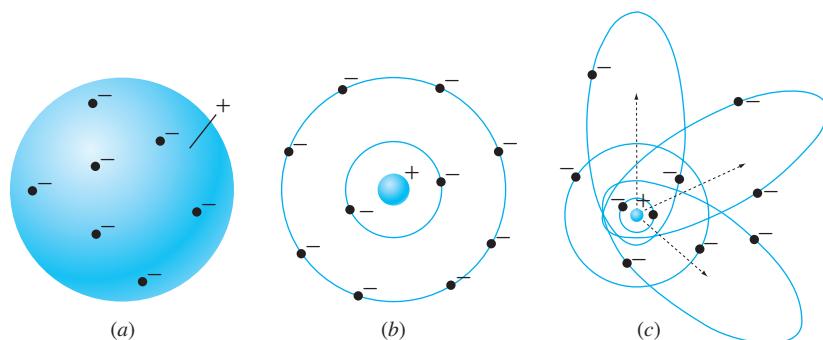


Figure 1.1 (a) The Thomson model of an atom, in which the positive charge is uniformly distributed in a sphere and the electrons are considered to be negative point charges embedded in it; (b) the Bohr model, in which the positive charge is concentrated in a small nucleus and the electrons orbit in circles; (c) the Wilson-Sommerfeld model, which is similar to the Bohr model except that it allows for elliptical orbits.

The energies Bohr predicted for electrons in atoms were in excellent agreement with the experimental results obtained from spectroscopic data.

In 1916, Wilson and Sommerfeld generalized the Bohr model to apply it to any physical system in which a particle's motion is periodic with time. This modification allows for the possibility of elliptical orbits, as shown in Figure 1.1c.

1.3 APPLICATION TO THE HYDROGEN ATOM

In this section, we briefly review the Bohr model of the hydrogen atom. The hydrogen atom is emphasized because *hydrogen-like impurities* are important in semiconductor devices, and these impurities can be treated in a manner analogous to the Bohr model. In the Supplement to Part 1, we will compare these results to those obtained using quantum mechanics as represented by Schrödinger's equation.

1.3.1 THE BOHR MODEL FOR THE HYDROGEN ATOM

We start with the Bohr model, in which the electrons revolve around the nucleus in circular paths. Because the mass of the nucleus is 1.67×10^{-27} kg, 1830 times that of the electron, the nucleus is considered to be fixed in space.

We consider as an example the neutral hydrogen atom, which has one orbiting electron, and we treat the electron and nucleus both as point charges. The coulomb force between two particles with charges Q_1 and Q_2 is

$$F = \frac{Q_1 Q_2}{4\pi\epsilon_0 r^2} = \frac{-q^2}{4\pi\epsilon_0 r^2} \quad (1.1)$$

where r is the distance between the two charges and $\epsilon_0 = 8.85 \times 10^{-12}$ farads/meter is the permittivity of free space (because there is only free space between the nucleus and the electron). The expression at the far right-hand side of Equation (1.1) is obtained by recognizing that the hydrogen nucleus has only one proton, so Q_1 is equal to $+q = 1.602 \times 10^{-19}$ Coulombs, the elemental charge, and the charge of the electron Q_2 is equal to $-q$. The resulting negative sign in Equation (1.1) indicates that the force is attractive.

We now have an expression for the attractive (centripetal) force between the two particles, and we recall from classical mechanics that the force F on a particle is equal to minus the gradient of the potential energy, or

$$F = -\nabla E_p = -\frac{dE_p}{dr} \quad (1.2)$$

In the last expression, the gradient is taken in the r direction, and E_p is the potential energy of the electron at position r . Equation (1.2) with the aid of (1.1) can be rewritten as

$$dE_p = dE_p(r) = -Fdr = \frac{q^2 dr}{4\pi\epsilon_0 r^2} \quad (1.3)$$

One can integrate both sides to obtain E_P , but there will be a constant of integration. The actual value of the potential energy is arbitrary (as is the choice of the constant), since the value of the potential energy depends entirely on one's choice of reference. We can choose a convenient reference by noting that the coulomb force at infinite distance is zero. It makes sense for this case, then, to choose $r = \infty$ as a reference point, so we define the potential energy at $r = \infty$ as the *vacuum level*, E_{vac} :

$$E_P(r = \infty) = E_{\text{vac}} \quad (1.4)$$

This is the energy required to free the electron from the influence of the nucleus, essentially by moving the electron infinitely far away from it. If the electron is infinitely far from the nucleus, it cannot really be considered part of the atom—it is now a free electron in vacuum.

Now we can solve Equation (1.3) for a given value of r :

$$\int_{E_P}^{E_{\text{vac}}} dE_P = \int_r^{\infty} \frac{q^2 dr}{4\pi\epsilon_0 r^2} \quad (1.5)$$

where E_P is the electron potential energy at some distance r from the nucleus. Integrating both sides and rearranging, we obtain

$$E_P = E_{\text{vac}} - \frac{q^2}{4\pi\epsilon_0 r} \quad (1.6)$$

Figure 1.2 shows a plot of the r dependence of E_P . From Equation (1.1), and since the force is equal to minus the gradient (slope) of the potential energy, we see that the force on the electron is directed toward the nucleus, or the coulomb force is centripetal. Since the nucleus is considered to be a point charge, E_P approaches negative infinity as r approaches zero. Since the radius of the nucleus is on the order of 10^{-14} m, however, and the radius of the smallest electron orbit is on the order of 10^{-10} m, the potential energy reaches a minimum near $r = 0$.

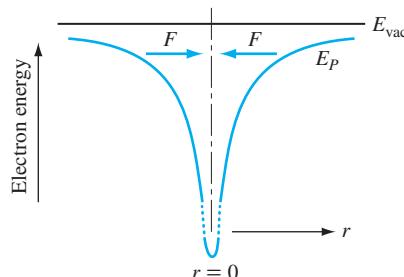


Figure 1.2 Potential energy diagram for an electron in the vicinity of a single positive point charge. The electron is considered to be a point charge.

Since the electron is revolving in a circle of radius r around the nucleus, we know from Newtonian mechanics that its centrifugal force is equal to

$$F = \frac{mv^2}{r} \quad (1.7)$$

For the atom to be stable, the net force on the electron must be zero. Equating our previous expression for the centripetal force due to the coulomb attraction [Equation (1.1)] to the centrifugal force [Equation (1.7)], we can write

$$\frac{mv^2}{r} - \frac{q^2}{4\pi\epsilon_0 r^2} = 0 \quad (1.8)$$

Bohr also postulated that the integral of the angular momentum around one complete orbit is an integer multiple of Planck's constant h :

$$\oint P_\theta d\theta = \int_0^{2\pi} mrvd\theta = nh \quad (1.9)$$

where n is an integer. Since the orbit is assumed circular in the Bohr model, r is a constant, and so are the potential energy E_P and the speed v . Therefore, the integral becomes

$$2\pi mrv = nh \quad (1.10)$$

There is a solution for each integer value of n , so we write

$$mv_n r_n = n \frac{h}{2\pi} = n\hbar \quad (1.11)$$

Here we have introduced a new symbol; it turns out that engineers and physicists (and now you) use the quantity $h/2\pi$ so much that there is a special character for it, \hbar , pronounced "h-bar." The subscripts n in Equation (1.11) indicate the particular orbital radius or speed associated with a specific quantum number n .

If we simultaneously solve Equations (1.8) and (1.11), we can derive an expression for the *Bohr radius of the nth state*, where by "state" we mean the properties associated with a particular value of n :

$$r_n = \frac{4\pi\epsilon_0 n^2 \hbar^2}{mq^2} \quad (1.12)$$

and the speed of the electron in that particular state is

$$v_n = \frac{q^2}{4\pi\epsilon_0 n \hbar} \quad (1.13)$$

Our primary goal, however, is to find the energies associated with these states. We know that the total energy of a system is equal to the kinetic energy plus the potential energy. The kinetic energy of the n th energy level is

$$E_{K_n} = \frac{1}{2}mv_n^2 = \frac{mq^4}{2(4\pi\epsilon_0)^2 n^2 \hbar^2} \quad (1.14)$$

For the n th energy level, we can find r_n from Equation (1.12) and use that in Equation (1.6) to write for the potential energy

$$E_{Pn} = E_{\text{vac}} - \frac{mq^4}{(4\pi\epsilon_0)^2 n^2 \hbar^2} \quad (1.15)$$

Thus, the total energy E_n is

$$E_n = E_{Kn} + E_{Pn} = E_{\text{vac}} - \frac{mq^4}{2(4\pi\epsilon_0)^2 n^2 \hbar^2} \quad (1.16)$$

We say that the energy is *quantized*. It can have only discrete values associated with the quantum number n . We note that $n = 1$ refers to the smallest radius and energy of the electron in the Bohr model, $n = 2$, the next larger values, etc.

EXAMPLE 1.1

Find the energies and radii for the first four orbits in the hydrogen atom.

■ **Solution**

$$\begin{aligned} E_n &= E_{\text{vac}} - \frac{mq^4}{2(4\pi\epsilon_0)^2 n^2 \hbar^2} \\ &= E_{\text{vac}} - \frac{(9.11 \times 10^{-31} \text{ kg})(1.60 \times 10^{-19} \text{ C})^4}{(2)(4)^2 (3.1416)^2 (8.85 \times 10^{-12} \text{ F/m})^2 (1.05 \times 10^{-34} \text{ J} \cdot \text{s})^2} \left(\frac{1}{n}\right)^2 \quad (1.17) \\ E_n &= E_{\text{vac}} - \left(\frac{1}{n}\right)^2 (2.18 \times 10^{-18} \text{ J}) \\ &= E_{\text{vac}} - \left(\frac{1}{n}\right)^2 (13.6 \text{ eV}) \end{aligned}$$

Here a new unit of energy is introduced, the *electron volt* (eV). The electron volt is defined as the amount of energy acquired by an electron when it is accelerated through 1 volt of electric potential. To convert between SI (International System) units (joules) and electron volts, use

$$1 \text{ eV} = 1.60 \times 10^{-19} \text{ joules}$$

Electron volts are *not* SI units, and therefore they must be used with care in calculations.

The Bohr radii can be calculated from Equation (1.12):

$$\begin{aligned} r_n &= \frac{4\pi\epsilon_0 n^2 \hbar^2}{mq^2} = \frac{(4)(3.1416)(8.85 \times 10^{-12} \text{ F/m})(1.05 \times 10^{-34} \text{ J} \cdot \text{s})^2}{(9.11 \times 10^{-31} \text{ kg})(1.60 \times 10^{-19} \text{ C})^2} \times n^2 \quad (1.18) \\ r_n &= 0.0526 n^2 \text{ nm} \end{aligned}$$

The energies and Bohr radii of the first four energy levels are given in Table 1.1. These energies and radii are plotted in Figures 1.3 and 1.4, respectively.

Table 1.1 The first four Bohr energies and orbital radii for the hydrogen atom

E_n	r_n
$E_1 = E_{\text{vac}} - 13.6 \text{ eV}$	$r_1 = 0.0526 \text{ nm}$
$E_2 = E_{\text{vac}} - 3.40 \text{ eV}$	$r_2 = 0.212 \text{ nm}$
$E_3 = E_{\text{vac}} - 1.51 \text{ eV}$	$r_3 = 0.477 \text{ nm}$
$E_4 = E_{\text{vac}} - 0.850 \text{ eV}$	$r_4 = 0.848 \text{ nm}$

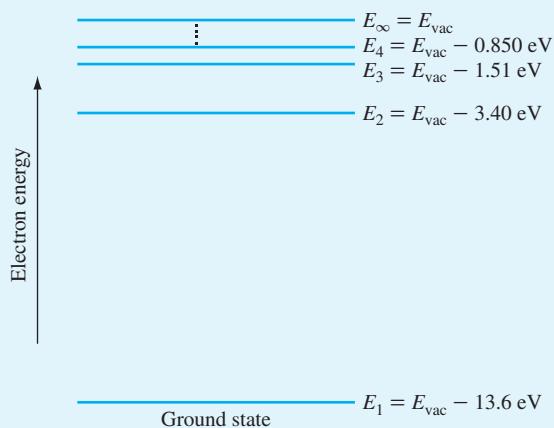


Figure 1.3 Allowed energies in the hydrogen atom. Higher energies occur increasingly close to each other, approaching the vacuum level.

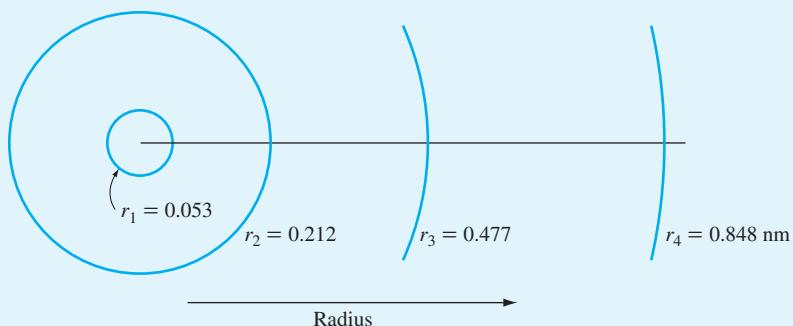


Figure 1.4 Radii of the first four electronic orbits of the hydrogen atom, according to the Bohr model.

There are several things to notice in this example. First, the differences between the vacuum energy level and the allowed energies vary as $1/n^2$. Thus the higher the quantum number, and therefore the energy, the closer together (in energy) the energy levels are. Second, the Bohr radius varies as n^2 . This means that the higher the energy level, the farther the electron is from the atomic nucleus. If the electron has energy greater than E_{vac} , the coulombic force is not enough to keep the electron bound to the atom. In this case, a hydrogen ion (H^+) is created as the electron leaves the atom.

Also, notice that we do not give a specific value for the energy of a particular state, but rather we express the energies as so many electron volts from some reference level (in this case E_{vac}). It is pointless to say, “This level is at 10 eV,” since 10 eV could be anywhere, depending on your choice of reference. This point cannot be emphasized enough. Potential energy and thus total energy are arbitrary. The kinetic energy, however, is not arbitrary.

The energy of a state must always be expressed as an energy difference—the difference between the energy of the state and some known reference, for example, $E_{\text{vac}} - E$.

Finally, it should be pointed out that although the number of possible states is infinite, once an electron occupies one of these states, the hydrogen atom becomes neutral and other electrons are not attracted.

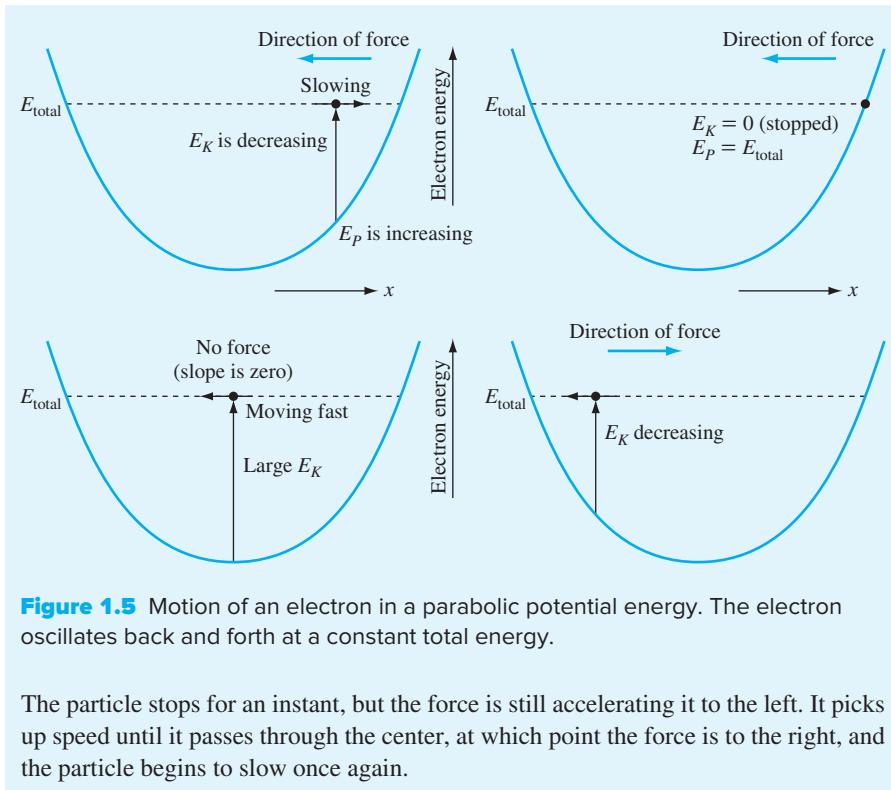
EXAMPLE 1.2

Consider a particle in a one-dimensional universe, oscillating in the parabolic potential energy shown in Figure 1.5. This represents an approximation to an electron in a modern quantum well laser.¹ The potential energy function is a parabola, and the particle is a simple harmonic oscillator. Explain the motion of the particle using the energy diagram, paying attention to where the kinetic energy is largest, where it is smallest, and the directions of the forces.

■ Solution

Conservation of energy dictates that the particle must remain at a constant energy. Thus, it oscillates back and forth at this particular energy. When the particle is in the center, it has the smallest potential energy and the largest kinetic energy, and thus the largest velocity. Recall that the force on the particle is equal to the negative of the slope of the potential energy [Equation (1.2)]. Therefore, as the particle travels through the center, e.g., to the right, the force is to the left (the slope is positive to right of center). Thus, the particle decelerates. It continues to slow as it moves to the right. The total energy is constant, but since the potential energy is increasing, the kinetic energy decreases. When the particle gets to the edge, it will have zero kinetic energy.

¹It is called a *quantum well* because the potential energy forms a “well” with quantized energy states.



1.3.2 APPLICATION TO MOLECULES: COVALENT BONDING

We now extend the Bohr model of the hydrogen atom to the hydrogen ion, H_2^+ , and the hydrogen molecule, H_2 . Figure 1.6a indicates the electron energy diagrams for two isolated hydrogen nuclei. By *isolated*, we mean that the nuclei are sufficiently far apart that they do not influence each other. The energy levels for an electron are the same for each nucleus, those calculated in the previous section. They are quantized as we saw.

When the nuclei are allowed to approach each other, an electron would be influenced by both nuclei according to Coulomb's law. The electron potential energy of a single electron (H_2^+ ion) at any point is now

$$E_P = E_{\text{vac}} - \frac{q^2}{4\pi\epsilon_0 r_1} - \frac{q^2}{4\pi\epsilon_0 r_2} \quad (1.19)$$

where r_1 and r_2 are the distances between the electron and each of the two nuclei, respectively. Figure 1.6b shows the allowed electron energy states, which are still quantized. Notice that an electron in the ground state (lowest energy state) would be bound to one of the nuclei, but an electron in an excited state could travel back and forth between the nuclei, in effect shared by the two atoms. Since

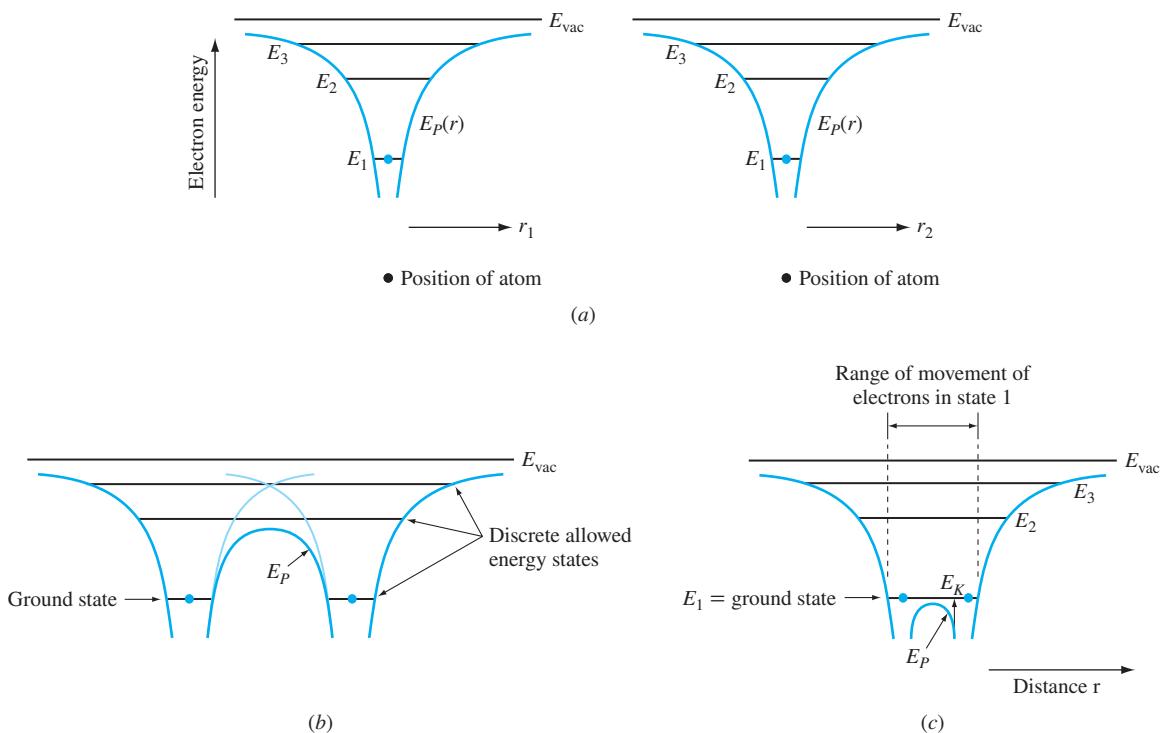


Figure 1.6 (a) Energy band diagram for two noninteracting hydrogen atoms; (b) as the nuclei are brought together, the upper energy levels merge, and electrons in those levels are shared between the atoms; (c) the nuclei are sufficiently close together that all energy levels are shared. Since the lowest level is usually the only occupied level for hydrogen, if it is occupied by two electrons the H_2 molecule is stable.

electrons tend to seek their lowest allowed energy, this condition of the electron being in one of the upper levels would not last long—the electron would quickly revert to the ground state.

Figure 1.6c shows the energy band diagram for the case where the separation is small enough that the potential energy maximum between the nuclei is below the ground state energy (E_1). In this situation, an electron in the ground state would be shared by the two nuclei, oscillating between the two positions at which $E = E_P$.

An interesting thing happens in this case, however. Since *each* nucleus has a ground state associated with it, it turns out that two electrons can occupy these ground states for a neutral H_2 molecule. Although the presence of the second electron will alter the electrostatic forces and therefore the energy band diagram slightly, the result is that both electrons will oscillate in the region indicated.

In the region between the nuclei, the kinetic energy ($E_K = E_1 - E_P$), and thus the velocity, is small. The electrons travel more slowly in this region, or on the average, the electrons spend most of their time between the two nuclei. The

electrons therefore create a negatively charged “electron cloud” in this region that tends to attract the two positive nuclei together. If the internuclear spacing is too small, however, the potential energy E_P decreases, which increases the kinetic energy E_K since total energy E is conserved. As the kinetic energy and therefore the electron speed increases, the electron cloud effect is reduced, lessening the attractive force. At a particular spacing, the electron-cloud–induced nuclear bonding is stable, and a stable H_2 molecule results. This mechanism is referred to as *covalent bonding*.

1.3.3 QUANTUM NUMBERS AND THE PAULI EXCLUSION PRINCIPLE

A basic premise of quantum mechanics is that the energies in an atom are quantized, or exist only at certain discrete values. We earlier introduced the quantum number n , called the *principal quantum number*. It describes the energy of an electron in an allowed state. As indicated in the Supplement to Part 1, from quantum mechanical considerations (e.g., Schrödinger’s equation) there are three other quantum numbers, including the azimuthal quantum number n_θ , which describes the ellipticity of an orbit. The last two describe the tilt of the orbits, and *spin*² of the electron. The energy is determined primarily by the principal quantum number. The other three quantum numbers can affect the energy, but only slightly.

The physical meanings of these quantum numbers are not essential to the understanding of transistors, but the *Pauli exclusion principle* is essential, because it governs which states may be occupied:

PAULI EXCLUSION PRINCIPLE

No two electrons in an interacting system can have the exact same set of quantum numbers.

For example, in the lowest energy orbit of an atom, $n = 1$. We know from freshman chemistry that this state can hold two electrons; those two electrons must have different spin quantum numbers, either $+\frac{1}{2}$ or $-\frac{1}{2}$.

In the $n = 2$ state, there are two possible orbital shapes. One orbit is spherically symmetric and holds two electrons of opposite spin (the “s” state). There are three elliptical orbits with the same shape but different orientations. Each of these can hold two electrons of opposite spin, bringing the maximum number

²Spin is a purely quantum-mechanical parameter and has no analog in classical mechanics. It can be crudely considered to result from considering the electron to be a spinning sphere. The charge in the spinning electron then produces a magnetic field. The magnitude of this spin magnetic field is quantized, and it can have two possible orientations with respect to an applied magnetic field.

of electrons in the second “shell” to eight. The periodic table is built on these quantum numbers.

In the following section, we will see how the Pauli exclusion principle affects the electron energies of solid materials such as silicon crystals.

1.3.4 COVALENT BONDING IN CRYSTALLINE SOLIDS

In this section we investigate the mechanisms by which silicon atoms bond to each other to form crystals. Silicon, by far the most important semiconductor material, has atomic number 14. It has two electrons in the first shell (which is full), eight in the second (also full), and four electrons in the half-full third shell. The third shell can hold up to eight electrons. There exist all of the higher shells, too, as for any atom, but they are empty except when the atom is in an excited state.

Energy Bands Silicon crystallizes into the so-called diamond structure in which each atom has four “nearest neighbors.” The arrangement of the atoms is discussed briefly in Section 1.8. For now, we merely note that within the crystal, each Si atom shares one of its four outer electrons with each of its (four) nearest neighbors. Neighboring atoms are bound together by two electrons as shown in the bonding description of Figure 1.7a, where the crystal is shown in two dimensions for clarity. In this bonding model, the crystal is held together by the electrostatic forces between the bonding electrons and the positive ion cores (covalent bonding).

Since the interatomic spacing is small (on the order of a quarter of a nanometer), an electron is influenced by all nearby nuclei as well as nearby electrons. Just as for the hydrogen molecule, the potential energy of an electron is the sum of the potential energies due to each neighboring atom, as shown in Figure 1.7b. Here we are plotting the potential energy E_P as a function of position, in this case in the x direction. Note that except near the surface, the potential energy E_P is a periodic function with the periodicity of the crystal lattice.

An important feature of Figure 1.7b is that the discrete states associated with the isolated atoms (Figure 1.6a) are now broadened into energy bands in the crystal. For a crystal with N atoms, each state originating from a single atom splits into a band of N discrete states. In a crystal containing billions of atoms, these states are infinitesimally close to each other in energy, and each band is often considered to be continuous in energy. We think of each shell of the discrete Si atoms as combining into a band of allowed states in a silicon crystal.

In Figure 1.7b, the valence and conduction bands, shown in the figure, are common to all the atoms in the crystal, while the lower energy states, corresponding to the inner shells, are highly localized near the atomic nuclei. As would be expected, these lower states are only slightly influenced by the presence of neighboring atoms and thus are spread only slightly in energy compared with the higher energy states. Electrons in these two lower energy shells are tightly bound to their respective atoms and play no role in the operation of devices. We will therefore not discuss them further.

The third shell of atomic silicon splits into two bands in crystalline Si. The lower of these bands, the *valence band*, contains four electrons. These valence

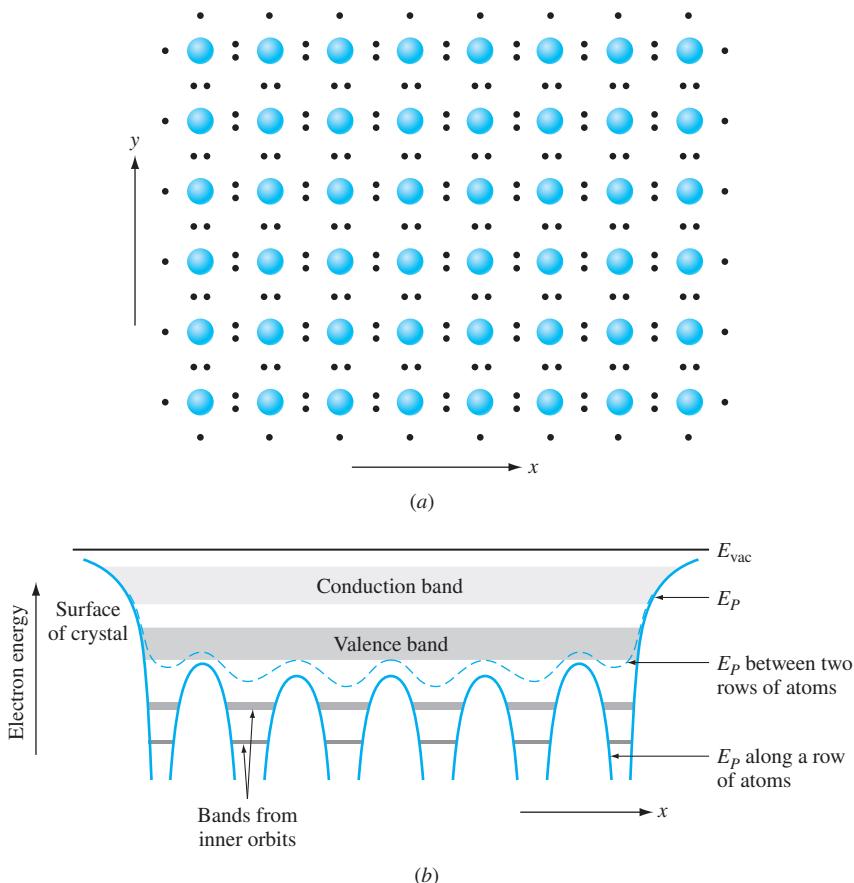


Figure 1.7 (a) Two-dimensional covalent bonding representation of a crystalline solid; (b) potential energy for an electron in that crystal along a row of atoms (solid line) and between rows (dashed line). In this representation the electron is considered a point charge.

electrons are responsible for covalent bonding in crystalline Si.³ The four vacant states in the third shell of atoms in Si form a band in crystalline Si called the *conduction band*. This conduction band is split off from the valence band by an energy of 1.12 eV at room temperature.

Note that in the regions of space between the atoms, the kinetic energies and thus the velocities of the electrons in the valence band are relatively small. The electrons, since they are moving slowly, spend much of their time in these regions between the atoms. Thus, an electron cloud exists between atoms. The attraction between these negative clouds and adjacent positive nuclei (covalent bonding) is what holds the crystal together.

³Recall that for hydrogen, it was the electrons in the first shell that took part in covalent bonding.

Ionic Bonding We have seen that silicon bonds covalently. Other elements in column IV of the periodic table can be expected to do the same. Apart from silicon, however, there are many other technologically interesting semiconductor materials, called *compound semiconductors*. Instead of every atom being of the same element, the crystal may consist of regular arrangements of different elements. One example is gallium arsenide (GaAs). This has the same general crystal structure as silicon, except that alternate atoms are gallium and the ones between them are arsenic. Gallium is in column III of the periodic table, and thus has three electrons in its outer shell. Arsenic, from column V, has five. Thus, when these two atoms are neighbors, they can each “fill” their outer shells with eight electrons by sharing electrons.

The electrons still spend most of their time in the region between the nuclei, but now there is a slight difference. Since the arsenic nucleus has greater positive charge, the electrons tend to be attracted toward the As side of the bond. This has, in essence, the effect of charging the As atom slightly negatively and ionizing the Ga atom slightly positively. The resulting coulombic force helps hold the crystal together. Indium phosphide (InP) is another example of a III-V semiconductor crystal. The III-V’s have a bonding character that is mostly covalent but partially ionic.

Cadmium telluride is an example of a II-VI semiconductor, and its bonding is even more ionic and less covalent. Table salt (NaCl) is an example of a I-VII crystal, although it is not a semiconductor. NaCl is considered to be primarily ionically bonded.

Electron Affinity, Ionization Potential, and Band Gap With the aid of Figure 1.8, we define some quantities useful in semiconductor electronics. The lowest energy in the conduction band is designated as E_C , the conduction band edge, while the highest energy in the valence band is denoted E_V , the valence band edge. We have not drawn the lower bands since they are not interesting to semiconductor device physics.

The *ionization energy* γ is defined as the minimum energy required to excite an electron from the top of the valence band in the crystal to the vacuum level.

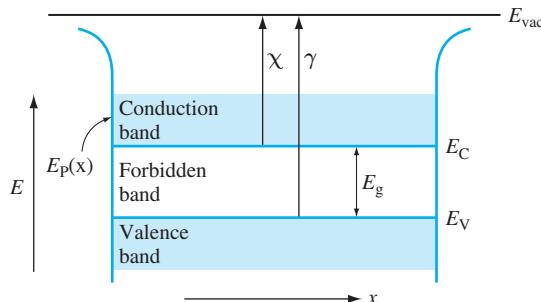


Figure 1.8 Definitions of vacuum energy E_{vac} , electron affinity χ , ionization energy γ , and the energy gap E_g .

Most of the electrons will be in the valence band, because those states are at lower energies than those in the conduction band. The electrons most likely to be ionized are those requiring the least amount of extra energy to leave the atom, and these are at the top of the valence band:

$$\gamma = E_{\text{vac}} - E_V \quad (1.20)$$

The *electron affinity* χ is defined as the energy difference between the vacuum level and the vacant state of lowest energy, in this case E_C :

$$\chi = E_{\text{vac}} - E_C \quad (1.21)$$

The energy gap E_g is the minimum energy required to excite an electron from the valence band to the conduction band

$$E_g = E_C - E_V \quad (1.22)$$

As can be seen from Figure 1.8,

$$\gamma = E_g + \chi \quad (1.23)$$

The electron affinity and forbidden band gap (normally referred to simply as *band gap*) are fundamental properties of a semiconductor. Values for some semiconductors of interest are shown in Table 1.2. These quantities will be used extensively in the description of device operation. Also included in Table 1.2 is an insulator, amorphous SiO₂, or silicon dioxide. It is included because its properties are important in some semiconductor devices.

It should be pointed out that while the band gap of a semiconductor can be measured accurately, this is not the case for the electron affinity. The band gap can be determined by the energy required to excite electrons from the valence band to the conduction band. This can be done by measuring the photoconductivity as a function of photon energy, as is discussed in Chapter 3. The photons penetrate into the bulk material where the band properties are well defined. To measure electron affinity, however, excited electrons must pass through the

Table 1.2 The electron affinity and band gap for some common materials at 300 K

Semiconductor	Electron affinity χ (eV)	Band gap E_g (eV)
Si	4.05	1.12
GaAs	4.07	1.43
Ge	4.0	0.67
GaP	4.30	2.25
AlSb	3.6	1.6
InP	4.4	1.35
SiC	4.0	2.2
GaN	4.1	3.437
GaSb	4.06	0.7
SiO ₂ (amorphous)	~0.9	~9

semiconductor surface to a collector in a vacuum. The band properties at the surface can differ from those in the bulk because of surface contamination or mechanical strain. As might be expected, there is some scatter in the published values for electron affinity. Those listed in Table 1.2 are considered reasonably reliable. Since SiO_2 is amorphous, the electron affinity and band gap are not well defined. The values given for SiO_2 are approximate.

Let us now discuss the occupancy of these energy bands. All things tend to seek their lowest energies, so in general one would expect to find most of the electrons in the valence band or lower at any given time. In fact, at absolute zero temperature, every electron is in the lowest possible state, implying that all the lowest allowed bands are full, up to and including the entire valence band.

At absolute zero (0 kelvins), every electron is in the lowest possible energy state. In a perfect semiconductor, every state in the valence band is occupied by an electron. Every state in the conduction band is empty.

Assuming the valence band to be filled, and no states in the forbidden band, the minimum energy that can be absorbed by a valence electron is E_g , enough to excite an electron from E_V to E_C . If the energy available is not enough to span the gap, the electron cannot absorb the energy—its final state would be a forbidden one. Thus, the electron will remain in the valence band. Note that current cannot flow at absolute zero—the electrons cannot move because all the states are filled; there is nowhere to go.

At room temperature, because of thermal agitation, a few⁴ electrons are excited into the conduction band, as seen in Figure 1.9. Each one eventually falls back down to a vacant state in the valence band, re-emitting the excess energy as heat or light. The average time an electron spends in the conduction band is called the “electron lifetime” or just “lifetime” and is on the order of 10^{-10} to 10^{-3} seconds, depending on the material.

Electrons in the conduction band are free to move around within the crystal. They travel at constant energy (between collisions), but now there are many empty states at the same energy into which an electron can move. This band is called the *conduction band* because the moving electrons carry current.

Notice that in Figure 1.9, we show only the top of the valence band and the bottom of the conduction band on the energy band diagram. Because everything tends to seek its lowest energy, and because the valence band is essentially full, we would expect to very rarely see any empty states in the valence band except near the very top. If an electron deep inside the valence band were excited into the conduction band, a valence electron of higher energy would immediately (within about 10^{-12} seconds) fall into the resultant vacant state. Similarly, if an electron were excited to an energy high up in the conduction band, it would very

⁴“Few” is relative—in silicon, the number is on the order of 10 billion electrons per cubic centimeter in the conduction band at room temperature, about 1 out of every 10^{15} electrons in the crystal.

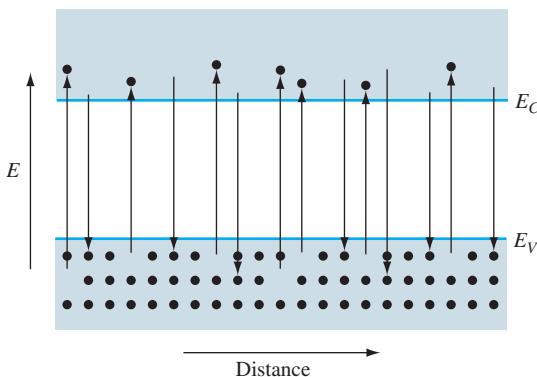


Figure 1.9 At room temperature, electrons are being excited up to the conduction band and relaxing back to the valence band. At any given moment there is some number of electrons in the conduction band.

quickly find a lower energy state. Therefore, most of the interesting activity is occurring near the top of the valence band and near the bottom of the conduction band.

At nonzero temperatures, there are a few empty states in the valence band. We call these empty states *holes*, and interestingly, they can move around, too. The state doesn't actually move, but if an electron adjacent to an empty state moves into that state, it leaves an empty state behind. From Figure 1.10 one can see that if an electron moves to a vacant state to the left, that has the same net effect as one hole moving one step to the right.

If, however, one electron represents a unit of current, that one hole could also be thought of as carrying current. If a negatively charged electron moves to the left, the current it “carries” goes to the right. If we think of the same current as

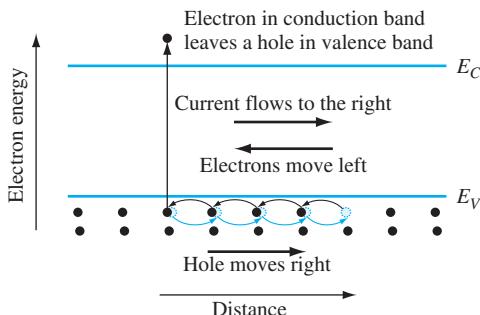


Figure 1.10 Movement of many electrons is treated as the movement of one positively charged “hole.”

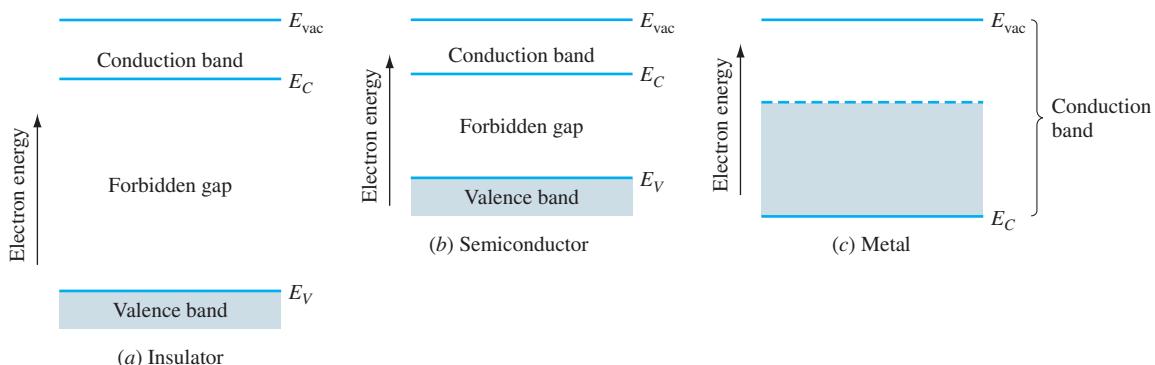


Figure 1.11 Energy band diagrams for (a) an insulator, (b) a semiconductor, and (c) a metal. The energies in the shaded regions are in general occupied.

resulting from the hole moving to the right, the hole must be positively charged. As we will see, it turns out to be very useful to think of holes as positively charged current carriers in semiconductors. The concept of holes is developed more completely in Chapter 3.

Polycrystalline and Amorphous Materials Polycrystalline materials have small regions (grains) of single-crystal material with different crystalline orientations. These grains have dimensions on the order of a few nanometers to a few millimeters. Because of the different crystalline orientations of the grains, the crystal periodicity at the grain boundaries is interrupted. This in turn affects the band structure near the grain boundaries.

Amorphous materials have some short-range order, but no long-range periodicity. Consequently, the band structure departs significantly from that of a crystalline material.

Insulators, Semiconductors, and Metals Silicon is called a semiconductor because it conducts better than an insulator but not as well as a metal. The reasons for this can be found, in part, from the energy band diagrams for these various materials, as seen in Figure 1.11. An insulator has a very wide forbidden gap,⁵ and thus few valence electrons have sufficient energy to become excited to the conduction band. Thus, it is unlikely even at room temperature for electrons to get excited into the conduction band, and insulators conduct poorly.

Compare this situation with a metal. In a metal, the highest occupied band, referred to as the conduction band, is only partially filled. Therefore, all of the electrons in this band can contribute to current, even at very low temperatures.

A semiconductor has an intermediate-size band gap, Figure 1.11b. At a finite temperature it has more electrons in the conduction band and holes in the valence

⁵The case is shown for a crystalline insulator. Many insulators (e.g., SiO_2) are amorphous, and the value of the forbidden gap is not well defined since a small concentration of states exist throughout the “forbidden band.”

band than an insulator, so it conducts better. On the other hand, it has nowhere near the conductivity of a metal. We will examine other important ramifications of semiconductors' energy band structure in future chapters.

1.4 WAVE-PARTICLE DUALITY

Classically, energy is considered to be transported either by waves (e.g., sound waves, water waves, electromagnetic waves) or by particles (e.g., electrons, bullets). However, particles such as electrons also exhibit wavelike behavior. For electrons, some of their behaviors are more easily explained by considering them to be waves, while other behaviors lend themselves to the particle description. In practice, the particle description is more convenient for solving some kinds of problems while the wave explanation is better for other problems. Both descriptions will be used to explain the operation of semiconductor devices such as transistors and lasers.

Similarly, what are classically considered to be waves can sometimes be considered to be particles. There are a number of experiments that demonstrate the wave-particle duality.⁶ Here we simply state the results important in the study of electron devices.

1. Classical waves have energies that are quantized. Each quantum of energy can be considered a particle. For electromagnetic radiation (e.g., light), these particles are called *photons*. For acoustic waves (e.g., sound), the particles are called *phonons*. Each such particle has energy

$$E = h\nu = \frac{h}{2\pi} \cdot 2\pi\nu = \hbar\omega \quad (1.24)$$

where h is Planck's constant and ν is the frequency of the classical wave. In this book we distinguish between photons and phonons by the expressions

$$E = h\nu = \hbar\omega_{\text{pht}}(\text{photons})$$

$$E = \hbar\omega_{\text{phn}}(\text{phonons})$$

where the subscripts pht and phn refer respectively to photons and phonons.

2. Classical particles can be considered to be waves possessing energy and wavelength. The electric field of an electromagnetic wave traveling in the x direction can be expressed by a simple sinusoidal function of amplitude A :

$$\mathcal{E}(x, t) = A \cos \left[2\pi \left(\frac{x}{\lambda} - \nu t \right) \right] = A \cos(Kx - \omega t)$$

where λ is the wavelength, t is time, and K is called the *wave vector*. Since the photon energy is $E_{\text{pht}} = h\nu = \hbar\omega_{\text{pht}}$, and $\hbar = h/2\pi$,

⁶The interested reader is referred to texts on atomic physics [1].

$$\mathcal{E}(x, t) = A \cos\left(Kx - \frac{E}{\hbar}t\right) \quad (1.25)$$

The wave vector K is also called the propagation constant. (In one-dimensional problems, it is often called the *wave number*. In three dimensions it is a vector, indicating direction.) It is related to the wavelength by

$$K = \frac{2\pi}{\lambda} \quad (1.26)$$

Matter can also be described using waves. An electron is an example of a particle that has mass but can be described by a wave as well. A matter wave is expressed in one dimension by a wave function, $\Psi(x, t)$

$$\Psi(x, t) = A \sin\left(Kx - \frac{E}{\hbar}t\right) \quad (1.27)$$

where the wavelength of the matter wave is

$$\lambda = \frac{2\pi}{K} \quad (1.28)$$

Matter waves are normally written as

$$\Psi(x, t) = A e^{i[Kx - (E/\hbar)t]} \quad (1.29)$$

and Equation (1.27) is the imaginary part of Equation (1.29).

These expressions are for waves in a vacuum where the amplitude A is constant with position. They will be modified slightly within a material, e.g., for electrons in a semiconductor.

1.5 THE WAVE FUNCTION

Equation (1.29) is an example of a one-dimensional wave function of a matter wave in vacuum. In general, the wave function is a function of three spatial dimensions and time; i.e.,

$$\Psi = \Psi(x, y, z, t) \quad (1.30)$$

For simplicity, unless the three-dimensional formulation is important for the concepts being introduced, the one-dimensional formulation will be used in this book.

1.5.1 PROBABILITY AND THE WAVE FUNCTION

A basic connection between the properties of the wave function $\Psi(x, t)$ and the behavior of the associated particle is the *probability density* $P(x, t)$. The probability density is given by

$$P(x, t) = \Psi^*(x, t)\Psi(x, t) \quad (1.31)$$

where $\Psi^*(x, t)$ is the complex conjugate of $\Psi(x, t)$. The probability that the particle will be found in region dx near the coordinate x at time t is

$$P(x, t) dx = \Psi^*(x, t) \Psi(x, t) dx \quad (1.32)$$

Since the particle must be somewhere, the probability of finding it in space is unity, or integrating over all space,

$$\int P(x, t) dx = \int \Psi^*(x, t) \Psi(x, t) dx = 1 \quad (1.33)$$

in which case the wave function is said to be *normalized*.

1.6 THE ELECTRON WAVE FUNCTION

As is discussed in the Supplement to Part 1, the wave function $\Psi(x, t)$ for an electron can be written as the product of a space-dependent part and a time-dependent part. In the Supplement to Part 1 it is demonstrated that for the potentials that do not vary with time, that is, no lattice vibrations (phonons), the time-dependent wave function is $e^{-j(E/\hbar)t}$, so that

$$\boxed{\Psi(x, t) = \psi(x) e^{-j(E/\hbar)t}} \quad (1.34)$$

where $\Psi(x, t)$ is the time-dependent wave function, $\psi(x)$ is the time-independent wave function, and E is the particle energy. The time-independent wave function $\psi(x)$ depends on the potential energy variation with position, $E_p(x)$. The time-independent wave function can be obtained from the time-independent Schrödinger's equation as indicated in the Supplement to Part 1:

$$\frac{-\hbar^2}{2m} \frac{d^2\psi(x)}{dx^2} + E_p(x)\psi(x) = E\psi(x) \quad (1.35)$$

Time-independent Schrödinger equation

Knowing the potential energy as a function of x , we can find $\psi(x)$ from Equation (1.35). Once the wave function is known, the behavior of the electron can be predicted. For example, in Example 1.1 we saw that a potential energy distribution that formed a potential well trapped an electron in a region of space. The probability of finding the electron in a particular region can be found by inserting the wave function into Equation (1.32).

For semiconductor devices, we are primarily interested in the behavior of electrons in semiconductor crystals. We will work up to this gradually.

1.6.1 THE FREE ELECTRON IN ONE DIMENSION

In the simplest model, we consider the (artificial) case of an electron of mass m_0 in a “crystal” in which the electron potential energy is a constant. Figure 1.12a shows the physical picture. Figure 1.12b shows the simplified energy diagram in the x direction. Note that there are still energy barriers at the surfaces—but we will simplify the problem by assuming the electron is nowhere near a surface, so that the problem has now been reduced to an electron in a universe of constant potential energy. This is called the *free-electron approximation*.

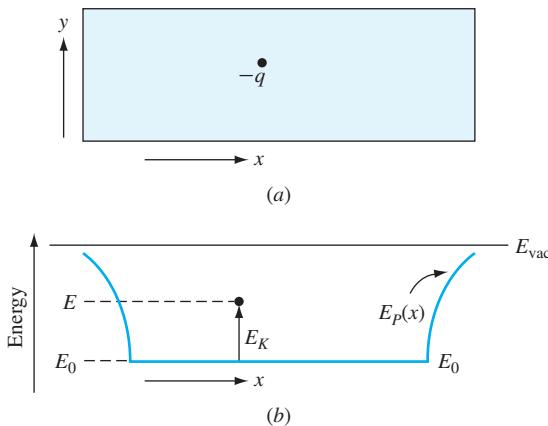


Figure 1.12 The free-electron model for an electron in a crystal; (a) the physical picture; (b) the potential is assumed constant everywhere inside the crystal.

Any electron, including this one, has a total energy $E = E_P + E_K$, which because of conservation of energy is a constant. That is, the electron must move horizontally on the energy diagram in Figure 1.12b. To change energies, the electron would have to give up energy or acquire it from somewhere—for example, by colliding with an atom or a photon. We'll keep the problem simple for the moment by making the further assumption that such collisions don't occur. The results then will be valid only during the time between collisions.

Since E_P is equal to some constant E_0 , we can write the time-independent Schrödinger's equation in one dimension [Equation (1.35)] as

$$-\frac{\hbar^2}{2m_0} \frac{d^2\psi(x)}{dx^2} + E_0\psi(x) = E\psi(x) \quad (1.36)$$

We have substituted the constant potential energy (independent of position) E_0 for $E_P(x)$. Both E_0 and the total energy E are constants, so rewriting this differential equation in the standard form results in

$$\frac{d^2\psi(x)}{dx^2} + \frac{2m_0}{\hbar^2}(E - E_0)\psi(x) = 0 \quad (1.37)$$

The solution to Equation (1.37) is

$$\psi(x) = Ae^{jKx} + Be^{-jKx} \quad (1.38)$$

or alternatively

$$\psi(x) = C \sin(Kx) + D \cos(Kx) \quad (1.39)$$

Here A and B or C and D are some constants to be determined from the boundary conditions, and

$$K = \sqrt{\frac{2m_0(E - E_0)}{\hbar^2}} = \sqrt{\frac{2m_0E_k}{\hbar^2}} \quad (1.40)$$

The total energy E minus the potential energy E_0 is just the kinetic energy E_K , as indicated in this equation. The time dependence of the wave function will be that shown in Equation (1.34). The total wave function for the free electron is therefore

$$\Psi(x, t) = A e^{j[Kx - (E/\hbar)t]} + B e^{j[-Kx - (E/\hbar)t]} \quad (1.41)$$

This equation has the form of two plane waves, one traveling in the $+x$ direction, and one traveling in the $-x$ direction. Certainly, the electron could be going either way. Also, note that the quantity E/\hbar appears where the angular frequency ω usually appears in an equation for an electromagnetic plane wave. It would be great if we could say that the electron “wave” has a frequency of $\omega = E/\hbar$ but, unfortunately, the value of E depends on the choice of potential energy reference. Since the reference energy is arbitrary, E/\hbar is not unique, and so the “frequency” of the electron is not a physically measurable quantity.

We have established that E/\hbar is not an easily interpreted quantity for the electron wave. What about K ? What is its meaning?

We consider the first term of Equation (1.41), $A e^{j[Kx - (E/\hbar)t]}$. If we consider a point of constant phase of this wave, for positive K the phase term $Kx - (E/\hbar)t$ remains constant with increasing time only if x also increases. In other words, the wave is going in the positive x direction if K is positive. Similarly for negative K , this term represents a wave traveling in the negative x direction. Equation (1.41) can be written as

$$\Psi(x, t) = A e^{j[Kx - (E/\hbar)t]} \quad (1.42)$$

and the direction of propagation is given by the sign of K . As indicated in connection with Equations (1.25) and (1.27), the quantity K is called the *wave vector*.

We use the wave vector to describe the wavelike properties of the electron. For example, the velocity of a point of constant phase of the wave is called the phase velocity and is given by

$$v_p = \frac{x}{t} = \frac{E}{\hbar K}$$

The phase velocity is not unique because the total energy E is dependent on the choice of potential energy reference.

The velocity associated with the center of mass of the particle is the “group velocity” v_g of the wave. The group velocity is given by

$$v_g = \frac{dx}{dt} = \frac{1}{\hbar} \frac{dE}{dK} \quad (1.43)$$

To further discuss the concept of wave vector we can rearrange Equation (1.40) as follows:

$$E = E_0 + \frac{\hbar^2 K^2}{2m_0} \quad (1.44)$$

This is the E - K relation for the free electron, shown in Figure 1.13. Note that the second term in Equation (1.44) is the kinetic energy (since $E = E_P + E_K = E_0 + E_K$) and looks a lot like $p^2/2m_0$, the kinetic energy of a classical particle. Here, though, the classical momentum p is replaced by $\hbar K$.

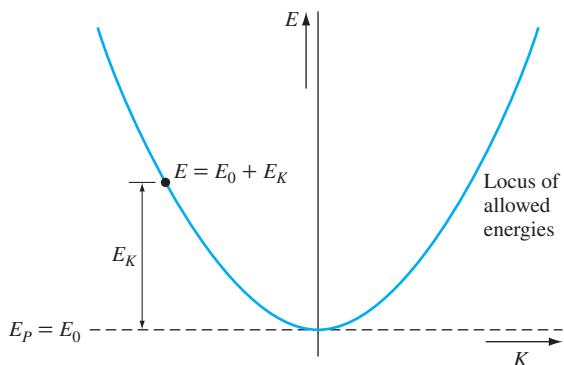


Figure 1.13 The E - K diagram for the free electron.

1.6.2 THE DE BROGLIE RELATIONSHIP

Since we can consider the free electron to be either a wave or a particle, it would be useful to have a way to connect the two different descriptions. Examining Equation (1.42) we see that at constant t , Ψ repeats itself whenever $Kx = n\pi$. That is, x increases by one wavelength λ for every integer n , or

$$K = \frac{2\pi}{\lambda} \quad (1.45)$$

If we take the derivative of Equation (1.44) we get

$$\frac{dE}{dK} = \frac{\hbar^2 K}{m_0}$$

Inserting this into Equation (1.43) and multiplying both sides by m_0 , we find that $m_0 v_g = \hbar K$, which is the classical momentum p of the free electron.

Since $K = 2\pi/\lambda$, and $\hbar = h/2\pi$, we obtain the de Broglie relationship

$$\lambda = \frac{h}{p} \quad (1.46)$$

It relates the wavelength (quantum mechanics) to the momentum of the particle (classical mechanics). It is important to remember that the de Broglie relationship holds *only* when the potential energy is constant over the path of the electron. It does not hold for electrons in solids.

We can get still more out of Equation (1.44). If we take the second derivative, we find that

$$\frac{\partial^2 E}{\partial K^2} = \frac{\hbar^2}{m_0} \quad (1.47)$$

This implies that the curvature (the second derivative) of the E - K locus is inversely proportional to the mass. For the parabolic E - K relation of the free electron, the curvature is constant, so the mass is constant, a reassuring result.

1.6.3 THE FREE ELECTRON IN THREE DIMENSIONS

If the potential energy is a constant in three-dimensional space [$E_P(x, y, z) = E_P(r) = E_0$], then the wave function becomes

$$\Psi(\vec{r}, t) = \Psi(x, y, z, t) = A e^{j[\vec{K} \cdot \vec{r} - (E/\hbar)t]} \quad (1.48)$$

where

$$\vec{K} = \hat{i} K_x + \hat{j} K_y + \hat{k} K_z \quad (1.49)$$

and

$$\vec{r} = \hat{i} x + \hat{j} y + \hat{k} z \quad (1.50)$$

Here x , y , and z are the coordinates of any position in space at which the wave function is being evaluated.

The magnitude of the wave vector \vec{K} in three dimensions is

$$|\vec{K}| = \sqrt{K^2} = \sqrt{\frac{2m}{\hbar^2}(E - E_0)} = \sqrt{\frac{2m}{\hbar^2}E_K} \quad (1.51)$$

or

$$E - E_0 = E_K = \frac{\hbar^2 K^2}{2m} \quad (1.52)$$

as we found for the free electron in one dimension.

For a three-dimensional system with constant potential energy, then,

1. There is one allowed energy “band”—the range of energies between E_0 and infinity represents the allowed energy states for the free electron.⁷ The energy band has a minimum at $E = E_0$, $K = 0$.
2. The velocity of the electron in a given direction is given by

$$v_x = \frac{1}{\hbar} \frac{\partial E}{\partial K_x} \quad v_y = \frac{1}{\hbar} \frac{\partial E}{\partial K_y} \quad v_z = \frac{1}{\hbar} \frac{\partial E}{\partial K_z} \quad (1.53)$$

3. From Equation (1.52), the E - K relation is a paraboloid in three dimensions. The curvature of the paraboloid is given by its second derivative $\partial^2 E / \partial K^2 = \partial^2 E_0 / \partial K^2 + \partial^2 E_K / \partial K^2$. For the free electron, E_0 is constant everywhere, so the curvature of the E - K curve is the same in any direction in K space (the total energy is also a constant):

$$\frac{\partial^2 E}{\partial K^2} = \frac{\partial^2 E_K}{\partial K^2} = \hbar^2 \frac{1}{m_0} \quad (1.54)$$

The mass m_0 is a constant for the free electron.

4. De Broglie’s relation $\lambda = h/p$ is valid for this case, and $p = mv$.

⁷Here we ignore the relativistic effects that limit the electron energy to mc^2 . Such high energies are not important in semiconductor electronics.

Although the free electron model is a poor approximation for electrons in semiconductors, the wave-particle duality discussed above leads to somewhat similar results for *quasi-free electrons* in real semiconductors, as discussed in the next section.

1.6.4 THE QUASI-FREE ELECTRON MODEL

Although the free-electron model ($E_P = \text{constant}$) is sometimes used for metals because of its simplicity, it isn't appropriate for semiconductors because it doesn't predict the existence of discrete energy bands. More realistic models that do predict multiple bands do so because they take into account the periodicity of the potential energy for an electron in a crystal lattice. Since the exact functional relation of the potential energy isn't accurately known, these models assume different periodic functions for the potential energy. Such a model is known as a *quasi-free electron model*. We examine one such case, starting with a one-dimensional universe and then extending the result to three dimensions.

Consider an electron deep within a solid, a one-dimensional crystal of lattice constant a . The corresponding potential energy function for the electron is shown in Figure 1.14, where the electron is considered to be a point charge. By assuming the electron is not close to any surface, we can neglect the effects of the surfaces, and the potential energy looks like an infinite, strictly periodic function.

Let the potential in the crystal be described by some periodic function $E_P(x)$. Since it is periodic in x with periodicity a , then $E_P(x) = E_P(x \pm na)$, where n is some integer. Then Schrödinger's equation [Equation (1.35)] becomes

$$\frac{d^2\psi(x)}{dx^2} + \frac{2m_0}{\hbar^2} [E - E_P(x)]\psi(x) = 0 \quad (1.55)$$

Compare this with Equation (1.37) for the electron in a constant potential.

Since we don't know the exact form of the potential energy $E_P(x)$, we can't solve Schrödinger's equation exactly. However, since we do know that the potential is periodic, we can use a theorem known as the *Bloch theorem* [2], which states that for an electron in a periodic potential, the time-independent wave function is

$$\psi(x) = U_K(x) e^{iKx} \quad (1.56)$$

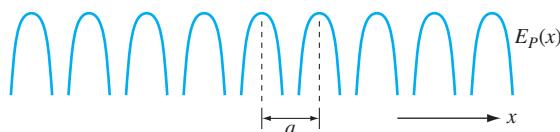


Figure 1.14 The electron potential energy in a crystal is a periodic function.

where $U_K(x)$ is some function that is also periodic in x with the periodicity of the crystal. Beyond that, we do not know the exact form of $U_K(x)$. The complete Bloch wave function is

$$\Psi(x, t) = U_K(x) e^{j[Kx - (E/\hbar)t]} \quad (1.57)$$

which represents a unit amplitude plane wave ($e^{j[Kx - (E/\hbar)t]}$), modulated by some periodic function $U_K(x)$ with period a . Since it is periodic,

$$U_K(x) = U_K(x + na) \quad (1.58)$$

where n is an integer. Even though we do not know everything about U_K , since it is periodic, we can deduce some things about the general characteristics of the E - K relationship.

For example, we know that the crystalline forces acting on the electron are independent of the direction of propagation (sign of K). Therefore

$$E(K) = E(-K) \quad (1.59)$$

That in turn implies that $E(K)$ has an extremum (either a relative maximum or minimum) at $K = 0$.

The time-independent wave function, Equation (1.56), can be multiplied by

$$1 = e^{j(2\pi nx/a)} e^{-j(2\pi nx/a)} \quad (1.60)$$

giving

$$\psi(x) = U_K(x) e^{-j(2\pi nx/a)} e^{jKx} e^{j(2\pi nx/a)} \quad (1.61)$$

We write this as

$$\psi(x) = U_K e^{-j(2\pi nx/a)} e^{j(K+2\pi n/a)x} \quad (1.62)$$

Observe that both U_K and $e^{-j(2\pi nx/a)}$ are periodic in x with the periodicity a of the lattice. We can therefore combine them into one Bloch modulation factor

$$U'_K = U_K e^{-j(2\pi nx/a)} \quad (1.63)$$

If we define

$$K' = K + \frac{2\pi n}{a} \quad (1.64)$$

we can rewrite Equation (1.62) as

$$\psi(x) = U'_K e^{jK'x} \quad (1.65)$$

Since Equation (1.65) is the same as Equation (1.62), and has the same periodicity, both equations must represent states corresponding to the same energy, or

$$E(K) = E\left(K + \frac{2\pi n}{a}\right) \quad (1.66)$$

Therefore, the $E(K)$ relation is periodic in K , with period $2\pi/a$.

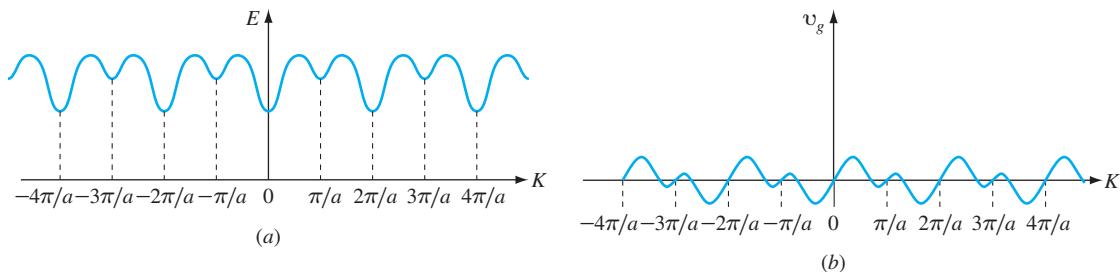


Figure 1.15 One possible E versus K diagram for the periodic potential. (a) E versus K ; (b) the group velocity v_g versus K .

A possible E - K relation illustrating these points is indicated in Figure 1.15a. We have shown that $E(K)$ is an even function [Equation (1.59)], and that $E(K) = E(K + 2\pi n/a)$, for $n = 0, \pm 1, \pm 2, \dots$. From this we can conclude that the E - K curve, whatever its shape actually is, is symmetric about the points $K = 0, \pm 2\pi/a, \pm 4\pi/a, \dots$, and thus it must have either equivalent maxima or equivalent minima at these points. In addition, because of this symmetry, the E - K relation must have other equivalent extrema midway between these values, or at $K = \pm\pi/a, \pm 3\pi/a, \dots$

From Figure 1.15a, we see that:

1. $E(K)$ is periodic in K space, with period $2\pi/a$.
2. Equivalent extrema in E exist at $K = 0, \pm 2\pi/a, \pm 4\pi/a, \dots$.
3. Equivalent extrema exist at $K = \pm\pi/a, \pm 3\pi/a, \dots$.
4. The slope of the E - K curve is zero at $K = 0, \pm\pi/a, \pm 2\pi/a, \pm 3\pi/a, \dots$.
5. The group velocity $v_g = (1/\hbar)dE/dK$ is periodic in K space with the same periodicity as the E - K curve. This is shown in Figure 1.15b.

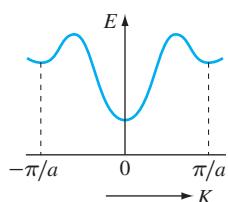


Figure 1.16 The reduced, or first Brillouin, zone.

In fact, it turns out that *all* of the measurable quantities are periodic with the same periodicity, $2\pi/a$ in K space. Since everything repeats, it is customary to consider just the region from $-\pi/a \leq K \leq \pi/a$. This is known as the *reduced zone* or the *first Brillouin zone* [3, 4], as shown in Figure 1.16 (compared with the extended zone of Figure 1.15). In fact, since E versus K is symmetric about $K = 0$, sometimes only half of the reduced zone is indicated.

For this example, a relative minimum was assumed to exist at $K = 0$ and thus also at $K = 2\pi n/a$. We could just as well have chosen a relative maximum, and there may be other extrema as well (we chose a case for which there is also a maximum in the interior of the zone). The exact shape and number of maxima and minima cannot be predicted from the general considerations presented here. The actual E - K diagram could be calculated in principle if $E_p(x)$ were known.

We now return to the E - K diagram of Figure 1.16. This is actually another form of the energy band diagram, since it shows the maximum and minimum energies of a particular band. Figure 1.16 is redrawn in Figure 1.17a to illustrate this point. The allowed range of energies is between E_2 and E_0 . In Figure 1.17b

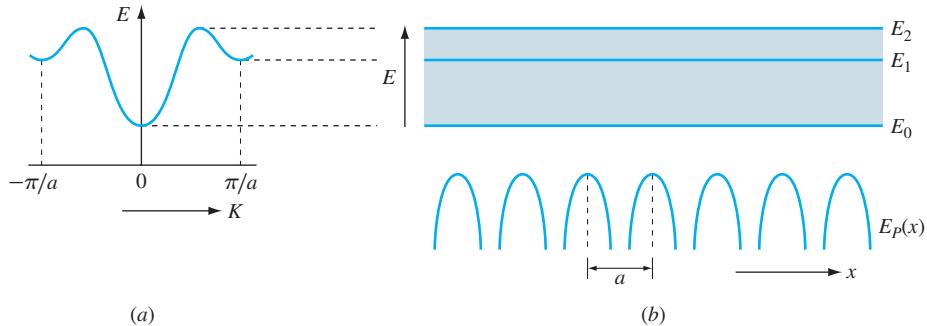


Figure 1.17 (a) The E - K diagram; (b) the corresponding energy band (E - x) diagram.

the energy band diagram in real space E - x is plotted, and the energies of the maxima and minima are plotted as functions of position in the crystal. The maxima and minima are the same everywhere in the crystal, as long as the electron is not near a surface. Also plotted against x is the potential energy function (the periodic function) for the electron, considered to be a point.

In a nearly empty band, electrons will tend to occupy the lower energy states and thus have properties associated with the minimum of the E - K plot in Figure 1.17a. Likewise, in a nearly filled band, the vacant states (holes) will be near the band maximum.

The E - K diagram and the regular energy band diagram complement each other. In the E - K diagram, no information about the position of a given electron is known, but the energy band diagram does not reveal anything about the wave vector (wavelength, direction of motion). In effect, the two descriptions, E - x and E - K , are Fourier transforms of each other. Both relationships are required for an understanding of electronic behavior in crystals, and we will use both types of diagrams.

Since the value of $U_K(x)$ changes periodically with position, Equation (1.56) for the wave function is a modulated plane wave, or Bloch wave function. Again, the wave vector K is related to the wavelength by

$$K = \frac{2\pi}{\lambda} \quad (1.67)$$

but this time

$$K \neq \sqrt{\frac{2m_0}{\hbar^2}(E - E_P(x))} \quad (1.68)$$

Since we assumed we knew nothing about the potential energy function $E_P(x)$ other than that it is periodic, no more is known about the exact form of $U_K(x)$ except that it is periodic.

We saw that for a free electron, the momentum $p = \hbar K$. This is not the case for the quasi-free electron model because the assumption of constant potential

energy everywhere does not apply. De Broglie's relation does not hold. We note, however, that $\hbar K$ is sometimes referred to as "crystal momentum" since it is analogous to classical momentum of the free electron.

In a three-dimensional crystal the potential energy $E_P(r)$ is periodic in all crystallographic directions and

$$\psi(r) = U_K(r) e^{i\vec{K} \cdot \vec{r}} \quad (1.69)$$

where $U_K(r)$ has the same periodicity as $E_P(r)$ and $|K| = 2\pi/\lambda$.

In the free-electron and quasi-free electron models, collisions were neglected, and so an electron particle could be represented by a single Bloch wave. In a real crystal, however, the electron mean free path is on the order of a few nanometers, and realistically such collisions must be considered. As a result, an electron particle must be considered as a superposition of wave functions. This topic is considered in more detail in the Supplement to Part 1.

1.6.5 REFLECTION AND TUNNELING

There is an interesting consequence to considering the electron to be a wave. Consider an electron wave approaching the potential barrier shown in Figure 1.18a. The electron has some energy lower than the height of the barrier. We expect the electron to be reflected from the barrier, but we cannot say that the wave stops abruptly at the barrier and turns around. Instead, the electron wave actually penetrates the barrier a short distance, as shown in the figure. This means that part of the electron charge is in the (classically) forbidden region. For example, if the barrier in the figure represents the surface of the crystal, some fraction of the electron charge is actually found outside the crystal.

This tunneling process is analogous to the penetration of electromagnetic waves a short distance (the *skin depth*) into a conductor. To reduce this penetration, *shielding* is employed to protect electronic systems. These systems are enclosed by a metal of high conductivity (e.g., copper or aluminum) and sufficient thickness to permit negligible electromagnetic penetration. There is an analogy for electrons: To prevent electrons from penetrating, a barrier must be of high enough potential and sufficient thickness. The barrier in Figure 1.18a is infinitely thick, so the electron penetrates a short distance but ultimately is reflected.

Next, suppose the barrier is very thin, as shown in Figure 1.18b. Since the electron wave penetrates the barrier a short distance, if the barrier is thinner than the penetration distance, part of the electron wave could be found on the other side. A physical electron cannot be divided in two, so it must end up either on one side or the other. The wave function of the electron extends through the barrier, however. Recall that the wave function multiplied by its complex conjugate is a probability density function [Equation (1.32)]. The quantum mechanical interpretation is that since the wave function extends into the allowed region on the far side of the barrier, there is some probability that the electron will "tunnel" through the barrier and end up on the other side. The probability is quite small, and usually the electron will be reflected from the barrier. Although the

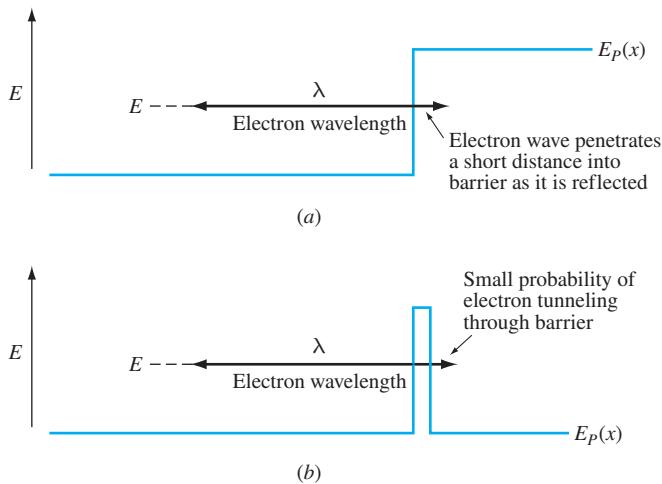


Figure 1.18 An electron wave is extended in space. (a) When the wave reflects from the potential barrier, the electron wave function extends a short distance into the forbidden region. Thus some fraction of the electron charge is found to the right of the barrier. (b) If the barrier is very thin, the electron wave function Ψ may extend all the way through it. Since the probability density $\Psi^*\Psi$ is not zero on the far side of the barrier, there is some (small) chance that the electron will cross through the barrier and emerge on the other side.

probability of a given electron tunneling through the barrier can be minuscule, the number of electrons striking the barrier can be large, with a resulting tunneling current that is not necessarily negligible.

The penetration of the electron into a barrier and the tunneling of electrons through thin potential barriers are both important quantum mechanical effects that very much affect the operation (and reliability) of semiconductor devices, as we will see throughout this book. These effects are discussed in more detail in the Supplement to Part 1 and in the chapters on device operation.

1.7 A FIRST LOOK AT OPTICAL EMISSION AND ABSORPTION

We have discussed electrons being in the conduction band or in the valence band of a semiconductor, and we noted that electrons can change energy states, moving from one band to the other. When an electron makes a transition, e.g., from the valence band to the conduction band, it has to acquire the extra energy from somewhere. The extra energy can come from vibration of the crystal lattice (phonons) or from an optical source (photons) or a combination of the two.

As indicated earlier, although we usually think of light as electromagnetic wave energy, light can also be thought of as consisting of particles or photons. The energy of a photon is related to the frequency of the light wave:

$$E = h\nu = \hbar\omega_{\text{ph}} \quad (1.70)$$

where h is Planck's constant and ν is the frequency.

Light energy must be absorbed or emitted in integer multiples of $h\nu$. That is, only an integral number of photons can be absorbed or emitted. For example, consider a beam of red light whose wavelength is 600 nm. The frequency of the light (wave) is given by:

$$\nu = \frac{c}{\lambda} \quad (1.71)$$

where c is the speed of light and λ is the wavelength. Thus for $\lambda = 600$ nm, the corresponding frequency is

$$\nu = \frac{3.0 \times 10^8 \text{ m/s}}{600 \times 10^{-9} \text{ m}} = 5 \times 10^{14} \text{ Hz} \quad (1.72)$$

or 500 terahertz. The energy of a single photon is then

$$\begin{aligned} E = h\nu &= (6.63 \times 10^{-34} \text{ J} \cdot \text{s})(5 \times 10^{14} \text{ s}^{-1}) = 3.3 \times 10^{-19} \text{ J} \\ &= 3.3 \times 10^{-19} \text{ J} \cdot \frac{1 \text{ eV}}{1.6 \times 10^{-19} \text{ J}} = 2.06 \text{ eV} \end{aligned} \quad (1.73)$$

Light whose wavelength is 600 nm can be emitted or absorbed only in multiples of 2.06 eV.

EXAMPLE 1.3

An electron in a hydrogen atom can switch from one allowed energy level to another by either absorbing or emitting energy. Scientists use the spectra of the light emitted by various materials to determine their energy structure. For each of the following spectral lines, determine the initial and final energy states.

Solution

- a. $\lambda_1 = 656.28$ nm (red). From Equations (1.70) and (1.71), the energy corresponding to this wavelength is

$$E = \frac{hc}{\lambda} = \frac{(6.63 \times 10^{-34} \text{ J} \cdot \text{s})(2.998 \times 10^8 \text{ m/s})}{656.28 \times 10^{-9} \text{ m}} = 3.02 \times 10^{-19} \text{ J} = 1.89 \text{ eV}$$

We have to try some combinations of allowed energy levels for the hydrogen atoms to find which difference corresponds to this value. From Table 1.1, the answer is $E_3 - E_2 = (E_{\text{vac}} - 1.51) - (E_{\text{vac}} - 3.4) = 1.89$ eV. Thus the initial state of the electron is state 3, and the final state is state 2.

- b. $\lambda_2 = 486.13 \text{ nm}$ (blue). The energy corresponding to this wavelength is

$$E = \frac{hc}{\lambda} = 2.55 \text{ eV}$$

This energy results only from a transition from E_4 to E_2 , or

$$E_4 - E_2 = (E_{\text{vac}} - 0.85) - (E_{\text{vac}} - 3.4) = 2.55 \text{ eV}$$

- c. $\lambda_3 = 434.05 \text{ nm}$ (violet). The energy corresponding to this wavelength is

$$E = 2.86 \text{ eV}$$

This corresponds to

$$E_5 - E_2 = (E_{\text{vac}} - 0.544) - (E_{\text{vac}} - 3.4) = 2.86 \text{ eV}$$

- d. Which energy level do all of these transitions have in common? This group of transitions is called the Balmer series. The Balmer series describes optical transitions ending on E_2 . There is another point to be made here. The energy of a photon depends on the wavelength (frequency) of the light ($E = h\nu$). A more energetic (more intense) beam of 600 nm would consist of more photons, but each photon still has to have 2.06 eV of energy. The power of a light beam shining on an object depends on both the energy per photon and the number of photons striking the object per second.

For example, suppose that a beam of photons of $\lambda = 600 \text{ nm}$ (2.06 eV) is incident on a semiconductor. Let the semiconductor have a band gap of 2.5 eV as shown in Figure 1.19a. This particular material cannot absorb a photon of 2.06 eV. Why not? Consider an electron near the top of the valence band. To absorb the photon's energy, the electron would have to end up in a new energy state 2.06 eV higher than where it started. There is no allowed energy state there, however, because it would be in the forbidden energy band. Thus, the material is transparent to this particular wavelength of light.⁸

In Figure 1.19b, a new material is chosen whose band gap is smaller than the energy of the photon. Now the photon's energy can be absorbed, in which case the energy of the photon is transferred to the electron and the photon is annihilated.

In optical emission, the process is reversed. An electron initially at a high energy state (i.e., in the conduction band) can “fall” down to the valence band—but it must release the extra energy. This energy could be heat or light, or some combination of the two. Here the law of conservation of energy is used to discuss emission and absorption. Another restriction, conservation of wave vector, analogous to the conservation of momentum in classical mechanics, must also be observed, as discussed in the Supplement to Part 1.

⁸It is, of course, possible for two photons to strike the electron simultaneously and thus give $2 \times 2.06 = 4.12 \text{ eV}$ to the electron. This is more than enough energy to excite an electron from the valence band to the conduction band, but three particles (two photons and one electron) colliding simultaneously is statistically unlikely.

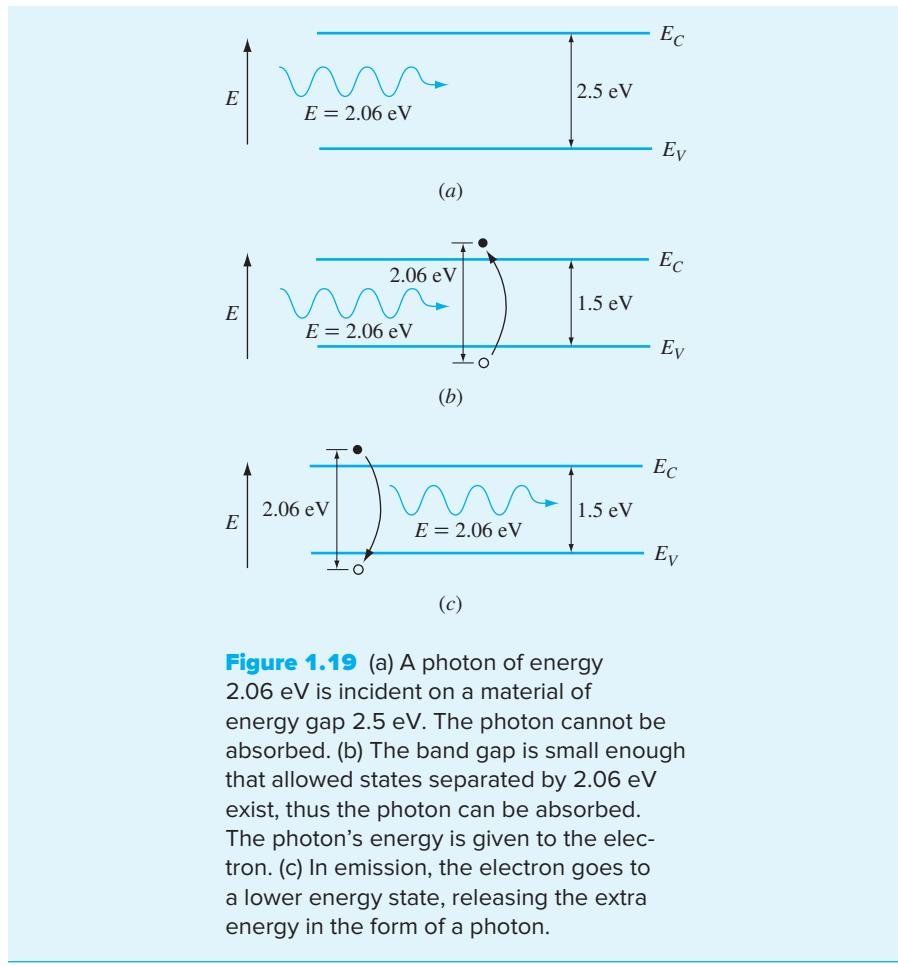


Figure 1.19 (a) A photon of energy 2.06 eV is incident on a material of energy gap 2.5 eV . The photon cannot be absorbed. (b) The band gap is small enough that allowed states separated by 2.06 eV exist, thus the photon can be absorbed. The photon's energy is given to the electron. (c) In emission, the electron goes to a lower energy state, releasing the extra energy in the form of a photon.

EXAMPLE 1.4

Light of $100 \mu\text{W}$ is incident on a photodetector.

- If the light is green (500 nm), how many photons are striking the surface per second?
- If the power remains at $100 \mu\text{W}$, but the wavelength of the light is changed to infrared ($\lambda = 1 \mu\text{m}$), now how many photons strike the surface per second?

■ Solution

- The energy per photon is $E = h\nu$. The frequency ν for the green light is given by

$$\nu = \frac{c}{\lambda} = \frac{2.99 \times 10^8 \text{ m/s}}{500 \times 10^{-9} \text{ m}} = 5.98 \times 10^{14} \text{ s}^{-1} = 598 \text{ THz}$$

The energy per photon is thus

$$\begin{aligned} E &= h\nu = (6.62 \times 10^{-34} \text{ J} \cdot \text{s})(5.98 \times 10^{14} \text{ s}^{-1}) \\ &= 3.95 \times 10^{-19} \text{ J} \cdot \left(\frac{1}{1.6 \times 10^{-19} \text{ J/eV}} \right) = 2.47 \text{ eV} \end{aligned}$$

Power is $P = E/t = 100 \mu\text{W}$, so

$$\begin{aligned} 100 \mu\text{W} &= (\text{energy per second}) \\ &= (\text{energy per photon}) \cdot (\text{number of photons striking surface/second}) \\ &= N \cdot 3.95 \times 10^{-19} \text{ J} \\ N &= \frac{100 \times 10^{-6} \text{ W}}{3.95 \times 10^{-19} \text{ J/photon}} = \frac{100 \times 10^{-6} \text{ J/s}}{3.95 \times 10^{-19} \text{ J/photon}} \\ &= 2.54 \times 10^{14} \text{ photons/s} \end{aligned}$$

b. For the infrared case, the energy per photon is

$$E = h\nu = \frac{hc}{\lambda} = \frac{(6.62 \times 10^{-34} \text{ J} \cdot \text{s})(2.99 \times 10^8 \text{ m/s})}{1 \times 10^{-6} \text{ m}} = 1.98 \times 10^{-19} \text{ J}$$

and the number of photons per second is

$$N = \frac{100 \times 10^{-6} \text{ W}}{1.98 \times 10^{-19} \text{ J/photon}} = \frac{100 \times 10^{-6} \text{ J/s}}{1.98 \times 10^{-19} \text{ J/photon}} = 5.05 \times 10^{14} \text{ photons/s}$$

Thus, the number of photons required to deliver a given amount of power depends on the wavelength of the light.

Because we often convert wavelength to energy in our calculations, it is helpful to multiply the constants together once and remember the result:

$$\begin{aligned} E &= h\nu = \frac{hc}{\lambda} = \frac{(6.62 \times 10^{-34} \text{ J} \cdot \text{s})(2.99 \times 10^8 \text{ m/s})}{\lambda} \cdot \frac{1 \text{ eV}}{1.6 \times 10^{-19} \text{ J}} \\ &= \frac{1.24 \times 10^{-6} \text{ eV}}{\lambda} \end{aligned}$$

or

$$E(\text{eV})\lambda(\mu\text{m}) = 1.24 \quad (1.74)$$

The energy per photon in eV times the photon wavelength in μm is 1.24. Thus, a photon whose wavelength is 500 nm (0.5 μm) has an energy of $1.24/0.5 = 2.48 \text{ eV}$.

EXAMPLE 1.5

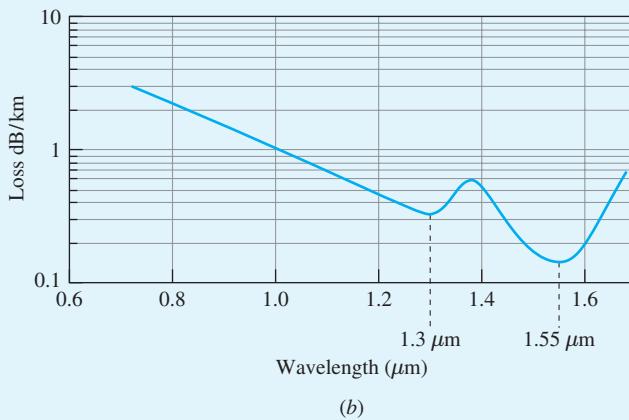
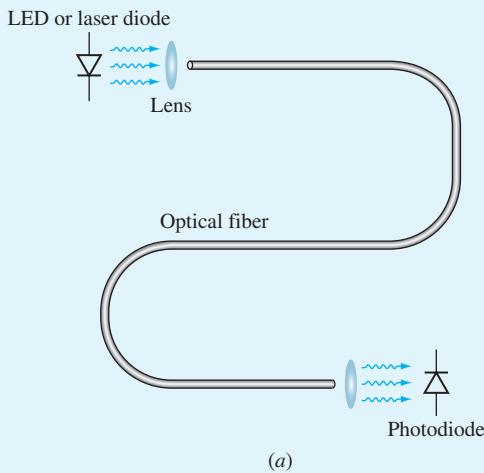


Figure 1.20 (a) A communication fiber optic link contains a light source, a fiber, and a photodetector. (b) Typical absorption spectrum for optical fiber.

An optical communication system is shown schematically in Figure 1.20a. It consists of a semiconductor laser diode optical source, a semiconductor photodetector, and a fiber optic cable to transmit the optical signal between source and detector. As indicated in Figure 1.20b, modern fiber has a minimum absorption coefficient at $\lambda = 1.55 \mu\text{m}$. Thus to obtain a maximum fiber length between repeaters, the laser should emit light at this wavelength.

- Assuming the laser emits photons having energy equal to its band gap, what must be its band gap?
- Since most photodetectors are sensitive to photons having energy greater than their band gaps, what is the maximum band gap of the detector?

Solution

- a. From [Equation (1.74)],

$$E(\text{eV}) = \frac{1.24}{\lambda(\mu\text{m})} = \frac{1.24}{1.55} = 0.80 \text{ eV}$$

The laser's band gap should be 0.8 eV.

- b. Similarly, the maximum band gap of the photodetector should be 0.8 eV.

Note: To obtain a material with a specific band gap, appropriate alloys of three or four materials are used. Choosing the ratios of the various elements in these alloys to produce the desired band gap is referred to as *band-gap engineering*.

1.8 CRYSTAL STRUCTURES, PLANES, AND DIRECTIONS

We have pointed out that silicon, gallium arsenide, and other semiconductors, as well as some metals, are crystals. Crystallography is of great interest to people who fabricate semiconductor devices. For the purposes of appreciating the operating physics of diodes and transistors, some minimal understanding of crystal planes and directions is useful.

Crystals are regular structures in which the atoms are arranged in a pattern that repeats throughout the material. For example, consider the crystal in Figure 1.21a. It is termed *simple cubic*. The atoms are arranged in cubes, with an atom at each vertex. All six faces are square with dimension a , called the *lattice constant*. There are several variations on the cubic structure. One is the face-centered cubic (FCC) lattice, Figure 1.21b. It has, in addition to the corner atoms, an atom in the center of each face of the cube. Common table salt (NaCl) has a face-centered cubic structure—alternate atoms are sodium and the ones in between are chlorine. The body-centered cubic structure, Figure 1.21c, has an atom in the center of each cube.

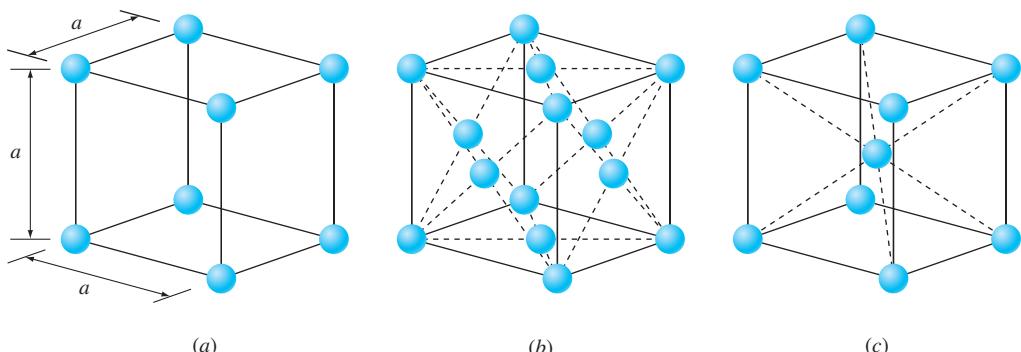


Figure 1.21 Cubic crystals: (a) simple cubic; (b) face-centered cubic, an atom in the center of every face, and (c) body-centered cubic.

Silicon's crystal structure is the diamond structure, which is actually a variation on the cubic structure. The cubic element has an atom in the center of every face, like the FCC lattice, but has in addition four internal atoms as shown in Figure 1.22a. In actuality, the diamond structure consists of two FCC lattices that interpenetrate.

GaAs is also a diamond lattice in the sense that the atoms are arranged in the same way as in Figure 1.22a. Here, however, half of the atoms are gallium and half are arsenic, as shown in Figure 1.22b. This results in the *zinc blende* structure, named after a mineral that exhibits this arrangement. It is interesting to note that the zinc blende structure is actually two interpenetrating FCC lattices; in GaAs, one FCC is of gallium and the other is of arsenic. As we will see later on in the book, the properties of semiconductor devices depend on the crystal plane of the surface of the semiconductor and in which crystallographic direction the electrons (or holes) travel.

There are three surfaces and three crystallographic directions that are the most important in semiconductors. These are shown in Figure 1.23 for a cubic lattice. The (100) plane is the plane of any one of the faces of the cube. The (110) plane includes two edges of the cubes and cuts the cube diagonally. The directions [100], [110], and [111] are perpendicular to the corresponding planes. Thus, any of the $\langle 100 \rangle$ directions⁹ is along a cube edge, any $\langle 110 \rangle$ direction is diagonal across the cube but parallel to one plane of the crystal, and the

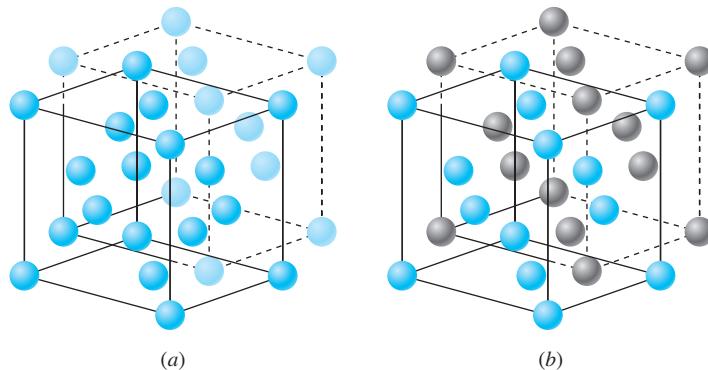


Figure 1.22 (a) The diamond structure consists of two interpenetrating FCC lattices. The second FCC cube is offset by one-quarter of the longest diagonal. The dashed lines indicate the part of the second FCC lattice that is outside the unit diamond cell. (b) A zinc blende material has the same structure, but two types of atoms. The black atoms are one type (for example, gallium) and the colored atoms are the other (arsenic).

⁹A specific plane specification is enclosed in parentheses, while equivalent planes, e.g., (100), (010), (001), (−100), etc., are enclosed in curly brackets, {100}. Similarly, specific directions are enclosed in brackets, [100], [010], etc., and equivalent directions are enclosed in angle brackets, $\langle 100 \rangle$.

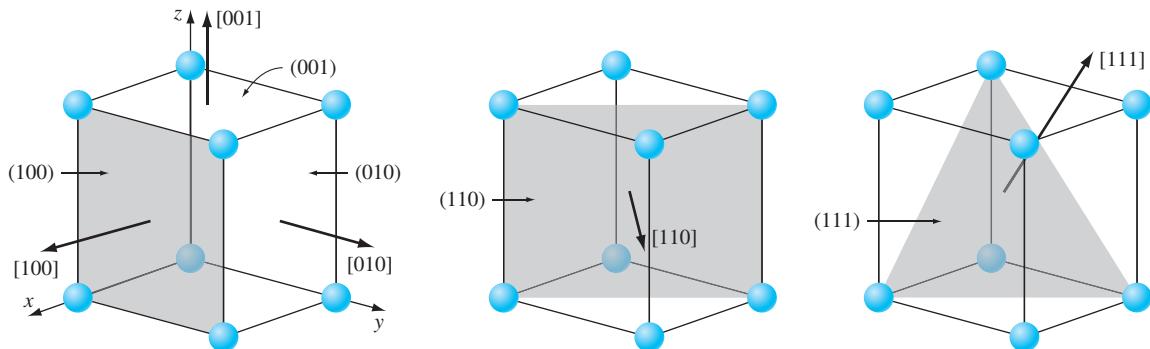


Figure 1.23 The three most important crystallographic planes (in parentheses) and the corresponding crystallographic directions (square brackets).

$\langle 111 \rangle$ direction cuts the cube along a diagonal. These directions and planes are important to those fabricating the devices, and familiarity with the terminology is useful.

1.9 SUMMARY

We have laid some groundwork for understanding the fundamental role of electrons in materials. From the Bohr model, we saw that the allowed energy states of an electron are actually quantized—only certain discrete values are allowed. This is a fundamental concept in the field of quantum mechanics, which will be discussed in more detail in the next chapter and throughout the book.

These states were determined analytically for an isolated hydrogen atom. We then discussed covalent bonding in molecules and in crystalline solids, in which electrons are shared equally between atoms to hold the solid together. In bonding that is partially ionic, the electrons are also shared, but in addition there is a coulombic attraction between nuclei that helps to hold the solid together.

When many atoms are brought close together as in a solid, the discrete allowed states smear into energy bands, in accordance with the Pauli exclusion principle. We saw that for a given material, there are certain ranges of energy that are allowed and other forbidden bands in which no states exist (at least not for the ideal crystal).

We also learned how to tell, from the energy band diagram, which electrons are bound to individual atoms—those whose bands do not extend as far as the next atom. Those electrons that are involved in covalent bonding occupy the valence band that extends across the crystal so that electrons can be shared. The higher energy conduction band also extends across the crystal and contains electrons that are quasi-free. These electrons can travel around the crystal easily, conducting electricity.

The concept of wave-particle duality was briefly discussed. What are classically thought of as particles (e.g., electrons) also possess wave properties

(e.g., wavelength), while what are classically considered waves (e.g., photons) can also be considered to be particles. Matter waves are characterized by a wave function, which can be calculated from Schrödinger's equation if the potential energy of the particle is known. Two examples were considered for an electron within a crystal: the free electron, in which the electron potential energy is considered to be constant, and the quasi-free electron, where the potential is a periodic function with position within the crystal. These two cases will be referred to repeatedly throughout this book.

The interactions of classical particles (electrons) and waves (light) were briefly discussed in terms of the conservation of energy associated with the interaction of the particles involved.

Because the electrical properties of semiconductors depend on their crystal structure, the structures of some semiconductors of interest (silicon and gallium arsenide) were briefly discussed.

Finally, we learned how to determine some electrical properties of material directly from the energy band diagram. A material with a wide band gap will generally not conduct well and can be an insulator. On the other hand, in a metal, the band containing the conduction electrons is partly full even at low temperatures, and these materials conduct very well. Semiconductors have intermediate band gaps, and their conductivity is between that of insulators and metals.

It may seem intuitive that electrical engineers should want materials that either conduct very well, or insulate very well. The usefulness of semiconductors is certainly not obvious from what we've learned so far. We will spend the next few chapters developing an understanding of the special properties that these intermediate band-gap energies confer. The rest of the book will be devoted to showing how to use these properties to make useful semiconductor devices.

1.10 REFERENCES

1. Robert Eisberg and Robert Resnick, *Quantum Physics of Atoms, Molecules, Solids, Nuclei and Particles*, 2nd ed., John Wiley and Sons, New York, 1985.
2. Charles Kittel, *Introduction to Solid State Physics*, 8th ed. John Wiley and Sons, New York, 2004.
3. R. A. Smith, *Wave Mechanics of Crystalline Solids*, Chap. 4, Chapman and Hall, London, 1961.
4. Leon Brillouin, *Wave Propagation in Periodic Structures*, 2nd ed., Dover, New York, 1953.

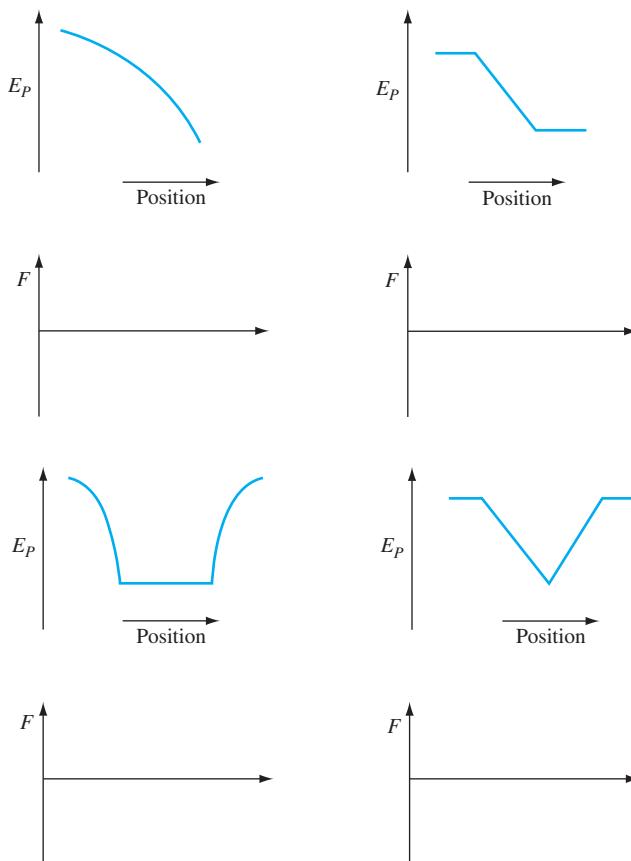
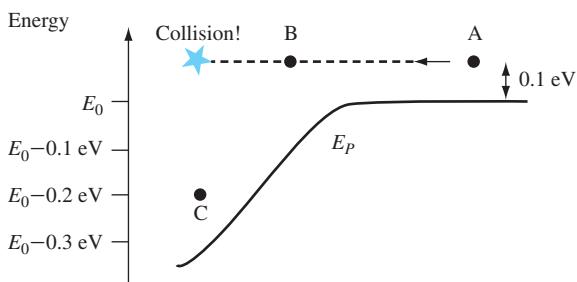
1.11 REVIEW QUESTIONS

1. Define an electron volt.
2. Define a photon.
3. State the Pauli exclusion principle.

4. Explain why the energy levels of an atom become energy bands in a solid.
5. What is the difference between the band structure of a crystalline insulator and that of a semiconductor?
6. What is the difference between the band structure of a semiconductor and that of a metal?
7. Explain why a semiconductor acts as an insulator at 0 K and why its conductivity increases with increasing temperature.
8. Define a hole in a semiconductor.
9. Indicate pictorially how a hole contributes to conduction.
10. Draw the potential energy picture of a metal. Explain qualitatively the existence of the potential barrier at the surface.
11. We saw that it is possible, from a quantum mechanical point of view, for an electron to tunnel through a thin barrier, even though such tunneling is forbidden classically. The thicker the barrier or the higher the barrier, the lower the probability of tunneling. A barrier on the order of tenths of a nanometer is thin enough for an electron to tunnel through. Comment on the probability of an entire person, composed of an astronomical number of electrons, protons, etc., tunneling through a brick wall.
12. Explain why an electron's energy increases when it absorbs a photon. What happens to the photon?

1.12 PROBLEMS

- 1.1 Show that Equation (1.6) follows from Equation (1.3).
- 1.2 Fill in the steps to derive Equations (1.12) and (1.13) from Equations (1.9) and (1.11).
- 1.3 Using the Bohr model, find the first three energy levels for a He^+ ion, which consists of two protons in the nucleus with a single electron orbiting it. What are the radii of the first three orbits?
- 1.4 For each of the potential energy distributions in Figure 1P.1, sketch the force on an electron in this region of space, paying attention to magnitude and sign.
- 1.5 Consider the electron in the energy diagram in Figure 1P.2. It is initially at point A, and moving to the left. At point C, it collides with an atom and loses some energy.
 - a. What is its initial potential energy?
 - b. What is its initial kinetic energy?
 - c. What is its initial acceleration?
 - d. At point B, what is its potential energy? Kinetic energy?
 - e. What are its potential and kinetic energies just before the collision?
 - f. What are its potential and kinetic energies just after the collision?

**Figure 1P.1****Figure 1P.2**

- g. Is the electron still moving after the collision?
- h. Where did the energy lost by the electron go?
- i. What is the direction of force on the electron after the collision?
- j. What is the direction of travel of the electron after the collision?

1.6 Find the kinetic energy of each of the following. Express all your answers in electron volts.

- An electron in the lowest energy level of a hydrogen atom according to the Bohr model
- A grain of pollen weighing 10 ng and drifting at 1 mm/s
- An ant weighing 3 mg and running at 1 cm/s

1.7 For each of the following semiconductors, indicate to what degree you expect covalent or ionic bonding, and why:

Silicon

Gallium nitride

Silicon carbide

Cadmium telluride

1.8 For each of the following semiconductors, sketch (to scale) the energy band diagrams:

- Indium phosphide $E_g = 1.35 \text{ eV}$, $\chi = 4.4 \text{ eV}$
- Germanium $E_g = 0.67 \text{ eV}$, $\chi = 4.0 \text{ eV}$
- Gallium nitride: $E_g = 3.437 \text{ eV}$, $\chi = 4.1 \text{ eV}$
- Amorphous silicon dioxide $E_g = 9 \text{ eV}$, $\chi = 0.9 \text{ eV}$

1.9 What minimum energy must an electron at the bottom of the conduction band in gallium phosphide gain to become free of the crystal? Repeat for an electron at the top of the valence band.

1.10 A nondegenerate semiconductor cannot conduct current at absolute zero (degeneracy will be discussed in Chapter 2). How much energy must at least one electron obtain in germanium before conduction is possible?

1.11 At room temperature in a cubic centimeter of Si there will be about 10 billion electrons in the conduction band.

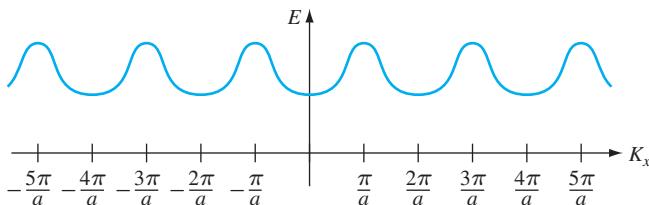
- How many holes are in the valence band?
- If electrons are constantly seeking lower energies and recombining with holes (empty states at lower energies), then how can the number 10 billion remain constant?

1.12 Suppose the electron in Figure 1.12 is traveling to the right at a constant energy. What happens to it when it approaches the surface of the material? Explain your answer, using the energy diagram.

1.13 Show that Equation (1.38) is a solution to Equation (1.37). What is the significance of the positive and negative values of K ?

- 1.14** a. Calculate the de Broglie wavelength of
- A free electron with 1 eV of kinetic energy
 - A free electron with 1 keV of kinetic energy
 - A grain of pollen weighing 10 ng and drifting at 1 cm/s
 - Yourself, walking at 1.5 m/s on your way to class

- b. What is the size of a typical atom? You begin to see why quantum mechanics and the wave description are not useful for large objects.
- 1.15** What is the wavelength of an electron at the bottom of the E - K relationship of Figure 1.13? What is its kinetic energy there?
- 1.16** The laws of classical physics apply to an object in motion if its dimensions are much larger than its de Broglie wavelength. A particle can be considered as a wave (Eq. 1.25), where the wave vector K is $K = 2\pi/\lambda$, and for a free particle (constant E_p) $\lambda = h/p$, where $P = mv$. Find the wavelength of a car of mass 3000 lb traveling at 60 miles per hour and of a free electron traveling at 10^6 cm/s. Note: The dimensions of joules are $\text{kg}\cdot\text{m}^2/\text{s}^2$.
- 1.17** Find the energy of a photon having wavelengths of (a) 1 cm, (b) 1 μm and (c) 1 nm.
- 1.18** An AM radio station radiates 100 KW of power at 1 MHz.
- What is the energy of each photon?
 - How many photons are radiated per second?
- 1.19** An electron is moving at a constant velocity (i.e., free electron) of 10^5 cm/s. Determine:
- Its momentum.
 - Its wave vector.
 - Its de Broglie wavelength.
 - Its kinetic energy.
- 1.20** Consider the E - K diagram shown in Figure 1P.3.
- Verify that it meets the required criteria:
 - $E(K)$ is periodic in K space with period $2\pi/a$.
 - Equivalent extrema exist at $K = 0, \pm 2\pi/a, \pm 4\pi/a, \dots$
 - Equivalent extrema exist at $K = \pm\pi/a, \pm 3\pi/a, \pm 5\pi/a, \dots$
 - The slope of the E - K curve is zero at $K = 0, \pm\pi/a, \pm 2\pi/a, \dots$
 - Indicate the first Brillouin zone.
 - Sketch the corresponding v_g - K diagram.
 - In what regions of the E - K diagram are electrons most likely to be found for this material?

**Figure 1P.3**

- 1.21** Explain the analogy between using a conductor thicker than the skin depth to shield a region of space from electromagnetic waves and the ability of an electron to penetrate a potential barrier.
- 1.22** The infinitely thick potential barrier of Figure 1.18a can be considered a crude approximation to the potential barrier at the surface of a semiconductor (see Figure 1.12).
- How, then, might you construct a thin potential barrier like that in Figure 1.18b? Thin potential barriers are used in a wide variety of semiconductor devices, including tunnel diodes, contacts, and field-effect transistors.
 - How would you construct a potential well (a thin region of lower potential energy bounded by a region of higher potential energy)? Potential wells are widely used in lasers, photodetectors, and heterojunction bipolar transistors.
- 1.23** a. From the Bohr model, what emission wavelength would you expect for a transition in hydrogen from E_3 to E_1 ? Transitions ending at E_1 are collectively called the Lyman series and are generally found in the ultraviolet region of the spectrum.
- b. What emission wavelength would you expect from a transition from E_3 to E_2 ? This is the first emission line in the Balmer series, and it is visible.
- 1.24** a. What wavelength of light should you shine on hydrogen to cause electrons to go from E_2 to E_4 by optical absorption?
- b. What would happen if you passed a beam of $\lambda = 520$ nm (bluish green) through a hydrogen gas? Explain your answer.
- 1.25** In discussing Figure 1.19a, we pointed out that, in a material with a band gap of 2.5 eV, an electron near the top of the valence band could not absorb a photon of energy 2.06 eV, since it would have to end up at a forbidden energy state.
- What about an electron deep in the valence band, more than 2.06 eV below the band edge E_V ? Why is it unlikely for this electron to absorb the photon?
 - Why is unlikely for a photon of 2.06 eV to be absorbed by an electron in the conduction band?
- 1.26** For a simple cubic crystalline structure of lattice constant $a = 0.40$ nm,
- How many atoms are there per unit volume? (*Hint:* An easy way to proceed is to calculate the volume of the unit cell and the number of atoms per unit cell. Although there are eight atoms involved in any given unit cell, each atom in the simple cubic structure is part of eight different cells, one corner of each. Thus, there are 8 atoms times 1/8 atom per corner or 1 atom per unit cell.)
 - How many atoms per unit area are there in the (100) plane? The (110) plane? The (111) plane?
 - What if the lattice is FCC instead (still with $a = 0.40$ nm)? Now how many atoms per unit volume are there?

2

CHAPTER

Homogeneous Semiconductors

2.1 INTRODUCTION AND PREVIEW

In this chapter, we extend our understanding of electron motion in solids to investigate some of the electronic properties of homogeneous semiconductors. Homogeneous semiconductors are those that consist of one uniform material; for example, pure (intrinsic) silicon, or silicon with impurities uniformly distributed. In later chapters, devices that combine different materials and nonuniformly distributed impurities are discussed.

Semiconductors conduct current in a unique manner. Instead of the current being carried by electrons as it is in metals, two types of particles or charge carriers can carry current in semiconductors: electrons and holes. The hole is an artificial concept, as indicated in Chapter 1, but it is very convenient for describing the electronic properties of semiconductors.

To discuss the behavior of electrons and holes, it is necessary to use quantum mechanics, since the motion of electrons in a semiconductor crystal cannot be accurately described by using conventional classical mechanics. Quantum mechanics is discussed somewhat further in the Supplement to Part 1.

As discussed in Chapter 1, an electron can be described by its wave function. The wave function is found by solving Schrödinger's equation [Equation (1.35)], for which one has to know the potential energy function E_P of the electron as a function of position. Unfortunately, this is not completely nor accurately known for an electron in a solid. Two approximations were discussed in Chapter 1: the free electron approximation, in which E_P was assumed constant with position, and the quasi-free electron approximation, in which E_P is a periodic function of position.

Fortunately, for the case of semiconductors it is possible to use "pseudo-classical mechanics" to describe the motion of most electrons of interest, which is to say those that carry electrical current. Pseudo-classical mechanics is a way of applying Newton's laws and other familiar equations, using "effective"

quantities that give the correct results. For example, we can use an effective mass for electrons, which, when inserted into classical equations such as $F = ma$, will predict the actual motion of the electron. Newton's law then becomes

$$F = m^*a \quad (2.1)$$

where m^* is referred to as the *effective mass* of the electron and takes into account the interaction of the electron with the forces resulting from the periodic nature of the lattice. The force F in Equation (2.1) is then the resulting *external* force acting on the electron. In this way, we can treat the electron as if it were a normal free particle, even though it is in reality nothing close to that. This approach works only up to a point, but is very convenient where applicable.

The electrons of interest in semiconductors usually are near the top or bottom of a band. Those near the bottom of a band behave as the “quasi-free” electrons discussed briefly in Chapter 1. It turns out that holes, found near the top of the valence band, can also be treated as particles.

The relationship between energy and wave vector K for electrons in crystals is discussed in this chapter. The wave vector arose when we chose to consider electrons as waves, but the energy–wave vector relationship allows us to determine some of the parameters needed to treat the electrons as particles.

The effects of the impurity and temperature dependencies of the quasi-free electron (and hole) concentrations and their distributions with energy are also discussed in this chapter.

2.2 PSEUDO-CLASSICAL MECHANICS FOR ELECTRONS IN CRYSTALS

We have seen that for electrons in crystals, we can't use classical mechanics, with familiar quantities such as mass, velocity, kinetic energy, and potential energy, and familiar equations such as $F = ma$. It would be convenient, however, to be able to use classical mechanics, as opposed to quantum mechanics, because of the simpler equations and because of the physical insight that classical mechanics provides. We show in this section that it is possible to define a pseudo-classical mechanics in which we use the classical equations, but the true electron mass is replaced by an effective mass m^* , which incorporates the electron interaction with the periodic potential of the crystal. This effective mass, when used in the classical equations, correctly predicts the behavior of the electron in a crystal provided the force F on the electron is taken as the external, or applied, force. We start with the easiest case, an electron in a one-dimensional crystal.

2.2.1 ONE-DIMENSIONAL CRYSTALS

The Free Electron In Chapter 1, the energy-wave vector relation for an electron in a one-dimensional crystal was discussed for two cases. In the free-electron approximation, the potential energy was assumed constant with the value E_0 . Then the E - K relation has the form

$$E = E_0 + \frac{\hbar^2 K^2}{2m_0} \quad (2.2)$$

where the second term is the kinetic energy

$$E_K = \frac{\hbar^2 K^2}{2m_0} \quad (2.3)$$

For this free-electron case, de Broglie's relation is valid and $m_0 v = p = h/\lambda$, where v is the electron velocity. The kinetic energy has the classical value

$$E_K = \frac{m_0 v^2}{2} = \frac{p^2}{2m_0}$$

For this case Newton's law, $F = m_0(dv/dt)$, is valid. Note that for the free electron case, there are no complicating additional forces from the crystal, and thus we can use the actual classical equations and the actual free electron mass.

The velocity of the electron in the crystal is given by the group velocity [Equation (1.43)]

$$v = v_g = \frac{1}{\hbar} \frac{dE}{dK} \quad (2.4)$$

and is proportional to the slope of the E - K diagram. From $F = m_0(dv/dt)$ and Equation (2.4), the electron mass m_0 can be expressed as

$$m_0 = \hbar^2 \left(\frac{d^2 E}{dK^2} \right)^{-1} \quad (2.5)$$

and since the E - K curve is parabolic, $d^2 E/dK^2$ is a constant and thus m_0 is constant in energy.

The Quasi-Free Electron Recall that the quasi-free electron model takes the electron to be in a region in a crystal whose potential energy is modeled as a periodic function extending to infinity in both directions (Figure 1.14). To obtain a physical picture of the behavior of quasi-free electrons, it is convenient to express their properties in forms analogous to those of a free electron, Equations (2.2) to (2.5).

To find the effective mass, we will examine the E - K diagram of a material. Consider first the case of an electron near the bottom of the conduction band as shown in Figure 1.16 and repeated in Figure 2.1a. We have chosen a material for which the bottom of the conduction band is at $K = 0$; GaAs is one such material. We will expand the E - K relation in a power series. Since we are considering an electron near the bottom of the band where $E = E_C$, we expand the energy about $K = 0$:

$$E = E_C + \left[\left(\frac{dE}{dK} \right) \Big|_{K=0} \right] K + \left[\left(\frac{1}{2} \frac{d^2 E}{dK^2} \right) \Big|_{K=0} \right] K^2 + \text{HOTs} \quad (2.6)$$

where HOTs refers to higher-order terms. The derivatives are evaluated at $K = 0$ where $dE/dK = 0$. Neglecting the HOTs gives

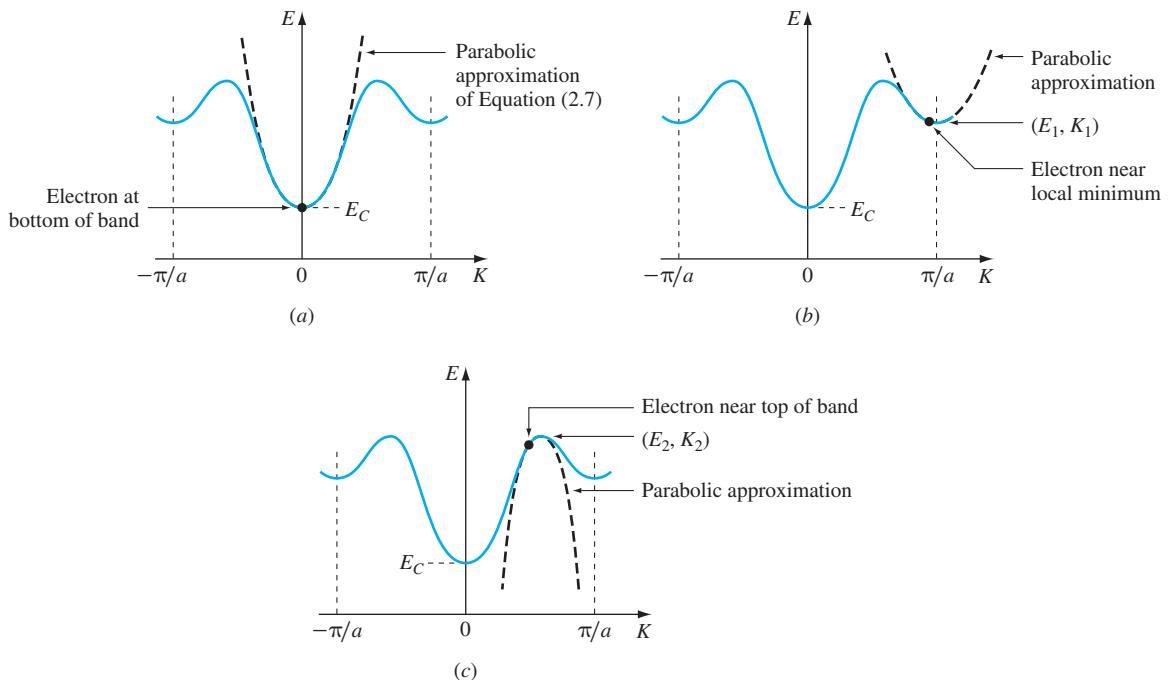


Figure 2.1 The E - K diagram for (a) an electron at the bottom of the conduction band at $K = 0$, where the velocity and kinetic energy are both zero and thus the total energy is equal to the potential energy; (b) an electron in a local minimum, where it has a different effective mass, and (c) an electron near the top of the band.

$$E = E_C + \left(\frac{1}{2} \frac{d^2 E}{dK^2} \right) K^2 \quad (2.7)$$

This is the expression for a parabola. We can justify neglecting the higher order terms provided we stay in a region where the E - K relation is parabolic, in this case near the bottom of the band.

For a free electron, the group velocity was given by

$$v = \frac{1}{\hbar} \frac{dE}{dK}$$

[Equation (2.4)]. Here, too, the group velocity has the same form. That is, the velocity of the electron in the crystal is the group velocity, given by

$$v = v_g = \frac{1}{\hbar} \frac{dE}{dK} \quad (2.8)$$

or the group velocity is proportional to the slope of the E - K diagram. From Figure 2.1a, an electron at the very bottom of the band ($K = 0$, $dE/dK = 0$) has zero velocity. Since the kinetic energy is that energy associated with motion, at $K = 0$ we can deduce that the kinetic energy $E_K = 0$. The total energy of an electron

is the sum of the kinetic energy and potential energy, so at the bottom of the conduction band, the total energy is the same as the potential energy E_P . In other words

$$E_P = E_C \quad (2.9)$$

Thus, the potential energy for an electron near the bottom of a band is the energy at the bottom of the band.

Compare Equation (2.7) for the electron in a one-dimensional crystal with Equation (2.2) for a free electron (both repeated here):

$$E = E_0 + \frac{\hbar^2}{2m_0} K^2 \quad \text{free electron} \quad (2.2)$$

$$E = E_C + \left(\frac{1}{2} \frac{d^2 E}{dK^2} \right) K^2 \quad \text{quasi-free electron} \quad (2.7)$$

For the free electron, the quantity E_0 is the potential energy and the second term in Equation (2.2) is its kinetic energy. For the electron near the bottom of the conduction band in a crystal [Equation (2.7)], we have substituted E_C for the potential energy ($E_P = E_C$), and thus the quantity $\frac{1}{2}(d^2E/dK^2)K^2$ is the kinetic energy, or the energy associated with the motion of the quasi-free electron. Therefore

$$E_K = \frac{1}{2} \frac{d^2 E}{dK^2} K^2 \quad (2.10)$$

Since the term $\frac{1}{2}(d^2E/dK^2)$ of Equation (2.10) is analogous to $\hbar^2/2m_0$ of Equation (2.2), we can define an effective mass to be

$$m^* = \hbar^2 \left[\frac{d^2 E}{dK^2} \right]^{-1} \quad (2.11)$$

analogous to the case of a free electron [Equation (2.5)],

$$m_0 = \hbar^2 \left[\frac{d^2 E}{dK^2} \right]^{-1} \quad (2.5)$$

The effective mass is inversely proportional to the curvature of the E - K diagram.

Recall that by neglecting the higher-order terms, we effectively restrict ourselves to regions where the E - K curve approximates a parabola, in this case near the bottom of the conduction band. In such a region, the effective mass is constant. Notice, however, that the E - K curve also approximates a parabola in the vicinity of $K = \pm\pi/a$, and the concept of constant effective mass can be used here also. Near this minimum (E_1, K_1), we can expand the energy about $K = K_1$ in a manner analogous to the above procedure. An electron in the parabolic region near this minimum has a different potential energy (where $v_g = 0$ and thus $E_K = 0$) at $E_P = E_1$, and

$$E_K = \frac{1}{2} \frac{d^2 E}{dK^2} \Big|_{K=K_1} (K - K_1)^2$$

and

$$m^* = \hbar^2 \left[\frac{d^2 E}{dK^2} \Big|_{K=K_1} \right]^{-1}$$

Since the curvature can be different here than at $K = 0$, however, in general so is the value of m^* .

Recall that for a classical particle of mass M , the kinetic energy is $E_K = Mv^2/2 = P^2/2M$, where $Mv = P$, the particle momentum. Similarly, for an electron near the minimum at $K = 0$ in a periodic crystal, from Equations (2.10) and (2.11),

$$E_K = \frac{m^* v^2}{2} = \frac{\hbar^2 K^2}{2m^*} \quad (2.12)$$

where v is the group velocity, and thus the quantity $\hbar K$ is often referred to as the *crystal momentum*, as pointed out in Chapter 1.

$$\text{Crystal momentum} = \hbar K$$

The kinetic energy near a minimum at K_1 becomes

$$E_K = \frac{\hbar^2 (K - K_1)^2}{2m^*} \quad (2.13)$$

where m^* is the electron effective mass at this minimum. Note that in this case the electron momentum is not the same as the crystal momentum, or $m^*v \neq \hbar K$. In an electron transition it is the crystal momentum $\hbar K$, rather than the particle momentum, that is conserved.

Now consider an electron near E_2 in Figure 2.1c. Note that the E - K diagram looks parabolic here also, although the curvature is negative. The slope of the E - K curve is zero at the top of the band, so from Equation (2.8) the electron velocity at E_2 is zero, and the kinetic energy is also zero. The potential energy of the electron near the local maximum at the top of the band is therefore E_2 . In that case, the total energy of the electron in Figure 2.1c is less than its potential energy. This implies that the kinetic energy is negative. In the parabolic region near a maximum, the curvature of the E - K curve is negative, or from Equation (2.11), m^* is actually negative. This may seem nonphysical, and we will return to address this later. For now we note that a negative effective mass producing a negative kinetic energy is consistent with the expression for E_K in Equation (2.12) or (2.13).

Between the regions near the extrema, the effective mass as defined by Equation (2.11) is a function of energy since the HOTs of Equation (2.6) cannot be neglected. Since the development here is based on neglecting the HOTs, the results are valid only in the parabolic regions where m^* is constant.

Next, we consider the force on an electron. Recall from classical mechanics

$$F = -\frac{dE_P}{dx} \quad (2.14)$$

In an analogous manner, in pseudo-classical mechanics we have E_P equal to E_C , for an electron near the bottom of the band, and

$$F = -\frac{dE_C}{dx} \quad (2.15)$$

An example of a case in which a force is applied is shown in Figure 2.2a. Here a voltage is applied across a semiconductor sample, generating an electric field \mathcal{E} . From classical mechanics, we know that $F = -q\mathcal{E}$. From Equation (2.15), we can see that if a force is applied, there must be a slope or gradient in the conduction band edge.¹ This idea is illustrated in Figure 2.2b. In this case, the electric field is in the negative x direction, and the force on the electrons accelerates them to the right ($F = -dE_C/dx$). Remember that by conservation of energy, the total energy (between collisions) is constant and the electrons travel horizontally on the energy band diagram. From the diagram, we observe that with increasing x , the potential energy (E_C) decreases and the kinetic energy (the difference between the total energy and the potential energy) increases. This is consistent with our expectation that as the electron is accelerated by the electric field, its kinetic energy increases.

As it travels, the electron will collide with atoms, defects, or impurities. Energy can be transferred to the other particle during the collision, in which case the electron loses some of its total energy, as shown in Figure 2.2b. The electron continues to be accelerated by the force of the applied electric field, however. The acceleration is

$$a = \frac{dv}{dt} \quad (2.16)$$

Since $F = m^*a$, near a minimum, where m^* is positive, the electron is accelerated in the direction of the applied force. Near a maximum, however, the effective mass is negative. This means that instead of the electron being accelerated, it is *decelerated* in the direction of the applied force. This is illustrated in Figure 2.2c. The potential energy is E_V for an electron near the top of the valence band, and thus the force on the electron is $F = -dE_V/dx$. If there are vacant states (holes) near the top of the valence band, an electron below E_V will move in the positive x direction (because its effective mass is negative here) through those empty states until it reaches E_V (zero velocity) or makes a collision. At that time the electron loses energy and again is forced to move in the positive x direction. Note that this electron movement is equivalent to holes moving in the negative x direction. This behavior results because the total force on the electron is the combination of the externally applied force plus the force exerted by all the atoms in the crystal. The effective mass accounts for the internal atomic forces

¹In the discussion of the band structure in a crystal, E_P was assumed to be a periodic function in space as indicated in Figure 1.14. If, however, an electric field exists in the semiconductor as shown in Figure 2.2, the periodicity of E_P is offset by the electric field, and thus the energy bands are somewhat altered—the band edges (E_C and E_V) are not well defined. Still, except for very high electric fields, to good approximation E_C and E_V can be considered to be well defined and tilted in space by the electric field as drawn in Figure 2.2b.

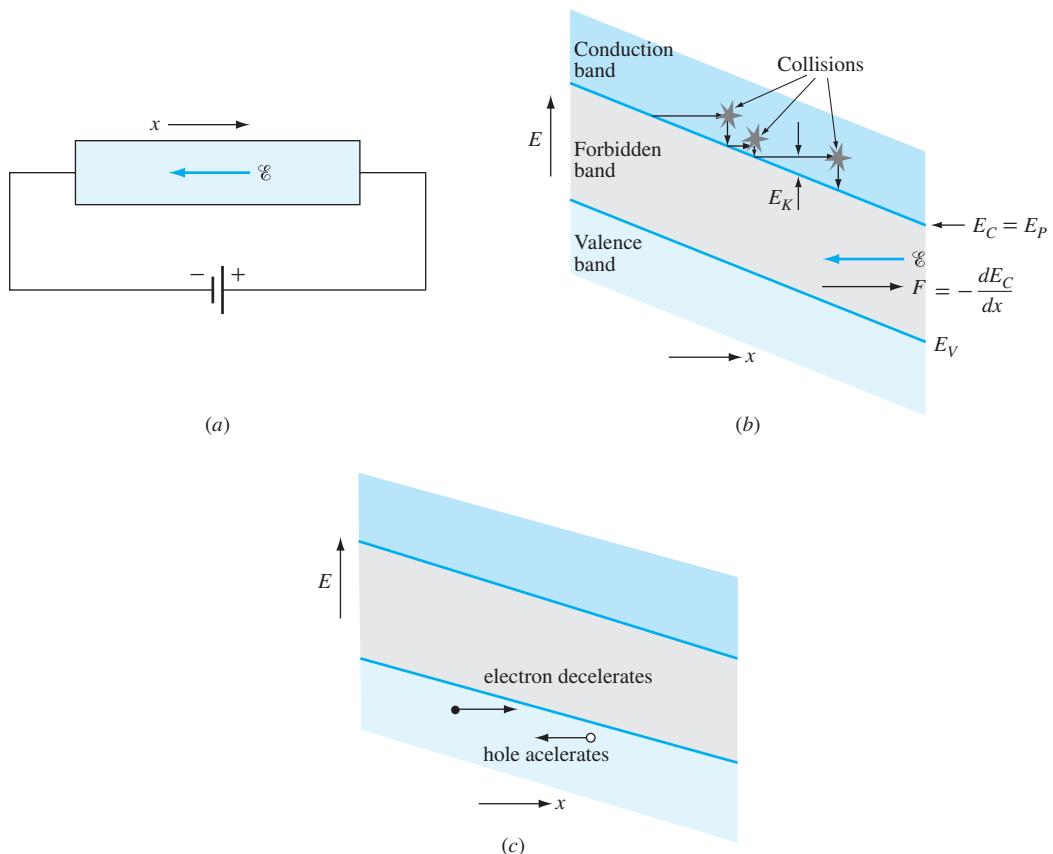


Figure 2.2 An external electric field is applied across a bar of semiconductor. (a) The physical picture; (b) the energy band diagram. Electrons in the conduction band are accelerated to the right; they travel at constant energy between collisions. In (c) electrons in the valence band are decelerated as they move to the right. This is equivalent to holes being accelerated to the left.

so that pseudo-classical mechanics can be used. In this case, the effects of the internal forces combine to accelerate the electron the “wrong” way.

2.2.2 THREE-DIMENSIONAL CRYSTALS

In pseudo-classical mechanics, we use the E - K diagram to find the effective mass, from which we predict the behavior of an electron, under an applied field (for example). Recall from Chapter 1 that we developed our understanding of the general shape and properties of the E - K diagram from a discussion of the Bloch wave function (i.e., it is periodic in K , the slope is zero at the center of the Brillouin zone, etc.). The discussion of the E - K diagram and the effective parameters used

in pseudo-classical mechanics in one dimension can be extended to two and three dimensions. We state here the results for three-dimensional crystals:

1. The Bloch wave in three dimensions is given by

$$\Psi(\vec{r}, t) = U_K(\vec{r}) e^{j[\vec{K} \cdot \vec{r} - (E/\hbar)t]} \\ \psi(\vec{r}) = U_K(\vec{r}) e^{j[\vec{K} \cdot \vec{r}]} \quad (2.17)$$

Where \vec{r} is a position vector and $U_K(\vec{r})$ is periodic in \vec{r} with the periodicity of the crystal.

2. The E - \vec{K} relation in any band is periodic in K space with the periods $2\pi/a, 2\pi/b$, and $2\pi/c$, where a, b , and c are the periodicities in the x, y , and z directions.
3. All of the information in the E - \vec{K} curve, and parameters derived from it, is contained in the reduced zone or the first Brillouin zone.
4. The (group) velocity of an electron in three dimensions is

$$\vec{v} = \frac{1}{\hbar} \left(\hat{i} \frac{\partial E}{\partial K_x} + \hat{j} \frac{\partial E}{\partial K_y} + \hat{k} \frac{\partial E}{\partial K_z} \right) \quad (2.18)$$

where \hat{i}, \hat{j} , and \hat{k} are the unit vectors in the x, y , and z directions respectively.

5. There are relative extrema in K space at $K = 0$ and at the edge of the reduced zone in every principal crystallographic direction. Stated differently, at the edge of the reduced zone (in three dimensions, the edge is a surface), the slope of the E - K curve is zero in every direction in K space. There is one exception in cases where two bands coincide at the boundary of the reduced zone. In that case, for cubic crystals the slopes of the two E - K curves do not each have to be zero, but their sum must be zero.
6. The E - \vec{K} relation is difficult to plot for a three-dimensional crystal; in practice, E is plotted as a function of K along the principal crystallographic directions. Because of symmetry about $K = 0$, it is necessary to plot E for only half of the Brillouin zone.
7. Near a relative extremum of a band, the effective mass m^* can be defined such that Newton's laws can be used. This effective mass is positive near a minimum and negative near a maximum.
8. Near an extremum, the curvature of the E - K plot may depend on the direction of K , and thus the direction in which the electron is traveling. In such a case, the effective mass is direction dependent:

$$m_x^* = \hbar^2 \left(\frac{\partial^2 E}{\partial K_x^2} \right)^{-1} \\ m_y^* = \hbar^2 \left(\frac{\partial^2 E}{\partial K_y^2} \right)^{-1} \\ m_z^* = \hbar^2 \left(\frac{\partial^2 E}{\partial K_z^2} \right)^{-1} \quad (2.19)$$

9. For a semiconductor having a cubic unit cell structure, for the extremum at $K = 0$, the curvature is direction independent and

$$m^* = m_x^* = m_y^* = m_z^* \quad (K = 0) \quad (2.20)$$

or the effective mass is a scalar.

2.3 CONDUCTION BAND STRUCTURE

We are now ready to examine the individual conduction band structures for some common semiconductor materials. The E - K relations for the conduction bands of GaAs, Si, and Ge are plotted in Figure 2.3, along the direction of K with the lowest minimum for that material.

The absolute minimum in the conduction band for GaAs occurs at $K = 0$. Because of the symmetry of its cubic structure, the curvature of the E - K plot near that minimum, and thus m^* , is the same for any direction of motion. For GaAs, then, the effective mass for an electron in the conduction band is a scalar [Equation (2.20)]. Its value has been measured to be $0.067m_0$, where m_0 is the rest mass of a free electron. GaAs is an example of a material with two bands having the same energy at the edge of the zone in the $\langle 100 \rangle$ directions²; thus, the E - K slope there is not zero, but rather the sum of the slopes is zero.

Silicon (Figure 2.3b), on the other hand, has conduction band minima in the $\langle 100 \rangle$ directions at a value of K of about 0.85 of the K value at the zone edge. This is the bottom of the conduction band, where most of the electrons of interest are found. Unfortunately, the effective mass is not a scalar here, because the curvature is different in different directions (in K space) at this minimum. For an electron traveling in a $[100]$ direction in silicon, there is one effective

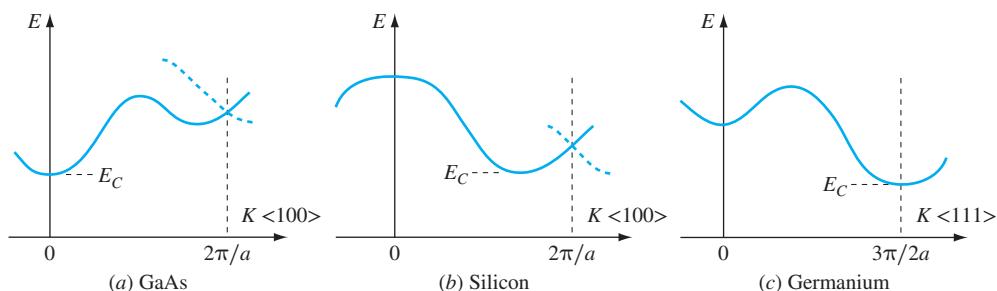


Figure 2.3 The E - K diagrams for three common semiconductors. The crystallographic direction is shown on the K axis. The slope of the E - K curve must be zero at the Brillouin zone edge, unless multiple bands coincide there. The bottom of the conduction band is designated E_C .

²The edge of the zone in diamond and zinc blende structures is at $K = 2\pi/a$. In the $\langle 100 \rangle$ directions the spacing between atomic planes is $a/2$. In the $\langle 111 \rangle$ directions the spacing is $2a/3$ where a is the lattice constant.

mass, called the longitudinal effective mass m_{\parallel}^* , and the two transverse effective masses, used for travel in the other two directions, are equal and denoted m_{\perp}^* .

In Figure 2.3c, we see that germanium has absolute minima in the conduction band at the zone edges in the $\langle 111 \rangle$ directions. The effective mass is not a scalar here, either. Another minimum at $K = 0$ exists at a higher energy. Electrons in this higher-energy minimum have scalar effective mass.

When the effective mass near the bottom of the conduction band is not a scalar, as in Si and Ge, its value depends on the direction in which the electron is traveling. Therefore, some sort of weighted average of m_{\parallel} , and m_{\perp} is required to obtain a value of m^* . The particular averaging method and the resulting values of m^* depend on the particular type of problem being solved. The two most common averaging techniques result in the *conductivity effective mass* m_{ce}^* for electrons, used for conductivity and related calculations, and the *density-of-states effective mass* m_{dse}^* for electron concentrations. Both of these averaging techniques are discussed in Online Module OM5. Table 2.1 lists the values for the effective mass components and average values for some common semiconductors. [1]

2.4 VALENCE BAND STRUCTURE

Although the semiconductor band structure varies from material to material for the conduction band, the valence bands are qualitatively similar for most semiconductors important in electronics. They also tend to be simpler than the conduction bands. The valence bands consist of three overlapping bands. These bands all have absolute maxima at $K = 0$ but with different curvatures, resulting in different effective masses, as shown in Figure 2.4.

Recall that the effective mass of an electron, being inversely proportional to the curvature of the E - K diagram, is negative near the top of the valence band. Furthermore, the valence band is normally almost entirely full of electrons, with just a few empty states near the top of the band. If we consider these empty states to be the holes, then we can consider the holes to have a *positive* charge along with a *positive* effective mass. Later in the chapter, it is shown that this move is legitimate.

Two of the bands have the same energy maxima at $K = 0$. The two bands h and l refer to *heavy holes* (smaller curvature) and *light holes* (higher curvature)

Table 2.1 Effective masses for electrons near the bottom of the conduction band in units of m_0 , the rest mass of the free electron at 300 K

Material	$m_{(K=0)}^*$	m_{\parallel}^*	m_{\perp}^*	m_{ce}^*	m_{dse}^*
Si	—	0.92	0.197	0.26	1.09
GaAs	0.067	—	—	0.067	0.067
Ge	—	1.64	0.082	0.12	0.56
InP	0.077	—	—	0.077	0.077

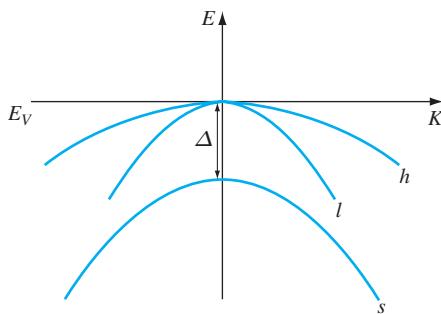


Figure 2.4 The E - K diagram for the valence band in most important semiconductors, indicating (h) the heavy hole band, (l) the light hole band, and (s) the split-off band. The top of the valence band is designated E_V .

respectively. A third band is split off by an energy Δ , due to spin-orbit coupling. This spin-orbit interaction results from magnetic forces that are influenced by the individual spins of the electrons and the magnetic fields resulting from their orbits. The split-off band is designated s .

The values of effective mass for holes associated with each of the bands, along with the split-off energy Δ , are given in Table 2.2. Also given are values for the conductivity effective mass and density-of-states effective mass for holes, m_{ch}^* and m_{dsh}^* , as discussed in Online Module OM5. Again, these result from two different ways of averaging, except that for electrons in the conduction bands, we were averaging the longitudinal and transverse effective masses. For holes in the valence band we combine the light hole and heavy hole. Normally, the split-off band is enough removed in energy from the other bands that it contains few holes and can be ignored. As for electrons, conductivity effective masses are used for calculations involving electrical conduction, while the density-of-states effective mass is used in calculations related to carrier concentrations.

Table 2.2 Effective masses for holes in the valence bands of several semiconductors (masses are given as multiples of the electron rest mass m_0)

Material	m_{lh}^*	m_{hh}^*	m_{sh}^*	Δ , eV	m_{ch}^*	m_{dsh}^*
Si	0.16	0.48	0.24	0.044	0.36	1.15
GaAs	0.082	0.45	0.15	0.34	0.34	0.48
Ge	0.044	0.28	0.08	0.29	0.21	0.292
InP	0.08	0.4	0.15	0.11	0.3	0.42

2.5 INTRINSIC SEMICONDUCTORS

A semiconductor is said to be intrinsic if it contains no impurities and no crystalline defects. As we will see, this implies that at equilibrium the concentration (number per unit volume) of electrons in the conduction band is equal to the concentration of holes in the valence band.

In an intrinsic semiconductor at absolute zero, electrons occupy all of the electronic energy states in the valence band, and all the states in the conduction band are empty. This is to be expected since, at absolute zero, every electron will be found at the lowest possible energy state.

At temperatures above absolute zero, the crystal, like all materials, emits blackbody radiation in the form of photons. The energy of the photons can be absorbed by an electron, giving it enough energy to move up to the conduction band. At higher temperatures, the electrons can also acquire some thermal energy, which is transferred to them from the crystal lattice. The atoms in the crystal lattice vibrate, and these lattice vibrations can be transmitted through the crystal as waves. These acoustic waves are called *phonons*, and associated with each phonon is an energy and a wave vector. Like electrons and photons, phonons can be treated as waves or as particles, as discussed in the Supplement to Part 1.

Electrons can be excited from the valence band to the conduction band by the absorption of a blackbody radiation photon, or more than one photon, or a combination of photon and phonon (see Prob. 2.7). This leaves an empty state or hole in the valence band. The electron is now quasi-free, and there is a quasi-free hole in the valence band. The electron and hole together are referred to as an *electron-hole pair*. When the electron-hole pair is created by absorption of phonons, we call the process *thermal generation*, since the phonons or lattice vibrations carry the thermal energy of the crystal. If a photon provides the energy, the process is termed *optical generation*.

Figure 2.5 indicates the generation process in physical space using a *bond diagram* (a) and on the *energy band diagram* (b). In the bond diagram (here represented in two dimensions), each silicon atom shares its four outer electrons (the valence electrons) with four neighboring silicon atoms. Thus, the electrons make up the covalent bonds between the atoms, and each atom has a complete outer shell of eight electrons.

If one electron in the valence band should acquire some extra energy (e.g., by absorbing photons) it may break out of the bond and go to a higher energy state in the conduction band. Here, the electron no longer contributes to the covalent bonding. This electron now can move about the crystal. If the electron, now in the conduction band, should move, then that moving charge represents current.

Similarly, the vacant state in the valence band can be considered a quasi-free hole, which is free to move throughout the crystal. If an electron from a nearby bond shifts into the empty state, then a new empty state appears in that nearby bond. The hole, in other words, has moved.

For every electron that is excited to the conduction band there is necessarily a hole left behind in the valence band. Therefore, in an intrinsic semiconductor

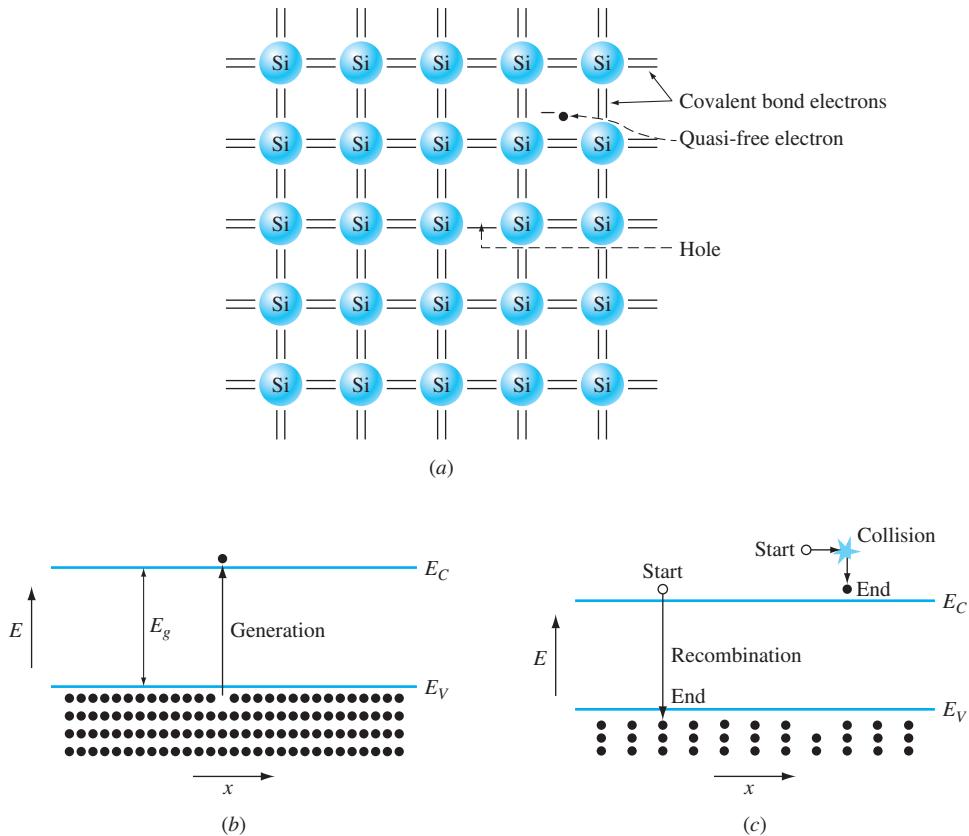


Figure 2.5 In generation, a valence electron acquires some extra energy and moves into the conduction band. (a) Physical picture or bond diagram; (b) energy band diagram. In recombination (c), an electron from the conduction band falls into a hole in the valence band, and both the conduction-band electron and the hole disappear. It is also possible for an electron to lose energy by colliding with something (c, right) in which case it may remain in the conduction band. For visual clarity, the excited electron and hole are indicated as being close to each other. In reality they are at least several lattice constants apart.

the equilibrium concentration n_0 of electrons in the conduction band is the same as the equilibrium concentration of holes p_0 in the valence band. That is, for intrinsic material

$$n_0 = p_0 = n_i \quad \text{intrinsic} \quad (2.21)$$

where the subscript 0 indicates the concentration at equilibrium and n_i is defined as the equilibrium concentration of electrons or holes in an intrinsic semiconductor.

Eventually a given electron will recombine if a hole is available. That is, the electron will move into the empty state in the valence band, filling the hole as shown on the left of Figure 2.5c. When recombination occurs, both the quasi-free

electron and the hole disappear. The energy lost by the recombining electron is given off as a photon (for example, in lasers), as phonons (heat), or a combination of the two.

When the electron returns to the valence band, there is no longer a quasi-free electron, and there is also no longer a hole. This process is called recombination. When recombination occurs, both the quasi-free electron and the hole disappear.

Electrons will remain in the conduction band an average time τ_n before recombining with holes in the valence band. This average time between generation and recombination is called the electron lifetime or carrier lifetime. (Electrons and holes are considered carriers because they carry current.) Typical values of τ range from 10^{-10} to 10^{-3} seconds at room temperature and are material dependent.

An electron in the conduction band can also lose energy by colliding with another particle (e.g., a phonon), as seen on the right-hand side of Figure 2.5c. In this case, the electron may move to a lower energy in the conduction band. The mean free time between collisions for the electron is appreciably shorter than its lifetime, the mean free time being about 10^{-13} to 10^{-12} seconds. Thus, the electron will make many collisions before recombining.

2.6 EXTRINSIC SEMICONDUCTORS

We have seen that in intrinsic semiconductors, the concentration of conduction band electrons is equal to the concentration of holes in the valence band. By adding *dopant* atoms, it is possible for the numbers of electrons and holes to be unequal, in which case the material is said to be *extrinsic*. A semiconductor is said to be extrinsic if $n_0 \neq p_0$. Extrinsic semiconductors are created by incorporating *impurity* atoms into the intrinsic material (a process called *doping*). The dopant atoms can be either donors or acceptors, as developed in the following sections.

If $n_0 > p_0$, a semiconductor is said to be n type, meaning current is carried predominantly by negatively charged electrons. If $p_0 > n_0$, the material is p type, and current is carried predominantly by positively charged holes.

2.6.1 DONORS

Assume that an atom with five electrons in its outer shell, such as phosphorus, is substituted for a Si atom in an otherwise pure crystal of silicon, as shown in Figure 2.6a for the bond representation. Silicon has a valence of four. The extra electron of the phosphorus atom is not needed for the covalent bonding, since there are enough other electrons to fill the bonds. Thus, the extra electron is more loosely bound to the phosphorus atom than the valence electrons. The extra electron can be easily “donated” to the conduction band, as we will see, and thus phosphorus is called a *donor* atom. The donated electron in the conduction band is now free to move about the crystal. For visual clarity it is indicated near the ionized donor in Figure 2.6a (and other bond diagrams), whereas in reality it is further from the donor ion than is a loosely bound electron.

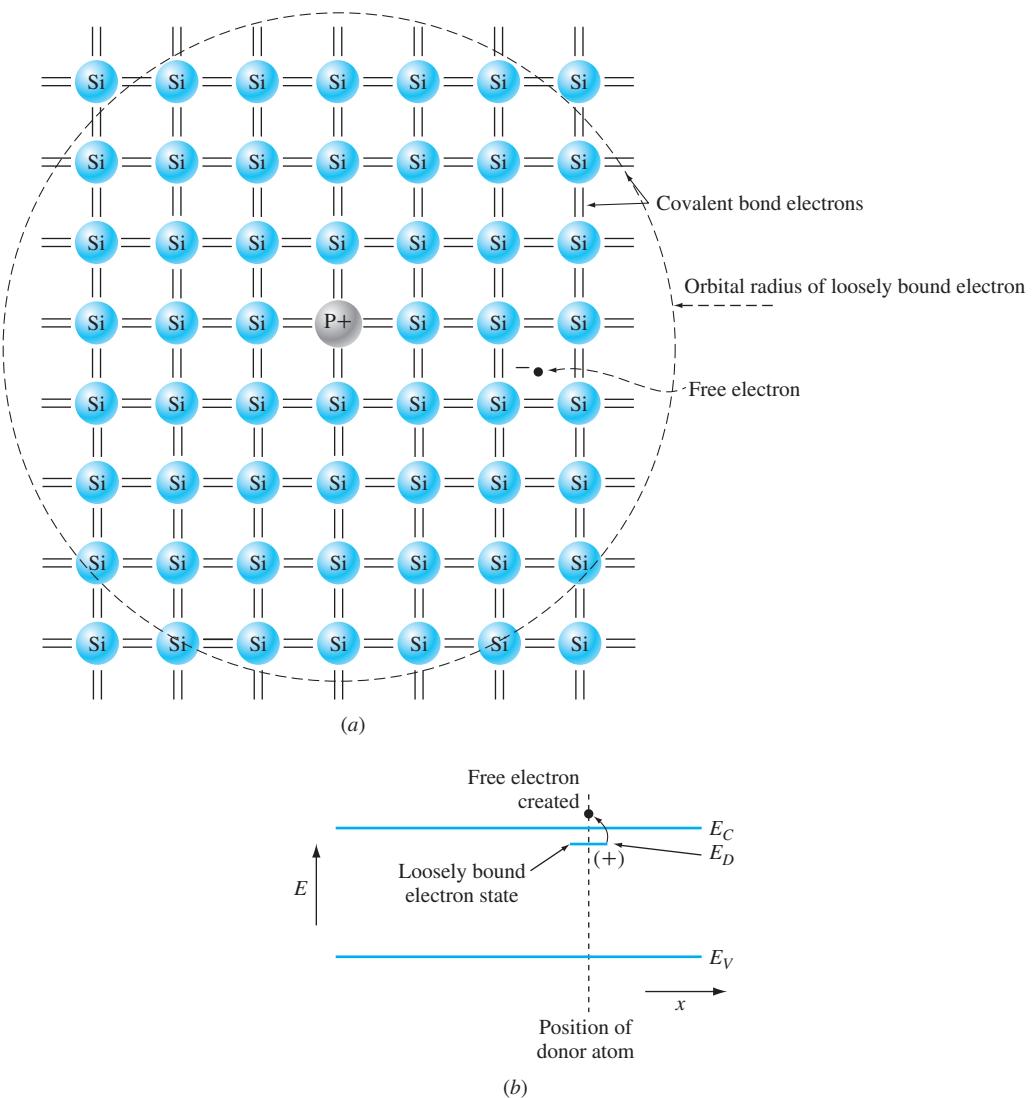


Figure 2.6 Donors in a silicon crystal: (a) bond diagram of the crystal; (b) energy band diagram of silicon doped with one phosphorus (donor) atom.

The energy band diagram for a Si crystal doped with a phosphorus donor is shown in Figure 2.6b. The phosphorus atom has a slightly different set of energy levels than the surrounding silicon atoms, and some of these are actually in silicon's forbidden gap. The lowest energy (ground) state of an electron associated with the donor atom exists only near that donor atom, so the donor state of energy E_D does not exist throughout the crystal. Also, the donor state is close to the conduction band edge. This indicates that it takes very little energy to excite

the donor's fifth electron into the conduction band. That is consistent with our previous idea that the electron is loosely bound to the donor atom because it is not participating in bonding. Notice that there is no hole created in this case. The phosphorus atom left behind is ionized positively, but it cannot move since it is "chained down" to the crystal by the four valence bonds and thus cannot contribute to current. The loosely bound electron (illustrated in Figure 2.6a) cannot contribute to current since it is (loosely) bound to the fixed donor ion. Only if it receives enough energy to excite it into the conduction band and thus becomes free of the ion (i.e., becomes quasi-free) can it contribute to current. Only the electron contributes to conduction, not the ion.

Now we examine donors more quantitatively and try to determine their energy levels. Earlier we said that the extra electrons contributed by the donor atoms appeared in the conduction band. When that is the case, the donor atom is positively ionized since it has lost one electron. The positive charge of the ion produces a coulombic force on the negative electron. This problem resembles the hydrogen atom problem, in which an electron is bound to the single-positive-charge hydrogen nucleus. From the Bohr model (as well as from quantum mechanics) for the hydrogen atom the n th energy level is given by

$$E_n = E_{\text{vac}} - \frac{m_0 q^4}{2(4\pi\epsilon_0)^2 \hbar^2 n^2} = E_{\text{vac}} - \frac{13.6}{n^2} \text{ eV} \quad (2.22)$$

and the Bohr radius (in quantum mechanical terms the most probable distance between the electron and the nucleus) is

$$r_n = \frac{4\pi\epsilon_0 n^2 \hbar^2}{m_0 q^4} = 0.053 n^2 \text{ nm} \quad (2.23)$$

For the hydrogen atom in vacuum, E_{vac} is the minimum total energy required to remove an electron from the influence of the hydrogen core (i.e., to create a free electron). A single phosphorus atom in space similarly has a set of discrete energy levels, as shown in Figure 2.7a, and an electron of energy E_{vac} or higher is free of the effect of the atom. Analogously, for a donor atom in a semiconductor, E_C is the total energy an electron must attain to be removed from the influence of the donor ion, Figure 2.7b. When the electron has energy $E > E_C$ it is free to move around the crystal in the conduction band. An electron at elevated energy states also has a large Bohr radius, so at sufficiently high energy the distance between the electron and bound donor ion is large enough that the coulombic force between them is negligible. In Figure 2.7b only one donor level is shown (the lowest one), but actually there are several allowed states for a donor in a semiconductor. For example, Figure 2.7c shows the band structure for a donor atom in GaAs. There are discrete states below the conduction band edge, analogous to the single hydrogen atom in vacuum. If any of these states is occupied, it will most likely be the lowest one. We call that level the donor ground state energy E_D .

To continue our analysis, we assume the following:

1. The average distance between the electron and the ion is large enough that the material between them more closely resembles the crystal than free

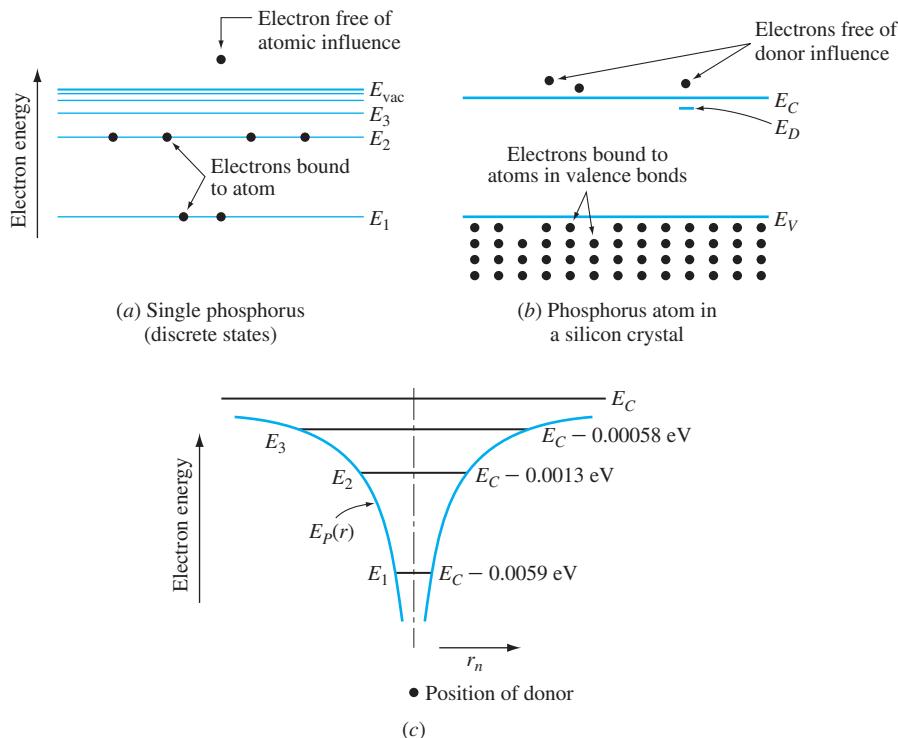


Figure 2.7 (a) The discrete energy states in a single phosphorus ion. (b) The energy band diagram for a semiconductor crystal containing a donor atom. For the discrete atom, an electron must have an energy equal to E_{vac} or higher to escape the influence of the nucleus. In the semiconductor, the electron must have an energy of E_C or greater to escape the influence of the donor ion. (c) Energy diagram for a donor in GaAs.

space. This permits us to replace the permittivity ϵ_0 of free space with the macroscopic permittivity ϵ of the material.

2. The electron in the bound state of the donor atom can be treated as though it has a conductivity effective mass m_{ce}^* equal to that of an electron near the bottom of the conduction band. (Although restricted to the region near the donor ion, the bound electron is moving in the periodic potential of the crystal, as are conduction electrons.)

Under these assumptions, Equations (2.22) and (2.23) become

$$E_n = E_C - \frac{m_{ce}^* q^4}{2(4\pi\epsilon)^2 n^2 \hbar^2} = E_C - \frac{13.6}{n^2} \left(\frac{m_{ce}^*/m_0}{\epsilon^2/\epsilon_0^2} \right) \text{ eV} \quad (2.24)$$

and

$$r_n = \frac{4\pi\epsilon n^2 \hbar^2}{m_{ce}^* q^2} = 0.053 \frac{\epsilon/\epsilon_0}{m_{ce}^*/m_0} n^2 \text{ nm} \quad (2.25)$$

We consider the case of GaAs in which an As atom (valence 5) is replaced by a tellurium atom (valence 6), and so there is an extra electron not required for bonding. For GaAs, $m_{ce}^*/m_0 = 0.067$ and $\epsilon/\epsilon_0 = 13.2$. This gives

$$E_n = E_C - \frac{0.0052}{n^2} \text{ eV} \quad (2.26)$$

$$r_n = 10.4n^2 \text{ nm} \quad (2.27)$$

We see that all values of r_n are large compared with the lattice constant of GaAs (0.565 nm) so assumption 1 is reasonably valid. That is, the lowest energy at which a bound electron can be found corresponds to an orbit with a radius of about 20 lattice constants. For visual clarity, in the bond diagram [Figure 2.6(a)] electron and donor are shown much closer than this.

The energy calculated for the ground state (the non-ionized donor atom) from Equation (2.26) is 0.0052 eV below the conduction band edge, which agrees reasonably well with experiment (0.0059 eV). Thus, assumption 1 can be taken to be reasonably accurate. Note that for the excited states, r_n is large enough that assumption 1 is quite good, and the energies calculated from Equation (2.26) agree very closely with experiment.

For the case of silicon (valence 4) doped with P (valence 5), the agreement is not as good. Using $m_{ce}^*/m_0 = 0.26$ and $\epsilon/\epsilon_0 = 11.8$ results in

$$E_n = E_C - \frac{0.026}{n^2} \text{ eV} \quad (2.28)$$

$$r_n = 2.4n^2 \text{ nm} \quad (2.29)$$

For the ground state, from Equation (2.28), $E_C - E_1 = 0.026$ eV, which is appreciably smaller than the measured value of 0.045 eV for phosphorus in silicon. This discrepancy is explained by noticing that $r_1 = 2.4$ nm [Equation (2.29)], which is not sufficiently large compared with silicon's lattice constant of 0.543 nm to justify assumption 1. Notice, however, that for all of the *excited* states, r_n is large enough to reasonably satisfy assumption 1, and for these states the model agrees quite well with experiment.

For semiconductors (or insulators) with large effective mass or small dielectric constant, however, Equations (2.24) and (2.25) are not valid, and donor states are at energies appreciably below the conduction band.

2.6.2 ACCEPTORS

Instead of doping a semiconductor with donors, one could add impurities that have fewer electrons in the outer shell than the replaced atoms. Consider the case of Si (four valence electrons) doped with boron, which has three valence electrons. Figure 2.8a shows how doping Si with boron will produce a p-type

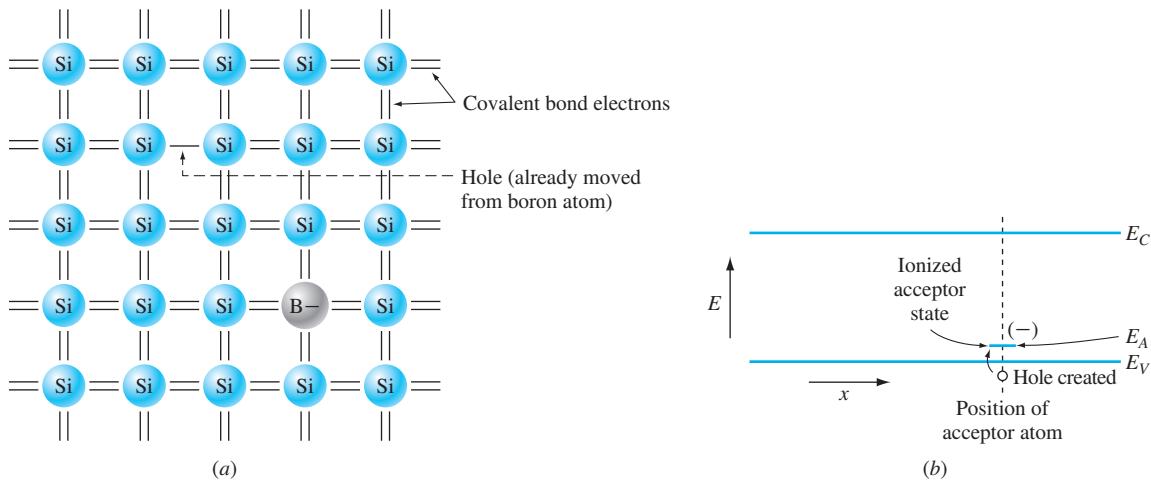


Figure 2.8 Acceptors in a semiconductor: (a) bond diagram; (b) energy band diagram. An electron is excited from the valence band to the acceptor state, leaving behind a quasi-free hole.

material. Boron is called an acceptor in silicon. We know that atoms in general like to fill their outer shells, even if it means sharing electrons. The silicon atoms near the boron atom would like to have eight electrons apiece, but they are collectively one electron short. If an electron from a nearby silicon atom should occupy that state, the bond becomes full, but the boron atom becomes negatively charged since it now has more electrons than protons. Since the electron most likely came from some nearby covalent bond, a hole was left behind. The energy band diagram for a semiconductor material doped with one acceptor is shown in Figure 2.8b. Note that in (a), the hole has already moved a distance from the boron atom.

It does not require much energy for an electron to be excited from the valence band to the acceptor state, at energy E_A . When that does happen, a hole is created in the valence band (a hole that is free to move around) and the acceptor atom becomes negatively ionized. Again, the acceptor atom is locked into the crystal and cannot move or carry current; only the hole can move.

For the acceptors, to find the energy of the acceptor states E_A , we look at the valence band rather than the conduction band. As indicated earlier, in most semiconductors of interest to electronics there are three valence bands, all of which have a maximum at $K = 0$. Two of these have maxima at $E = E_V$, while the third is split off by an energy Δ . Following a similar procedure to the calculations we used for the donor states, and making the same assumptions, for the bound states of holes we obtain

$$E_n = E_V + \frac{13.6}{n^2} \left(\frac{m_{ch}^*/m_0}{\epsilon^2/\epsilon_0^2} \right) \text{eV} \quad (2.30)$$

$$r_n = 0.053 n^2 \left(\frac{\epsilon}{\epsilon_0} \right) \left(\frac{m_0}{m_{ch}^*} \right) \text{nm} \quad (2.31)$$

where m_{ch}^* is the conductivity effective mass for holes. By “bound state for a hole,” we mean that the acceptor atom is occupied by a hole (is un-ionized).

Recall that earlier we indicated that if a III-V material such as GaAs is the host material, it could be made n type by substituting atoms from the sixth column of the periodic table (e.g., tellurium) for arsenic atoms (five valence electrons). [GaAs could also be made n type by substituting atoms from the fourth column, such as silicon, for the gallium atoms (three valence electrons). In either case there is one more electron added than is needed to complete the valence bond.] Similarly, we can obtain p-type GaAs by replacing column III (gallium), with column II (zinc), or by replacing column V, arsenic, with column IV, silicon. The system is then one electron short of the number needed to complete the covalent bonding, and that atom becomes an acceptor.

In GaAs, column IV elements such as germanium and silicon can be either donors or acceptors, depending on which type of atom is replaced. These are called *amphoteric impurities*. Silicon preferentially occupies Ga sites and thus is normally a donor.

2.7 THE CONCEPT OF HOLES

2.7.1 HOLE CHARGE

We mentioned earlier that states in the valence band that are *not* occupied by electrons are referred to as holes, and that these holes can move around in the valence band. In this section, we will develop this idea further and show more rigorously that holes can be thought of as carrying current and possessing a positive charge. We approach this discussion by considering the current.

Consider an n-type semiconductor with an electric field applied. There are n electrons per unit volume in the conduction band, and p holes per unit volume in the valence band ($n > p$, for n type). The charge density in the conduction band is therefore $-qn$, the charge per electron times the density of electrons. The electron current density (amperes per unit area) is

$$J = -qn\langle v \rangle \quad (2.32)$$

where $\langle v \rangle$ is the average velocity of the electrons. We could also have written Equation (2.32) in the form

$$J = -\frac{q}{\text{volume}} \sum_i v_i \quad (2.33)$$

where the volume is the volume of the crystal and the summation is taken over the velocities of the individual electrons.

Equation (2.33) is also valid for electrons in the valence band, but the valence band is essentially filled. A completely filled valence band is shown for the one-dimensional case in the E - K diagram in Figure 2.9a. Recall that the velocity of an electron is given by

$$v = \frac{1}{\hbar} \frac{dE}{dK} \quad (2.34)$$

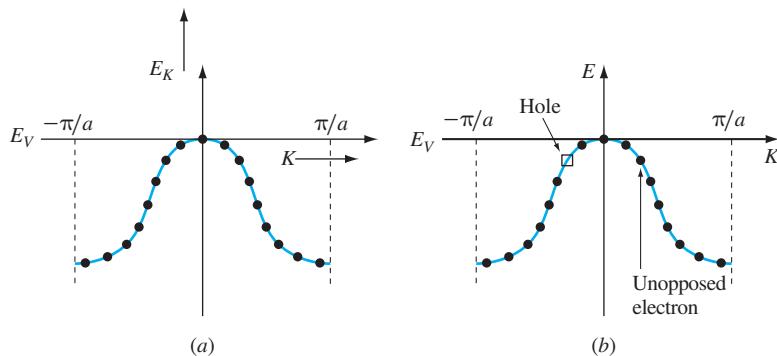


Figure 2.9 States in the valence band on the E - K diagram. (a) All states are full; (b) one state is empty (hole), meaning there is an electron with no opposing electron at the same energy but opposite K vector and opposite velocity.

From the figure, the E - K diagram is symmetrical, so for each electron with velocity v_i there is an opposing electron with velocity $-v_i$. Thus for electrons in the valence band, the total current density $J = 0$ when every state is filled. Note that the electrons are moving about in real space, but at a given energy, at every value of E there are two states in the figure, with opposite velocities.

At temperatures above absolute zero, the valence band is largely occupied by electrons, but there is still a significant population of holes (for example in p-type material the holes represent 10^{-8} to 10^{-4} of the valence band population of electrons). Most of the electrons in the valence band are opposed by electrons with opposite velocity, but not all, as shown in Figure 2.9b. It is therefore convenient to consider the unopposed electrons:

$$J = \frac{-q}{\text{volume}} \sum_i v_{ui} \quad (2.35)$$

where the subscript u refers to the unopposed electrons. Equation (2.35) can be rewritten

$$J = \frac{-q}{\text{volume}} \sum_i [v_i - v_{hi}] = \frac{-q}{\text{volume}} [\sum_i v_i - \sum_i v_{hi}] \quad (2.36)$$

where the first term in the brackets is the summation over all the electrons in the (full) band while the second (v_{hi}) is the summation over the vacant states (holes), $\sum_i v_{ui} = \sum_i (v_i - v_{hi})$.

Since

$$\sum_i v_i = 0 \quad (2.37)$$

it follows that

$$J = +\frac{q}{\text{volume}} \sum_i v_{hi} \quad (2.38)$$

If we consider the holes to be particles, from Equation (2.38) we see that they must be each considered to have a charge of $+q$. Thus for a hole density p , the total hole current density is

$$J = +qp\langle v_h \rangle \quad (2.39)$$

It therefore makes no difference whether we consider the current to be due to all the electrons in the valence band of charge $-q$ or whether we consider the current to be due to the number of holes of charge $+q$. Generally we talk about the holes when discussing conduction current in the valence band.

2.8 EFFECTIVE MASS OF ELECTRONS AND HOLES

The effective mass of electrons and holes can be measured using the Hall effect. Consider a semiconductor with a current flowing to the right (electrons flowing to the left), as shown in Figure 2.10. A magnetic field is also applied such that the direction of \vec{B} is into the page. The Lorentz force, $F = Q[\vec{v} \times \vec{B}]$ on an electron is given by

$$\vec{F}_e = -q[\vec{v}_e \times \vec{B}] = m_e^* \vec{a}_e \quad (2.40)$$

where \vec{F}_e is the force on the electron, \vec{v}_e the electron velocity, \vec{a}_e its acceleration, and m_e^* its effective mass. The negative sign comes from the negative charge of the electron. For an n-type semiconductor (Figure 2.10a), current is carried by electrons in the conduction band. Their velocity is to the left. Since $\vec{v}_e \times \vec{B}$ is directed downward in this example, the force on the electron is directed upward. Since m_e^* is positive, the electron is accelerated upward. Thus, the top of the semiconductor is negative with respect to the bottom and the generated voltage, V , is negative. From the measured value of V , along with the conductivity and dimensions of the semiconductor, the effective mass can be calculated.

Earlier we showed that when an electron is near the top of the valence band, its effective mass is negative, since the curvature of the E - K diagram is negative in that region. We now show that if we consider the vacant states to be holes with positive charge, they must also be considered to have a positive effective mass, equal in magnitude to the (negative) effective mass of the unopposed electron. We do this by considering a p-type semiconductor with a current flowing to the right (electrons flowing to the left), as shown in Figure 2.10b. Let us think of the current as being carried by unopposed electrons in the valence band. A magnetic field is again applied such that the direction of \vec{B} is into the page. The Lorentz force on an unopposed electron is given by

$$F_{ue} = -q[\vec{v}_{ue} \times \vec{B}] = m_{ue}^* \vec{a}_{ue}$$

where the subscript “ue” refers to the unopposed electron in the valence band. Since $\vec{v}_{ue} \times \vec{B}$ is directed downward in this example, the force on the electron is directed upward. The unopposed electron is necessarily near the top of the valence band (because the states at lower energies are very likely to be occupied),

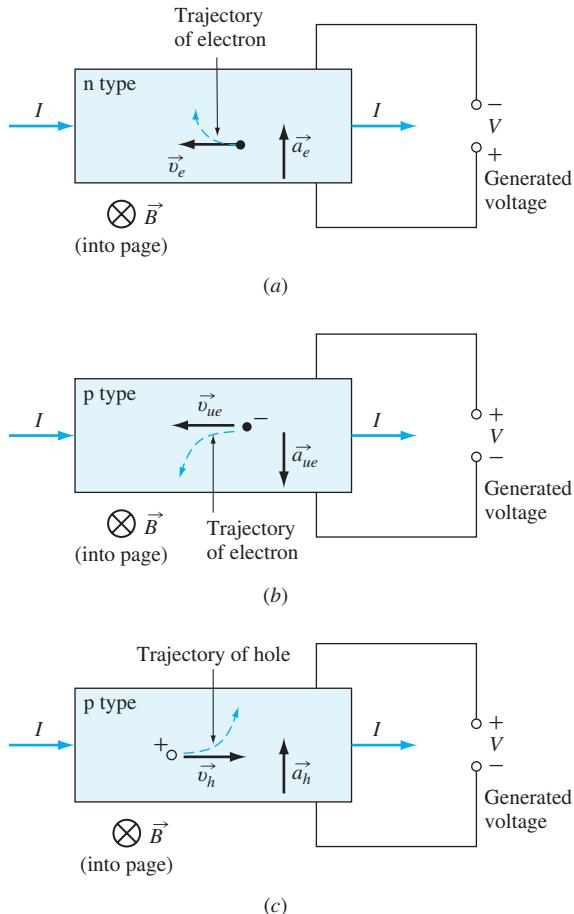


Figure 2.10 Lorenz force on a carrier in a semiconductor. (a) Electron in the conduction band (negative charge, positive effective mass). (b) Unopposed electron in the valence band (negative charge, negative effective mass). (c) Hole in the valence band (positive charge, positive effective mass).

so its effective mass m_{ue}^* is negative. This means that the electron is, in fact, accelerated *downward*, the result being that the bottom of the sample becomes negatively charged. Now the generated voltage on the sample is positive, as shown in the figure.

If, instead of considering the electrons, we consider a hole as shown in Figure 2.10c, the Lorentz force is

$$\vec{F}_p = +q[\vec{v}_h \times \vec{B}] = m_h^* \vec{a}_h \quad (2.41)$$

where m_h^* and a_h are the effective mass and acceleration of the hole. The holes, charged positively, have a velocity \vec{v}_h to the right because the current is flowing to the right. The net Lorentz force is then of the same magnitude and direction as for the unopposed electron. For the same physical result, that is, a sample whose top is positively charged with respect to its bottom, each hole must have actually gone upward, implying that its effective mass is in fact positive, but equal in magnitude to that of the electron in the opposing state.

We conclude, therefore, that in a nearly filled band (the valence band), the vacant states can be treated as particles, called holes. These holes have positive charge and positive effective mass and thus can respond to external forces. The effective mass of a given hole is equal in magnitude but opposite in sign to that of an electron near the top of the valence band (i.e., positive).

2.9 DENSITY-OF-STATES FUNCTIONS FOR ELECTRONS IN BANDS

We have established that there are bands of energy states that electrons can occupy in a semiconductor. We have also pointed out that most electrons of interest will be near the bottom of the conduction band, and that most holes will be near the top of the valence band.

To describe current flow in semiconductors, however, we need more information. We need to know the density of quasi-free electrons and of holes and their distributions with energy more precisely. It is not evident yet, but in most electronic devices, the device current is determined by the number of electrons or holes that have enough energy to surmount various energy barriers that are built into the devices. To determine the current-voltage relations in these devices, then, we have to know the distribution of electrons (and holes) with energy.

There are two things that go into that determination. The first is that we must know how the available states are distributed in energy. The second factor we need to know is the probability that a state at a given energy is occupied. For example, we expect intuitively that the higher the energy state, the less likely it is to be occupied.

We begin with the distribution of states in the next section, and in the following section we discuss the probability of occupancy.

2.9.1 DENSITY OF STATES AND DENSITY-OF-STATES EFFECTIVE MASS

We begin by examining the density-of-states function, which is derived in Online Module OM5. For a free electron ($E_p = E_0$ and is constant), the available states are distributed in energy according to

$$S(E) = \frac{1}{2\pi^2} \left(\frac{2m_0}{\hbar^2} \right)^{3/2} \sqrt{E - E_0} \quad (2.42)$$

where $S(E)$, the *density-of-states function*, is the number of states per unit volume per unit energy. We would like to use an analogous formula to describe the density of states in a semiconductor. We confine ourselves to the parabolic

regions of the E - K diagram, where the effective mass is constant. We can then write for an electron near the bottom of the conduction band

$$\begin{aligned} S(E) &= \frac{1}{2\pi^2} \left(\frac{2m_{dse}^*}{\hbar^2} \right)^{3/2} \sqrt{E - E_C} \\ &= \frac{1}{2\pi^2} \left(\frac{2m_{dse}^*}{\hbar^2} \right)^{3/2} \sqrt{E_K} \quad \text{conduction band} \end{aligned} \quad (2.43)$$

where m_{dse}^* is referred to as the density-of-states effective mass for electrons. The kinetic energy E_K is the difference between the total energy E and the potential energy E_C for an electron in the parabolic region near the bottom of the conduction band: $E_K = E - E_C$.

Near the bottom of the conduction band of a material such as GaAs, m_e^* is a scalar. As we mentioned earlier, however, and discuss in some detail in Online Module OM5, for a material like Si, in each equivalent minimum the effective mass is direction dependent, so some type of average of longitudinal and transverse electron masses ($m_{||}^*$ and m_{\perp}^* respectively) must be used, and the number of equivalent minima must be taken into account to arrive at a value of m_{dse}^* . The details of this are also handled in Online Module OM5. Values for some semiconductors of interest were given in Table 2.1.

Similarly, in the parabolic region near the top of the valence band, the density of states function is

$$S(E) = \frac{1}{2\pi^2} \left(\frac{2m_{dsh}^*}{\hbar^2} \right)^{3/2} \sqrt{E_V - E} \quad \text{valence band} \quad (2.44)$$

This equation is for the density of states of electrons or *holes*, the valence band edge E_V is the potential energy for holes in the valence band, and m_{dsh}^* is some combination of the light hole and heavy hole effective masses as discussed in Online Module OM5. Some results were listed in Table 2.2.

The density-of-states functions for electrons in the conduction band and in the valence band are plotted schematically in Figure 2.11, where the $S(E)$ plots are superimposed on the energy band diagram (energy versus position). Note that there are no states for electrons or holes in the forbidden band (for purely intrinsic material), so $S(E)$ is zero there. We emphasize that Equations (2.43) and (2.44) for $S(E)$ are valid only in the parabolic regions of the E - K curves.

2.10 FERMI-DIRAC STATISTICS

Now that we know how the available states are distributed in energy, we need to examine the probability that a given energy state is occupied. Given one state at energy E_a , let $f(E_a)$ be the probability that an electron occupies this particular state. If there are S_a states per unit volume at energy E_a , then the electron concentration n_a at E_a is

$$n_a(E_a) = S_a f(E_a) \quad (2.45)$$

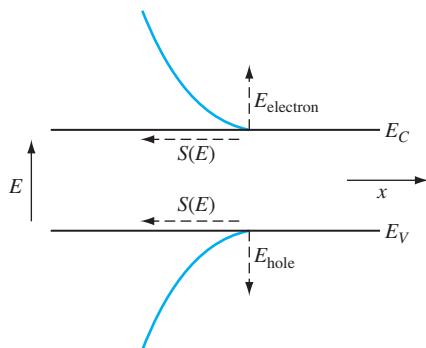


Figure 2.11 The density-of-states functions for electrons in the conduction band and the valence band. The density-of-states-versus-energy plot is superimposed on the energy band diagram (energy versus position x).

We can generalize for states distributed in energy:

$$n(E) = S(E)f(E) \quad (2.46)$$

where $n(E)$ is the number of electrons per unit volume per unit energy. In a small energy range dE ,

$$n(E)dE = S(E)f(E)dE \quad (2.47)$$

In a given band, then, at equilibrium, the total number of electrons per unit volume, n_0 , in the entire energy band is

$$n_0 = \int_{\text{band}} S(E)f(E)dE \quad (2.48)$$

where the integration is taken over the total band of allowed energies.

To determine n_0 exactly, both $S(E)$ and $f(E)$ must be known. In Section 2.9.1, we gave equations for $S(E)$, but they are valid only near the band extrema, because these equations involve the density-of-states effective mass. As long as we consider carriers near the band extrema, $S(E)$ is known. We then turn to the probability of occupancy $f(E)$. The term $f(E)$ can be determined for all energies, but it has a different form depending on whether the electrons under consideration are free to move through the crystal (as within a band) or bound to localized states (as for impurities).

The derivation for $f(E)$ is somewhat involved; we state and discuss the results in the following sections. [2]

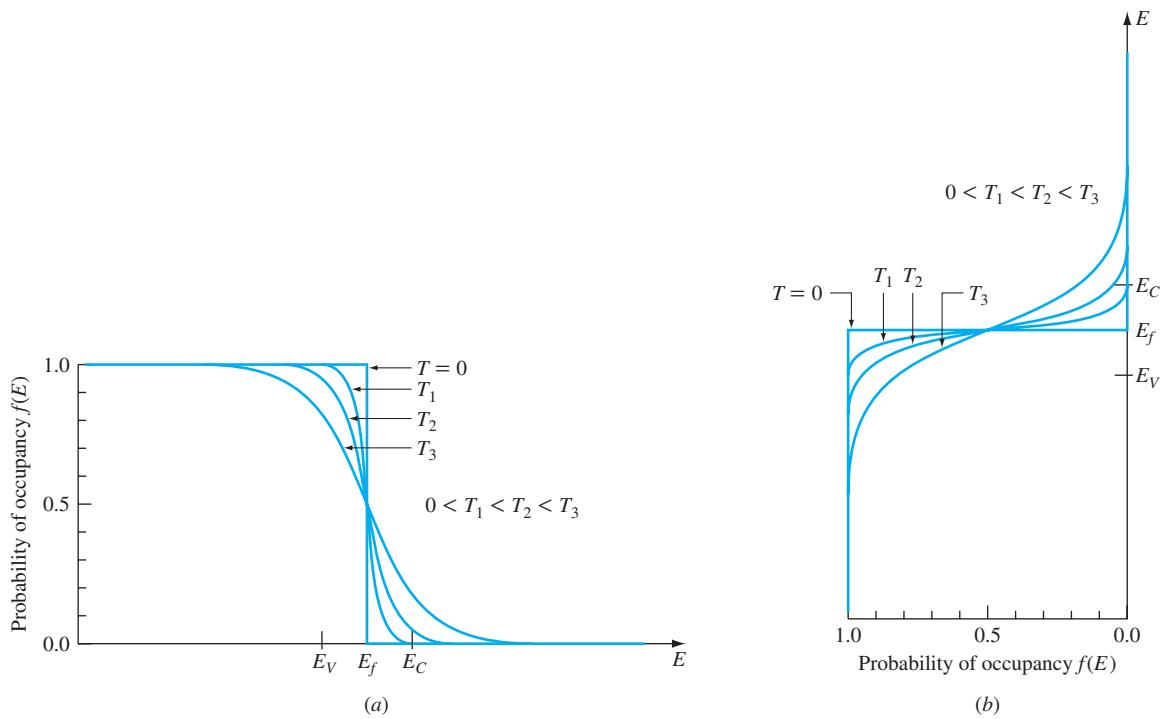


Figure 2.12 The Fermi-Dirac distribution function gives the probability of occupancy of an energy state E if the state exists.

2.10.1 FERMI-DIRAC STATISTICS FOR ELECTRONS AND HOLES IN BANDS

The probability that an electron occupies a given state at energy E in an allowed band is given by

$$f(E) = \frac{1}{1 + e^{(E - E_f)/kT}} \quad \text{Fermi-Dirac probability function} \quad (2.49)$$

where the energy E_f is a reference energy called the *Fermi energy* or *Fermi level*, and Equation (2.49) is known as the *Fermi-Dirac probability function*. Figure 2.12a shows the Fermi-Dirac distribution as a function of energy for several temperatures. We also indicate possible values for E_C and E_V . In Figure 2.12b we have rotated by 90° the conventional plot of the dependent variable on the y axis, independent variable on the x axis. We do this to make this plot agree with other plots, such as energy band diagrams, in which energy is plotted vertically. Probability increases to the left on the plot in Figure 2.12b.

The Fermi level or Fermi energy is chosen to be the particular energy level for which $f(E) = \frac{1}{2}$, or the probability of occupancy of a state at that energy

(if the state exists) is 50 percent. States at energies below the Fermi level are more likely to be occupied than empty. States at energies above the Fermi level have a probability of occupancy less than $\frac{1}{2}$, meaning they are more likely to be empty than full.

Figure 2.12 shows the Fermi-Dirac function for several different temperatures. At absolute zero, every electron is at its lowest possible energy. In a semiconductor, that means that every state in the valence band is occupied. The probability of occupancy of any state in the valence band is therefore unity, and the probability of occupancy of any state in the conduction band is zero. The Fermi level happens to be very close to the middle of the forbidden gap for intrinsic materials. Note that the probability of occupancy of a state at the Fermi level is $\frac{1}{2}$, but that there are, in fact, no states there.

Now consider a somewhat higher temperature, T_1 . Since the crystal has some thermal energy, some electrons have enough energy to be in the conduction band. The probability of occupancy of a state near the bottom of the conduction band is less than $\frac{1}{2}$, but it is not zero, either. This means that at any given time, a given state is *probably* empty, but there is a small probability that an electron occupies it. The higher the energy, the less likely the state is to be occupied. Similarly, a state near the top of the valence band is *probably* occupied. Some of these states will be empty, indicating that there are some holes. The states most likely to be unoccupied in the valence band are those near the top, meaning that most of the holes will be found near the top of the valence band.

The probability of a state being occupied by a *hole* is one minus the probability of a state being occupied by an electron, since the state must either be occupied or not. Thus the Fermi-Dirac distribution for holes is

$$f_p(E) = 1 - f(E) \quad (2.50)$$

From Equation (2.49), this gives

$$f_p(E) = 1 - \frac{1}{1 + e^{(E - E_f)/kT}} \quad (2.51)$$

or

$$f_p(E) = \frac{1}{1 + e^{(E_f - E)/kT}} \quad (2.52)$$

If in Equation (2.49) the term $e^{(E - E_f)/kT} \gg 1$, then $f(E) \ll 1$ and

$$f(E) \approx e^{-(E - E_f)/kT} \quad (2.53)$$

Similarly, in Equation (2.52), for $f_p(E) \ll 1$,

$$f_p(E) \approx e^{-(E_f - E)/kT} \quad (2.54)$$

Equations (2.53) and (2.54) are referred to as *the Boltzmann approximations to the Fermi-Dirac probability function*, or simply the *Boltzmann probability function*.

EXAMPLE 2.1

When can the Boltzmann approximation safely be used?

Solution

The expression $f(E) = e^{-(E-E_f)/kT}$ is approximately equal to the true probability

$$f(E) = \frac{1}{1 + e^{(E-E_f)/kT}}$$

if the quantity $e^{(E-E_f)/kT}$ is large compared with unity. Let us take “large compared with” to mean greater by a factor of 10. Then $e^{(E-E_f)/kT} \approx 10$ and $(E - E_f) > kT \ln(10)$. Thus for the Boltzmann approximation to apply, we require that $(E - E_f) > kT \ln(10) = 2.3kT$. That is, if we are calculating the probability of occupancy of a state that is greater than $2.3kT$ away from the Fermi level, we can use the simpler, approximate form [Equation (2.53) for electrons or Equation (2.54) for holes].

Returning to Figure 2.12, we see that with increasing temperature, the probability of occupancy of a given state in the conduction band increases. Likewise, the probability of occupancy of a given state in the valence band decreases. This means that at higher temperatures, there are more electrons and more holes available to carry current.

EXAMPLE 2.2

Estimate the probability of occupancy of a state at the bottom of the conduction band in intrinsic Si at room temperature.

Solution

The band gap of silicon at room temperature is 1.12 eV. The electron is at the bottom of the conduction band, or at $E = E_C$. As we will show later, the Fermi level is approximately at midgap for intrinsic materials, so $E_C - E_f \approx E_g/2 = 0.56$ eV. From Equation (2.49), we have

$$\begin{aligned} f(E) &= \frac{1}{1 + e^{(E-E_f)/kT}} = \frac{1}{1 + e^{(E_c-E_f)/kT}} \\ &= \frac{1}{1 + e^{(0.56 \text{ eV})/(0.026 \text{ eV})}} = \frac{1}{1 + 2.26 \times 10^9} = 4.4 \times 10^{-10} \end{aligned}$$

or about one state in two billion is occupied. Note that in this example the Boltzmann probability distribution function is valid.

The Fermi-Dirac probability function for electrons in bound states is somewhat different, but for most of the cases in this book, we will assume that all donors and acceptors are ionized. This assumption is justified in Online Module OM2.

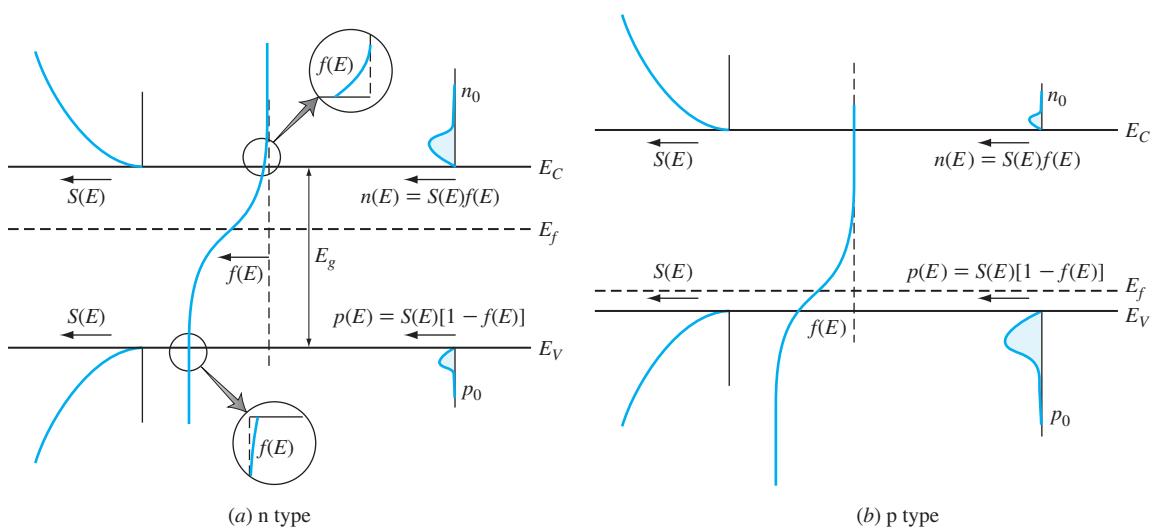


Figure 2.13 The distribution of the electrons near the bottom of the conduction band, $n(E)$, is the product of the density-of-states distribution $S(E)$ times the probability of occupancy of states $f(E)$ at a particular energy. The distribution of holes near the top of the valence band $p(E)$ is the product of the density-of-states distribution times the probability of vacancy of states at a particular energy. (a) n type, (b) p type.

2.11 ELECTRON AND HOLE DISTRIBUTIONS WITH ENERGY

We now know the distribution of available states $S(E)$ near the bottom of the conduction band and near the top of the valence band, and we know the probability of occupancy $f(E)$ of a state of energy E . In this section, we will put this information together to find the distribution of electrons with energy in the conduction band and the distribution of holes with energy in the valence band. The total concentrations of electrons in the conduction band and holes in the valence band are then obtained.

Figure 2.13 shows a plot of the energy band diagram of a semiconductor. Part (a) depicts the case of an n-type semiconductor and (b), p-type. Plots for $S(E)$ and $f(E)$ are superimposed onto the energy band diagram. The electron and hole distribution functions, $n(E)$ and $p(E)$, are also shown shaded, where

$$\begin{aligned} n(E) &= S(E)f(E) && \text{conduction band} \\ p(E) &= S(E)[1 - f(E)] = S(E)f_p(E) && \text{valence band} \end{aligned}$$

The area under the respective curves represents the total electron concentration n_0 in the conduction band and hole concentration in the valence band p_0 . The shaded area for electrons is larger than for holes in n-type material because the Fermi level is closer to the conduction band than to the valence band. In p-type material, the hole concentration is larger than the electron concentration

because the Fermi level is closer to the valence band. For visual clarity the area representing the relative values for n_0 and p_0 are greatly exaggerated in the figure. As will be shown, for n -type silicon, with a donor concentration of 10^{17} cm^{-3} , $n_0 \approx 10^{17} \text{ cm}^{-3}$ and $p_0 \approx 10^3 \text{ cm}^{-3}$, a difference too large to be shown in a single figure.

The total number of electrons in the conduction band at equilibrium is obtained by integrating over the band:

$$n_0 = \int_{E_C}^{\text{top}} S(E)f(E)dE = \int_{E_C}^{\text{top}} n(E)dE \quad (2.55)$$

where n_0 is the equilibrium concentration of electrons in the conduction band and the integration is taken from the bottom of the conduction band E_C to the top of the band.

Similarly, the equilibrium hole distribution function in the valence band becomes

$$p_0 = \int_{\text{bottom}}^{E_V} S(E)f_p(E)dE = \int_{\text{bottom}}^{E_V} p(E)dE \quad (2.56)$$

where the integration is taken over the range of valence band energies.

Recall that most of the carriers are concentrated near the extrema of their respective bands. This is fortunate, because these are also the regions where the effective mass is constant, and thus where $S(E)$ is known. Expressions for $S(E)$ were given by Equations (2.43) and (2.44), and are repeated here:

$$S(E) = \frac{1}{2\pi^2} \left(\frac{2m_{dse}^*}{\hbar^2} \right)^{3/2} \sqrt{E - E_C} \quad \text{electrons} \quad (2.43)$$

$$S(E) = \frac{1}{2\pi^2} \left(\frac{2m_{sh}^*}{\hbar^2} \right)^{3/2} \sqrt{E_V - E} \quad \text{holes} \quad (2.44)$$

Note that the effective masses in the preceding equations are the “density of states” effective masses. Further, the energies at the top of the conduction band and the bottom of the valence band are not accurately known, meaning Equations (2.55) and (2.56) cannot be solved exactly.

We will now solve Equations (2.55) and (2.56) (approximately) for *nondegenerate semiconductors*. By definition, a semiconductor is said to be nondegenerate if the probability of any state in the conduction band being occupied by an electron is small and the probability of any state in the valence band being occupied by a hole is small.³ If by “small” we take a probability of 10 percent, then this implies that the Fermi level is at least $2.3kT$ below the conduction band

³Conversely, a semiconductor is said to be degenerate if the Boltzmann approximation is not valid for one of the bands. An alternative definition of a degenerate semiconductor is that the Fermi level is inside the conduction band (degenerate, n type) or inside the valence band (degenerate, p type).

edge and at least $2.3kT$ above the valence band edge. When $E_C - E_f > 2.3kT$ and $E_f - E_V > 2.3kT$, then the Boltzmann approximation to the Fermi-Dirac probability function is valid and can be used in both bands.

To solve the equations we need to make two more simplifications:

1. Assume $S(E)$ has the form given in Equations (2.43) and (2.44) anywhere that $E \geq E_C$ or $E \leq E_V$ respectively. The density of states function $S(E)$ is zero in the forbidden gap.
2. The upper limit of integration in the conduction band is extended to $+\infty$ and the lower limit of integration in the valence band is extended to $-\infty$.

The last assumption can be justified since the probability factors $f(E)$ and $f_p(E)$ decrease rapidly (exponentially) with energy away from the band edges. In those regions, the product of the density of states and the probability will be so small it contributes little to the integration. The major contribution to the integration will be within the first few kT of the band edges, where Equations (2.43) and (2.44) are valid (and the density-of-states effective masses are known).

We can now write Equations (2.55) and (2.56) as

$$n_0 = \frac{1}{2\pi^2} \left(\frac{2m_{dse}^*}{\hbar^2} \right)^{3/2} \int_{E_C}^{\infty} \sqrt{E - E_C} e^{-[(E - E_f)/kT]} dE \quad (2.57)$$

and similarly

$$p_0 = \frac{1}{2\pi^2} \left(\frac{2m_{dsh}^*}{\hbar^2} \right)^{3/2} \int_{-\infty}^{E_V} \sqrt{E_V - E} e^{-[(E_f - E)/kT]} dE \quad (2.58)$$

The results are

$$n_0 = N_C e^{-[(E_C - E_f)/kT]} \quad (2.59)$$

$$p_0 = N_V e^{-[(E_f - E_V)/kT]} \quad (2.60)$$

where

$$N_C = 2 \left(\frac{m_{dse}^* k T}{2\pi \hbar^2} \right)^{3/2} \quad (2.61)$$

$$N_V = 2 \left(\frac{m_{dsh}^* k T}{2\pi \hbar^2} \right)^{3/2} \quad (2.62)$$

Here N_C is called the *effective density of states in the conduction band*. If there were N_C states all located at $E = E_C$, the probability of one state being occupied would be $e^{-[(E_C - E_f)/kT]}$ and the density of electrons in the conduction band, n_0 , would be given by Equation (2.59). Likewise, N_V is referred to as the effective density of states in the valence band.

The effective densities of states N_C and N_V can be expressed for a nondegenerate semiconductor as

$$N_C = 2.54 \times 10^{19} \left(\frac{m_{dse}^*}{m_0} \right)^{3/2} \left(\frac{T}{300} \right)^{3/2} \text{ cm}^{-3} \quad (2.63)$$

$$N_V = 2.54 \times 10^{19} \left(\frac{m_{dsh}^*}{m_0} \right)^{3/2} \left(\frac{T}{300} \right)^{3/2} \text{ cm}^{-3} \quad (2.64)$$

These forms are useful for comparing materials or temperatures. For example, for a material of smaller effective mass, the effective density of states is correspondingly reduced.

We can derive a very useful relation by multiplying Equations (2.59) and (2.60):

$$n_0 p_0 = N_C N_V e^{-[(E_C - E_f)/kT]} e^{-[(E_f - E_V)/kT]} = N_C N_V e^{-(E_g/kT)} = n_i^2 \quad (2.65)$$

Where $E_g = E_C - E_V$ or

$$n_0 p_0 = n_i^2 \quad \text{nondegenerate semiconductors} \quad (2.66)$$

The quantity n_i is the electron concentration in the conduction band for intrinsic material. This shows that the electron-hole product at equilibrium is strongly dependent on temperature and band gap. Equation (2.66) holds for all nondegenerate semiconductors at equilibrium, but for intrinsic materials it is also true that

$$n_0 = p_0 = n_i \quad \text{intrinsic} \quad (2.22)$$

Equation (2.66) is often referred to as the *law of mass action* and is valid only for nondegenerate semiconductors. It is analogous to a similar relation in chemistry.

EXAMPLE 2.3

Find the intrinsic concentration n_i for the case of Si ($n_0 = p_0 = n_i$) at room temperature (300 K). The effective masses are⁴ $m_{dse}^* = 1.09m_0$ and $m_{dsh}^* = 1.15m_0$.

Solution

From Equation (2.65),

$$n_i = \sqrt{N_C N_V} e^{-E_g/2kT} \quad (2.67)$$

From Equations (2.63) and (2.64), we find that at 300 K, $N_C = 2.86 \times 10^{19} \text{ cm}^{-3}$ and $N_V = 3.10 \times 10^{19} \text{ cm}^{-3}$.

Since the result depends strongly (exponentially) on $E_g/2kT$, we use the accurate values of $E_g = 1.1242 \text{ eV}$ for Si at $T = 300 \text{ K}$ and $kT = 0.02586 \text{ eV}$ to avoid roundoff error. Then

$$\begin{aligned} n_i(300 \text{ K}) &= \sqrt{(2.86 \times 10^{19})(3.01 \times 10^{19})} e^{-1.1242/(2 \times 0.02586)} \\ &= 1.08 \times 10^{10} \text{ electrons/cm}^3 \end{aligned}$$

⁴ The parameters for Si used here are from reference [1]. In some of the earlier literature slightly different values are used for effective masses, resulting in $n_i \approx 1.5 \times 10^{10} \text{ cm}^{-3}$.

From Equations (2.59) and (2.60), we can write for an intrinsic semiconductor

$$n_0 = p_0 = n_i = N_C e^{-(E_C - E_i)/kT} = N_V e^{-(E_i - E_V)/kT} \quad (2.68)$$

where E_i is defined as the Fermi energy for an intrinsic semiconductor. Then

$$\frac{N_C}{N_V} = e^{(E_C - E_i)/kT_e - (E_i - E_V)/kT} = e^{(E_C + E_V - 2E_i)/kT} \quad (2.69)$$

Solving for E_i gives

$$E_i = \frac{(E_C + E_V)}{2} + \frac{kT}{2} \ln \frac{N_V}{N_C} \quad (2.70)$$

Where $(E_C + E_V)/2$ is the energy at midgap.

We can use this to locate the intrinsic Fermi level E_i . From Equations (2.61) and (2.62),

$$\frac{N_V}{N_C} = \left(\frac{m_{dsh}^*}{m_{dse}^*} \right)^{3/2} \quad (2.71)$$

and

$$E_i = \left(\frac{E_C + E_V}{2} \right) + \frac{3}{4} kT \ln \left(\frac{m_{dsh}^*}{m_{dse}^*} \right) \quad (2.72)$$

This means that E_i is offset from midgap by the term $\frac{3}{4}kT \ln(m_{dsh}^*/m_{dse}^*)$, which is usually small.

EXAMPLE 2.4

Find the energy by which E_i is offset from midgap for Si at room temperature.

■ Solution

From Equation (2.72),

$$E_i - E_{\text{midgap}} = \frac{3}{4} kT \ln \left(\frac{m_{dsh}^*}{m_{dse}^*} \right) = \frac{3}{4} (0.026) \ln \left(\frac{1.15}{1.09} \right) = 1.05 \text{ meV}$$

which is small compared with $E_g = 1.12 \text{ eV}$.

It is often convenient to use alternative expressions to those of Equations (2.59) and (2.60) for n_0 and p_0 . Since for an intrinsic semiconductor

$$n_0 = n_i = N_C e^{-(E_C - E_i)/kT}$$

$$p_0 = n_i = N_V e^{-(E_i - E_V)/kT}$$

(2.73)

we can write

$$N_C = n_i e^{(E_C - E_i)/kT}$$

$$N_V = n_i e^{(E_i - E_V)/kT}$$

(2.74)

and (2.59) and (2.60) become

$$\begin{aligned} n_0 &= n_i e^{(E_f - E_i)/kT} \\ p_0 &= n_i e^{(E_i - E_f)/kT} \end{aligned} \quad (2.75)$$

We emphasize that the preceding equations are valid only for nondegenerate semiconductors (E_f more than $2.3kT$ away from either band edge).

When the material has donors or acceptors, they contribute to the overall concentration of electrons and holes. For example, if a given semiconductor is doped with N_D donors per cubic centimeter, more electrons are added to the conduction band. We can show that this has the effect of raising the Fermi level. Rearranging Equation (2.75) gives

$$E_f - E_i = kT \ln\left(\frac{n_0}{n_i}\right) \quad (2.76)$$

As n_0 increases, so does $E_f - E_i$.

When a sample is doped with N_D donors per cubic centimeter, it is not necessarily the case that $n_0 = N_D + n_i$. This is because when the donors are ionized, and their electrons are given to the conduction band, some of those electrons will fall to the valence band and recombine with holes. This is illustrated in Figure 2.14. This also implies that the number of holes in n-type material is something less than the intrinsic value p_i . Typically, however, the doping concentration is much greater than the intrinsic concentration. This means that the number of holes available for recombination is small and *most* of the donated electrons remain in the conduction band. Thus for an n-type semiconductor, as long as $N_D \gg n_i$, we can make the approximation that

$$n_0 \approx N_D \quad \text{when } N_D \gg n_i \quad (2.77)$$

Similarly, if a p-type sample is doped with N_A acceptors, some electrons from the conduction band may fall into some of the acceptor states, but if $N_A \gg n_i$, the effect on the concentration of holes is small and

$$p_0 \approx N_A \quad \text{when } N_A \gg n_i \quad (2.78)$$

Whether the material is doped n type or p type, however, as long as the doping level is not degenerate, it is still true that

$$n_0 p_0 = n_i^2$$

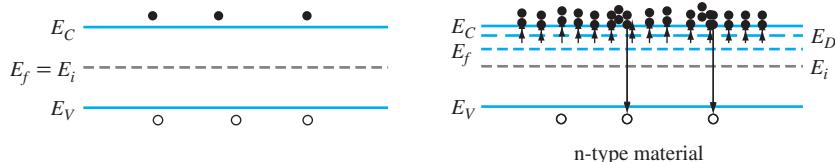


Figure 2.14 Electrons and holes in intrinsic and doped n-type material. In the doped case, some of the electrons from the donors recombine with holes.

EXAMPLE 2.5

Find the electron and hole concentrations in GaAs doped with $N_A = 10^{16} \text{ cm}^{-3}$, and locate the Fermi level. The band gap of GaAs is 1.43 eV.

Solution

We first find n_i for GaAs from Equation (2.67), for which we need to know the effective densities of states N_C and N_V . Assuming room temperature, and using the values of the density-of-states effective masses for GaAs from Tables 2.1 and 2.2, we have from Equations (2.63) and (2.64)

$$\begin{aligned} N_C &= 2.54 \times 10^{19} \left(\frac{m_{dse}^*}{m_0} \right)^{3/2} \left(\frac{T}{300} \right)^{3/2} \text{ cm}^{-3} \\ &= 2.54 \times 10^{19} \left(\frac{(0.067m_0)}{m_0} \right)^{3/2} (1)^{3/2} \\ &= 4.4 \times 10^{17} \text{ cm}^{-3} \end{aligned}$$

Similarly,

$$\begin{aligned} N_V &= 2.54 \times 10^{19} \left(\frac{m_{dsh}^*}{m_0} \right)^{3/2} \left(\frac{T}{300} \right)^{3/2} \text{ cm}^{-3} \\ &= 2.54 \times 10^{19} (0.48)^{3/2} \\ &= 8.4 \times 10^{18} \text{ cm}^{-3} \end{aligned}$$

Therefore, from Equation (2.67),

$$\begin{aligned} n_i &= \sqrt{N_C N_V} e^{-(E_g/2kT)} \\ &= \sqrt{(4.4 \times 10^{17})(8.4 \times 10^{18})} e^{-[1.43/2(0.026)]} \\ &= 2.2 \times 10^6 \text{ cm}^{-3} \end{aligned}$$

Since the doping concentration N_A is significantly larger than n_i , we can assume $p_0 \approx N_A$:

$$p_0 = N_A = 1 \times 10^{16} \text{ cm}^{-3}$$

We can locate the Fermi level by rearranging Equation (2.60) to give

$$E_f - E_V = -kT \ln \left(\frac{p_0}{N_V} \right) = -0.026 \text{ eV} \cdot \ln \left(\frac{10^{16}}{8.4 \times 10^{18}} \right) = 0.17 \text{ eV}$$

Note that we must express the Fermi level relative to some reference position; in this case we have chosen the valence band edge.

The energy band diagram is drawn in Figure 2.15. This tells us that the Fermi level is below the intrinsic level, consistent with our knowledge that the material is p type, having more holes than electrons.

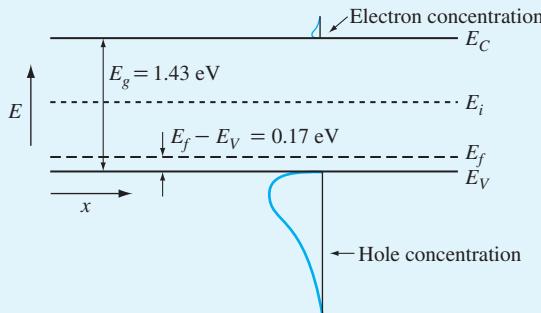


Figure 2.15 Energy band diagram for the p-type GaAs of Example 2.5.

We indicated that these equations apply only when the material is nondegenerate, so we should check that. The material is nondegenerate if the Fermi level is at least $2.3kT$ above the valence band edge. Since $2.3kT = 0.06 \text{ eV}$, and $E_f - E_V = 0.17 \text{ eV}$, the Fermi level is sufficiently removed from the valence band edge that the semiconductor is nondegenerate.

Finally, we can find the concentration of electrons from Equation (2.66):

$$n_0 = \frac{n_i^2}{p_0} = \frac{(2.2 \times 10^6)^2}{10^{16}} = 4.8 \times 10^{-4} \text{ cm}^{-3}$$

It is common to have materials that contain both acceptors and donors. For example, a diode consists of a p-type region and an n-type region that abut. To fabricate these, one might start with a p-type material and add extra donors to part of it. This is called *compensation*. If the number of donors exceeds the number of acceptors, the material becomes n type. One has to include the effects of both dopant types in computing the numbers of electrons and holes. For example, when both donors and acceptors are present, some of the electrons from donor states will fall into acceptor states, ionizing both dopants but not contributing any charge carriers. Thus, assuming all donors and acceptors are ionized,

$$\text{If } N_D > N_A, \text{ then } n_0 = N_D - N_A \text{ and the material is n type} \quad (2.79)$$

and

$$\text{If } N_A > N_D, \text{ then } p_0 = N_A - N_D \text{ and the material is p type} \quad (2.80)$$

Both of these equations apply only when $(N_D - N_A)$ or $(N_A - N_D) \gg n_i$.

EXAMPLE 2.6

A sample of silicon is doped everywhere with a background concentration of $N_A = 4 \times 10^{16} \text{ cm}^{-3}$. Then 10^{17} cm^{-3} donors are added. Find the room temperature concentrations of electrons and holes in the original and final materials and draw the energy band diagram for each.

Solution

Let us begin with the p-type material. We have $N_A = 4 \times 10^{16} \text{ cm}^{-3}$, and we found earlier that for silicon $n_i = 1.08 \times 10^{10} \text{ cm}^{-3}$. We have therefore satisfied the condition that $N_A \gg n_i$, and so the hole concentration in the p-type material is $p_0 = N_A = 4 \times 10^{16} \text{ cm}^{-3}$. We can solve Equation (2.60) to find the Fermi level:

$$\begin{aligned} E_f - E_V &= -kT \ln\left(\frac{p_0}{N_V}\right) = -0.026 \text{ eV} \cdot \ln\left(\frac{4 \times 10^{16}}{3.1 \times 10^{19}}\right) \\ &= 0.17 \text{ eV} \end{aligned}$$

This is greater than $2.3kT = 0.06 \text{ eV}$, so the material is nondegenerate and our use of Equation (2.60) is valid.

The concentration of the electrons in the p-type material is

$$n_0 = \frac{n_i^2}{p_0} = \frac{(1.08 \times 10^{10})^2}{4 \times 10^{16}} = 2.9 \times 10^3 \text{ cm}^{-3} \quad \text{p type}$$

For the n-type material, there are both donors and acceptors. Again assuming room temperature and nondegeneracy, we can write

$$n_0 = N_D - N_A = 10^{17} - 4 \times 10^{16} = 6 \times 10^{16} \text{ cm}^{-3}$$

From Equation (2.59) the Fermi level is located at

$$\begin{aligned} E_C - E_f &= -kT \ln\left(\frac{n_0}{N_C}\right) = -0.026 \text{ eV} \cdot \ln\left(\frac{6 \times 10^{16}}{2.86 \times 10^{19}}\right) \\ &= 0.16 \text{ eV} \end{aligned}$$

and the material is thus nondegenerate. The minority carrier concentration is then found from $n_0 p_0 = n_i^2$, giving

$$p_0 = \frac{n_i^2}{n_0} = \frac{(1.08 \times 10^{10})^2}{6 \times 10^{16}} = 1.9 \times 10^{-3} \text{ cm}^{-3}$$

The energy band diagrams for these two materials are shown in Figure 2.16. The p-type material is shown in part (a) and the n-type material is in part (b). Notice the positions of the Fermi level in each case and the relative sizes of the carrier distributions (for clarity, not shown to scale). Also note that in the p-type material, the acceptor energy level E_A is shown, but there are no donors. In the n-type material, both E_A and E_D exist.

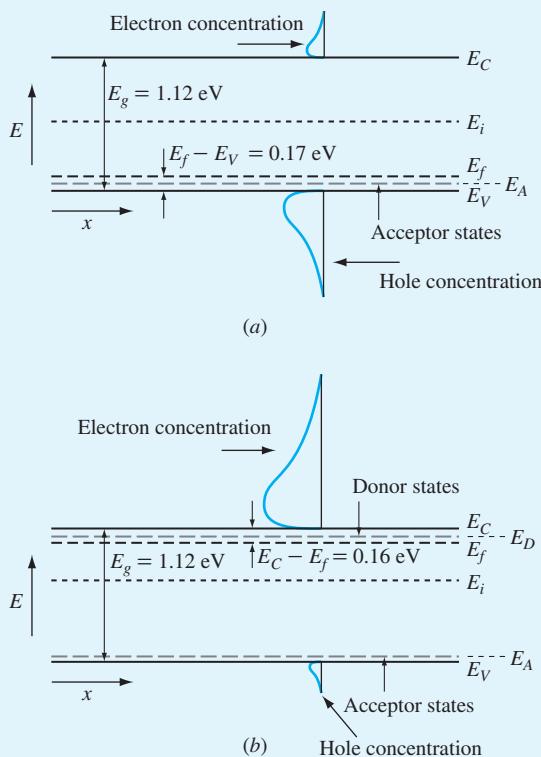


Figure 2.16 The energy band diagrams for Example 2.6: (a) uncompensated p-type material; (b) compensating donors are added to make the material net n type.

EXAMPLE 2.7

In Example 2.6, for the case of a semiconductor with both donors and acceptors and $N_D > N_A$, we used the relation that $n_0 = N_D - N_A$. We wish to justify this. To do this we use the condition that the crystal is electrically neutral.

For the case of $N_D > N_A$ and electrical neutrality, the charge density ρ is zero. Thus the sum of the negative charges (n_0 and the ionized acceptors N_A) must be equal to the sum of the positive charges (p_0 and N_D).

$$\rho = q(p_0 - n_0 + N_D - N_A) = 0$$

Since $n_0 p_0 = n_i^2$, we have $n_0 = \frac{n_i^2}{p_0}$

$$\rho = q \left(\frac{n_i^2}{n_0} - n_0 + N_D - N_A \right) = 0$$

Solving for n_0 ,

$$n_0^2 = n(N_D - N_A) - n_i^2$$

and we obtain

$$n_0 = \frac{1}{2} [N_D - N_A + \sqrt{(N_D - N_A)^2 + 4n_i^2}]$$

As an example, let $N_D = 10^{17} \text{ cm}^{-3}$ and $N_A = 10^{16} \text{ cm}^{-3}$. Then $(N_D - N_A) = 9 \times 10^{16} \text{ cm}^{-3}$.

Since $(N_D - N_A)^2 \gg 4n_i^2$ ($n_i = 1.08 \times 10^{10} \text{ cm}^{-3}$), to good approximation

$$n_0 = N_D - N_A = 9 \times 10^{16} \text{ cm}^{-3}$$

and

$$p_0 = \frac{n_i^2}{n_0} = \frac{(1.08 \times 10^{10})^2}{9 \times 10^{16}} = 1.3 \times 10^3 \text{ cm}^{-3}$$

For practical cases, to good approximation for $N_D > N_A$, $n_0 = N_D - N_A$ and for $N_A > N_D$, $p_0 = N_A - N_D$.

Now let us go back and examine more closely the expression for the electron density distribution function with energy $n(E)$ for electrons in the conduction band. From Equation (2.46), we know that

$$n(E) = S(E)f(E)$$

where, from Equation (2.43),

$$S(E) = \frac{1}{2\pi^2} \left(\frac{2m_{dse}^*}{\hbar^2} \right)^{3/2} \sqrt{E - E_C}$$

We also recall that for a nondegenerate semiconductor, the probability of occupancy $f(E)$ of a state of energy E can be expressed as

$$f(E) = e^{-[(E - E_f)/kT]} \quad (2.81)$$

But, since $E - E_f = (E - E_C) + (E_C - E_f) = E_K + (E_C - E_f)$, we can write

$$n(E) = \frac{1}{2\pi^2} \left(\frac{2m_{dse}^*}{\hbar^2} \right)^{3/2} e^{-[(E_C - E_f)/kT]} \sqrt{E_K} e^{-E_K/kT} \quad (2.82)$$

Where $E_K = (E - E_C)$ is the kinetic energy of the electron at energy E .

By letting a constant C be defined as

$$C = \frac{1}{2\pi^2} \left(\frac{2m_{dse}^*}{\hbar^2} \right)^{3/2} e^{-[(E_C - E_f)/kT]} \quad (2.83)$$

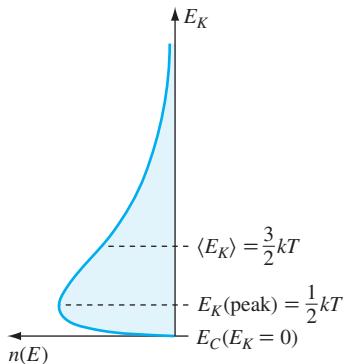


Figure 2.17 The electron distribution function $n(E)$ as a function of energy (energy on the vertical axis).

we can simplify this to

$$n(E_K) = C\sqrt{E_K}e^{-E_K/kT} \quad (2.84)$$

This function was shown schematically in Figure 2.13, and it is shown enlarged in Figure 2.17. We see that with increasing E_K , the electron concentration $n(E_K)$ increases, reaches a maximum, and then decreases approximately exponentially.

EXAMPLE 2.8

- Find the kinetic energy at which the peak electron concentration occurs, E_K (peak).
- Find the average kinetic energy E_K for an electron in the conduction band.
- Find the wavelength of an electron in the conduction band of GaAs having this average kinetic energy.

Solution

- Since $n(E_K)$ reaches a peak at $E_K(\text{peak})$, we know that the derivative is zero. From Equation (2.84),

$$\frac{dn(E_K)}{dE_K} = 0 = C \left[\sqrt{E_K} \left(-\frac{1}{kT} e^{-E_K/kT} + e^{-E_K/kT} \left(\frac{1}{2\sqrt{E_K}} \right) \right) \right]$$

which gives the result

$$n(E_K(\text{peak})) = \frac{1}{2}kT$$

or, in other words, the kinetic energy for the maximum value of $n(E) = \frac{1}{2}kT$ above the conduction band edge. At room temperature this is $\frac{1}{2}(0.026)$ eV = 0.013 eV.

- b. We can find the average value of the kinetic energy from $n(E_K)$, the energy distribution function for electrons in the conduction band,

$$\langle E_K \rangle = \frac{\int_0^{\infty} E_K n(E_K) dE_K}{\int_0^{\infty} n(E_K) dE_K}$$

Substituting for $n(E_K)$ from Equation (2.84) gives

$$\langle E_K \rangle = \frac{\int_0^{\infty} E_K C E_K^{1/2} e^{-E_K/kT} dE_K}{\int_0^{\infty} C E_K^{1/2} e^{-E_K/kT} dE_K} = \frac{\int_0^{\infty} E_K^{3/2} e^{-E_K/kT} dE_K}{\int_0^{\infty} E_K^{1/2} e^{-E_K/kT} dE_K}$$

Evaluating the integrals we get

$$\langle E_K \rangle = \frac{3}{2} kT$$

This is a familiar number from chemistry and physics classes. The key points here are:

- For E_K greater than about $3kT$, $S(E)$ varies slowly compared with $f(E)$.
- To good approximation, we can consider $n(E_K)$ to decrease exponentially with E_K .

- c. To find the wavelength, from Chapter 1 we know that the wave vector

$$K = \sqrt{\frac{2m^*}{\hbar^2} E_K} = \frac{2\pi}{\lambda}$$

This implies that the wavelength of an electron with the average kinetic energy of $\frac{3}{2}kT$ is

$$\lambda = \frac{2\pi}{K} = \frac{2\pi}{\sqrt{\frac{2m^*}{\hbar^2} E_K}} = \frac{2\pi\hbar}{\sqrt{2m^* \left(\frac{3}{2}kT\right)}}$$

Since for electrons in GaAs, $m^* = m_{dse}^* = 0.067m_0$,

$$\lambda = \frac{2\pi(1.055 \times 10^{-34} \text{J} \cdot \text{s})}{\sqrt{2(0.067)(9.11 \times 10^{-31} \text{kg}) \left[\frac{3}{2}(1.38 \times 10^{-23} \text{J/K})(300 \text{K}) \right]}} = 24 \text{ nm}$$

The electron wavelength is on the order of 40 lattice constants. We confirm again that the electron is spread out over the crystal, as indicated earlier.

2.12 TEMPERATURE DEPENDENCE OF CARRIER CONCENTRATIONS IN NONDEGENERATE SEMICONDUCTORS

In discussing the equilibrium carrier concentrations up to this point, we assumed that the semiconductor was at or near room temperature (300 K). Therefore, the band gap of Si was assumed to be constant (1.12 eV) and the intrinsic carrier concentration is $1.08 \times 10^{10} \text{ cm}^{-3}$ (300 K). It was further assumed that $N_D \gg n_i$

or $N_A \gg n_i$ and that all donors and acceptors were ionized such that at equilibrium the majority carrier concentration is

$$n_0 = N_D - N_A \quad N_D > N_A \quad \text{n type}$$

$$p_0 = N_A - N_D \quad N_A > N_D \quad \text{p type}$$

The minority carrier concentration was then determined from the relation

$$n_0 p_0 = n_i^2 \quad (2.66)$$

These assumptions, however, must be modified at high temperatures, where thermally generated electron-hole pairs contribute to the carrier concentration, and at low temperatures, where the impurities are not completely ionized.

2.12.1 CARRIER CONCENTRATIONS AT HIGH TEMPERATURES

In Section 2.11 we assumed that all of the donors had given up their extra electrons to the conduction band and that all the acceptors had likewise become occupied by electrons from the valence band. In the Supplement to Part 1, it is shown that in Si these assumptions are valid above about 200 K. The concentration of electrons still bound to donors or holes still bound to acceptors is small compared with the impurity concentration and thus, to good approximation, all impurities can be considered to be ionized. Further, since for semiconductors used in devices the impurity concentration is normally much larger than n_i , the contribution of thermally generated majority carriers can be neglected. At high enough temperatures, however, n_i is comparable to the net doping level and must be considered.

The intrinsic carrier concentration n_i can be expressed [Equation (2.67)] as

$$n_i = \sqrt{N_C N_V} e^{-E_g/2kT} \quad (2.85)$$

and so should be highly temperature dependent. The temperature dependence of N_C , N_V , and E_g must be considered when calculating $n_i(T)$. The temperature dependence of N_C and N_V was discussed in Section 2.11 (equations repeated here).

$$N_C = 2.54 \times 10^{19} \left(\frac{m_{dse}^*}{m_0} \right)^{3/2} \left(\frac{T}{300} \right)^{3/2} \text{ cm}^{-3} \quad (2.63)$$

$$N_V = 2.54 \times 10^{19} \left(\frac{m_{dsh}^*}{m_0} \right)^{3/2} \left(\frac{T}{300} \right)^{3/2} \text{ cm}^{-3} \quad (2.64)$$

The temperature dependence of E_g for Si can be expressed

$$E_g(T) = 1.170 - \frac{4.73 \times 10^{-4} T^2}{T + 636} \text{ eV} \quad (2.86)$$

This is plotted in Figure 2.18 (solid line).

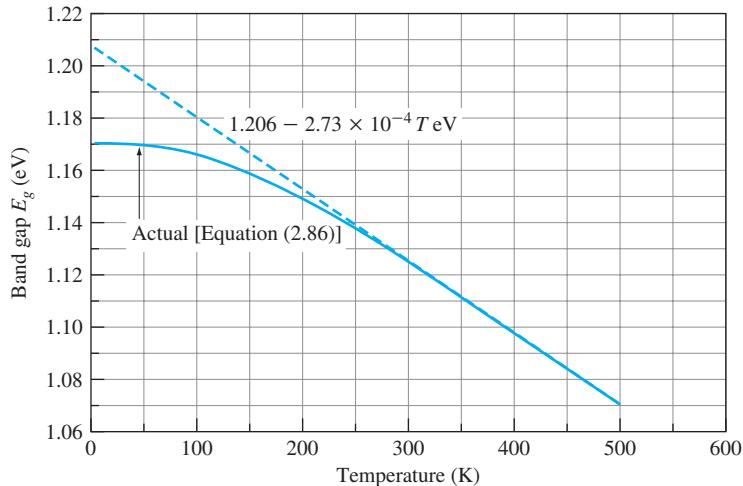


Figure 2.18 Energy band-gap dependence of silicon on temperature.

The dashed line in Figure 2.18 shows that above 250 K, E_g can be approximated by the simpler straight-line expression

$$E_g = 1.206 - 2.73 \times 10^{-4} T \text{ eV} \quad (T > 250 \text{ K}) \quad (2.87)$$

Next, let us discuss the intrinsic carrier concentration. For silicon, n_i can be expressed as [1]

$$n_i = 5.71 \times 10^{19} \left(\frac{T}{300} \right)^{2.365} e^{-(6733/T)} \text{ cm}^{-3} \quad (2.88)$$

At high enough temperatures, thermally generated carriers can contribute to the concentration of majority carriers. To find out at what temperatures this effect becomes important, we find the electron and hole concentrations employing the technique of space charge neutrality. That is, the number of positive charges must be equal to the number of negative charges in any macroscopic region. Electrons and ionized acceptors are negatively charged, while holes and ionized donors are positively charged. Thus for a uniformly doped semiconductor, at equilibrium, for space charge neutrality and assuming all impurities are ionized,

$$p_0 + N_D = n_0 + N_A \quad (2.89)$$

or

$$p_0 - n_0 + N_D - N_A = 0 \quad \text{space-charge neutrality} \quad (2.90)$$

For nondegeneracy such that

$$n_0 p_0 = n_i^2$$

Equation (2.90) can be expressed

$$\frac{n_i^2}{n_0} - n_0 + N_D - N_A = 0 \quad (2.91)$$

or

$$n_0^2 - n_0(N_D - N_A) - n_i^2 = 0 \quad (2.92)$$

Solving for n_0 ,

$$n_0 = \frac{N_D - N_A}{2} + \left[\left(\frac{N_D - N_A}{2} \right)^2 + n_i^2 \right]^{1/2} \quad (2.93)$$

The hole concentration can then be found from Equation (2.66):

$$p_0 = \frac{n_i^2}{n_0}$$

Note that for $(N_D - N_A) \gg n_i$ (the case assumed in Section 2.11, where high-temperature effects are not significant), Equation (2.93) reduces to our previous result of

$$n_0 = N_D - N_A \quad (N_D - N_A) \gg n_i \quad (2.94)$$

Figure 2.19 shows the variation of n_0 with temperature, with net donor doping as a parameter, above 200 K, where complete ionization is a good approximation. It can be seen that for n_i greater than about 20 percent of $(N_D - N_A)$, n_0 increases appreciably with increasing temperature.

Similarly, for a nondegenerate p-type semiconductor,

$$p_0 = \frac{N_A - N_D}{2} + \left[\left(\frac{N_A - N_D}{2} \right)^2 + n_i^2 \right]^{1/2} \quad (2.95)$$

and $n_0 = n_i^2/p_0$.

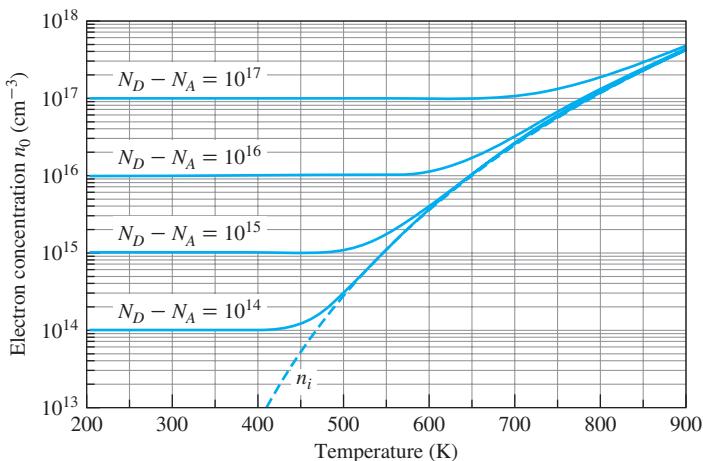


Figure 2.19 Plot of electron concentration n_0 as a function of temperature in n-type silicon for four values of net doping. Also indicated is the temperature dependence of n_i .

EXAMPLE 2.9

For $N_D \gg N_A$, find N_D such that n_0 is 10 percent greater than N_D .

Solution

From Equation (2.93), neglecting N_A since it is small,

$$n_0 = 1.1N_D = \frac{N_D}{2} + \left[\left(\frac{N_D}{2} \right)^2 + n_i^2 \right]^{1/2}$$

$$N_D(1.1 - 0.5) = \left[\frac{N_D^2}{4} + n_i^2 \right]^{1/2}$$

$$0.6N_D = N_D \left[\frac{1}{4} + \frac{n_i^2}{N_D^2} \right]^{1/2}$$

$$(0.6)^2 = 0.36 = \frac{1}{4} + \frac{n_i^2}{N_D^2}$$

$$N_D = \frac{n_i}{\sqrt{0.11}} = 3.0n_i$$

From Figure 2.19, we see that n_0 depends on N_D as well as temperature. It is convenient to have a normalized plot in which N_D is incorporated. Assuming $N_D \gg N_A$, from Equation (2.93)

$$n_0 = \frac{N_D}{2} + \left[\left(\frac{N_D}{2} \right)^2 + n_i^2 \right]^{1/2} \quad (2.96)$$

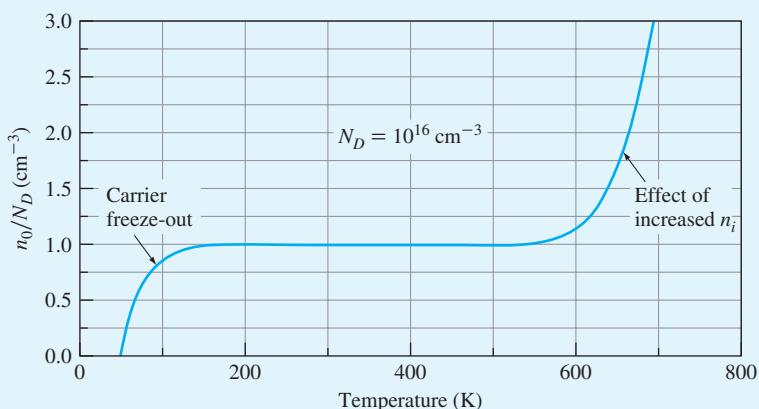


Figure 2.20 Normalized plot of $\frac{n_0}{N_D}$ as a function of temperature. This plot is for $N_D = 10^{16} \text{ cm}^{-3}$.

which can be expressed as

$$\frac{n_0}{N_D} = \frac{1}{2} \left\{ 1 + \left[1 + \left(\frac{2n_i}{N_D} \right)^2 \right]^{1/2} \right\} \quad (2.97)$$

For the temperature dependence of n_i given by Equation (2.88), n_0/N_D is plotted as a function of temperature in Figure 2.20 on a linear scale for $N_D = 10^{16} \text{ cm}^{-3}$. As discussed next, at low temperatures $n_0 < N_D$, since some electrons are still bound to donor atoms in this case. This is also shown in Figure 2.20.

2.12.2 CARRIER CONCENTRATIONS AT LOW TEMPERATURES (CARRIER FREEZE-OUT)

Electrons bound to donor atoms or holes bound to acceptor atoms are said to be “frozen out.” Carrier freeze-out, which can be significant at low temperatures, will affect the carrier concentrations since the dopants are not all ionized under these conditions. The equilibrium concentration of carriers frozen out is discussed in Online Module OM3. Here, we point out that, to find the equilibrium carrier concentrations at low temperatures (below about 200 K), we find the number of dopants that are not ionized. Each donor that is not ionized contains an electron that would otherwise be in the conduction band, so the concentration of “missing” electrons n_D is equal to the concentration of un-ionized donors. Similarly there are p_A holes still attached to the acceptors. The space charge neutrality equation [Equation (2.90)] becomes

$$p_0 + p_A - n_0 - n_D + N_D - N_A = 0 \quad (2.98)$$

The result for a doping concentration of $N_D = 10^{16} \text{ cm}^{-3}$ of phosphorus in silicon is shown at the low-temperature end of Figure 2.20.

We point out that, with increased doping, a decreasing fraction of the carriers will be frozen out. In Si doped greater than about $4 \times 10^{18} \text{ cm}^{-3}$ (degenerately doped), the ground states of the dopant atoms overlap. In this case there are no bound states, and no carrier freeze-out occurs.

2.13 DEGENERATE SEMICONDUCTORS

Up to now we have discussed the properties of semiconductors with small doping levels (i.e., nondegenerate semiconductors). The values for the semiconductor band gaps we have discussed are for intrinsic materials at equilibrium. It was assumed that this value of E_g is independent of doping levels for nondegenerate semiconductors. For most electron devices, however, there are at least some regions that are degenerately doped. The high doping levels change the electronic properties of the materials enough to significantly affect some device characteristics. A primary effect is known as *impurity-induced band-gap narrowing*.

We will discuss this effect first, then discuss a related effect, impurity-induced *apparent band-gap narrowing*, and then examine the effect on equilibrium carrier concentrations. As before, we will discuss the band gap at equilibrium.⁵ The dependence of E_g on doping concentration has a minor effect on the electrical characteristics of semiconductor diodes and field-effect transistors, but it has a major effect on the electrical properties of bipolar junction transistors, as discussed in Chapter 9.

2.13.1 IMPURITY-INDUCED BAND-GAP NARROWING

In Chapter 1, we showed that as the atoms in a crystal get closer together, their electronic orbits overlap in space, and the discrete energy levels of the individual atoms spread out in energy and form bands. A similar situation exists for impurities in semiconductors—the discrete energy levels of the dopants can smear into *minibands*. This happens when the doping concentration is high enough that the dopant atoms are near enough to each other that they can influence one another. This is reflected in the energy band diagram of Figure 2.21. On the left is a non-degenerately doped n-type semiconductor with no acceptors. The donor atoms are sparse enough that they act as isolated atoms in a sea of silicon. Under heavy enough doping, shown on the right, the discrete donor states spread into a band.

The dominant effect behind the minibands is shown schematically in Figure 2.22. Here we use the Bohr model and consider two donor atoms that are close enough together that their first excited states overlap. If the first excited states overlap, then so do all of the higher-energy impurity states. All excited states overlap in energy with each other, and they also overlap with the conduction band of the host silicon. This implies that the bottom of the conduction band, E_C , is actually lowered from its intrinsic value E_{C0} to the bottom of the overlapping impurity bands, as shown in Figure 2.21. The result is that the band gap is reduced. [4]

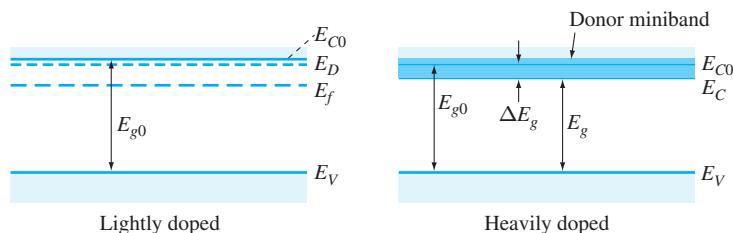


Figure 2.21 Under high doping concentrations, the formerly discrete donor levels smear into a band, effectively narrowing the band gap by an amount ΔE_g .

⁵In some nonequilibrium cases, the electrostatic energy associated with electron-hole interactions can further reduce the band gap. [3]

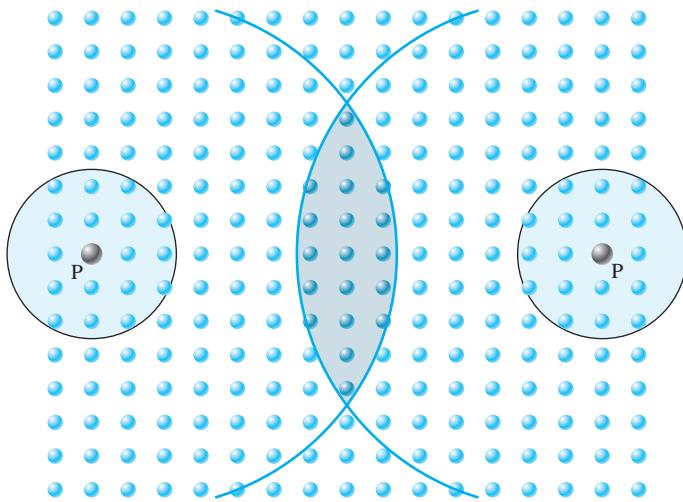


Figure 2.22 The states for the higher donor levels can overlap if the doping concentration is high enough (dopant atoms close enough together).

As the donor concentration increases, the shift in the conduction band, ΔE_C , also increases, where

$$\Delta E_C = E_{C0} - E_C \quad (2.99)$$

This increase in ΔE_C with increasing doping N_D is largely due to an increase in the number of spatially overlapping equi-energy impurity states and also from the broadening in energy of these impurity bands. At a sufficiently high doping level, even the ground states overlap. At that point, ΔE_C continues to increase, but now the downward shift of the conduction band edge results primarily from the broadening of the ground state band. For Si the ground states of the donors overlap at an impurity concentration of about $4 \times 10^{18} \text{ cm}^{-3}$. This concentration is high compared with some other materials. For semiconductors with small electron conductivity effective mass, the ground states overlap at much smaller doping levels. For example, GaAs with $m^* = m_{ee}^* = 0.067m_0$, the donor ground states overlap at $N_D \approx 2 \times 10^{16} \text{ cm}^{-3}$.

Normally, for device considerations, it is more useful to talk in terms of the reduction in the band gap, rather than that of the conduction band edge. We thus define the *impurity-induced band-gap reduction* ΔE_g , where for n-type material,

$$\Delta E_g = E_{g0} - E_g(N_D) \quad \text{n-type} \quad (2.100)$$

where E_{g0} is the band gap of intrinsic material and $E_g(N_D)$ is the band gap for a donor concentration of N_D . The band-gap narrowing for n-type Si has been fit to the empirical expression [5]

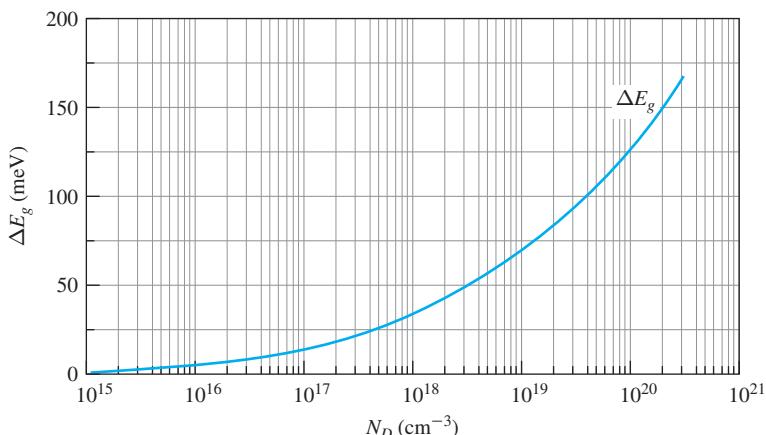


Figure 2.23 Reduction of room-temperature band gap ΔE_g as a function of donor density in phosphorus-doped silicon.

$$\Delta E_g = [(4.372 \times 10^{-11} N_D^{0.5})^{-4} + (1.272 \times 10^{-6} N_D^{0.25})^{-4}]^{-1/4} \text{ eV} \quad (2.101)$$

where N_D is the donor concentration in cm^{-3} . This expression is plotted in Figure 2.23.

From Figure 2.23 we see that the band-gap reduction is negligible for $N_D < 10^{16} \text{ cm}^{-3}$ and increases to 128 meV at $N_D = 10^{20} \text{ cm}^{-3}$. For $N_D \ll N_C$, the concentration of states in this impurity band is small and can be neglected. Then for N_D less than about 10^{17} cm^{-3} , the electron concentration in the impurity band can be neglected, and to good approximation

$$n_0 = N_C e^{-(E_{C0} - E_f)/kT}$$

2.13.2 APPARENT BAND-GAP NARROWING

We have seen how in n-type material the conduction band edge, and thus the band gap, decrease with increased donor doping. As the doping increases, however, the Fermi level moves closer to the band edge. At some doping level the Fermi level will actually cross into the altered conduction band.

In this case, it is usually the energy of the Fermi level that is of more interest than that of the conduction band edge. The energy of the Fermi level is, however, difficult to calculate at high doping levels. Many of the excited donor states (or minibands) are occupied, and their energy dependencies are not well understood.

Figure 2.24 indicates the energy band diagram for a nondegenerate and a degenerate semiconductor. The band gap of the nondegenerate material is E_{g0} , and that of the impurity-induced band gap is E_g , while the apparent band gap is E_g^* . The nondegenerate conduction band edge is E_{C0} , the impurity-induced

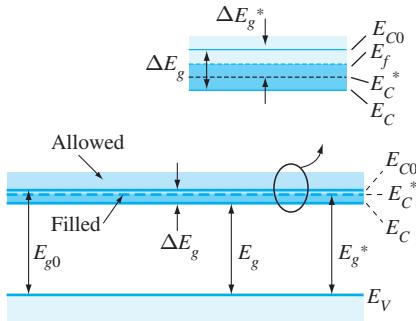


Figure 2.24 Energy band diagram of degenerate n-type semiconductor compared to that of a nondegenerate semiconductor.

conduction band edge is E_C , and the apparent conduction band edge is E_C^* . In general, the location of E_f is not precisely known, and the apparent band gap narrowing is determined from device measurements.

There is more than one way to define apparent band-gap narrowing. A common definition involves adapting the equilibrium electron-hole product from the value used for a nondegenerate semiconductor. Adapting Equation (2.65), repeated here

$$n_0 p_0 = N_C N_V e^{-\frac{E_{g0}}{kT}} = n_t^2 \quad (2.65)$$

to

$$n_0 p_0 = N_C N_V e^{-\frac{(E_{g0} - \Delta E_g^*)}{kT}} = n_t^2 e^{\frac{\Delta E_g^*}{kT}} \quad (2.102)$$

This definition of ΔE_g^* will be important later for describing the variation of current gain (β) in bipolar junction transistors, where the term ΔE_g^* takes into account all of the heavy doping effects that influence β .⁶ It is difficult to determine ΔE_g^* experimentally, and there is considerable scatter in the reported data. One relation reported of ΔE_g^* on doping level is [6]

$$\Delta E_g^* = 6.92 \left[\ln \frac{N}{1.3 \times 10^{17}} + \sqrt{\left(\ln \frac{N}{1.3 \times 10^{17}} \right)^2 + 0.5} \right] \text{ meV} \quad (2.103)$$

This relation is plotted in Figure 2.25 as a function of doping. Equation (2.103) is valid for both n-type and p-type silicon where N represents N_D in n-type silicon and N_A in p-type silicon.

⁶A more intuitive definition might be the difference in energy between the valence band edge and the Fermi level, and indeed that definition is useful in spectroscopy. For bipolar transistors, however, the definition used here is more useful.

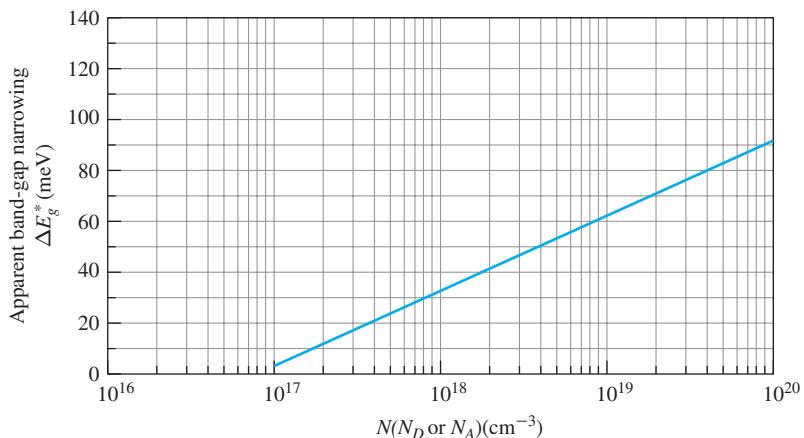


Figure 2.25 Plot of apparent band-gap narrowing for degenerate n-type and p-type silicon.

EXAMPLE 2.10

A sample of room-temperature silicon is doped with $N_D = 4 \times 10^{19} \text{ cm}^{-3}$ and $N_A = 10^{15} \text{ cm}^{-3}$ acceptors. Find n_0 , and draw the complete energy band diagram.

Solution

At room temperature we can assume all the donors are ionized, and since $N_A \ll N_D$, we can neglect the acceptors and take $n_0 \approx N_D$ or $4 \times 10^{19} \text{ cm}^{-3}$. From Figure 2.23 [or Equation (2.101)], we see that the conduction band edge has been lowered by about 100 meV. Since $N_A = 10^{15}$, we can neglect any band-gap narrowing due to acceptors. Thus the impurity-induced band-gap narrowing is 100 meV.

The apparent band-gap narrowing for this degenerately doped material is, from Figure 2.25 [or Equation (2.103)], about 80 meV. The resulting energy band diagram is shown in Figure 2.26.

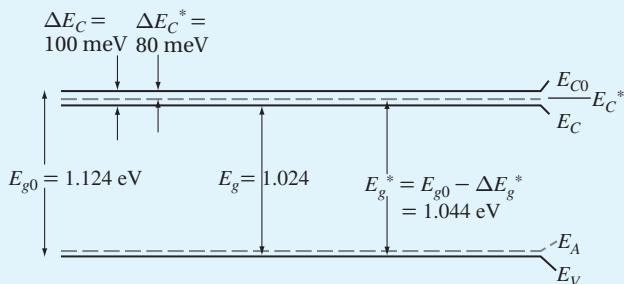


Figure 2.26 Resultant energy band diagram for Example 2.10.

2.14 SUMMARY

In this chapter, we saw that we could define a carrier effective mass in such a way that if it were used in classical mechanical equations (pseudo-classical mechanics), the results would predict what actually happens to the electron or hole in response to an external force. Electrons were found to have negative effective mass in some cases (e.g., near the top of the valence band), but in those cases it makes more sense to talk about holes having positive effective mass and positive charge $+q$. Electrons near the bottom of the conduction band can be treated as particles of charge $-q$ and with a positive effective mass.

Most of the electrons available for conduction are found near the bottom of the conduction band, and similarly most of the holes are found near the top of the valence band. This is fortuitous, since those happen to be the regions where the pseudo-classical mechanics applies.

For a one-dimensional crystal, near the conduction band minimum E_C at $K = 0$, the electron energy can be expressed

$$E = E_C + \left(\frac{1}{2} \frac{d^2 E}{dK^2} \right) K^2$$

where E_C is its potential energy and its kinetic energy is

$$E_K = \frac{1}{2} \frac{d^2 E}{dK^2} K^2$$

The electron (group) velocity is

$$v = v_g = \frac{1}{\hbar} \frac{dE}{dK}$$

and the effective mass m^* is related to the curvature of the E - K curve:

$$m^* = \hbar^2 \left(\frac{d^2 E}{dK^2} \right)^{-1}$$

Analogous expressions hold for holes in the valence band.

For a three-dimensional semiconductor having a single minimum at $K = 0$ (e.g., GaAs, InP), the electron effective mass is a scalar and single-valued. For cases in which the effective mass is not a scalar or single-valued, we differentiate between two different averages for effective mass. One is the density-of-states effective mass m_{ds}^* , used in calculations involving the density-of-states function. The other is the conductivity effective mass m_c^* . For other processes (e.g., tunneling, discussed in the Supplement to Part 1) different effective masses are used. It is important to recognize which effective mass to use in any given calculation.

The values of electron and hole density-of-states effective masses and conductivity effective masses are given in the following table.

Material	m_{dse}^*/m_0	m_{dsh}^*/m_0	m_{ce}^*/m_0	m_{ch}^*/m_0
Si	1.09	1.15	0.26	0.36
GaAs	0.067	0.48	0.067	0.34
Ge	0.56	0.29	0.12	0.21
InP	0.077	0.42	0.077	0.30

We also saw how the electrons and holes are distributed in energy. This result comes from two parameters:

1. The density of states $S(E)$ in each band, which varies parabolically with energy (near the band edges) for electrons in the conduction band and for holes in the valence band, where

$$S(E) = \frac{1}{2\pi^2} \left(\frac{2m_{ds}^*}{\hbar^2} \right) \sqrt{|E_K|}$$

2. The probability $f(E) = 1/(1 + e^{(E-E_f)/kT})$ of occupancy of a given state of energy E , where E_f is the Fermi energy.

The electron concentration distribution function with energy is then

$$n(E) = S(E)f(E)$$

2.14.1 NONDEGENERATE SEMICONDUCTORS

A material is defined as nondegenerate if the Fermi level is inside the forbidden gap and at least $2.3kT$ away from either band edge.

Substitutional impurities can affect the carrier concentrations in semiconductors. Donors have more electrons available than required for bonding, and these electrons are easily excited into the conduction band. Acceptors have too few electrons, and electrons from the valence band are easily excited into these acceptor states, leaving holes in the valence band.

To find the concentrations of carriers in a semiconductor, the steps are:

1. Assume all donors and acceptors are ionized (valid at room temperature as long as the net doping is sufficiently large that $|N_D - N_A| \gg n_i$, the intrinsic concentration). Under these conditions, in n-type material $n_0 = N_D - N_A$, and in p-type material $p_0 = N_A - N_D$. This gives us the majority carrier concentration.
2. To find the minority carrier concentration, if the material is nondegenerate, we can use

$$n_0 p_0 = n_i^2 = N_C N_V e^{-E_g/kT}$$

where n_i is the electron concentration of intrinsic material and

$$N_C = 2.54 \times 10^{19} \left(\frac{m_{dse}^*}{m_0} \right)^{3/2} \left(\frac{T}{300} \right)^{3/2} \text{cm}^{-3}$$

$$N_V = 2.54 \times 10^{19} \left(\frac{m_{dsh}^*}{m_0} \right)^{3/2} \left(\frac{T}{300} \right)^{3/2} \text{cm}^{-3}$$

The equilibrium carrier concentrations are

$$n_0 = N_C e^{-[(E_C - E_f)/kT]}$$

$$p_0 = N_V e^{-[(E_f - E_V)/kT]}$$

or alternatively

$$n_0 = n_i e^{-[(E_f - E_i)/kT]}$$

$$p_0 = n_i e^{-[(E_i - E_f)/kT]}$$

At high temperatures, thermally excited electron-hole pairs contribute to the carrier concentrations. At very low temperatures, the carrier concentrations are reduced as a result of electrons becoming bound to donor atoms or holes to acceptor atoms.

2.14.2 DEGENERATE SEMICONDUCTORS

If the material is sufficiently degenerate, i.e., the Fermi level is within the conduction band (n type) or within the valence band (p type), we use

$$n_0 p_0 = n_i^2 e^{\frac{\Delta E_g^*}{kT}} \quad \text{degenerate n or p type}$$

Where ΔE_g^* is the apparent or effective impurity-induced band-gap shrinkage.

2.15 REFERENCES

1. M. A. Green, “Intrinsic concentration, effective density of states and effective mass in silicon,” *J. Appl. Phys.*, 67, pp. 2944–2954, 1990.
2. J. S. Blakemore, *Semiconductor Statistics*, App. A, Dover Publications, Toronto, 1987.
3. H. P. D. Lanyon and R. A. Tuft, “Bandgap narrowing in heavily doped silicon,” *IEDM Technical Digest*, p. 316, 1978.
4. Yaun Taur and Tak H. Ning, *Fundamentals of Modern VLSI Devices*, 2nd ed., Chap. 2, Cambridge University Press, Cambridge, 2009.
5. S. Sokolic and S. Amon, “Modelling heavy doping effects for low temperature device simulations,” *Journal de Physique IV*, Colloque C6, pp. C6-133–C6-138, 1994.
6. D. B. M. Klassen, J. W. Slotboom, and H. C. deGraaff, “Unified apparent bandgap narrowing in n- and p-type silicon,” *Solid State Electron.*, 35, pp. 125–129, 1992.

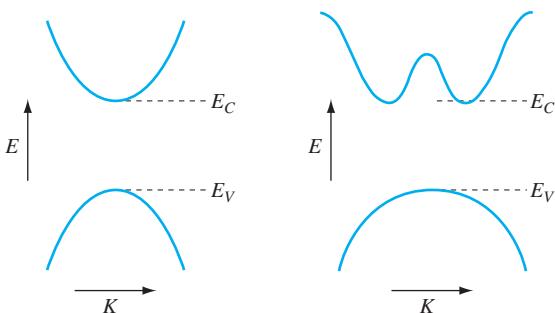
2.16 REVIEW QUESTIONS

1. What is the distinction between an intrinsic and an extrinsic semiconductor?
2.
 - a. Define intrinsic concentration of holes.
 - b. What is the relationship between this density and the intrinsic concentration of electrons?
 - c. What do these equal at 0 K?
3. Define:
 - a. Donor.
 - b. Acceptor.
4. A semiconductor is doped with both donors and acceptors of concentrations N_D and N_A respectively. Write the equation or equations from which to determine the equilibrium electron and hole concentrations n_0 and p_0 .
5. Explain physically the meaning of the following statement: An electron and a hole recombine and disappear.
6. Explain when and why the Boltzmann approximation can be used.
7. Explain what is meant by a distribution function. Use as an example the distribution in age of people in the United States.
8. Plot the Fermi-Dirac distribution function $f(E)$ versus energy E for $T = 0\text{ K}$ and $T = 2500\text{ K}$. What are the meanings of these plots?
9. The electron energy distribution function is given by the product of two factors. What is the interpretation to be given to each of these factors?
10. Sketch the energy band diagrams for (a) an intrinsic, (b) an n-type, and (c) a p-type semiconductor. Indicate the positions of the Fermi level, the conduction band edge, the valence band edge, the donor level, and the acceptor level, and label both axes.
11. Explain what it means for a semiconductor to be degenerate.
12. Under what conditions do discrete donor states expand into minibands?
13. Draw an energy band diagram for a degenerately doped p-type semiconductor. Indicate the band-gap narrowing and the apparent band-gap narrowing.

2.17 PROBLEMS

Assume that all impurities are ionized and unless otherwise specified, assume $T = 300\text{ K}$. Physical constants for some common semiconductors can be found in Appendix A.

- 2.1 Of the two materials whose E - K diagrams are shown in Figure P2.1, which will have the lower effective mass for electrons? Which will have the lower effective mass for holes? Explain how you arrived at your conclusion.

**Figure P2.1**

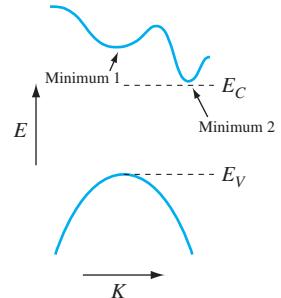
- 2.2** Assume a material has an E - K diagram given by

$$E_{K(\text{Conduction})} = E_C + E_1 \cos^2\left(\frac{Ka}{2}\right)$$

$$E_{K(\text{Valence})} = E_V - E_2 \sin^2\left(\frac{Ka}{2}\right)$$

Let $a = 0.4$ nm, $E_1 = 4$ eV, and $E_2 = 3$ eV.

- a. Sketch the E - K diagram for the first Brillouin zone. Label the axes completely.
 - b. What is the effective mass for an electron near the bottom of the conduction band?
 - c. What is the effective mass for holes near the top of the valence band?
- 2.3** Is it safe to say that the effective mass of an electron in a material is always less than that in vacuum? Is it safe to say that the effective mass of an electron is always less than the effective mass of a hole in the same material?
- 2.4** Consider the hypothetical semiconductor whose E - K diagram is shown in Figure P2.2. Of the two different effective masses for electrons in the conduction band, which is larger, that for electrons in minimum 1 or for electrons in minimum 2? Which effective mass will be exhibited by the largest number of electrons and why?
- 2.5** Consider the energy band diagram in Figure P2.3.
- a. Find the electric field, and express the result in V/cm. In what direction does it point?
 - b. Find the force on the electron. In what direction is the electron accelerated?
 - c. Find the force on the hole. In what direction is the hole accelerated?

**Figure P2.2**

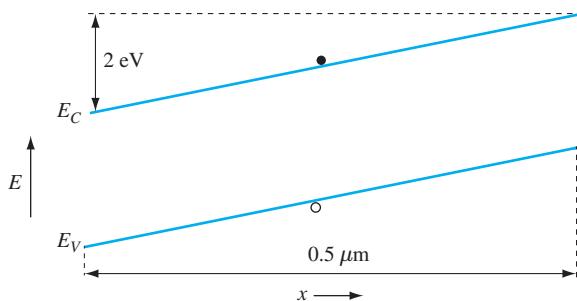


Figure P2.3

- 2.6** Consider the three electrons A, B, and C approaching a change in potential energy (a potential barrier) shown in Figure P2.4. Assume all three are initially moving to the right.
- Which electron initially has the greatest *kinetic energy*?
 - Which electron initially has the smallest *potential energy*?
 - What is the direction of the force at the barrier?
 - Will electron C experience a force greater than, equal to, or less than the force experienced by electron B in the region of the barrier?
 - Which electron(s) will end up with the same kinetic energy as it (they) started with?
 - Which electron will end up with the largest kinetic energy?
 - Which electron will end up with the smallest potential energy?
 - Recalling that $F = -q\mathcal{E}$, where \mathcal{E} is the electric field and q is the magnitude of the electron charge, indicate in which region(s) an electric field exists.
 - In the region(s) you indicated, in which direction does the electric field vector point?

C ●

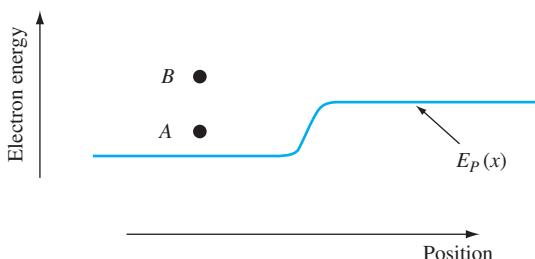


Figure P2.4

- 2.7** For an optical emission to occur, four things are needed:

- An electron must be at an elevated energy state.
- There must be an empty state at a lower energy for the electron to go to.
- Energy must be conserved (energy of electron before emission = energy of electron after emission + energy of photon).
- K must be conserved.

It turns out that photons have negligible K compared with electrons of interest in semiconductor electronics. That implies an electron making an optical transition must end up at the (almost) same value of K as when it started (the energy can change, since the energy difference goes to the photon). On the basis of your expectation of where electrons and holes are likely to be found, and the four conditions above, which material will be a more efficient light emitter, InAs or GaP? Their E - K diagrams are shown in Figure P2.5. Explain your reasoning. Gallium phosphide is an indirect gap material, meaning the conduction band minimum is not at the same value of K as at the maximum in the valence band. Indium arsenide is a direct gap material.

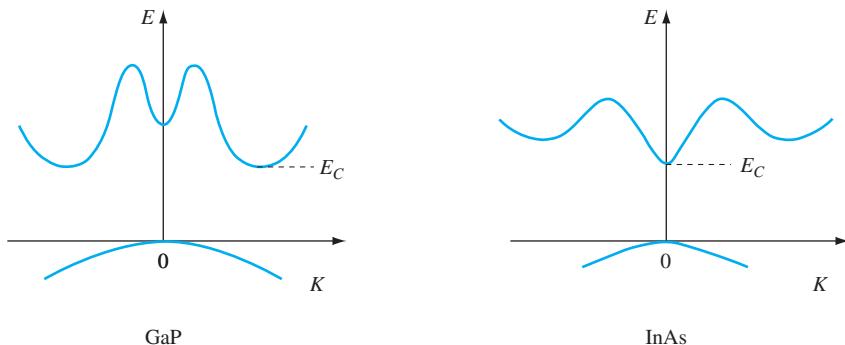


Figure P2.5

- 2.8** Consider the energy band diagram in Figure P2.6.

- What is the potential energy of the electron?
- What is the kinetic energy of the electron?
- What is the potential energy of the hole?

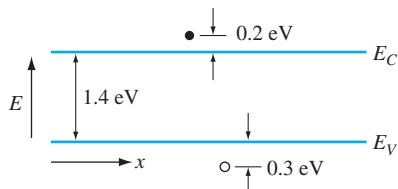


Figure P2.6

- d. What is the kinetic energy of the hole?
- e. The vertical axis is energy—meaning the electron energy. What direction represents increasing hole energy, up or down?
- 2.9** A sample of GaAs is subjected to an electric field of 150 V/cm.
- What is the velocity of an electron in the conduction band if it is initially at rest and is accelerated by this field for a period of 0.2 ps? (Use the conductivity effective mass since this problem is related to conduction.)
 - What is the terminal velocity for a hole under the same conditions?
- 2.10** Near the bottom of the conduction band in a semiconductor, the electron energy can be expressed as $E = E_C + AK^2$, where A is independent of K and is positive.
- Find an expression for the electron effective mass in terms of A .
 - Find an expression for the electron velocity as a function of E .
- 2.11** Consider a semiconductor material whose three-dimensional E - K relationship has the same curvature in the x and z directions but a different curvature in y . Let us consider the x and y directions, and let the relationship be $E(x, y) = AK_x^2 + BK_y^2$.
- Using a computer program such as MATLAB, produce a mesh plot of this function for $B = 2A$. Rotate the perspective to get an understanding of the shape of this function.
 - In which direction is the effective mass smaller?
 - What determines which effective mass a given electron actually has in this crystal?
- 2.12** For each of the following, identify whether the impurity will produce n-type or p-type material or be amphoteric:
- Germanium in zinc selenide.
 - Carbon in gallium phosphide.
 - Zinc in silicon.
 - Phosphorous replacing an aluminum atom in aluminum arsenide.
- 2.13** Show that the difference between the intrinsic level and midgap is small, using InP as an example. Compare the result with that of silicon.
- 2.14** What is the probability that a state at the conduction band edge of intrinsic InP is occupied at room temperature?
- 2.15** The donor ground states for tellurium in GaAs are 5.9 meV below the conduction band. (There are two of them because of spin.) At room temperature, what is the probability that a given ground state is occupied if the Fermi level is 0.030 eV below E_C ?
- 2.16** What is the probability of occupancy of a state 0.1 eV above the Fermi level? Below the Fermi level? Repeat for 0.2 eV above and below E_f .

- 2.17** Verify the definition of nondegeneracy; show that if the Fermi level is $2.3kT$ below the conduction band edge, the probability of occupancy of the lowest energy state in the conduction band is 10 percent.
- 2.18** Calculate the effective densities of states for electrons and holes at room temperature for
- Silicon.
 - GaAs.
 - Ge.
- 2.19** How different are the effective density of states N_C for electrons and the density of states N_V for holes in germanium? Express the result as a ratio.
- 2.20** Show that Equation (2.52) follows from Equation (2.51), that is:

$$f_p(E) = 1 - \frac{1}{1 + e^{(E - E_f)/kT}}$$

becomes

$$f_p(E) = \frac{1}{1 + e^{(E_f - E)/kT}}$$

- 2.21** Recall that the equations $n_0 = N_C e^{-[(E_C - E_f)/kT]}$ and $p_0 = N_V e^{-[(E_f - E_v)/kT]}$ are valid only for nondegenerately doped semiconductors. One reason is that the Boltzmann approximation is valid only for nondegenerate semiconductors. What is the other reason?
- 2.22** a. Silicon is doped with $N_D = 2 \times 10^{16} \text{ cm}^{-3}$. Find n_0 and p_0 , and locate the Fermi level. Draw the energy band diagram.
 b. A new batch of silicon is doped with boron to $N_A = 8 \times 10^{14} \text{ cm}^{-3}$. Find n_0 and p_0 and locate the Fermi level. Draw the energy band diagram.
- 2.23** Silicon is doped with $N_A = 6.0 \times 10^{16} \text{ cm}^{-3}$. Find n_0 and p_0 . Draw the energy band diagram indicating the position of the Fermi level.
- 2.24** InP is doped with $N_D = 2.0 \times 10^{15} \text{ cm}^{-3}$ and $N_A = 10^{14} \text{ cm}^{-3}$. Draw the energy band diagram and calculate the location of the Fermi level.
- 2.25** a. Calculate the intrinsic carrier concentration for Ge at room temperature.
 b. Make a semilogarithmic plot of n_i versus temperature over the range 200 to 500 K. The intrinsic carrier concentration should be on the vertical (logarithmic) axis. Add curves for silicon and GaAs.
- 2.26** Ge is doped with $4.5 \times 10^{15} \text{ Sb/cm}^{-3}$. Assuming Sb to be a shallow donor, find n_0 and p_0 at room temperature. Draw the energy band diagram. Why is the assumption of the donor being shallow important? What would change if the donor state were known to be deep in the forbidden gap?

- 2.27** A diode is made by using an n-type Si substrate ($N_D = 4.0 \times 10^{15} \text{ cm}^{-3}$) and then adding acceptors to a concentration of $N_A = 5.3 \times 10^{17} \text{ cm}^{-3}$ to one region. What are p_0 and n_0 in the p-type region?
- 2.28** A sample of GaAs is doped with $N_A = 7.0 \times 10^{16} \text{ cm}^{-3}$ and $N_D = 2.0 \times 10^{15} \text{ cm}^{-3}$.
- Find n_0 and p_0 .
 - Locate E_f .
 - Sketch the energy band diagram. Label both axes and identify E_C , E_V , E_f , E_i , and E_{vac} on your drawing. Indicate the electron affinity and the band gap.
- 2.29** A sample of GaN is doped with $N_D = 10^{18} \text{ cm}^{-3}$ and $N_A = 2 \times 10^{16} \text{ cm}^{-3}$.
- Find n_0 and p_0 .
 - Locate E_f .
 - Sketch the energy band diagram.
- 2.30** A sample of GaAs is doped with $N_D = 4.0 \times 10^{16} \text{ cm}^{-3}$ and $N_A = 8.5 \times 10^{15} \text{ cm}^{-3}$. Find the equilibrium concentrations of electrons and holes, and locate the Fermi level.
- 2.31** How heavily would you need to dope silicon with donors to violate the assumption of nondegenerate doping? How many acceptors would be needed to just cause a degenerately doped type material?
- 2.32** What donor concentration is required to elevate the Fermi level in Si to $2.3kT$ above E_i ? What acceptor concentration will lower E_f $2.3kT$ below E_i ?
- 2.33** Given that for a particular sample of silicon $N_D = 10^{15} \text{ cm}^{-3}$ and $n_0 = 10^7 \text{ cm}^{-3}$, find N_A and p_0 .
- 2.34** Manganese makes a donor trap state 0.53 eV below the conduction band edge in silicon. If the silicon is doped with $N_D = 1.0 \times 10^{16} \text{ cm}^{-3}$, what is the probability of occupancy of the trap state? Assume the concentration of Mn is small enough not to affect the overall doping and that the trap state is single-valued.
- 2.35** A sample of Si doped with phosphorus has its Fermi level 0.15 eV below E_C , and the donor ground state is 0.045 eV below E_C . Find n_0 .
- 2.36** The probability of occupancy of an energy state at the conduction band edge, E_C , of Si is 5.0×10^{-2} :
- Is this Si n type, p type, or intrinsic?
 - Find $N_D - N_A$.
- 2.37** The effective density-of-states function N_C for a given semiconductor is 10^{18} cm^{-3} . What is the electron density-of-states effective mass?
- 2.38** Given two semiconductors A and B, let them have the same density-of-states effective masses. Let $E_g = 1 \text{ eV}$ for A and 2 eV for B. Find the ratios of the intrinsic carrier concentrations.

- 2.39** Semiconductor devices for many applications must be able to withstand and operate over a wide range of temperatures, to operate from Antarctica or deep space to a tropical climate in a hot truck. For example, military specifications for semiconductor devices cover the range -55°C to $+150^{\circ}\text{C}$. Repeat Problem 2.27 at these two temperature extremes.
- 2.40** In a nondegenerate semiconductor, the electron distribution peaks at $E_C + kT/2$ as indicated in Figure 2.17, and the average electron kinetic energy is at $E_C + 3kT/2$. For intrinsic material, find the probability that a state at these two energies is occupied. Repeat for $E_C + 10kT$.
- 2.41** Calculate the electron concentrations in Si and GaAs if the Fermi level is 0.3 eV above the valence band edge.
- 2.42** Complete the mathematical steps in Example 2.8 and verify the results.
- 2.43** At what doping concentration does apparent band-gap narrowing become significant in n-type Si? (We will define significant as a narrowing of 0.03 eV.)
- 2.44** Suppose you were to dope some silicon with $N_D = N_A = 5 \times 10^{16} \text{ cm}^{-3}$.
- Where do you expect the Fermi level to be?
 - What do you expect n_0 and p_0 to be?
 - Verify (or adjust) your expectations by performing the calculations.
- What do you think is going on physically? What is happening to the electrons in the donor and acceptor states?
- 2.45** Plot and compare n_i for Si as a function of temperature from 250 to 500 K as calculated from Equation (2.88) with that calculated from Equations (2.85), (2.86), (2.63), and (2.64). Neglect the small temperature dependences of the density-of-states effective masses.
- 2.46** A silicon sample doped with $N_A = 10^{14}$ and $N_D = 10^{15} \text{ cm}^{-3}$ is at a temperature of 600 K. Find n_0 and p_0 .
- 2.47** The sample of Problem 2.46 is cooled to 250 K. What are the equilibrium carrier densities n_0 and p_0 now?
- 2.48** Find the apparent band-gap narrowing for silicon with $N_D = 2 \times 10^{19} \text{ cm}^{-3}$.
- 2.49** A sample of silicon is degenerately doped with $N_D = 6.0 \times 10^{18} \text{ cm}^{-3}$. Find the electron and hole concentrations.
- 2.50** Find n_0 and p_0 for silicon with $N_A = 1.5 \times 10^{19}$.
- 2.51** In this chapter we differentiated between conductivity effective mass and density-of-states effective mass. This problem illustrates the calculation of the conductivity of a semiconductor by using the conductivity effective mass. (A more thorough discussion appears in Online Module OM5.) The current density is the current passing a plane of unit area per second. Consider an n-type semiconductor. The current density is

$$J = -qn_0\langle v \rangle = -\sigma\mathcal{E}$$

where n_0 is the electron concentration, $\langle v \rangle$ is the average electron velocity, and σ is the conductivity. The negative sign indicates that the current

is in the opposite direction of the electron velocity since the electron is charged negatively. Since this is an n-type semiconductor, $n_0 = N_D$, and the conductivity can be expressed as

$$\sigma = -\frac{qN_D\langle v \rangle}{\mathcal{E}}$$

Find the conductivity of a semiconductor with $N_D = 10^{16} \text{ cm}^{-3}$ if the average time between collisions of electrons is 10^{-13} seconds.

CHAPTER 3

Current Flow in Homogeneous Semiconductors

3.1 INTRODUCTION

In Chapter 2 we learned how to determine how many carriers are available for conduction in a given semiconductor—holes in the valence band and electrons in the conduction band. In this chapter, we consider the mechanisms by which current flows.

There are two basic mechanisms by which charge carriers move in semiconductors: *drift* and *diffusion*. We begin with the more intuitive drift current, which results when the electrons and holes are in an electric field. The diffusion current, which is discussed later, arises when there is a variation in the concentration of carriers with position.

3.2 DRIFT CURRENT

A semiconductor at equilibrium has a net total current of zero. Although the electrons and holes are moving because of their thermal energy, the direction of movement is completely random, so all the currents contributed by the individual charge carriers add up to zero. Consider a single electron in the conduction band. Its path when there is no applied field is shown in Figure 3.1a. After each collision, the new direction is random, so that, averaged over time, the electron makes no overall progress in any particular direction and the net current is zero.

Recall from Figure 2.2, however, that when an external electric field is applied to a semiconductor, the negatively charged electrons are accelerated and the positively charged holes are accelerated in the opposite direction. Figure 3.1b shows the progress of a particular electron under the influence of a high electric field. Averaged over a time long enough to include many collisions, there is a

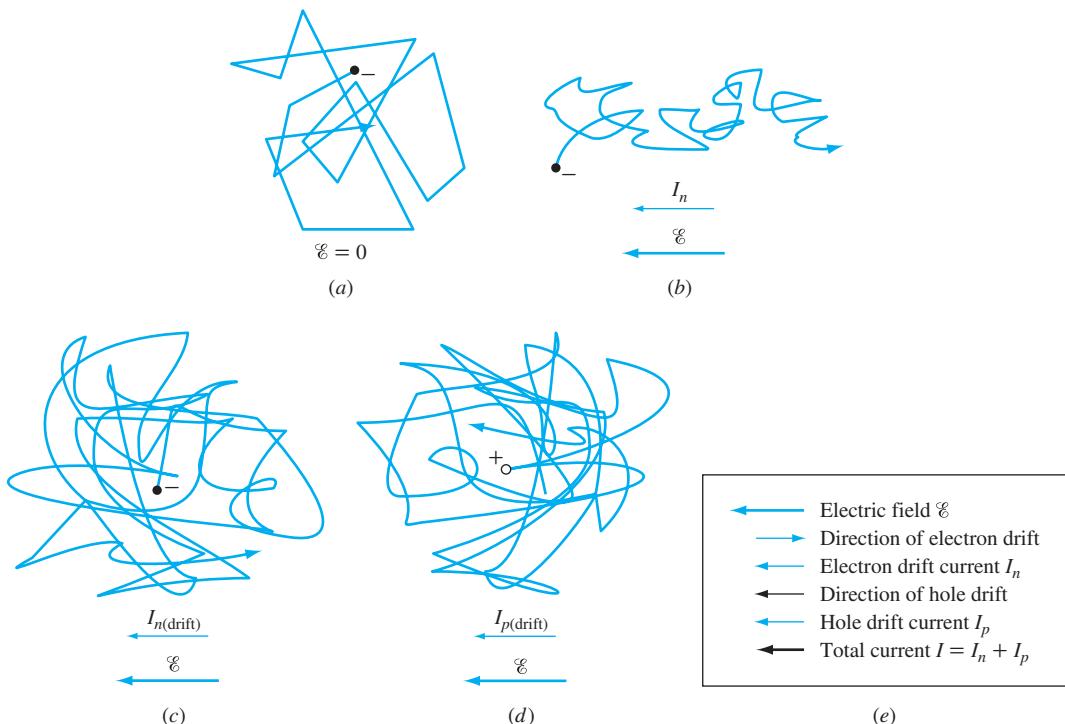


Figure 3.1 The motion of an electron in a crystal. The electron changes direction randomly whenever it makes a collision. (a) Under no applied field there is no net progress in any particular direction. (b) When a field is applied, the electron tends to drift in some particular direction. A trajectory such as this would be found only under very high fields. (c) Under low fields, the drift velocity is much smaller than the thermal speed. It takes many collisions before any appreciable progress is made. (d) A hole undergoes similar motion, but being positively charged, is accelerated in the opposite direction. (e) Electrons and holes drift in opposite directions, but the resulting currents are in the same direction.

general tendency for the electron to drift toward the right. Any net motion of a charge results in electric current, so this motion is termed “drift current.”

For more typical fields in a bulk semiconductor, the velocity induced by the applied field is actually quite small compared with the thermal speed as indicated in Figure 3.1c. For a field on the order of 10 V/cm, the average drift velocity is on the order of 10^4 cm/s. Compare this with the average electron speed between collisions (thermal speed) on the order of 10^7 cm/s at room temperature.

Electric current is defined as the amount of charge crossing some plane per unit time. The charge for electrons is negative; in Figures 3.1b and c, therefore, the electron current $I_{n(\text{drift})}$ is going to the left. We are using *drift* in the subscript to differentiate this current mechanism, driven by the presence of an electric field, from the diffusion current discussed later. Holes can also travel in these types of paths, as shown in Figure 3.1d. The overall drift of the holes is in the

opposite direction to that of the electron. This is because the holes are positively charged, so the same field accelerates them in the opposite direction to the electrons. The current the holes produce is, however, in the same direction as their motion, and the hole current $I_{p(\text{drift})}$ is also to the left, as shown in Figure 3.1e. The total drift current is the sum of these two:

$$I_{(\text{drift})} = I_{n(\text{drift})} + I_{p(\text{drift})} \quad (3.1)$$

Current I is a convenient quantity to define for a wire, but for a semiconductor, it is often more useful to talk about current density J , which is the amount of charge crossing a plane of unit area per unit time. That is,

$$J = \frac{I}{\text{area}} \quad (3.2)$$

We now derive an expression for current density. We recall from Ohm's law that the resistance R of a uniform sample of length L and cross-sectional area A (Figure 3.2) is

$$R = \frac{V}{I} = \frac{\rho L}{A} \quad (3.3)$$

where V is the voltage and ρ is referred to as *resistivity* and has common dimensions of ohm-centimeters ($\Omega \cdot \text{cm}$). Since the semiconductor is uniform, $V = \mathcal{E}L$, where \mathcal{E} is the electric field and L is the sample length. Using $I = JA$, we can express Equation (3.3) as

$$\frac{\mathcal{E}L}{JA} = \frac{\rho L}{A} \quad (3.4)$$

or

$$J_{(\text{drift})} = \frac{\mathcal{E}}{\rho} = \sigma \mathcal{E} \quad (3.5)$$

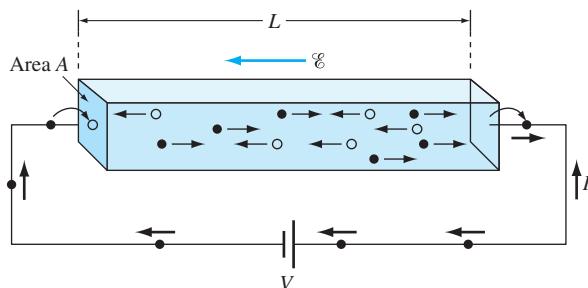


Figure 3.2 Current is carried by electrons in a wire, but in a semiconductor current can be carried by both electrons and holes.

where $\sigma = 1/\rho$ and is referred to as conductivity, with dimensions of reciprocal ohm-centimeters $[(\Omega \cdot \text{cm})^{-1}]$ or siemens per centimeter (S/cm), and we have added the subscript *drift* again since there will also be a diffusion current density component discussed later. Equation (3.5) is often referred to as *Ohm's law in point form*.

Since, in a semiconductor, electrons and holes can carry current, the total current density is

$$J_{(\text{drift})} = J_{n(\text{drift})} + J_{p(\text{drift})} \quad (3.6)$$

where $J_{n(\text{drift})} = \sigma_n \mathcal{E}$ and $J_{p(\text{drift})} = \sigma_p \mathcal{E}$. Here σ_n is the conductivity due to electrons and σ_p the conductivity due to holes. In an n-type semiconductor, electrons carry most of the current, since there are more electrons than holes, while in p-type material most of the current is carried by holes.

Before we continue, we have an observation to make about Figure 3.2. Both holes and electrons carry the current in the semiconductor, but in the wire, the current (the same amount) is carried by electrons alone. What happens to the holes? At the left end of the semiconductor bar, we see electrons arriving at the semiconductor–wire boundary from the wire, and at the same time, holes are arriving there from the bar. The electron, in effect, enters the semiconductor at the ohmic contact between wire and semiconductor and annihilates the hole. An opposing electron cancels every hole that gets lost. The hole is annihilated, but remember, it was only an empty state to begin with. When the electron moves into the state, the hole is lost but the electron keeps going.

We next obtain expressions for σ_n and σ_p . We first consider hole conduction.

Consider a p-type semiconductor of hole concentration p as shown schematically in Figure 3.3. We have used the more general notation p instead of the equilibrium concentration p_0 , because once a field is applied the sample is no longer at equilibrium and in some cases, $p \neq p_0$. Here an electric field is applied

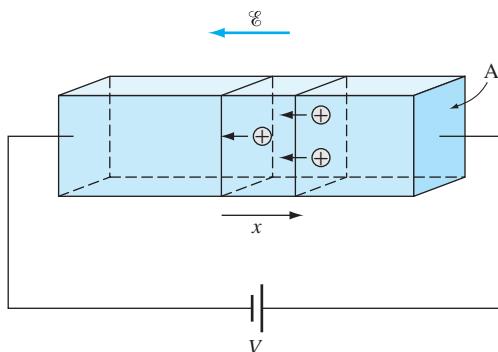


Figure 3.3 We find the conductivity of a sample by considering the flow of (in this case) holes across an area A .

in the negative x direction. The force on the holes, $F = +q\mathcal{E}$, will move them in the direction of the field. Assume that all the holes have an average field-induced (drift) velocity v_{dp} . The total charge crossing the area A in time dt is equal to $dQ = qpAv_{dp} dt$. Since current is defined as the charge passing a plane in unit time,

$$I_p = \frac{dQ}{dt} = qpAv_{dp} \quad (3.7)$$

The hole current density is $J_p = I_p/A$. But from Equations (3.7), (3.2), and (3.5),

$$J_{p(\text{drift})} = qp v_{dp} = \sigma_p \mathcal{E} \quad (3.8)$$

and

$$\sigma_p = qp \frac{v_{dp}}{\mathcal{E}} = qp \mu_p \quad (3.9)$$

where μ_p is called the *hole mobility* and is the hole drift velocity per unit field:

$$\mu_p = \frac{v_{dp}}{\mathcal{E}} \quad (3.10)$$

The mobility is a measure of how easily a carrier moves in a particular material. It is normally expressed in centimeters squared per volt-second ($\text{cm}^2/\text{V} \cdot \text{s}$). Similarly the electron mobility is

$$\mu_n = -\frac{v_{dn}}{\mathcal{E}} \quad (3.11)$$

and the electron conductivity σ_n is

$$\sigma_n = qn\mu_n \quad (3.12)$$

Finally, the total drift current density is

$$J_{(\text{drift})} = J_{n(\text{drift})} + J_{p(\text{drift})} = (q\mu_n n + q\mu_p p)\mathcal{E} = \sigma\mathcal{E} \quad (3.13)$$

and the total conductivity is

$$\sigma = qn\mu_n + qp\mu_p \quad (3.14)$$

3.3 CARRIER MOBILITY

The electron and hole mobilities are dependent on the impurity concentrations of donors and acceptors, on temperature, and on whether the carriers are minority carriers or majority carriers. (Majority carriers are electrons in n-type material and holes in p-type material.) The physics of these effects are discussed in the next section, but here we give some results for silicon at room temperature.

First, the mobility varies with the doping concentration. While there is considerable scatter in the experimental data,¹ for uncompensated material, the mobility in silicon is often characterized by the empirical formula:

$$\mu = \mu_0 + \frac{\mu_1}{1 + \left(\frac{N}{N_{\text{ref}}}\right)^\alpha} \quad (3.15)$$

where N is the doping concentration (either N_D or N_A) and μ_0 , μ_1 , N_{ref} and α are fitting parameters.

The fitting parameters depend on whether the carriers of interest are majority or minority carriers.² At room temperature Equation (3.15) becomes:

Majority carriers [1]:

$$\mu_n(N_D) = 65 + \frac{1265}{1 + \left(\frac{N_D}{8.5 \times 10^{16}}\right)^{0.72}} \quad (3.16)$$

$$\mu_p(N_A) = 48 + \frac{447}{1 + \left(\frac{N_A}{6.3 \times 10^{16}}\right)^{0.76}} \quad (3.17)$$

Minority carriers [2–4]:

$$\mu_n(N_A) = 232 + \frac{1180}{1 + \left(\frac{N_A}{8 \times 10^{16}}\right)^{0.9}} \quad (3.18)$$

$$\mu_p(N_D) = 130 + \frac{370}{1 + \left(\frac{N_D}{8 \times 10^{17}}\right)^{1.25}} \quad (3.19)$$

These equations apply only to silicon, and only under low field. (We will address high-field effects in Section 3.3.5.) They are plotted in Figure 3.4. From these plots we can see:

- At low impurity concentrations, majority carrier and minority carrier electron mobilities approach the same values: $\mu_n \approx 1330 \text{ cm}^2/\text{V} \cdot \text{s}$.
- A similar result holds for holes: $\mu_p \approx 495 \text{ cm}^2/\text{V} \cdot \text{s}$.
- Electron and hole mobilities (both majority carrier and minority carrier) reduce monotonically with increasing impurity concentration.
- For a given doping level, minority carrier mobilities for electrons and holes are greater than corresponding majority carrier mobilities.
- These fractional differences increase with increased doping.

¹The majority carrier mobilities have been extensively measured in uncompensated semiconductors. Minority carrier mobility is much more difficult to measure, and fewer results have been reported.

²It may surprise the student to know that in some devices, such as bipolar junction transistors, the minority carriers influence the device operation more than the majority carriers. In some field-effect transistors, the majority carriers are more important.

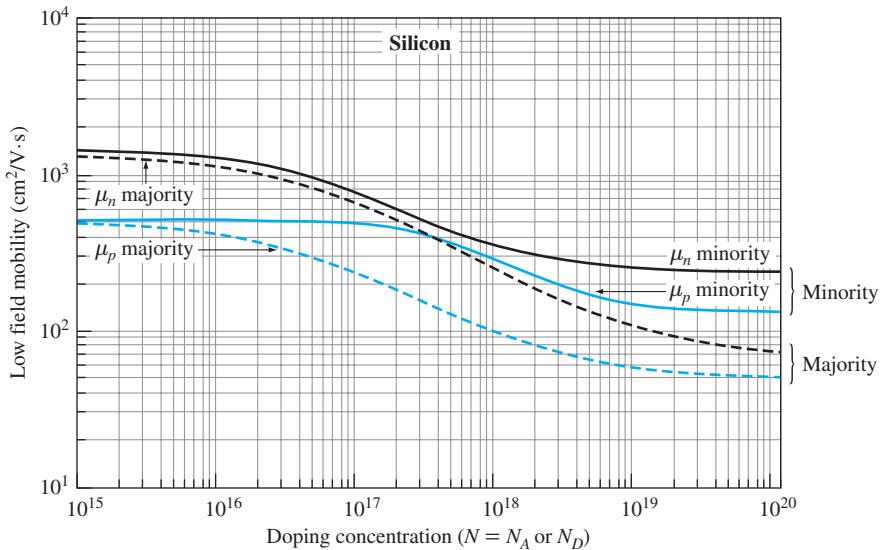


Figure 3.4 Room temperature majority and minority carrier mobility as functions of doping in p-type and n-type silicon. Solid lines: minority carriers; dashed lines: majority carriers.

EXAMPLE 3.1

Compute the room temperature resistivity of intrinsic silicon, and compare that with uncompensated n- and p-type silicon doped with 10^{17} cm^{-3} donors (or acceptors). Assume $n = n_0$ and $p = p_0$.

Solution

The resistivity of a sample is given by the reciprocal of the conductivity:

$$\rho = \frac{1}{\sigma}$$

and from Equation (3.14),

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

- a. *Intrinsic silicon:* We already know the room temperature equilibrium carrier concentrations for intrinsic silicon are $n_0 = p_0 = n_i = 1.08 \times 10^{10} \text{ cm}^{-3}$.

From Figure 3.4 we find the mobilities for electrons and holes for undoped (intrinsic) silicon:

$$\mu_n = 1330 \frac{\text{cm}^2}{\text{V} \cdot \text{s}}$$

$$\mu_p = 495 \frac{\text{cm}^2}{\text{V} \cdot \text{s}}$$

We can find the conductivity:

$$\begin{aligned}\sigma &= q(\mu_n n_0 + \mu_p p_0) = q(\mu_n n_0 + \mu_p p_0) \\ &= 1.6 \times 10^{-19} \text{ C} \left(1330 \frac{\text{cm}^2}{\text{V} \cdot \text{s}} \cdot 1.08 \times 10^{10} \text{ cm}^{-3} + 495 \frac{\text{cm}^2}{\text{V} \cdot \text{s}} \cdot 1.08 \times 10^{10} \text{ cm}^{-3} \right) \\ &= 3.15 \times 10^{-6} (\Omega \cdot \text{cm})^{-1}\end{aligned}$$

From the conductivity, we can find the resistivity:

$$\rho = \frac{1}{\sigma} = \frac{1}{3.15 \times 10^{-6}} = 3.17 \times 10^5 \Omega \cdot \text{cm}$$

- b. *n-type silicon:* We expect, since this material is more heavily doped, and since it has many electrons to carry current, that the result will be higher conductivity or lower resistivity than that of intrinsic material. Since $N_D \gg n_i$ and since virtually all impurities are ionized,

$$\begin{aligned}n_0 &= N_D = 10^{17} \text{ cm}^{-3} \\ p_0 &= \frac{n_i^2}{n_0} = \frac{(1.08 \times 10^{10} \text{ cm}^{-3})^2}{10^{17} \text{ cm}^{-3}} = 1.17 \times 10^3 \text{ cm}^{-3}\end{aligned}$$

From Figure 3.4, for a donor concentration of 10^{17} , we find the electron mobility and hole mobilities to be:

$$\begin{aligned}\mu_n &\approx 650 \frac{\text{cm}^2}{\text{V} \cdot \text{s}} && \text{majority carrier mobility} \\ \mu_p &\approx 480 \frac{\text{cm}^2}{\text{V} \cdot \text{s}} && \text{minority carrier mobility}\end{aligned}$$

Using these values, we obtain

$$\begin{aligned}\sigma &= q(\mu_n n_0 + \mu_p p_0) \\ &= 1.6 \times 10^{-19} \text{ C} \left(650 \frac{\text{cm}^2}{\text{V} \cdot \text{s}} \cdot 10^{17} \text{ cm}^{-3} + 480 \frac{\text{cm}^2}{\text{V} \cdot \text{s}} \cdot 1.17 \times 10^3 \text{ cm}^{-3} \right) \\ &= 1.6 \times 10^{-19} (6.5 \times 10^{19} + 5.62 \times 10^5) = 10.4 (\Omega \cdot \text{cm})^{-1}\end{aligned}$$

Note that the second term, the hole term, is negligible in this case. The resistivity is then

$$\rho = \frac{1}{\sigma} = \frac{1}{10.4} = 0.096 \Omega \cdot \text{cm}$$

The doped Si is considerably more conductive than the intrinsic material.

- c. *Same doping level as the previous example but p type instead of n type:* It is worthwhile to start a problem by thinking ahead to what kind of result we might expect. We expect the p-type material to be more conductive than the intrinsic, again because it has more carriers. On the other hand, holes are generally less

mobile than electrons, so the p-type material may not be as conductive as the equivalent n-type. We find the carrier concentrations first:

$$p_0 = N_A = 10^{17} \text{ cm}^{-3}$$

$$n_0 = \frac{n_i^2}{p_0} = 1.17 \times 10^3 \text{ cm}^{-3}$$

Since

$$\mu_n = 750 \frac{\text{cm}^2}{\text{V} \cdot \text{s}} \quad \text{minority carriers}$$

$$\mu_p = 230 \frac{\text{cm}^2}{\text{V} \cdot \text{s}} \quad \text{majority carriers}$$

Then

$$\begin{aligned} \sigma &= q(\mu_n n_0 + \mu_p p_0) \\ &= 1.6 \times 10^{-19} \text{ C} \left(750 \frac{\text{cm}^2}{\text{V} \cdot \text{s}} \cdot 1.17 \times 10^3 \text{ cm}^{-3} + 230 \frac{\text{cm}^2}{\text{V} \cdot \text{s}} \cdot 10^{17} \text{ cm}^{-3} \right) \\ &= 3.7 (\Omega \cdot \text{cm})^{-1} \end{aligned}$$

In this case, the electrons are so few they contribute negligibly to conduction.

The resistivity for the p-type sample is

$$\rho = \frac{1}{\sigma} = \frac{1}{3.7 (\Omega \cdot \text{cm})^{-1}} = 0.27 \Omega \cdot \text{cm}$$

The p-type material is less conductive than the n-type sample at the same doping levels because the current is carried primarily by holes and the holes are less mobile than the electrons.

3.3.1 CARRIER SCATTERING

Recall that by definition [Equations (3.10) and (3.11)], mobility is dependent on the drift velocity:

$$\mu = \left| \frac{v_d}{\mathcal{E}} \right|$$

The carrier drift velocity is influenced by scattering events, i.e., change in direction and/or energy of a carrier by collisions with a particle. Minority carrier mobility is dependent on these scattering mechanisms. As discussed later, the drift velocity is also dependent on the time the carrier spends in impurity states. This impurity band conduction, along with scattering, contributes to majority carrier scattering. Since both majority and minority carrier mobilities are influenced by scattering, we will examine the physics of scattering first.

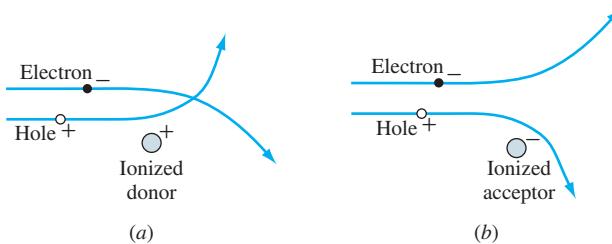


Figure 3.5 (a) An electron approaching an ionized donor is deflected toward it, but a hole is deflected away from the donor. (b) Electrons deflect away from the negatively charged ionized acceptors, but holes deflect toward them.

Carriers can be scattered by interactions (collisions) with “particles” such as ionized impurity atoms and by phonons. The drift velocity and thus the mobility are dependent on the mean free time between collisions.

Ionized Impurity Scattering The scattering effect of the impurities can be understood by recognizing that the impurities (donors and acceptors) are typically ionized, so they are charged. The Coulombic forces will deflect an electron or hole approaching the ionized impurity. This is known as *ionized impurity scattering* and is shown schematically in Figure 3.5. The situation is shown for a positive ion (a) and a negative ion (b). Holes are deflected away from the positive ion and toward the negative ion; electrons are deflected oppositely. The amount of deflection depends on the speed of the carrier and its proximity to the ion. It follows, then, that the more heavily a material is doped, the higher the probability that a carrier will collide with an ion in a given time and the smaller will be the mean free time between collisions and the smaller the mobility.

Furthermore, since the ionized donors and acceptors both scatter carriers approximately equally, the mobility due to ionized impurity scattering in compensated material depends on the *total* doping concentration $N_D + N_A$. Although Figure 3.4 is for uncompensated material, in practice the minority carrier mobilities are often estimated from it since usually $N_D \gg N_A$ or $N_A \gg N_D$.

Lattice (Phonon) Scattering Another influence on the carrier mobility results from lattice scattering, often called phonon scattering. We know that at any temperature the vibrating atoms create pressure (acoustic) waves in the crystal. As discussed in the Supplement to Part 1, these pressure waves are called *phonons*. Like electrons, phonons can be considered to be particles (wave-particle duality), each with energy

$$E_{\text{phonon}} = \hbar\omega \quad (3.20)$$

and wave vector

$$K = \frac{2\pi}{\lambda} \quad (3.21)$$

where ω is the angular frequency of the lattice vibration and λ is the corresponding wavelength. Thus, considering phonons as particles, a phonon can collide with an electron (or hole) and scatter it. Phonon energies extend over a small range—usually less than 0.1 eV. The energy distribution of phonons and the concentration of phonons at a given energy depend on the amplitude of the pressure wave, and thus on temperature. With increasing temperature, the stronger lattice vibrations cause an increase in the concentration of phonons and thus increased scattering or reduced mobility.

3.3.2 SCATTERING MOBILITY

Let us explore how the scattering affects the mobility. We start by considering the force on an electron due to the electric field:

$$F = -q\mathcal{E} = m_{ce}^* a = m_{ce}^* \frac{dv}{dt} \quad (3.22)$$

where a is the electron acceleration due to the field \mathcal{E} and v is the electron velocity. Note that, since we are considering conduction, we use the conductivity effective mass for electrons. Here we consider the mobility associated with a single scattering mechanism, e.g., ionized impurity scattering.

From Equation (3.22) we can write

$$dv = -\frac{q\mathcal{E}}{m_{ce}^*} dt \quad (3.23)$$

Integrating both sides of Equation (3.23) (between collisions) gives

$$v(t) - v(t_0) = -\frac{q\mathcal{E}}{m_{ce}^*}(t - t_0) \quad (3.24)$$

where t_0 is the time of the previous collision. This gives us the velocity reached by a particular electron just before its next collision at time t . If we consider a large number of collisions and take the average, we get

$$v_{dn} = \langle v(t) - v(t_0) \rangle = -\frac{q\mathcal{E}}{m_{ce}^*} \langle t - t_0 \rangle \quad (3.25)$$

We define the drift velocity of electrons as $v_{dn} = \langle v(t) - v(t_0) \rangle$. We also define the mean free time between collisions as

$$\bar{t}_n = \langle t - t_0 \rangle \quad (3.26)$$

Then the drift velocity for electrons can be written

$$v_{dn} = -\frac{q\mathcal{E}}{m_{ce}^*} \bar{t}_n \quad (3.27)$$

From Equations (3.11) and (3.27)

$$\mu_n = \frac{q\bar{t}_n}{m_{ce}^*} \quad (3.28)$$

For holes, a similar calculation with charge $+q$ and conductivity effective mass m_{ch}^* gives

$$v_{dp} = \frac{q\bar{t}_p}{m_{ch}^*} \mathcal{E} = \mu_p \mathcal{E} \quad (3.29)$$

$$\mu_p = \frac{q\bar{t}_p}{m_{ch}^*} \quad (3.30)$$

where μ_p is the hole mobility.

From the above it can be seen that the carrier mobility for any given scattering mechanism depends on the carrier conductivity effective mass and on the scattering time associated with that scattering mechanism. The scattering time is dependent on the frequency of carrier collisions with other “particles.”

Because the scattering times for carrier-ion and for carrier-phonon collisions are independent, by Matthiessen’s rule,

$$\frac{1}{t_n} = \frac{1}{t_{nii}} + \frac{1}{t_{nl}} \quad (3.31)$$

and

$$\frac{1}{\mu_n} = \frac{1}{\mu_{nii}} + \frac{1}{\mu_{nl}} \quad (3.32)$$

where the subscripts ii and l refer to ionized impurity and lattice (phonon) scattering respectively.

EXAMPLE 3.2

Find the mean free times between scattering, \bar{t}_n and \bar{t}_p , for intrinsic Si at room temperature.

Solution

For intrinsic Si, $N_D = N_A = 0$. Then from Equation (3.28), $\bar{t}_n = (m_{ce}^* \mu_n)/q$. From Figure 3.4, $\mu_n = 1330 \text{ cm}^2/\text{V} \cdot \text{s} = 0.133 \text{ m}^2/\text{V} \cdot \text{s}$. The conductivity effective mass for electrons is, from Table 2.1, $m_{ce}^* = 0.26 m_0 = 0.26 \times 9.11 \times 10^{-31} \text{ kg}$.

The mean scattering time is

$$\bar{t}_n = \frac{0.26 \times 9.11 \times 10^{-31} \times 0.133}{1.60 \times 10^{-19}} = 2 \times 10^{-13} \text{ s}$$

Similarly for holes, $\mu_p = 495 \text{ cm}^2/\text{V} \cdot \text{s}$, $m_{ch}^* = 0.36 m_0$, and $\bar{t}_p = 1 \times 10^{-13} \text{ s}$

3.3.3 IMPURITY BAND MOBILITY

The above two scattering mechanisms, ionized impurity scattering and phonon scattering, apply to both minority carriers and majority carriers. There is an additional scattering mechanism associated with majority carriers traveling within

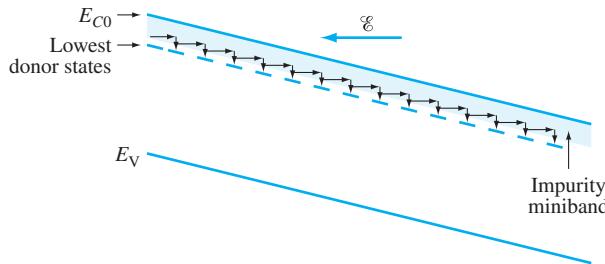


Figure 3.6 Majority carriers (in this case, electrons) drift in an n-type semiconductor. Electrons spend some time in the impurity band, where their mobility is reduced relative to that in the conduction band.

the dopant impurity states. As discussed in Chapter 2, with increased doping the states associated with the dopant atoms overlap in space and energy to form an impurity band extending from the bottom of the normal conduction band E_{C0} into the normally forbidden band. This impurity bandwidth increases with increased doping. This is illustrated for electrons in an uncompensated ($N_A = 0$) n-type semiconductor by Figure 3.6.

Earlier we discussed the deflection of electrons in the conduction band by ionized impurities. Electrons in the impurity band, however, are in states associated with the donor atoms. These electrons can be temporarily captured by a donor state, as shown in the figure. These electrons are repeatedly captured and re-released, slowing down their progress. Additionally, some of the electrons in the miniband are in orbit around the dopant atom. While they can move from one dopant atom to another since these orbits overlap, they cannot do this as readily as a Bloch wave (that is, as they would in the conduction band). With increased donor concentration, this impurity band becomes wider (in energy and in concentration of states). A greater fraction of the electrons travel in the impurity band where their drift velocity and thus mobility are reduced [Equation (3.11)].

We emphasize that in *uncompensated* material this mechanism (the slowing of majority carriers by the donor or acceptor states) is important only for majority carriers. In the n-type semiconductor of Figure 3.6, holes travel only in the normal valence band. In compensated material, however, in which both donors and acceptors are present, electrons will travel in the donor band while holes travel in analogous acceptor bands if the minority doping is great enough. The minority carrier hole (electron) mobility is then a function of the acceptor (donor) concentration.

Note that the (majority) electron mobility μ_n associated with this mechanism of impurity band scattering is dependent on N_D but independent of N_A , while μ_p is a function of N_A and independent of N_D . This is in contrast to ionized impurity scattering which, as we mentioned before, is a function of $N_D + N_A$ for both majority and minority carriers.

We emphasize that Figure 3.4 reflects measurements in uncompensated material. In compensated semiconductors, minority carrier current will be partially carried in the impurity bands, which will reduce the minority carrier mobilities from their values in Figure 3.4. Often however, $N_D \gg N_A$ or $N_A \gg N_D$, in which case the minority carrier band transport does not have a large effect on mobility.

From Figure 3.4, it can be seen that for small impurity concentrations such that the impurity band transport effect is negligible, the majority and minority carrier mobilities approach each other for electrons and for holes. With increasing doping level, this effect becomes more prominent and the majority carrier mobility becomes increasingly smaller than the minority carrier mobility.

In this section, we developed the concept of mobility using silicon. For comparison, the low-field doping dependence of majority carrier mobilities is plotted in Figure 3.7 for GaAs and Ge. Minority carrier mobility data are not available for these materials. It is seen that the low-field electron mobility of GaAs is appreciably larger than for Si or Ge, primarily because of its small effective mass.

3.3.4 TEMPERATURE DEPENDENCE OF MOBILITY

The mobility is also sensitive to temperature. With increasing temperature, the carrier mobility due to lattice scattering decreases because the increased phonon concentration causes increased scattering. For silicon, the mobility due to lattice scattering varies as $T^{-2.6}$ for electrons and as $T^{-2.3}$ for holes.

The effect of ionized impurity scattering, however, decreases with increasing temperature because the average thermal speeds of the carriers are increased. Thus, the carriers spend less time near an ionized impurity as they pass and the scattering effect of the ions is thus reduced, or the mobility due to ionized impurity scattering increases with increasing temperature.

These two effects operate simultaneously on the carriers. The temperature dependence of mobility for a typical semiconductor is shown schematically in Figure 3.8. At lower temperatures, ionized impurity scattering dominates, while at higher temperatures, phonon scattering dominates.

3.3.5 HIGH-FIELD EFFECTS

We assumed in the previous section that the velocity imparted by the applied electric field (drift velocity) was small compared with the thermal speed, or that the field had negligible effect on the scattering time \bar{t} . This assumption of constant \bar{t} results in the drift velocity being proportional to the field, or

$$v_d = \frac{q\bar{t}}{m_c^*} \mathcal{E} \quad \text{low field} \quad (3.33)$$

At high enough fields, however, the mean free time between collisions \bar{t} decreases with increasing \mathcal{E} , resulting in reduced mobility. The dependence of v_d on \mathcal{E} has been measured experimentally and is shown in Figure 3.9 for electrons

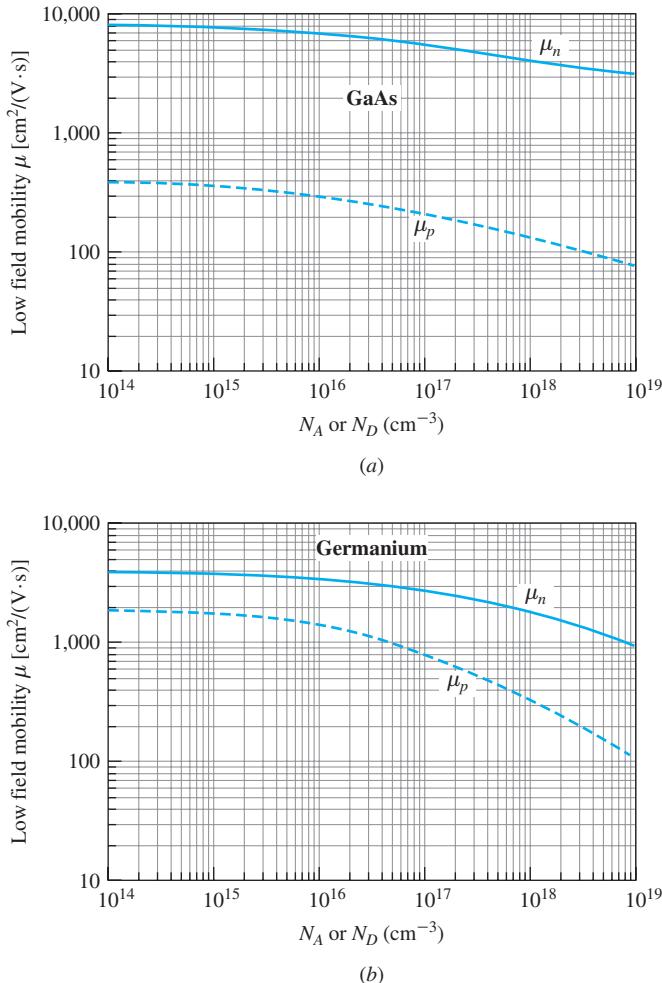


Figure 3.7 Room temperature electron and hole majority carrier mobilities (low field) as functions of uncompensated dopant concentration N in (a) GaAs and (b) Ge.

and holes in high-purity bulk Si and Ge, and for electrons in GaAs at room temperature.

At low fields, the drift velocity v_d is proportional to \mathcal{E} , so μ is constant. This value of μ is called the low-field mobility μ_{lf} , and that is what was plotted in Figures 3.4 and 3.7. As the field is increased, v_d increases sublinearly and appears to approach a limiting velocity v_{sat} . The value of v_{sat} is on the order of 1×10^7 cm/s for both electrons and holes in Si. It is on the order of 6×10^6 cm/s for electrons in GaAs and for both electrons and holes in Ge.

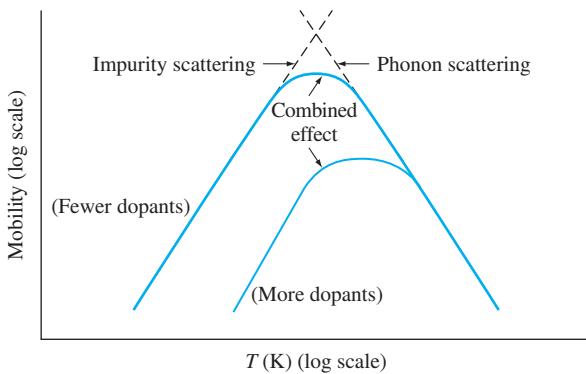


Figure 3.8 Mobility as a function of temperature. At low temperatures, impurity scattering dominates, but at high temperatures, lattice vibrations dominate.

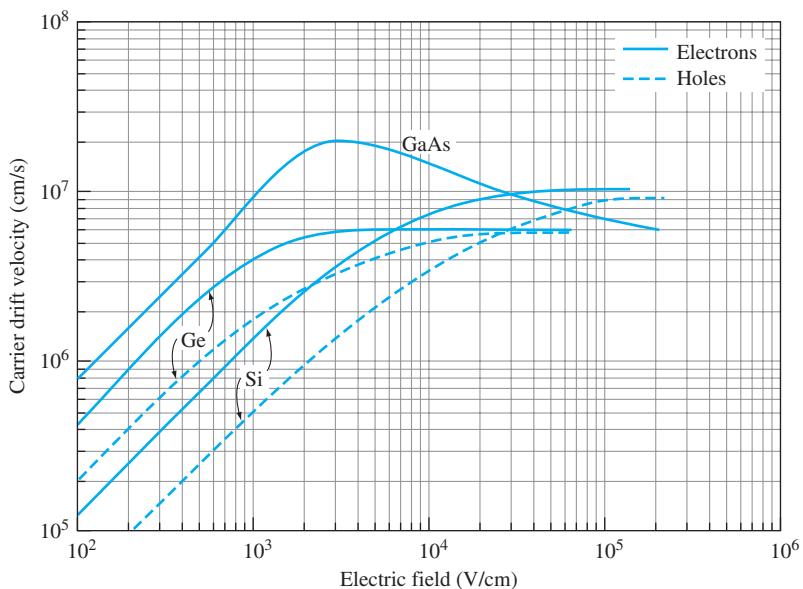


Figure 3.9 The experimentally measured dependence of the drift velocity on the applied field.

This velocity saturation results from a process called *optical phonon scattering*. An optical phonon, as discussed in the Supplement to Part 1, refers to a lattice vibration in which adjacent planes of atoms vibrate out of phase. At high fields, carriers are accelerated enough to gain sufficient kinetic energy between collisions that, when they collide with the lattice, they can impart enough energy to create an optical phonon. The probability that a carrier having this high an

energy will create a phonon is quite high, and thus the maximum carrier velocity is limited by the relation

$$\frac{m^* v_{\max}^2}{2} \approx E_{\text{phonon (opt.)}} \quad (3.34)$$

where $E_{\text{phonon (opt.)}}$ is the optical phonon energy and m^* is the carrier effective mass in the direction of \mathcal{E} . The value of $E_{\text{phonon (opt.)}}$ is 0.063 eV for Si and 0.034 eV for GaAs and Ge.

EXAMPLE 3.3

Estimate the value of the saturation velocity for electrons in Si. Make the following assumptions:

1. After each collision, an electron loses all of its kinetic energy.
2. Electrons gain the kinetic energy $E_{\text{phonon (opt.)}}$ without intermediate collisions; i.e., the creation of optical phonons is the only scattering mechanism.
3. Between $E_K = 0$ and $E_K = m^* v_{\max}^2 / 2$, the electrons are in the parabolic region of the E - K curve and so the concept of constant effective mass is valid.

Solution

From assumptions (1) and (2) and Equation (3.34), the average velocity is $v_{\max}/2$, or

$$v_{\text{sat}} = \frac{v_{\max}}{2} = \sqrt{\frac{E_{\text{phonon (opt.)}}}{2m_{ce}^*}} \quad (3.35)$$

where we use m_{ce}^* for m^* , since velocity is related to conduction.

For Si, $E_{\text{phonon (opt.)}} = 0.063$ eV and $m_{ce}^* = 0.26m_0$, and

$$v_{\text{sat}} = \sqrt{\frac{0.063 \times 1.6 \times 10^{-19}}{2 \times 0.26 \times 9.11 \times 10^{-31}}} = 1.46 \times 10^7 \text{ cm/s}$$

This value is somewhat above the experimental value on the order of 1×10^7 cm/s. The discrepancy results primarily from assumption 2. Since electrons do make intermediate collisions, the average saturation velocity is less than that calculated above.

The data in Figure 3.9 were measured on high-purity materials at room temperature. For more highly doped semiconductors, the scattering time is reduced, with a corresponding reduction in the low-field velocity and the saturation velocity. In doped materials, the mean free time between collisions \bar{t} is smaller, and assumption 2 becomes less accurate; i.e., intermediate collisions are increasingly important. Because of the dependence of \bar{t} on impurity concentration and temperature, v_{sat} also depends on these parameters.

Notice from Figure 3.9 that the drift velocity v_d increases monotonically with \mathcal{E} for electrons and holes in Si and Ge. For electrons in GaAs, however, v_d increases, reaches a maximum, and then decreases, approaching v_{sat} . This is

called velocity overshoot. It can be explained with the aid of the E - K relation for electrons in GaAs (Figure 2.3a). At low field, most of the electrons are in the minimum at $K = 0$, where $m^* = 0.067m_0$. When they collide, they are scattered back into this same minimum. At higher fields, they can gain enough kinetic energy between collisions to be scattered into the higher-energy minima, where the effective mass has a higher value, $m^* \approx 0.55m_0$. This reduces the drift velocity [Equation (3.27)] and the corresponding saturation velocity [Equation (3.35)].

This velocity saturation effect has a significant influence on the performance of transistors and cannot be ignored.

3.4 DIFFUSION CURRENT

We discussed earlier how applying an electric field resulted in drift current. Now we look into a second type of current in a semiconductor, the diffusion current. Diffusion is the process of mobile particles moving from regions of high concentrations to regions of low concentrations. This diffusion results from the random (thermal) motion of particles (often referred to as *random walk*).

As an example, consider a tray of ping-pong balls in which someone has very carefully placed all the balls in a neat group in the center of the tray. If we set the tray to vibrating (analogous to giving the lattice thermal energy, e.g., phonons) then the balls do not stay neatly grouped for very long. They diffuse and will eventually be uniformly distributed throughout the tray.

This same diffusion process occurs for mobile electrons and holes, and since they are charged particles, their motion results in current, referred to as *diffusion current*.

Consider again a bar of semiconductor material of cross-sectional area A at room temperature, as shown in Figure 3.10a. There is no electric field. We assume that at some time t the electrons are distributed nonuniformly with position, as shown in Figure 3.10b. Since the temperature of the semiconductor is nonzero, each electron has some thermal energy, and therefore some speed. The direction of the electron's motion, however, is random since there is no electric field. Each electron can go either to the right or to the left, with 50 percent probability either way.

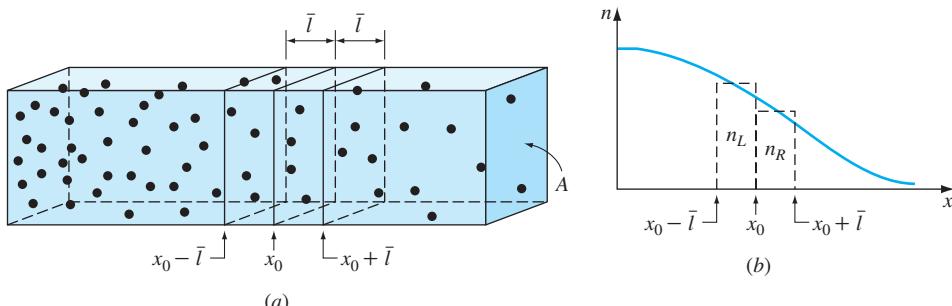


Figure 3.10 Diffusion. (a) A sample of semiconductor with a nonuniform distribution of electrons; (b) the electron density distribution.

Consider the plane $x = x_0$. The electrons pass this plane in either direction. On the left of the plane, there is a high concentration n_L of electrons, with half traveling to the right and half traveling to the left. On the right, there is a smaller concentration n_R of electrons, and half of those are going to the left. Thus, more electrons cross the $x = x_0$ plane from left to right than from right to left. There is a net flux of electrons to the right, and thus net electron diffusion current to the left. (Holes diffuse in exactly the same manner,³ always from regions of high concentration to regions of low concentration. The hole diffusion current is in the same direction as the hole flux because of the positive charge of the holes.)

We next calculate the diffusion current density. We start by finding the electron flux density, which is the number of electrons crossing a unit area at x_0 per unit time. Let \bar{t} be the mean free time between collisions and \bar{l} be the electron mean free path, which is the average distance a carrier progresses in its mean free time.⁴ We consider two regions, each of volume $\bar{l}A$, on either side of x_0 . On the average, half the electrons in this volume will arrive at the plane x_0 without colliding (the other half are going the other way), and they'll arrive there in time \bar{t} . In time \bar{t} , one-half of the electrons on the left will cross x_0 to the right and one-half of the electrons on the right will cross x_0 to the left. The net number of electrons crossing x_0 in time \bar{t} is one-half the difference in the number on either side within a distance \bar{l} from x_0 . The electron flux density is therefore

$$F_n = \frac{(n_L - n_R)\bar{l}}{2\bar{t}} \quad (3.36)$$

where n_L and n_R are the electron densities in the left and right regions respectively. If the change in n is small in distance \bar{l} , we can write

$$(n_L - n_R) \approx -\frac{dn}{dx} \quad (3.37)$$

and the electron flux density is

$$F_n = \frac{-\bar{l}^2}{2\bar{t}} \frac{dn(x)}{dx} = -D_n \frac{dn(x)}{dx} \quad (3.38)$$

where

$$D_n \equiv \frac{\bar{l}^2}{2\bar{t}} \quad (3.39)$$

The quantity D_n is called the *diffusion coefficient* for electrons. The minus sign in Equation (3.38) indicates that the electron flux density is in the direction of decreasing n , that is, always toward regions of lower concentration.

³And so do any other particles with random motion, for example, the ping-pong balls mentioned earlier or impurity atoms in a crystal.

⁴We saw that \bar{t} is on the order of 10^{-13} s and thermal speed \bar{v} is about 10^7 cm/s at room temperature. Then to a reasonable approximation, $\bar{l} = \bar{v}\bar{t}$ and \bar{l} is on the order of 10 nm.

The electron diffusion *current density* is equal to the flux density multiplied by the charge of the electron:

$$J_{n(\text{diff})} = -q F_n = q D_n \frac{dn(x)}{dx} \quad (3.40)$$

Similarly for holes,

$$J_{p(\text{diff})} = +q F_p = -q D_p \frac{dp(x)}{dx} \quad (3.41)$$

Note that, for diffusion, there is no force at work—the only thing causing the currents is the thermal energy of the carriers and the variation in concentration.

As will be discussed shortly, in many devices $dn(x)/dx$ and $dp(x)/dx$ vary with position and thus $J_{n(\text{diff})}$ and $J_{p(\text{diff})}$ [Equations (3.40) and (3.41)] also vary with position. To keep the total current constant, the drift currents change accordingly. This topic is discussed further in Section 3.7, where we consider the *continuity equations*.

We note that both diffusion coefficient and mobility are measures of how easily the particles move through a material, and one might suspect that they are related. In the next chapter, we will derive this relation, known as the *Einstein relation*. We state the results here

$$\frac{D_n}{\mu_n} = \frac{kT}{q} \quad (3.42)$$

$$\frac{D_p}{\mu_p} = \frac{kT}{q} \quad (3.43)$$

From the above, the diffusion coefficients are proportional to the mobilities. Figure 3.11 shows plots of room temperature minority carrier and majority carrier diffusion coefficients for electrons and holes in Si as functions of doping. (Compare with Figure 3.4 for mobilities.)

It is possible to apply a force to the electrons and holes as well, by applying an electric field. For the case in which electrons are subject to both an electric field and a variation in concentration, the total electron current is

$$J_n = J_{n(\text{drift})} + J_{n(\text{diff})} = q \mu_n n(x) \mathcal{E} + q D_n \frac{dn(x)}{dx} \quad (3.44)$$

and the hole current also has drift and diffusion components:

$$J_p = J_{p(\text{drift})} + J_{p(\text{diff})} = q \mu_p p(x) \mathcal{E} - q D_p \frac{dp(x)}{dx} \quad (3.45)$$

Using the Einstein relation, we can express Equations (3.44) and (3.45) as

$$J_n = q \mu_n \left[n \mathcal{E} + \frac{kT}{q} \frac{dn}{dx} \right] \quad (3.46)$$

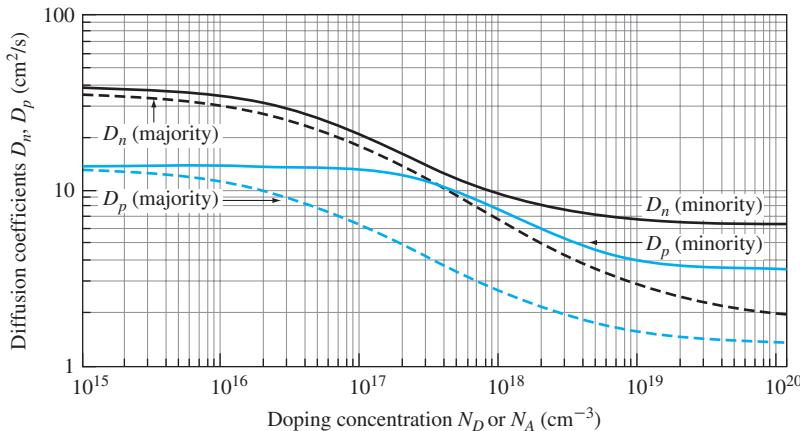


Figure 3.11 Room temperature diffusion coefficients for electrons and holes as a function of doping concentration for silicon.

$$J_p = q\mu_p \left[p\mathcal{E} - \frac{kT}{q} \frac{dp}{dx} \right] \quad (3.47)$$

The two components of each of these are the drift and the diffusion currents. The total current is

$$\begin{aligned} J &= J_{n(\text{drift})} + J_{p(\text{drift})} + J_{n(\text{diff})} + J_{p(\text{diff})} \\ &= q\mu_n n\mathcal{E} + q\mu_p p\mathcal{E} + qD_n \frac{dn}{dx} - qD_p \frac{dp}{dx} \end{aligned} \quad (3.48)$$

where in p-type material, μ_n and D_n are the minority carrier mobilities and diffusion coefficients, and μ_p and D_p are their majority carrier counterparts. The opposite is the case for n-type material.

In many devices (e.g., resistors), there are regions where both n and p are uniform in position (i.e., $dn/dx = dp/dx = 0$). In that case Equation (3.48) becomes

$$J = (q\mu_n n + q\mu_p p)\mathcal{E} = (\sigma_n + \sigma_p)\mathcal{E} = \sigma\mathcal{E} \quad (3.49)$$

which is Ohm's law in point form as expressed in Equation (3.5).

3.5 CARRIER GENERATION AND RECOMBINATION

At a finite temperature, there are always some electrons in the conduction band and some holes in the valence band. These carriers arise from ionized impurities and from the excitation of electrons from valence to conduction band, which create electron-hole pairs. Near and above room temperature, to good

approximation, all impurities are ionized, and so here we consider only those processes involving electron-hole pairs. By definition, a process of creating electron-hole pairs or exciting an electron from valence band to conduction band is referred to as *generation*. A process by which an electron from the conduction band is moved to the valence band, thus annihilating an electron-hole pair, is called *recombination*.

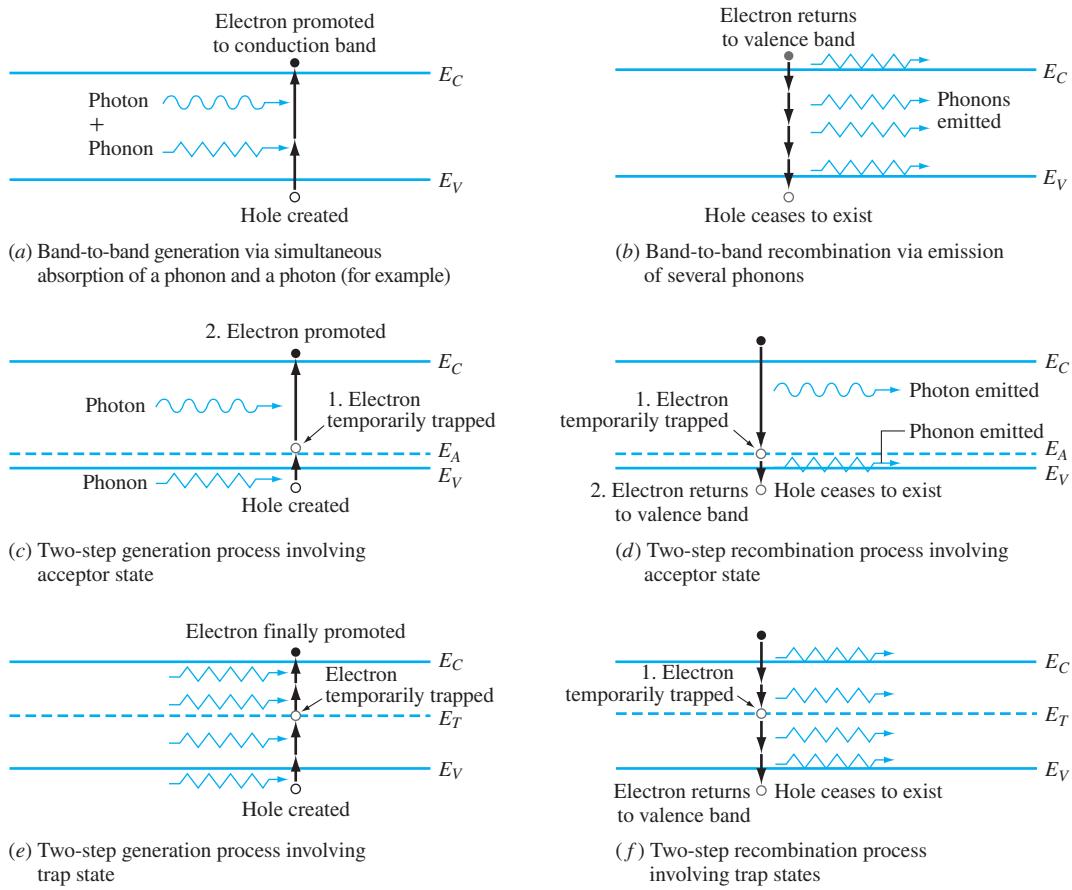


Figure 3.12 Various generation and recombination processes. (a) An electron-hole pair is generated when an electron absorbs (in this case) a phonon plus a photon. This generation could also occur by the absorption of a single photon. The photons and phonons are absorbed simultaneously. (b) Band-to-band recombination via the emission of multiple phonons. (c) A two-step generation process, in which, for example, the electron absorbs a phonon to promote it to the acceptor state, then in the next step it absorbs a photon to go to the conduction band. (d) A typical recombination event in p-type material involves emission of a photon to take the electron temporarily to the acceptor level, then the subsequent emission of the phonon returns it to the valence band, annihilating a hole. (e) and (f) Recombination and generation via trap states.

At equilibrium, generation and recombination in a semiconductor occur at equal rates, and thus the equilibrium concentration of electrons and holes (n_0 and p_0) is constant. However, semiconductor devices operate under nonequilibrium conditions, in which case this is not necessarily true, and $n \neq n_0$ and $p \neq p_0$. In many devices the processes of generation and recombination are important in determining their electrical characteristics. There are a number of physical processes involved in generation and recombination. Here we discuss the more common processes.

3.5.1 BAND-TO-BAND GENERATION AND RECOMBINATION

An electron can be excited from valence band to conduction band by the absorption of a photon having energy greater than the band gap (e.g., GaAs), by the simultaneous absorption of a photon and a phonon (e.g., Si). The process of band-to-band transitions is illustrated in Figure 3.12a. In this process an electron-hole pair is generated. Figure 3.12b illustrates the inverse process. When an electron recombines with a hole, then a photon, a photon plus a phonon, or multiple phonons (as illustrated) are generated. In either case, whether generation or recombination, energy and wave vector (crystal momentum) must be conserved.

3.5.2 TWO-STEP PROCESSES

Generation and recombination can also occur by a two-step process involving electronic states within the forbidden band. Figure 3.12c indicates the case for generation in p-type GaAs. Phonons or photons excite electrons from the valence band to acceptor levels. Photons can then excite the electrons from acceptor levels to the conduction band. Figure 3.12d indicates the recombination process. The transitions from conduction band to acceptor states emit photons while the transition from acceptor to valence band emits phonons.

Generation and recombination in Si is often by a two-step process [5], involving an energy level, called a *trap level*, of energy E_T , near the center of the forbidden band as indicated in Figures 3.12e and f.

3.6 OPTICAL PROCESSES IN SEMICONDUCTORS

Until now, we have been primarily considering semiconductors for which the carrier concentrations have their equilibrium values. In devices, however, often carrier concentrations have nonequilibrium values. For example, when a semiconductor is exposed to and absorbs light, (for example in a solar cell), extra electrons and holes are produced in excess of their equilibrium concentrations. In this section, we discuss the optical processes of absorption and emission in semiconductors and the effects of these excess carrier concentrations on current.

*3.6.1 ABSORPTION

In Chapter 1, we discussed absorption of photons by an atom. A single atom has discrete energy states, and for a photon to be absorbed, its energy must be equal to the difference in energy between two allowed states. In a semiconductor, the process is similar, except that, instead of isolated energy states, there exist allowed energy bands. Thus, there will be a range of photon energies that can be absorbed. We will see that there are further restrictions, however, based on the E - K diagrams.

Consider a semiconductor material whose energy band diagram is shown in Figure 3.13a. The bottom of the conduction band and the top of the valence band are shown. At equilibrium, there are n_0 electrons in the conduction band and p_0 holes in the valence band. Electron-hole pairs are constantly being generated thermally, and electrons are constantly recombining with holes as they seek lower energies. At thermal equilibrium, these two processes happen at exactly the same rate.

Now assume a photon of energy

$$E = h\nu \quad (3.50)$$

arrives at the semiconductor as in Figure 3.13b, where ν is the frequency of the light and h is Planck's constant. We already know that if the photon is to be absorbed, by conservation of energy the energy of the *system* must be the same before and after the absorption event. Before the collision of the electron and the photon, their combined energy is

$$E_{\text{electron+photon}} = E_0 + h\nu \quad (3.51)$$

where E_0 is the initial electron energy. When the photon collides with the electron, it is possible that the electron will absorb the energy of the photon. When the

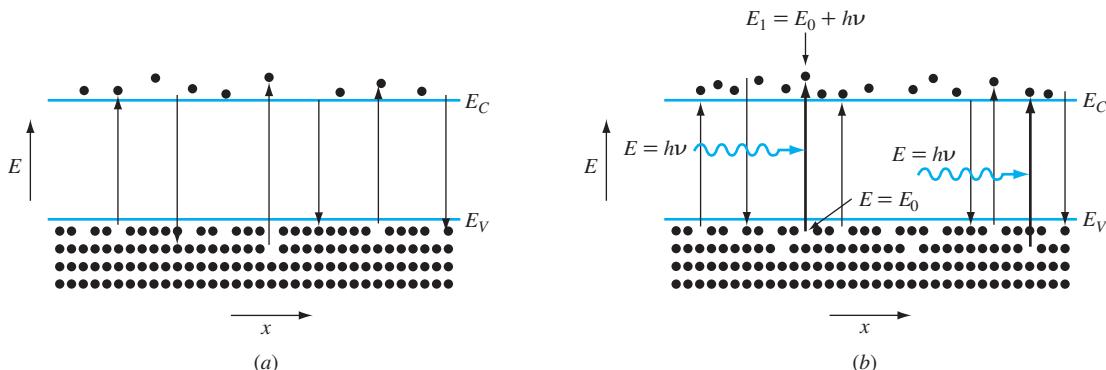


Figure 3.13 (a) At equilibrium, electrons and holes are generated and destroyed at equal rates, thus maintaining some constant equilibrium n_0 and p_0 . (b) When light shines on the sample, the photons can be absorbed, producing extra electron-hole pairs.

photon energy is transferred to the electron, the photon is annihilated. Therefore, after the event the electron, by conservation of energy, now has energy

$$E_1 = E_0 + h\nu \quad (3.52)$$

where $E_1 > E_C$.

Notice that if a photon of energy $E = h\nu$ less than the band gap is incident, that photon cannot be absorbed. For it to be absorbed, the electron would have to end up at an energy state somewhere inside the forbidden gap. Except through the help of traps or impurities that might provide the occasional state within the forbidden gap, this is not possible. Therefore, light of photon energy less than the band gap is not absorbed, and the semiconductor appears transparent to that radiation.

We have not told the whole story, however. Aside from conservation of energy, there is also the law of conservation of wave vector K (analogous to conservation of momentum in classical mechanics).⁵ Photons of interest in semiconductor electronics⁶ have wave vectors that are small compared with those at the edge of the Brillouin zone, and for many cases they can be considered to be essentially zero. Therefore, when the electron makes the energy transition, it must do so at virtually constant K , which is to say it must make a (nearly) vertical transition on the E - K diagrams. This is shown in Figure 3.14.

The particular semiconductor whose E - K diagram is shown in Figure 3.14a is called a direct gap material. *Direct* means that the minimum in the conduction band is at the same value of K as the maximum in the valence band, here at $K = 0$. Thus, the transition results from the direct interaction of a photon with an electron. Direct gap materials are generally efficient emitters and absorbers of optical energy because it is easy for electrons to move between the conduction band and valence band without having to acquire or give off K . GaAs and InP are good examples of direct gap materials.

An indirect gap material, on the other hand, is one like that shown in Figure 3.14b. In an indirect material, the minimum in the conduction band is not at the same value of K as the top of the valence band. An electron cannot go from one band to the other simply by absorbing a photon of energy close to the band gap because the photon cannot supply adequate wave vector. The electron needs to acquire both energy and wave vector to make the transition in indirect materials. Examples of indirect materials are Si, Ge, and GaP.

This raises an interesting point. Silicon is the material most commonly used in photodetectors and solar cells; yet it is an indirect gap material and thus not an efficient absorber of light. Furthermore, its primary band gap is 1.12 eV, which corresponds to a photon wavelength of just over 1 μm —in the

⁵As indicated earlier, the quantity $\hbar K$ is often referred to as *crystal momentum* (or sometimes just *momentum*).

⁶By this we mean radiation in the ultraviolet to infrared range—not X-rays.

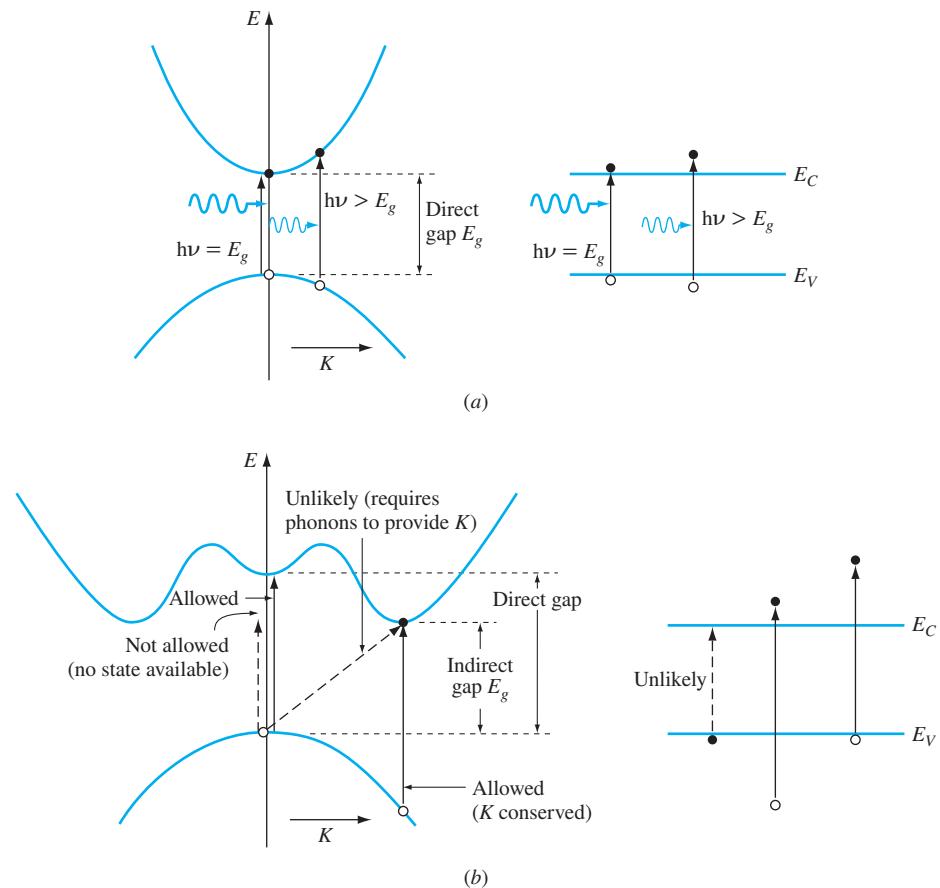


Figure 3.14 For absorption to occur, K must be conserved as well as E . (a) A direct gap semiconductor; on the left is the E - K diagram, and on the right the conventional energy band diagram. (b) An indirect gap material (so called because conduction band minimum and the valence band maximum do not occur at the same value of K and thus the photon-electron interaction is indirect).

near infrared. According to Figure 3.14b it should not be able to absorb light at all until considerably higher energies. Photons of energy less than that of the direct band gap can be absorbed, however, via a three-particle collision involving an electron, a photon, and a phonon. Phonons have adequate K vector but have little energy, while the opposite is true for photons. If a photon and a phonon collide with an electron at the same time, the electron can get both enough energy and enough K to get across the forbidden gap. Such a three-body collision is statistically unlikely, and as a result, silicon is not a very efficient absorber at wavelengths near the band gap. Still, it is widely used because of its low cost.

EXAMPLE 3.4

Verify that photons of interest have negligible wave vector compared with that of electrons at the edge of the Brillouin zone.

■ Solution

We consider a photon of wavelength 620 nm (energy 2 eV) (orange). It has wave vector

$$K_{\text{photon}} = \frac{2\pi}{\lambda} = \frac{6.28}{620 \times 10^{-9} \text{ m}} = 10.1 \times 10^6 \text{ (m)}^{-1}$$

We recall from Chapter 2 that an electron at the edge of the first Brillouin zone has a wave vector of π/a . For a lattice constant of $a = 0.5 \text{ nm}$,

$$K_e = \frac{\pi}{a} = \frac{3.14}{0.5 \times 10^{-9} \text{ m}} = 6.28 \times 10^9 \text{ m}^{-1}$$

Then the ratio of the K vectors is

$$\frac{K_{\text{photon}}}{K_e} = \frac{10.1 \times 10^6}{6.28 \times 10^9} = 1.6 \times 10^{-3}$$

Thus, the K vector of the photon is about 1/1000 that of the electron. Note, however, that high-energy (X-ray) photons can have wave vectors comparable to those of electrons at the edge of the Brillouin zone.

When light shines on a semiconductor, equilibrium is disturbed. When a given photon is absorbed, an extra electron (beyond the equilibrium number) is produced in the conduction band, and an extra hole is simultaneously produced in the valence band. An electron-hole pair is thus produced for every photon that is absorbed. These electrons and holes are termed *excess carriers*—excess above the equilibrium concentrations. The excess electron concentration is Δn and the excess hole concentration is Δp . The total carrier concentrations are the equilibrium values plus the excess concentrations:

$$\begin{aligned} n &= n_0 + \Delta n \\ p &= p_0 + \Delta p \end{aligned} \tag{3.53}$$

These excess carriers can diffuse if they are not uniformly distributed, and they can drift if there is an electric field. Even if the drift and diffusion currents are zero (or add to zero), as long as there are excess carriers, the material is not at equilibrium. We will see shortly that if the light is turned off, the excess carriers disappear with time and the material returns to equilibrium. Before we discuss that, however, we will examine optical emission in semiconductors.

3.6.2 EMISSION

The inverse of the optical process of absorption is optical emission. In this process, an electron initially in a state of energy E_1 makes the transition to a state of

lower energy E_0 as shown in Figure 3.15. By conservation of energy, the excess energy must be given off either as a photon (light) or phonons (heat) or both.

In the absence of phonon creation, the energy of the photon emitted must be equal to the energy lost by the electron. Therefore, we expect to see emission only at photon energies equal to the difference between allowed electron energies. We consider the case of a p-type direct gap semiconductor as shown in Figure 3.16a. At energies greater than that of the gap, emission is possible, but K must still be conserved. Therefore, optical transitions must still occur vertically on the E - K diagram. For energies close to, but greater than, the band gap, optical emission can occur only in direct gap semiconductors. Optical transitions are also possible, however, from conduction band to acceptor level, producing lower-energy photons. In this case, the accompanying transition from E_A to valence band produces phonons. For narrow-gap materials (e.g., InSb, $E_g = 0.17$ eV), band-to-band transitions are the more probable. For wider-gap materials (e.g., GaAs, $E_g = 1.43$ eV) conduction band-to-acceptor state transitions are more probable.

An indirect energy band diagram for a material such as GaP is shown in Figure 3.16b. Notice that one transition is marked as allowed but unlikely. This is because, although K would be conserved, the probability of a hole existing at this low energy is remote. The other transition shown is not permitted since K is

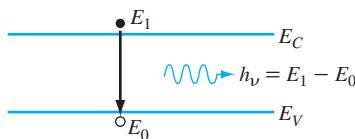


Figure 3.15 Optical emission.

The electron loses energy, giving off the excess as a photon of $E = h\nu$.

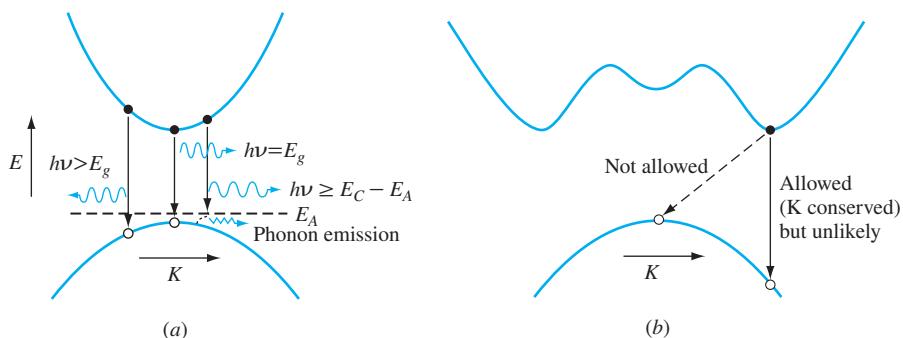


Figure 3.16 Emission on the E - K diagram. Both K and E must be conserved. (a) A direct gap material; (b) an indirect semiconductor.

not conserved. As a result, indirect gap materials are less efficient light emitters than direct gap materials.⁷

3.7 CONTINUITY EQUATIONS

The continuity equations are mathematical statements of the conservation of particles. They are fundamental to an understanding of the variations of carrier concentrations (electrons and holes) with time and position and their effects on the electrical and electro-optical characteristics, both static and transient. For simplicity, we consider the time dependence of the carrier concentrations in the x direction for a one-dimensional case. It is thus assumed that at a given time the carrier concentrations are uniform in the y and z directions.

In a metal wire, we find the current by examining the number of electrons entering or leaving a volume per unit time. In a semiconductor, both electrons and holes can enter a volume at one end and leave at the other, and we must account for both. Carriers may also pile up or disappear (recombine) in the volume. Therefore their numbers leaving at any given time may not be the same as the numbers entering. The continuity equation takes into account all the sources and sinks of electrons and holes.

Consider the semiconductor differential volume of unit cross-sectional area and length dx , as shown in Figure 3.17. An electron flux density F_n , composed of drift and diffusion, is flowing into the volume, and the flux density flowing out will in general be different. The difference is made up by the generation and recombination and the pile-up or decrease of carriers in the volume. In this differential volume, then, the rate of increase in conduction band electron concentration is

$$\frac{\partial n}{\partial t} dx = -\frac{\partial F_n}{\partial x} dx + (G_n - R_n) dx \quad (3.54)$$

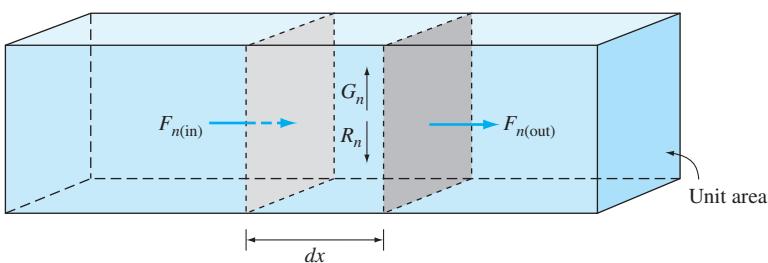


Figure 3.17 The geometry for determining the continuity equation. The rate at which carriers accumulate in the incremental volume depends on the incoming and outgoing currents as well as the recombination and generation within the region dx .

⁷There are exceptions. For example, by doping an indirect gap semiconductor appropriately, producing *isoelectronic traps* as discussed in Online Module OM11, substantial light emission from some indirect gap materials (e.g., GaP) can be obtained.

where G_n is the electron generation rate and R_n is the electron recombination rate (number of carriers per unit volume per unit time). The term $(\partial F_n / \partial x)dx$ represents the difference in electron flux at either end of the region dx .

We saw there are several mechanisms for generation and recombination. Let G_{th} be the thermal (phonon-induced) generation rate, G_{op} be the optical (photon induced) rate and G_{other} be the generation rate due to other processes (e.g., trapping, detrapping). Then the total electron generation rate is

$$G_n = G_{n(\text{th})} + G_{n(\text{op})} + G_{n(\text{other})} \quad (3.55)$$

measured in electrons generated per unit volume per second. Since the electron flux density is related to the current density by $F_n = -J_n/q$, Equation (3.54) becomes, after the dx terms are canceled,

$$\frac{\partial n}{\partial t} = \frac{1}{q} \frac{\partial J_n}{\partial x} + (G_n - R_n) \quad \text{continuity equation for electrons} \quad (3.56)$$

This equation states that the rate of increase in electron density at a point is equal to the increase in n due to unequal electron currents, plus the net electron generation rate $G_n - R_n$. We reiterate that J_n is the total electron current density and in general consists of drift and diffusion current. Equation (3.56) is called the *continuity equation for electrons*. Similarly, the continuity equation for holes is

$$\frac{\partial p}{\partial t} = -\frac{1}{q} \frac{\partial J_p}{\partial x} + (G_p - R_p) \quad \text{continuity equation for holes} \quad (3.57)$$

Next, returning to Equation (3.56), we recall that $n = n_0 + \Delta n$ [Equation (3.53)], and that the equilibrium electron concentration n_0 is time-independent, or $\partial n_0 / \partial t = 0$. Thus

$$\frac{\partial n}{\partial t} = \frac{\partial n_0}{\partial t} + \frac{\partial \Delta n}{\partial t} = \frac{\partial \Delta n}{\partial t} \quad (3.58)$$

Further, neglecting other processes,

$$G_n = G_{\text{th}} + G_{\text{op}} \quad (3.59)$$

Recombination also occurs. An electron has some lifetime associated with it, which is the average time an electron spends in the conduction band, before recombining. We will discuss the lifetimes in detail in Section 3.8, but for now, for a well-defined⁸ minority carrier (electron) lifetime τ_n , the recombination rate is proportional to the number of electrons available for recombination (meaning electrons in the conduction band) by

$$R_n = \frac{n}{\tau_n} = \frac{n_0}{\tau_n} + \frac{\Delta n}{\tau_n} \quad (3.60)$$

⁸By *well-defined* we mean that the lifetime is the average time that a minority carrier spends in its respective band, $\bar{\tau}_n = \tau_n$ or $\bar{\tau}_p = \tau_p$. This is usually the case for semiconductors used in electronic devices.

Substituting these results back into Equation (3.56) gives

$$\frac{\partial n}{\partial t} = \frac{\partial \Delta n}{\partial t} = \frac{1}{q} \frac{\partial J_n}{\partial x} + \left(G_{\text{th}} + G_{\text{op}} - \frac{n_0}{\tau_n} - \frac{\Delta n}{\tau_n} \right) \quad (3.61)$$

We will consider the special case of equilibrium. Equilibrium means that there is no net current ($J = 0$), there are no external fields applied, there are no temperature gradients, and no light is shining on the sample. Equilibrium therefore means there are no excess carriers. For equilibrium, then, $\Delta n = 0$, $J_n = 0$, and $G_{\text{op}} = 0$. Thus the left-hand side of Equation (3.61) is equal to zero, resulting in

$$G_{\text{th}} = \frac{n_0}{\tau_n} \quad (3.62)$$

Equation (3.61) then becomes, with the aid of Equation (3.62),

$$\boxed{\frac{\partial n}{\partial t} = \frac{\partial \Delta n}{\partial t} = \frac{1}{q} \left(\frac{\partial J_n}{\partial x} \right) + \left(G_{\text{op}} - \frac{\Delta n}{\tau_n} \right)} \quad (3.63)$$

Similarly for holes,

$$\boxed{\frac{\partial p}{\partial t} = \frac{\partial \Delta p}{\partial t} = -\frac{1}{q} \left(\frac{\partial J_p}{\partial x} \right) + \left(G_{\text{op}} - \frac{\Delta p}{\tau_p} \right)} \quad (3.64)$$

where, in general, G_{op} , Δn , and Δp are functions of position.

The current density consists of drift and diffusion

$$\begin{aligned} J_n &= J_{n(\text{drift})} + J_{n(\text{diff})} = qn\mu_n \mathcal{E} + qD_n \frac{dn}{dx} \\ J_p &= J_{p(\text{drift})} + J_{p(\text{diff})} = qp\mu_p \mathcal{E} - qD_p \frac{dp}{dx} \end{aligned} \quad (3.65)$$

and the continuity equations become

$$\boxed{\begin{aligned} \frac{\partial n}{\partial t} &= \frac{\partial \Delta n}{\partial t} = n\mu_n \frac{\partial \mathcal{E}}{\partial x} + \mu_n \mathcal{E} \frac{\partial n}{\partial x} + D_n \frac{\partial^2 n}{\partial x^2} + G_{\text{op}} - \frac{\Delta n}{\tau_n} \\ \frac{\partial p}{\partial t} &= \frac{\partial \Delta p}{\partial t} = -p\mu_p \frac{\partial \mathcal{E}}{\partial x} - \mu_p \mathcal{E} \frac{\partial p}{\partial x} + D_p \frac{\partial^2 p}{\partial x^2} + G_{\text{op}} - \frac{\Delta p}{\tau_p} \end{aligned}} \quad (3.66)$$

The use of these continuity equations will be illustrated in the next two sections for simple one-dimensional problems. In the more complicated cases, such as transistor analysis, the continuity equations in two or three (spatial) dimensions are often required. Often these cannot be solved in closed form, but are solved numerically in device simulators.

In three dimensions the continuity equations are

$$\frac{\partial n}{\partial t} = \frac{1}{q} \nabla \cdot J_n + (G_n - R_n)$$

$$\frac{\partial p}{\partial t} = -\frac{1}{q} \nabla \cdot J_p + (G_p - R_p)$$

3.8 MINORITY CARRIER LIFETIME

We have seen that electrons and holes can be generated thermally or optically and that there is also a restoring force: recombination. When electrons are found at elevated energies (e.g., in the conduction band), they tend to seek lower energies, and after some time they will recombine with holes in the valence band. In this section, we are interested in the time that takes.

We define the *minority carrier lifetime* τ as a measure of the average time, (τ_n) , an electron spends in the conduction band in a p-type semiconductor or a hole (τ_p) spends in the valence band in an n-type semiconductor. One method used to determine the lifetime is to measure the time dependence of photoconductivity when a semiconductor is illuminated.

We consider a uniform semiconductor sample connected in the circuit of Figure 3.18. The semiconductor is thin, and the wavelength of the light is chosen such that the absorption coefficient is small—this way the illumination can be considered uniform throughout the sample. If light pulses are applied, they create electron-hole pairs through absorption. The absorption creates excess carriers, and since there are more carriers to conduct, the conductivity increases. The resulting variation in current with time is measured with an oscilloscope as a time variation of voltage across the load resistance R_L . If the value of R_L is chosen to be small compared with the resistance R_S of the semiconductor, then the measured current is V_A/R_S . Recalling that the resistance of a semiconductor sample is $R_S = \rho L/A$, and $\sigma = 1/\rho$, we have

$$i(t) = \frac{V_A}{\rho(t)L} = \frac{V_A A \sigma(t)}{L} \quad (3.67)$$

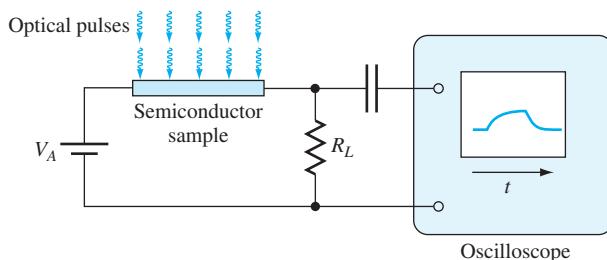


Figure 3.18 Schematic of a circuit used to measure minority carrier lifetime in semiconductors.

The time-dependent voltage is then proportional to $i(t)$, which in turn is proportional to the conductivity $\sigma(t)$, which changes in time as the number of carriers changes.

Recall that the conductivity of a semiconductor sample is given by

$$\sigma = \sigma_n + \sigma_p = q(\mu_n n + \mu_p p) \quad (3.14)$$

where

$$\begin{aligned} n &= n_0 + \Delta n \\ p &= p_0 + \Delta p \end{aligned} \quad (3.53)$$

Again, n_0 and p_0 are the equilibrium concentrations of electrons and holes while Δn and Δp are the excess (in this case, photoinduced) concentrations. We can combine Equations (3.14) and (3.53) to obtain

$$\sigma = q(\mu_n n_0 + \mu_n \Delta n + \mu_p p_0 + \mu_p \Delta p) \quad (3.68)$$

or

$$\sigma = \sigma_0 + \sigma_{pc} \quad (3.69)$$

Here the conductivity in the dark is

$$\sigma_0 = q(\mu_n n_0 + \mu_p p_0) \quad (3.70)$$

and the photoconductivity is

$$\sigma_{pc} = q(\mu_n \Delta n + \mu_p \Delta p) \quad (3.71)$$

As an example, we consider a p-type direct gap semiconductor in which recombination is band to band; i.e., electrons in the conduction band recombine directly with holes in the valence band. This is shown schematically at the right-hand end of Figure 3.19.

Since each photon creates an electron-hole pair, and since each recombining electron annihilates an electron-hole pair, the concentration of excess electrons

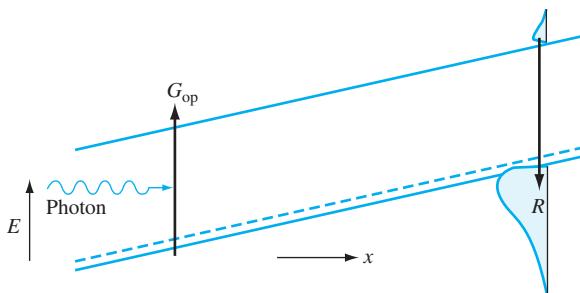


Figure 3.19 Energy band diagram of the semiconductor of Figure 3.18, under electrical bias and optical illumination. The combination rate R and the optical generation rate G_{op} are illustrated.

and holes is the same, or $\Delta n = \Delta p$. The rates at which carriers are being generated may not be the same as the rate at which they recombine, in which case carriers can accumulate or deplete.

The time dependence of the photocurrent is proportional to the time dependence of Δn and Δp . Since $\Delta n = \Delta p$, we solve the continuity equation [Equation (3.63) or (3.66)] for Δn . Because the semiconductor is uniform and uniformly illuminated, \mathcal{E} is constant and therefore $d\mathcal{E}/dx$, dn/dx , and thus $\partial J_n/\partial x$ are all equal to zero. Then

$$\frac{d\Delta n}{dt} = G_{\text{op}} - \frac{\Delta n}{\tau_n} \quad (3.72)$$

where the partial derivative is replaced by the derivative. Equation (3.72) must be solved for Δn with the initial conditions appropriate to the particular problem.

The photoconductivity is determined from Equation (3.71) or, since $\Delta p = \Delta n$,

$$\sigma_{pc}(t) = q\Delta n(t)(\mu_n + \mu_p)$$

Then from Equation (3.67), the time dependence of the current depends on $\Delta n(t)$. We will start with the case in which the semiconductor is initially at equilibrium, and the light is then turned on at $t = 0$.

3.8.1 RISE TIME

Before the light is turned on, the semiconductor is at equilibrium. Thus at $t = 0$, $\Delta n = 0$. In this case, the solution to Equation (3.72) for $t > 0$ is

$$\Delta n = G_{\text{op}}\tau_n(1 - e^{-t/\tau_n}) \quad (3.73)$$

This is an increasing function with time and increases with some characteristic time τ_n .

Figure 3.20a shows a plot of $\Delta n(t)$ for this case. We see that Δn increases with a time constant τ_n and reaches a maximum of

$$\Delta n_{\text{max}} = G_{\text{op}}\tau_n \quad (3.74)$$

The maximum is proportional to G_{op} , which is related to the light intensity and absorption rate, and to τ_n . When the light is turned on, it takes some time for the steady-state excess carrier concentration to be established (the rise time).

3.8.2 FALL TIME

Now let us assume the light is extinguished at some time $t = t_0$. After t_0 , the optical generation rate goes to zero. Inserting $G_{\text{op}} = 0$ into Equation (3.72) results in the solution

$$\Delta n(t) = \Delta n(t_0) e^{-(t - t_0)/\tau_n} \quad (3.75)$$

or Δn decays with the same time constant τ_n . For an n-type semiconductor, with similar approximations, Δn and τ_n are replaced by Δp and τ_p respectively.

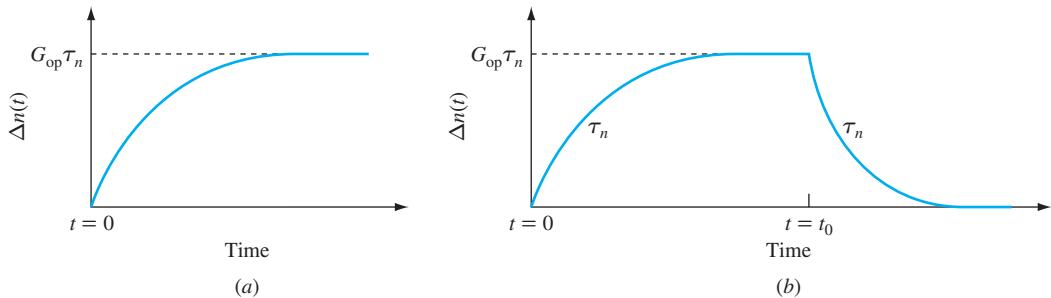


Figure 3.20 Variation of excess carriers in a semiconductor under pulsed illumination. (a) When the light is turned on, the excess carrier concentration increases exponentially. For the complete pulse (b), the rise and fall time constants are equal to the minority carrier lifetimes.

From Equation (3.60), the recombination rate is given by $R = n/\tau_n$. For a p-type direct gap semiconductor with $n \ll N_A$, which is usually the case,

$$R = n\beta p = n\beta N_A$$

where β is the probability that a given electron will recombine with a hole in unit time. Then

$$\tau_n = \frac{1}{\beta N_A}$$

or τ_n varies inversely with N_A . The heavier the doping, the shorter the lifetime. In a p-type indirect gap semiconductor (e.g., Si), however, recombination is via trap states within the forbidden band. Thus, R is limited by the density of the trap states N_T .

We have seen that while the light is on, the excess carrier concentration (and along with it the photoconductivity and photocurrent) reaches some maximum value that depends on G_{op} , which is proportional to the light intensity. The maximum current is also proportional to the minority carrier lifetime, from Equations (3.67), (3.71), and (3.74). Thus in a photodetector, the sensitivity increases as the carrier lifetime gets longer. The rise and fall times are also proportional to τ_n , however, so for high speed of response, one sacrifices sensitivity.

EXAMPLE 3.5

Show that for a p-type semiconductor the average time an electron spends in the conduction band is equal to the time constant τ_n .

Solution

Consider the photoconductivity experiment of Figure 3.18, and suppose that at $t = t_0 = 0$ the illumination is turned off. Then Equation (3.75) becomes

$$\Delta n(t) = \Delta n(0)e^{-t/\tau_n}$$

The average time \bar{t}_n that an electron spends in the conduction band is given by

$$\bar{t}_n = \frac{\int_0^\infty t \frac{dn}{dt} dt}{\int_0^\infty \frac{dn}{dt} dt}$$

We can substitute for dn/dt by using $dn/dt = d\Delta n/dt$. We then have that

$$\frac{dn}{dt} = \frac{d\Delta n}{dt} = \Delta n(0) \left(\frac{-1}{\tau_n} \right) e^{-t/\tau_n},$$

and thus in the denominator

$$\int_0^\infty \frac{dn}{dt} dt = \Delta n(0) \left[\frac{-1}{\tau_n} \right] \int e^{-t/\tau_n} dt = \Delta n(0)$$

For the integral in the numerator,

$$\int_0^\infty t \frac{dn}{dt} dt = \int_0^\infty t \frac{d\Delta n}{dt} dt = \Delta n(0) \tau_n$$

Thus

$$\bar{t}_n = \frac{\int_0^\infty t \frac{dn}{dt} dt}{\int_0^\infty \frac{dn}{dt} dt} = \frac{\Delta n(0) \tau_n}{\Delta n(0)} = \tau_n$$

As indicated above, the minority carrier lifetime is the average time an electron spends in the conduction band in p-type material or, in n-type material, the average time a hole spends in the valence band. Minority carrier lifetime has been measured as a function of doping in uncompensated Si; there is considerable scatter in the data. At high doping levels where the lifetimes are short, minority carrier lifetimes become difficult to measure. At low doping, the lifetime is strongly dependent on the trap concentration. Minority carrier lifetimes have been fitted to the empirical expressions [3, 4]

$$\tau_n = [3.45 \times 10^{-12} N_A + 9.5 \times 10^{-32} N_A^2]^{-1} \quad (3.76)$$

$$\tau_p = [7.8 \times 10^{-13} N_D + 1.8 \times 10^{-31} N_D^2]^{-1} \quad (3.77)$$

where N_A and N_D are expressed in cm^{-3} and τ_n and τ_p are in seconds. These lifetimes are plotted against doping level from Equations (3.76) and (3.77) in

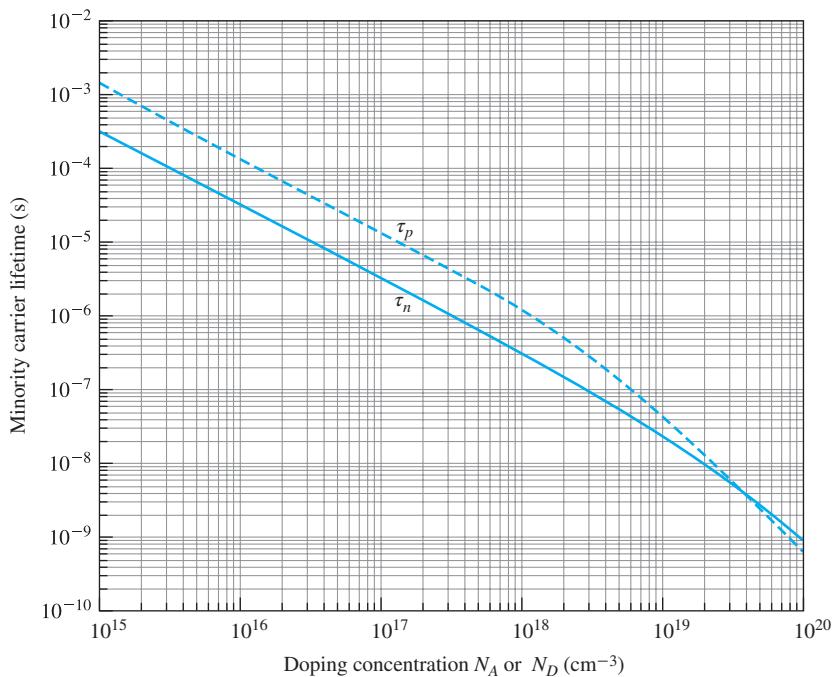


Figure 3.21 Plot of minority carrier lifetime in uncompensated high quality Si as a function of doping concentration N_A or N_D .

Figure 3.21. We can see that for high-purity Si, the minority carrier lifetimes are in the millisecond range.⁹ They decrease with increased doping and for a concentration of 10^{20} cm^{-3} are on the order of a nanosecond.

3.9 MINORITY CARRIER DIFFUSION LENGTHS

We have seen that recombination causes the carriers to have a finite lifetime. During that lifetime, they can also diffuse, since they will always have some kinetic energy. In this section, we will see how far a carrier diffuses, on the average, before it recombines.

Here we consider a p-type semiconductor that is illuminated from one side as shown in Figure 3.22a. The illumination is steady state, with photons of absorption coefficient high enough that for practical purposes they all can be considered to be absorbed at the surface. We also assume that there is no electric field in the semiconductor. The excess electrons generated at the surface then

⁹There is considerable scatter in the measured lifetimes of lightly doped Si because of the wide variations in trap concentrations. The plots in Figure 3.21 represent the highest measured values (lowest trap densities).

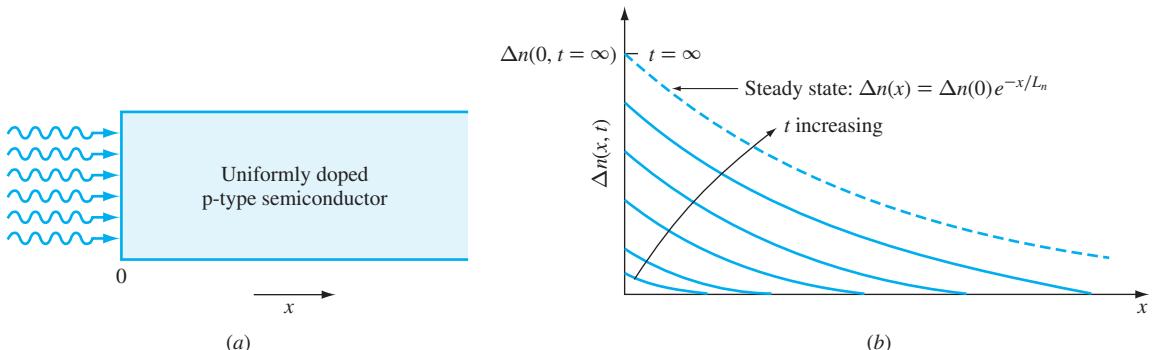


Figure 3.22 (a) Illustration of minority carrier diffusion in a surface-illuminated p-type semiconductor. The absorption is assumed to occur at the surface. (b) Plots of the excess minority carrier concentration as a function of distance into the bar with increasing time. As the excess carriers are generated at the surface, they diffuse to regions of lower concentration, where they recombine.

will diffuse into the regions of lower electron concentration—deeper into the semiconductor. As they penetrate, they will recombine with holes. Figure 3.22b indicates the variation of the excess electrons as a function of position with time as a parameter. With increasing time after the light is turned on, the electron concentration at the surface increases. The electrons diffuse into the bulk and recombine. The steady-state condition is reached when the generation rate at the surface is equal to the recombination rate in the bulk.

We can determine Δn as a function of x and t by solving the continuity equation for electrons [Equation (3.63)]. This can be simplified for the steady-state condition since then $\partial n/\partial t = 0$. Further, since all excess electrons are generated at the surface, for $x > 0$, $G_{op} = 0$, and since $\mathcal{E} = 0$, $J_{n(drift)} = 0$. Equation (3.63) then becomes

$$\frac{dJ_n}{dx} = q \frac{\Delta n}{\tau_n} \quad (3.78)$$

where the derivative replaces the partial derivative because there is no variation with time. Since

$$J_{n(diff)} = qD_n \frac{dn}{dx} = qD_n \left(\frac{dn_0}{dx} + \frac{d\Delta n}{dx} \right)$$

and n_0 is constant, Equation (3.78) becomes

$$qD_n \frac{d^2(\Delta n)}{dx^2} = q \left(\frac{\Delta n}{\tau_n} \right) \quad (3.79)$$

or

$$\frac{d^2(\Delta n)}{dx^2} = + \frac{\Delta n}{D_n \tau_n} \quad (3.80)$$

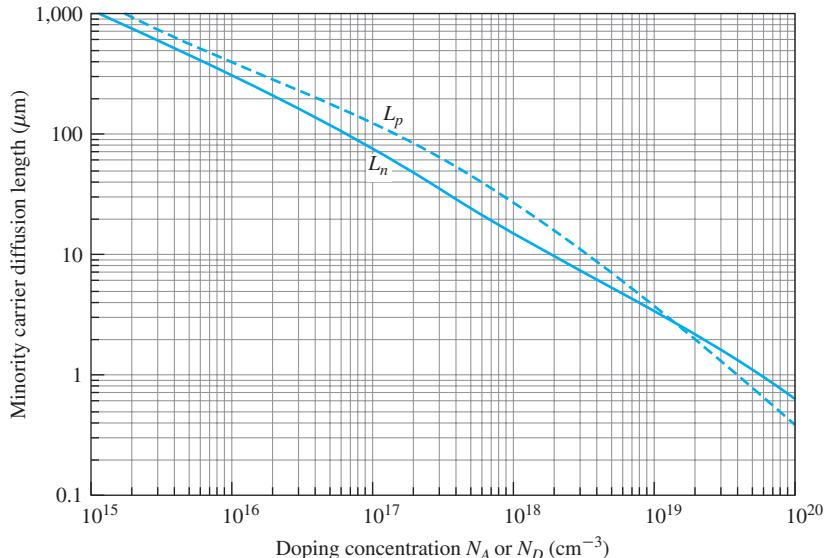


Figure 3.23 Minority carrier diffusion lengths L_n and L_p as functions of impurity concentration N_A or N_D in uncompensated high-quality Si.

This is a second-order differential equation, which can be solved with the boundary conditions $\Delta n = \Delta n(0)$ at $x = 0$ and $\Delta n = 0$ at $x = \infty$. The solution is

$$\Delta n(x) = \Delta n(0) e^{-x/\sqrt{D_n \tau_n}} = \Delta n(0) e^{-x/L_n} \quad (3.81)$$

Here we have introduced a new quantity called the diffusion length:

$$L_n = \sqrt{D_n \tau_n} \quad (3.82)$$

It is the average distance an electron diffuses before it recombines. Similarly the diffusion length for holes is

$$L_p = \sqrt{D_p \tau_p} \quad (3.83)$$

In more heavily doped Si, both D_n and τ_n are reduced below their intrinsic values, and thus so is L_n . The minority carrier diffusion lengths for Si can be calculated from Equations (3.82) and (3.83), using the minority carrier diffusion coefficients as expressed by the Einstein relation $D/\mu = kT/q$ and Equations (3.18) and (3.19), as plotted in Figure 3.11. The minority carrier lifetimes come from Equations (3.76) and (3.77). These minority carrier diffusion lengths L_n and L_p are plotted as functions of doping in Figure 3.23 for Si. It can be seen that the diffusion lengths reduce from about 1 mm at 10^{15} impurities per cm^3 to less than 1 μm at 10^{20} cm^{-3} .

The diffusion lengths in direct gap semiconductors are appreciably smaller than this, since their carrier lifetimes are much less.

3.10 QUASI FERMI LEVELS

For a semiconductor at equilibrium, the Fermi level is constant and its value can be used as a measure of the equilibrium concentrations of electrons and holes. In this section, we will introduce the *quasi Fermi level*, which can be used as a reference to find the concentrations of electrons and holes when a material is *not* at equilibrium.

Recall that for a nondegenerate semiconductor

$$\begin{aligned} n_0 &= N_C e^{-(E_C - E_f)/kT} = n_i e^{(E_f - E_i)/kT} \\ p_0 &= N_V e^{-(E_f - E_v)/kT} = n_i e^{(E_i - E_f)/kT} \end{aligned} \quad \text{equilibrium} \quad (3.84)$$

It is useful to have similar expressions for electrons and holes for the non-equilibrium case. We can write analogues to Equations (3.84):

$$\begin{aligned} n &= N_C e^{-(E_C - E_{fn})/kT} = n_i e^{(E_{fn} - E_i)/kT} \\ p &= N_V e^{-(E_{fp} - E_v)/kT} = n_i e^{(E_i - E_{fp})/kT} \end{aligned} \quad \text{nonequilibrium} \quad (3.85)$$

where n and p are respectively the total electron and hole concentrations, including the equilibrium and excess carriers. Equation (3.85) in effect defines the quasi Fermi levels for electrons, E_{fn} , and for holes, E_{fp} .

Solving for the quasi Fermi levels, we find that Equation (3.85) becomes

$$\begin{aligned} E_{fn} &= E_C - kT \ln \frac{N_C}{n} = E_i + kT \ln \frac{n}{n_i} \\ E_{fp} &= E_V + kT \ln \frac{N_V}{p} = E_i - kT \ln \frac{p}{n_i} \end{aligned} \quad (3.86)$$

As an example, consider the illuminated p-type semiconductor of Figure 3.22 in steady state. At the dark end of the bar, $n = n_0$ and $p = p_0$, the material is in equilibrium, and the Fermi level is defined. At equilibrium $E_{fn} = E_{fp} = E_f$, as shown in Figure 3.24. Toward the illuminated end of the sample, both n and p increase from their equilibrium values, and the quasi Fermi levels separate as indicated. Since the material is p type, $p_0 \gg n_0$, and since $\Delta n = \Delta p$ at any position, there are more holes than electrons, so the addition of a fixed number of excess carriers has a smaller fractional effect on the total concentration of holes than on the electrons. Thus the change in E_{fp} relative to E_f is less than for E_{fn} .

Just as we need to know the equilibrium electron and hole concentrations to determine the Fermi energy, to obtain the quasi Fermi levels we must know the electron and hole concentrations in the nonequilibrium case. The use of quasi Fermi levels on an energy band diagram is a convenient way to represent the electron and hole concentrations as a function of position.

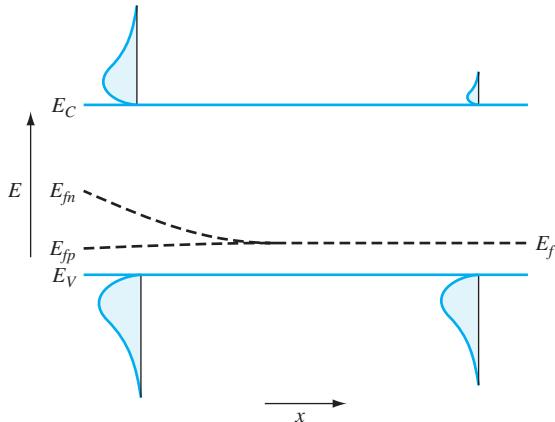


Figure 3.24 Illustration of quasi Fermi levels for electrons and holes for the steady-state nonequilibrium case of Figure 3.22, with $\mathcal{E} = 0$.

We emphasize that, just as Equation (3.84) is valid for nondegenerate semiconductors, Equations (3.85) and (3.86) are valid only for small Δn such that the quasi Fermi levels are greater than about $2.3 kT$ from their respective band edges.

In a nondegenerate semiconductor, the current densities J_n and J_p can be expressed in terms of the gradient of the quasi Fermi levels E_{fn} and E_{fp} respectively. The electron current can be expressed as in Equation (3.46):

$$J_n = q \mu_n \left[n(x) \mathcal{E}(x) + \frac{kT}{q} \frac{dn}{dx} \right] \quad (3.87)$$

But from Equation (3.85),

$$\begin{aligned} \frac{dn}{dx} &= \frac{d}{dx} [N_C e^{-(E_C - E_{fn})/kT}] \\ &= \frac{1}{kT} N_C e^{-(E_C - E_{fn})/kT} \left[\frac{dE_{fn}(x)}{dx} - \frac{dE_C(x)}{dx} \right] \\ &= \frac{n(x)}{kT} \left[\frac{dE_{fn}(x)}{dx} - \frac{dE_C(x)}{dx} \right] \end{aligned} \quad (3.88)$$

Since $dE_C(x)/dx = q\mathcal{E}(x)$, however, Equation (3.87) becomes

$$J_n(x) = \mu_n(x) n(x) \frac{dE_{fn}}{dx} \quad (3.89)$$

which can be expressed as

$$J_n(x) = q \mu_n(x) n(x) \frac{d}{dx} \left(\frac{E_{fn}}{q} \right) = \sigma_n(x) \frac{d}{dx} \left(\frac{E_{fn}}{q} \right) \quad (3.90)$$

This resembles Ohm's law in point form, with \mathcal{E} replaced by $d/dx(dE_{fn}/q)$.

Similarly,

$$J_p(x) = \sigma_p(x) \frac{d}{dx} \left(\frac{E_{fp}}{q} \right) \quad (3.91)$$

The above relations for J_n and J_p are valid for any combinations of drift and diffusion currents. For $dn/dx = dp/dx = 0$, there is no diffusion current and

$$\frac{d}{dx} \left(\frac{E_{fn}}{q} \right) = \frac{d}{dx} \left(\frac{E_{fp}}{q} \right) = \mathcal{E} \quad \text{no diffusion}$$

It should be emphasized again that the preceding use of quasi Fermi levels is valid only for nondegenerate semiconductors.

3.11 SUMMARY

We discussed two mechanisms of current flow in semiconductors. Current flows by carrier drift in an electric field, and by diffusion in the presence of a carrier concentration gradient.

In general, current flow is by a combination of drift and diffusion. The electron and hole current densities are

$$J_n = q\mu_n n \mathcal{E} + qD_n \frac{dn}{dx} = q\mu_n \left[n\mathcal{E} + \frac{kT}{q} \frac{dn}{dx} \right]$$

$$J_p = q\mu_p p \mathcal{E} - qD_p \frac{dp}{dx} = q\mu_p \left[p\mathcal{E} - \frac{kT}{q} \frac{dp}{dx} \right]$$

In the above, the mobilities and diffusion coefficients decrease with increasing doping concentrations. For either carrier type their values are larger if they are minority carriers than if they are majority carriers. The fractional difference increases with increased doping. For uncompensated Si, the values of μ can be found from Figure 3.4 and the values of D can be found from Figure 3.11. Mobilities and diffusion coefficients are related by the Einstein relation $D/\mu = kT/q$.

The carrier mobility is proportional to the mean free time between collisions, \bar{t}

$$\mu = \frac{q\bar{t}}{m^*}$$

At low fields, the mobility is independent of field. At high fields, it decreases with increasing field.

Illuminating a semiconductor with photons of energy greater than the band gap produces hole-electron pairs. Since photons have little wave vector, this produces a near-vertical transition in the electron $E-K$ diagram for a direct gap semiconductor. For an indirect gap semiconductor a phonon is required to furnish the change in wave vector required by the electron transition, making optical transitions less probable in indirect materials.

Minority carrier lifetimes and diffusion lengths are material dependent, but in any given material, both decrease with increasing impurity concentrations.

The continuity equations for electrons and for holes were introduced:

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

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

In three dimensions they become

$$\frac{\partial n}{\partial t} = \frac{1}{q} \nabla \cdot J_n + (G_n - R_n)$$

$$\frac{\partial p}{\partial t} = -\frac{1}{q} \nabla \cdot J_p + (G_p - R_p)$$

These equations are mathematical statements of the conservation of particles. While they are quite general, they can be solved in closed form for only relatively simple cases.

The minority carrier lifetime was defined as the average time τ_n an electron spends in the conduction band in a p-type semiconductor before recombining with a hole, or the average time τ_p a hole spends in the valence band in an n-type semiconductor. Analogously, the minority carrier diffusion lengths L_n and L_p are defined as the average distance a minority carrier will diffuse before recombining.

$$L_n = \sqrt{D_n \tau_n}$$

$$L_p = \sqrt{D_p \tau_p}$$

For a well-defined carrier lifetime, the continuity equations become, for electrons,

$$\frac{\partial n}{\partial t} = \frac{\partial \Delta n}{\partial t} = \frac{1}{q} \left(\frac{\partial J_{n(\text{drift})}}{\partial x} + \frac{\partial J_{n(\text{diff})}}{\partial x} \right) + \left(G_{\text{op}} - \frac{\Delta n}{\tau_n} \right)$$

and similarly for holes,

$$\frac{\partial p}{\partial t} = \frac{\partial \Delta p}{\partial t} = \frac{-1}{q} \left(\frac{\partial J_{p(\text{drift})}}{\partial x} + \frac{\partial J_{p(\text{diff})}}{\partial x} \right) + \left(G_{\text{op}} - \frac{\Delta p}{\tau_p} \right)$$

Just as the Fermi level is a useful concept for describing the electron and hole concentrations in semiconductors at equilibrium, the quasi Fermi levels E_{fn} and E_{fp} are useful for nonequilibrium cases. For nondegenerate semiconductors

$$n = N_C e^{-(E_c - E_{fn})/kT} = n_i e^{(E_{fn} - E_i)/kT}$$

$$p = N_V e^{-(E_{fp} - E_v)/kT} = n_i e^{(E_i - E_{fp})/kT}$$

From knowledge of the quasi Fermi levels, the electron and hole currents can be expressed

$$J_n(x) = \sigma_n(x) \frac{d}{dx} \left(\frac{E_{fn}}{q} \right)$$

$$J_p(x) = \sigma_p(x) \frac{d}{dx} \left(\frac{E_{fp}}{q} \right)$$

3.12 REFERENCES

1. D. M. Caughey and R. Thomas, “Carrier mobilities in silicon empirically related to doping and field,” *Proc. IEEE*, 55, pp. 2192–2193, 1967.
2. J. del Alamo, S. Swirhun, and R. M. Swanson, “Measuring and modeling minority carrier transport in heavily doped silicon,” *Solid State Electronics*, 28, pp. 47–54, 1985.
3. J. del Alamo, S. Swirhun, and R. M. Swanson, “Simultaneous measurement of hole lifetime, hole mobility and bandgap narrowing in heavily doped n-type silicon,” *IEDM Technical Digest*, pp. 290–293, 1985.
4. S. E. Swirhun, Y.-H. Kwark, and R. M. Swanson, “Measurement of electron lifetime, electron mobility and band gap narrowing in heavily doped p-type silicon,” *IEDM Technical Digest*, pp. 24–27, 1986.
5. Chi-Tang Sah, Robert W. Noyce, and William Shockley, “Carrier generation and recombination in P-N junctions and P-N junction characteristics,” *Proc. IRE*, 45, pp. 1228–1242, 1957.

3.13 REVIEW QUESTIONS

1. What are the two basic mechanisms by which current flows in semiconductors? Explain the physics of each.
2. What is the difference between current and current density?
3. Define conductivity. Give its dimensions.
4. Explain how both electrons and holes contribute to conductivity.
5. There are no holes in a metal. Explain what happens to the holes carrying current in a semiconductor when that current reaches and continues on into the wire connecting the semiconductor with the rest of the circuit.
6. Define mobility. Give its dimensions.
7. Explain why the majority and minority carriers have different mobilities.
8. Why is a doped semiconductor more conductive than an intrinsic one?
9. What is ionized impurity scattering? Is it stronger under low doping or high doping? Under low temperature or high temperature? Why?
10. What is phonon scattering? Is it stronger under low temperature or high temperature? Why?

11. Explain how the presence of an impurity miniband affects the mobility of majority carriers. Under what doping conditions (heavy, light, any) will this effect be appreciable?
12. What is velocity saturation? What causes it?
13. Electrons always diffuse to regions of lower concentration. Is the same true for holes?
14. What is a trap state?
15. Why are direct gap materials better light emitters and absorbers than indirect gap materials?
16. Do photons have large K vector or small K vector? What does that imply about optical transitions from one band to the other on the E - K diagram?
17. What are excess carriers?
18. Define carrier lifetime.
19. Why is the carrier lifetime of GaAs much less than that of Si?
20. Explain in words the meaning of the continuity equations. What are the various terms and what do they mean?
21. Explain in words how the continuity equation can be used to find the carrier lifetimes. You may wish to use Figure 3.17 in your explanation.
22. Define equilibrium.
23. What is the difference between dark conductivity and photoconductivity?
24. What is a quasi Fermi level? How is it different from the Fermi level?

3.14 PROBLEMS

- 3.1 Consider an electron with a kinetic energy of $kT/2$. Draw to scale an energy band diagram (assume silicon) and indicate where the electron will be. What is its thermal (instantaneous) speed? How does your result compare to typical saturation velocities in semiconductors?
- 3.2 Calculate conductivity for silicon at room temperature under each of the following doping conditions:
 - a. $N_D = 10^{18} \text{ cm}^{-3}$
 - b. $N_A = 5 \times 10^{16} \text{ cm}^{-3}$
 - c. $N_D = 2 \times 10^{19} \text{ cm}^{-3}$
- 3.3 Calculate the resistivity for each case in problem 3.2.
- 3.4 A lightly doped Si sample ($N_D = 10^{14} \text{ cm}^{-3}$) is heated from 300 to 400 K. Is its resistivity expected to increase or decrease? Explain your answer. What happens to this trend when the semiconductor is very heavily doped? Why?
- 3.5 Germanium is an interesting semiconductor because it has a small band gap ($E_g = 0.67 \text{ eV}$). (In fact, for a while it was not considered to be a

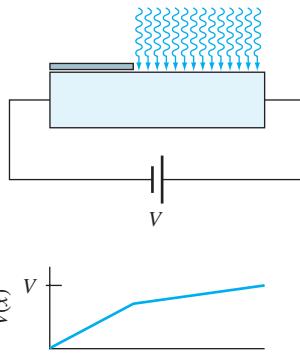
semiconductor but was classified as a metal. Now it is a semiconductor again.) As a result, it has a higher intrinsic concentration n_i than either silicon or GaAs. Do you expect the conductivity of intrinsic germanium to be less than or greater than that of intrinsic silicon? How about compared with GaAs? Why?

- 3.6 Consider a sample of silicon doped with $N_D = 5 \times 10^{19} \text{ cm}^{-3}$ having dimensions $0.10 \mu\text{m} \times 0.85 \mu\text{m} \times 65 \text{ nm}$. One volt is applied across the thinnest dimension. Find the current produced, and the corresponding current density. These dimensions could represent the current-conducting channel of a field-effect transistor.
- 3.7 Calculate the hole drift velocity for $N_A - N_D = 9.9 \times 10^{16} \text{ cm}^{-3}$ in a bar of Si of cross-sectional area 1.0 mm^2 for a current of 25 mA.
- 3.8 A silicon bar of 0.2 cm^2 cross section is 1 cm long. Find the resistance of the bar for (a) intrinsic Si, (b) for $N_D = 10^{15} \text{ cm}^{-3}$, $N_A = 0$ and (c) $N_A = 10^{16} \text{ cm}^{-3}$, $N_D = 0$.
- 3.9 A bar of silicon of 1 cm length is uniformly doped with 10^{18} phosphorus atoms. A voltage of 2.0 V is applied across the bar. Find (a) the electron drift current and (b) the hole drift current.
- 3.10 Explain how an electron can “collide” with a crystal defect. *Hint:* the normal periodic potential of the lattice is disturbed by a crystalline defect. A poor-quality crystal will have lower carrier mobilities than a good one.
- 3.11 Explain why, for a given noncompensated semiconductor with a given doping level N , the electron mobility is larger in p-type Si than in n-type Si. Refer to Figure 3.4.
- 3.12 With reduced N in Problem 3.11, the electron mobilities approach each other. Explain.
- 3.13 Compare the mean free time between collisions for electrons and for holes in intrinsic GaAs. How do these values compare with those for silicon?
- 3.14 Using the results of 3.13, find an average gain in kinetic energy between collisions for electrons and holes for an applied field of 100 V/cm.
- 3.15 A bar of silicon of 1 cm in length is uniformly doped with $1.0 \times 10^{16} \text{ cm}^{-3}$ phosphorus atoms. A voltage of 2.0 V is applied across the bar. Find:
 - a. the electron drift current density
 - b. the hole drift current density
- 3.16 Estimate the saturation velocity of electrons in intrinsic GaAs. How does your estimate compare with the experimental data?
- 3.17 The electron velocity in Si has its saturation value ($v_{\text{sat}} \approx 1 \times 10^7 \text{ cm/s}$) over the range of 5×10^4 to $2 \times 10^5 \text{ V/cm}$. Plot the mobility-field ($\mu - \mathcal{E}$) and $\bar{t} - \mathcal{E}$ relations over this range of fields.
- 3.18 An n-type sample of silicon is doped with $N_D = 5.5 \times 10^{18} \text{ cm}^{-3}$ and $N_A = 10^{16}$. Over a length of $1.5 \mu\text{m}$ the excess hole concentration

increases linearly from 10^{13} cm^{-3} to 10^{15} cm^{-3} . Calculate the hole diffusion current density.

- 3.19. The excess electron concentration in a semiconductor varies as $\Delta n(x) = 10^{20} e^{-\frac{1\mu\text{m}}{x}}$. Find the electron diffusion current at $x = 0.5 \mu\text{m}$. Let $D_n = 35 \text{ cm}^2/\text{s}$. Sketch the carrier concentration as a function of position, and label the directions of electron flow and electron current.
- 3.20 A sample of silicon is doped such that the electron concentration varies linearly across the sample. The sample is 0.5 mm thick. The donor concentration varies from $N_D = 0$ at $x = 0$ to $N_D = 10^{16} \text{ cm}^{-3}$ at $x = 0.5 \mu\text{m}$.
- Write equations for $n(x)$ and $p(x)$.
 - Find the electron diffusion current density.
 - Find the hole diffusion current density at $x = 0$ and $x = 0.5 \mu\text{m}$. Can the minority carriers contribute significant diffusion current?
 - Find an expression for $E_C(x) - E_f$ as a function of x . Sketch the energy band diagram. (The Fermi level is constant at equilibrium, so draw the Fermi level as a flat line and adjust $E_C - E_f$ appropriately.)
 - At equilibrium, the total current must be zero. Show that there must therefore be an internal electric field present in this sample. (The field is generated by the variation in doping, as we will see in the next chapter.)
- 3.21 Comment on the probability of absorption (zero, low, medium, high) by a photon of $\lambda = 600 \text{ nm}$ (red) by the following materials:
- Si
Ge
GaAs
InAs
SiC
GaN
CdS
- 3.22 A sample of intrinsic GaN is illuminated with a 10 mW/cm^2 beam of light at the blue edge of what people can see, $\lambda = 300 \text{ nm}$. The electron lifetime is 2 ps. The electron mobility for intrinsic GaN is about $400 \text{ cm}^2/\text{V} \cdot \text{s}$ and hole mobility is about $100 \text{ cm}^2/\text{V} \cdot \text{s}$.
- What is the number of photons arriving at the semiconductor surface per sec? (Recall energy = power \times time.)
 - Verify that photons of this energy can be absorbed.
 - Assuming every photon is absorbed and creates an electron-hole pair, and assuming the GaN sample is 1 mm thick, what is the optical generation rate?
 - What are the equilibrium electron and hole densities (in the dark)?
 - What are the excess carrier concentrations when the light is on?

- f. What are the recombination rates for electrons and for holes when the light is off? When the light is on?
- g. What are the steady-state carrier densities n and p ?
- h. How much does the conductivity of this sample change compared with its dark value?
- i. Suppose the power level is kept the same, but the wavelength of the light is shifted to the red edge of human vision ($\lambda = 700 \text{ nm}$). What is the generation rate now?
- 3.23** Solve Equation (3.72) to prove that Equation (3.73) is a solution under the appropriate initial conditions. Repeat for Equation (3.75).
- 3.24** A direct gap semiconductor sample is illuminated at one end with light of $\lambda = 500 \text{ nm}$ (green), with an intensity of 5.0 mW/cm^2 . The area of the illuminated surface is 0.80 cm^2 . Assume the carrier lifetimes are 10 ns .
- Find the number of photons striking the sample per second.
 - If every photon is absorbed uniformly (with x) within $1 \mu\text{m}$ of the surface, what are the excess carrier concentrations Δn and Δp in this region?
- 3.25** As we saw, it is possible to produce a nonuniform spatial distribution of charge carriers by shining a light on one part of a semiconductor, producing more electrons and holes in that part than elsewhere. What might be another way to produce more electrons and holes in one location, without applying an electric field? Would there be a net diffusion current?
- 3.26** Consider a bar of semiconductor illuminated as shown in Figure P3.1.
- Sketch the concentrations of electrons and holes as functions of position.
 - In which direction(s) will the electrons diffuse? Holes?
 - In what directions do the electron and hole diffusion currents go?
 - Explain why a plot of voltage versus position along the length of the sample is as shown in the figure.

**Figure P3.1**

- 3.27** a. Find the conductivity (in the dark) of a sample of GaAs doped with $N_A = 8.5 \times 10^{16} \text{ cm}^{-3}$.
- b. If the sample is illuminated such that the excess electron concentration is 10^{17} cm^{-3} , what is the excess hole concentration?
- c. What is the conductivity of this sample when the light is on?
- 3.28** Find the quasi-Fermi levels for electrons and holes for an illuminated silicon sample with $N_A = 7.0 \times 10^{17}$ and $\Delta n = \Delta p = 7.0 \times 10^{16} \text{ cm}^{-3}$. Compare these locations with the Fermi levels when the light is off.
- 3.29** A sample of InP is doped such that $E_f - E_V = 0.3 \text{ eV}$. It is also illuminated such that $\Delta n = \Delta p = 10^3 \text{ cm}^{-3}$. Find the quasi Fermi levels and sketch the energy band diagram. Repeat for $\Delta n = \Delta p = 10^{10}$.

Nonhomogeneous Semiconductors

So far, we have examined *homogeneous* semiconductors. By homogeneous we mean that the entire semiconductor is made of the same material and doped uniformly. Examples of the other case, nonhomogeneous semiconductors, are semiconductors whose doping varies with position or where the semiconductor composition varies. A classic example of the former is the pn junction diode, which is a junction between n-type and p-type material. We will lead up to the diode gradually, starting with a semiconductor in which the doping level is nonuniform with position. We then discuss the case in which the composition of the semiconductor itself is graded.

4.1 CONSTANCY OF THE FERMI LEVEL AT EQUILIBRIUM

First, we will establish a very important result:

In a system at equilibrium, the Fermi level is at constant energy.

To show this, we consider two materials (two semiconductors for this example) in intimate contact as indicated in Figure 4.1a. The energy band diagram is shown, with the vacuum level as a reference. In this example, the two semiconductors have different band gaps (meaning they are not the same material), and they are doped differently as well. Assume these materials have different Fermi levels as indicated in (b). When the materials are first joined, they are in a state we call *electrical neutrality*, as we shall see later. This state is not stable because there are electrons at higher energies in A than in B, and we expect that these will flow from A to available lower energies in B until the system of joined materials reaches some kind of equilibrium. At equilibrium, the total current must be zero.

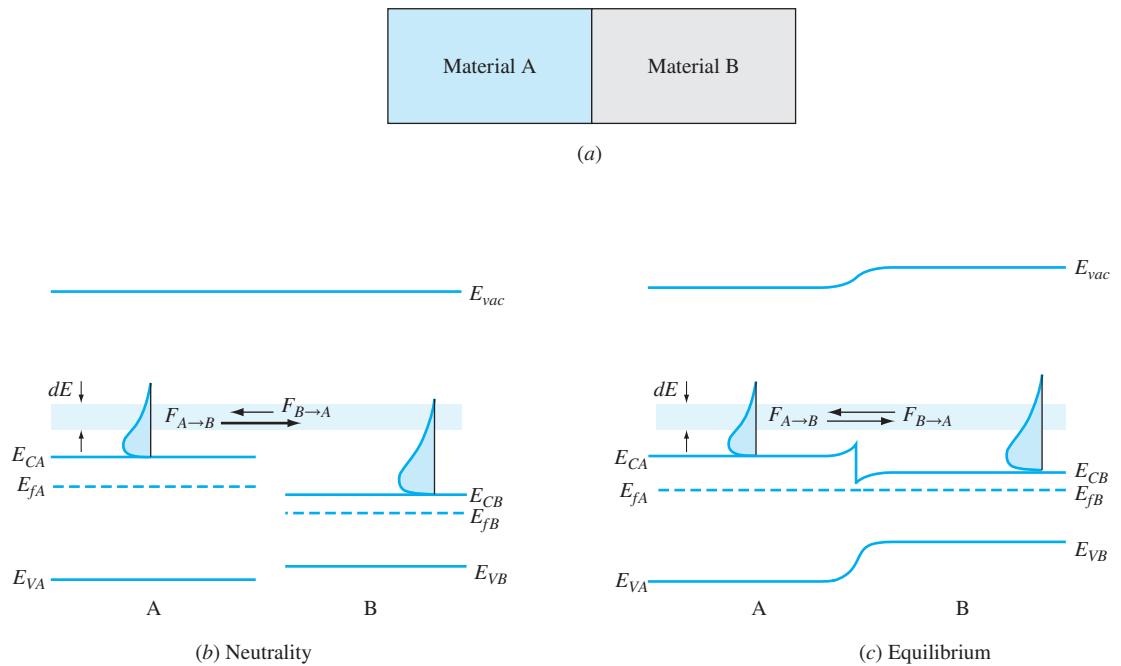


Figure 4.1 (a) When two different materials are in contact at equilibrium the net current must be zero. (b) When the materials are first joined, they are both electrically neutral. In the text it is shown that at equilibrium, the Fermi levels must be equal, as shown in (c).

Therefore, if carriers are moving from A to B, the same number of carriers must be moving from B to A.

To find the number of carriers available in A, let the density-of-states function in material A be $S_A(E)$, and the probability of occupancy be $f_A(E)$. The electron concentration at some energy E is $S_A(E)f_A(E)$. Similarly, $S_B(E)$ is the density-of-states function at energy E in material B and $f_B(E)$ is the Fermi probability at E in B. Let $F_{A \rightarrow B}$ be the rate of electron transfer (electron flux) from A to B. The movement of electrons requires electrons at a given state and nearby empty states to be at the same energy. For carriers to move from A to B, there must be empty states in B. The concentration of the empty states in B is $S_B(E)(1 - f_B(E))$. We can see from part (b) of the figure that, as drawn, in a given energy range dE (shaded in the figure) the concentration of electrons in material A is much higher than in material B. This implies that there are many empty states in B for electrons to go to.

If we consider the transfer flux of carriers in some small energy range dE , then

$$F_{A \rightarrow B} = CS_A(E)f_A(E)S_B(E)(1 - f_B(E))dE \quad (4.1)$$

where C is a constant. Similarly, if $F_{B \rightarrow A}$ is the transfer flux from B to A,

$$F_{B \rightarrow A} = C S_B(E) f_B(E) S_A(E) (1 - f_A(E)) dE \quad (4.2)$$

At equilibrium, however, the two fluxes must be equal (no current), so $F_{A \rightarrow B} = F_{B \rightarrow A}$ and

$$S_A(E) f_A(E) S_B(E) (1 - f_B(E)) = S_B(E) f_B(E) S_A(E) (1 - f_A(E)) \quad (4.3)$$

which reduces to

$$f_A(E) = f_B(E) \quad (4.4)$$

The probability of occupancy is the Fermi function, so we can substitute Equation (2.49):

$$\begin{aligned} f_A(E) &= \frac{1}{1 + e^{(E - E_{fA})/kT}} \\ f_B(E) &= \frac{1}{1 + e^{(E - E_{fB})/kT}} \end{aligned} \quad (4.5)$$

into Equation (4.4), producing

$$E_{fA} = E_{fB} \quad (4.6)$$

Thus, the Fermi levels are equal at equilibrium. This is shown in Figure 4.1c. The position of the Fermi level with respect to the conduction and valance band edges within a given material is a function of the doping, so if the Fermi levels are adjusted so that they line up in the figure, the entire energy band diagram must adjust along with it.

While the preceding discussion shows that at equilibrium the Fermi levels of two dissimilar materials are equal, it can be generalized to state that the Fermi level is constant throughout a system at equilibrium. The materials can be semiconductors, nondegenerate or degenerate, or insulators, metals, or alloys. This important result will be used extensively in describing the operation of electronic devices.

4.2 GRADED DOPING

Up to now we have always taken the net doping concentration in a semiconductor—either N_D or N_A —to be a constant with position. This is often not the case in semiconductor devices. We now consider the case of a semiconductor with nonuniform doping. First, we investigate the influence of the doping profile on the equilibrium energy band diagram and how this, in turn, affects carrier concentrations.

Many semiconductors used in devices use compensated materials, or those in which both acceptors and donors are present. The quantity of interest in computing electron and hole concentrations, locating Fermi levels, and drawing energy band diagrams is actually the *net* doping concentration:

$$N'_D = N_D - N_A \quad \text{n type} \quad (4.7)$$

or

$$N'_A = N_A - N_D \quad \text{p type} \quad (4.8)$$

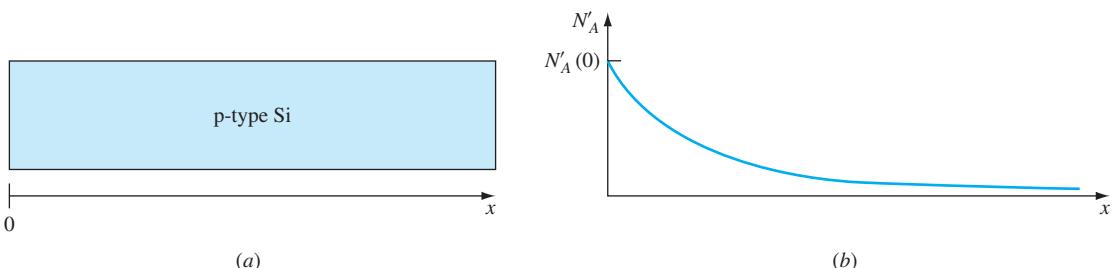


Figure 4.2 Nonuniformly doped semiconductor. (a) The bar of semiconductor material; (b) the net acceptor concentration as a function of distance.

As an example of graded doping, we choose the case of compensated p-type Si shown in Figure 4.2 in which the net acceptor concentration, $N'_A = N_A - N_D$, decreases with position from left to right. In this discussion, we assume that the material is everywhere nondegenerate and band-gap narrowing is therefore negligible, or that the band gap is independent of doping. This situation is typical in the base region of an npn bipolar transistor, for example.

We know from Chapter 2 that, where the net acceptor doping is greater, there will be more holes. In the sample in Figure 4.2, we expect the Fermi level to be close to the valence band edge near the left end where the doping is heavy, and we expect the Fermi level to be closer to midgap at the right end where the material is more lightly doped. We therefore might expect an energy band diagram similar to that in Figure 4.3, in which electrical neutrality is assumed to exist in every macroscopic region. By electrical neutrality, we mean that in any macroscopic region, the concentration of negative charge in the region equals the positive charge concentration, as shown in part (a) of the figure. In part (b), the vacuum level E_{vac} and the acceptor state energy level E_A are shown, and the electron affinity χ is also indicated.

Assuming that all the impurities are ionized, the hole concentration is equal to the net acceptor concentration, or

$$p(x) = N'_A(x) = N_V e^{-(E_f - E_V)/kT} \quad (4.9)$$

Since the left side of Equation (4.9) is a function of x , the right side must be also. The quantities N_V , k , and T are all constants, so it follows that $E_f - E_V$ varies with position. But recall that the Fermi level is the energy level at which the Fermi probability function is $\frac{1}{2}$. The Fermi function is independent of the distribution of states, the doping concentrations, etc., and depends solely on temperature. We know, however, that at equilibrium the Fermi level is at constant energy independent of position. Therefore, we can anticipate that the edge of the valence band will shift in energy as the doping varies. We also now realize that Figure 4.3b is wrong. We now need to determine how to correct it.

Returning to Figure 4.3b, the case of the assumed neutrality, we notice that the electric field

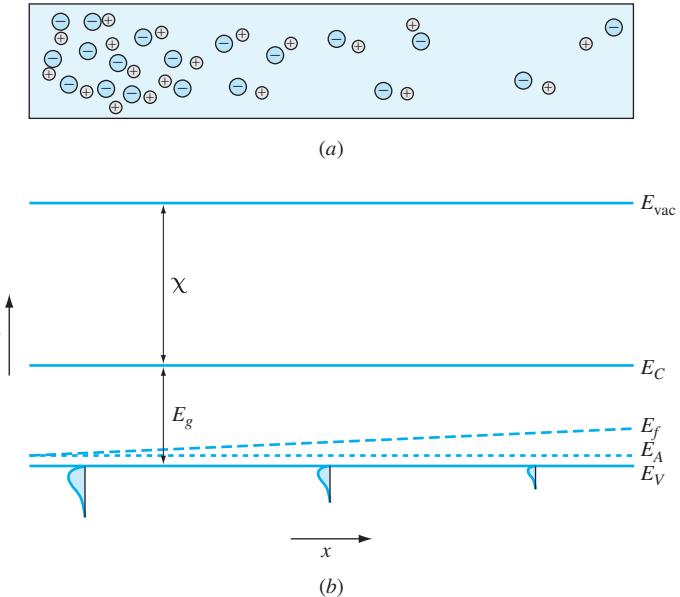


Figure 4.3 (a) The energy band diagram before equilibrium for the nonuniformly doped semiconductor. (b) Neutrality exists in every macroscopic region.

$$\mathcal{E} = \frac{1}{q} \frac{dE_{\text{vac}}}{dx} \quad (4.10)$$

is zero but that there is a gradient in the concentration of holes. We know from Chapter 3 that holes, being mobile, will diffuse to regions of lower concentration, in this case to the right. There is also a gradient in the concentration of ionized acceptors from Figure 4.2. The acceptor ions are negatively charged, but they can't diffuse because they are locked into the crystal lattice. Therefore, there is a net diffusion of positive charges (holes) from left to right, leaving uncompensated negative ions on the left. This is shown in Figure 4.4a. The number of positive charges (holes) per unit volume at the right-hand end of the sample then increases, leaving the negatively charged ions on the left. This separation of charges, however, creates an electric field in the semiconductor, and we no longer have a condition of electrical neutrality. The polarity of the field, though, is such that it tends to move the holes back toward the left. Thus, the diffusion of carriers to the right produces an electric field that tends to move them back to the left.

We can determine what the final situation at equilibrium will be. We know that at equilibrium, the net hole current density is zero, and that therefore the drift current due to the built-in electric field must balance the diffusion current caused by the concentration gradient:

$$J_p = J_{p(\text{drift})} + J_{p(\text{diff})} = 0 = q\mu_p p_0(x)\mathcal{E} - qD_p \frac{dp_0(x)}{dx} \quad (4.11)$$

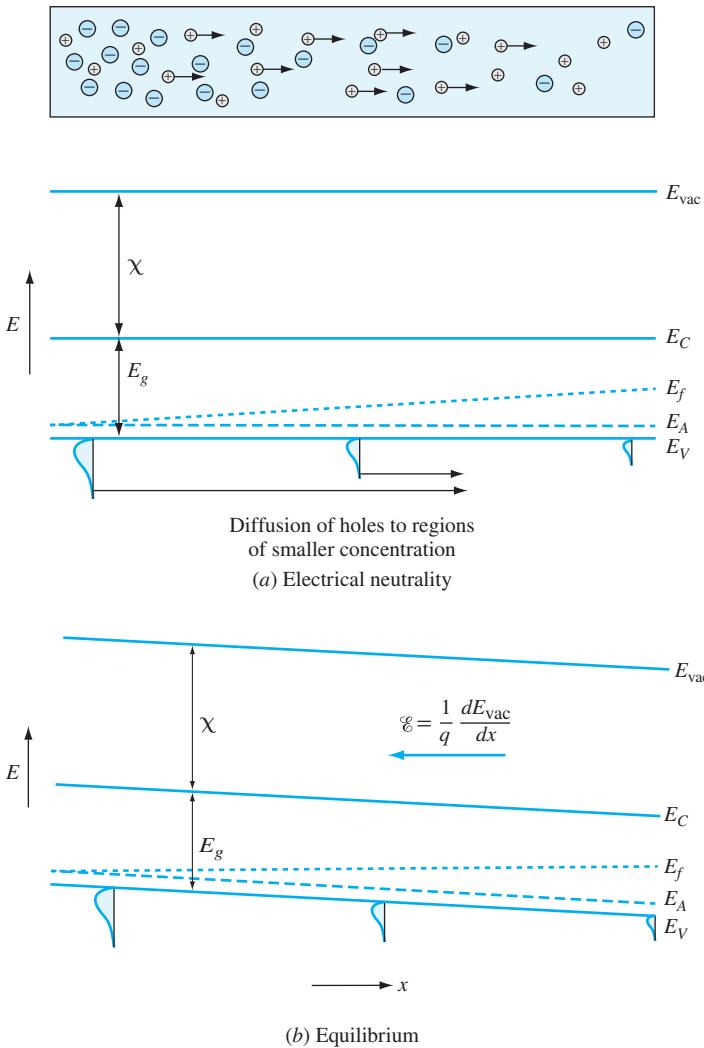


Figure 4.4 (a) The holes diffuse to regions of lower concentration, leaving behind ionized acceptors. This sets up an electric field. (b) The correct energy band diagram for a nonuniformly doped semiconductor at equilibrium.

where the hole concentration $p = p_0$ at equilibrium is a function of position. Because a relatively small shift in charge is required to establish this field, the approximation $p(x) = p_0(x) = N'_A(x)$ is still good, and Equation (4.11) becomes, with the aid of Equation (4.9),

$$\begin{aligned} q\mu_p p_0(x)\mathcal{E} &= -q \frac{D_p N_V}{kT} e^{-(E_f - E_V)/kT} \left[\frac{d(E_f - E_V)}{dx} \right] \\ &= -q \frac{D_p}{kT} p_0(x) \frac{d(E_f - E_V)}{dx} \end{aligned} \quad (4.12)$$

Since the Fermi level is a constant, $dE_f/dx = 0$, and

$$\mu_p \mathcal{E} = \frac{D_p}{kT} \frac{dE_V}{dx} \quad (4.13)$$

Next, since the semiconductor species (e.g., Si) is constant, so are the electron affinity, band-gap and ionization potential. Thus, E_V and E_C are parallel to E_{vac} . This means, from Equation (4.10),

$$\mathcal{E} = \frac{1}{q} \frac{dE_{\text{vac}}}{dx} = \frac{1}{q} \frac{dE_C}{dx} = \frac{1}{q} \frac{dE_V}{dx} \quad (4.14)$$

The electric field is therefore proportional to the slopes of the valence band edge and conduction band edge. The corrected energy band diagram at equilibrium is shown in Figure 4.4b.

Note that at equilibrium, there is *not* a state of electrical neutrality—to establish equilibrium the positive and negative charges were separated to some extent. However, since a small charge transfer is required to establish equilibrium, this region is normally referred to as a *quasi-neutral region*, or often as a *neutral region*.

To summarize, a convenient method to draw an energy band diagram at equilibrium is:

1. Assume electrical neutrality in every macroscopic region.
2. Using the vacuum level as a reference, i.e., with E_{vac} constant with position, draw the energy band diagram taking the electron affinity and band gap into account.
3. From a knowledge of the net doping, find the Fermi level with respect to the appropriate band edge (E_V for p-type, E_C for n-type material).
4. Adjust (tilt) the neutrality band diagram such that the Fermi level is at constant energy, keeping electron affinity and band gap constant.

We emphasize that the electric field created by the graded doping is built-in, not applied. It arises naturally when the doping concentration is not uniform, because the charged carriers diffuse. The semiconductor can still be at equilibrium as long as no electric field is applied *externally*. The total current is still zero at equilibrium.

Figure 4.5 shows all of the electron and hole fluxes in this sample, and the resulting currents. Electrons diffuse to regions of lower concentration (in this case to the left), resulting in electron diffusion current to the right. Electrons drift in the direction opposite to the electric field (to the right), resulting in a net electron drift current to the left. Convince yourself that the directions of the arrows in the figure are correct for holes.

Note that because of the built-in field, there is also a built-in voltage across this sample. The left side is at a higher electron potential energy than the right, meaning that there is a net voltage difference V_{bi} across the sample.¹ That does not imply that a nonuniformly doped semiconductor can be a battery, however.

¹The q in the figure is needed because the figure shows energy in eV and the voltage V_{bi} is in volts. The factor of q keeps the units correct.

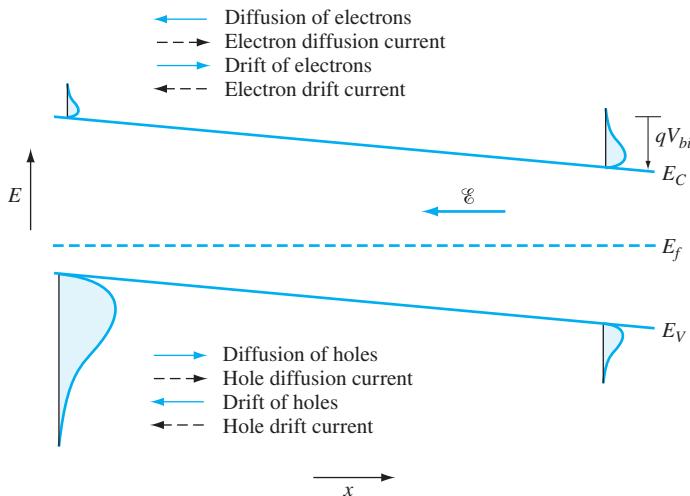


Figure 4.5 The currents that arise in the nonuniformly doped semiconductor must still sum to zero at equilibrium.

Figure 4.6 shows why. Suppose we connect a metal wire from one end of the semiconductor bar to the other. The system is still at equilibrium because we have not applied any external voltages. The energy band diagram for this situation shows that the vacuum level is the same everywhere in the wire. In the semiconductor, the vacuum level, and along with it the band edges, bend to meet the vacuum level of the metal. The band bending is greater at one end of the semiconductor bar than the other. Recalling that where the bands bend there is an electric field, you immediately deduce that there are some new built-in electric fields at the junctions between the semiconductor and the metal. The built-in voltages add up to zero around the circuit in the figure (Kirchhoff's voltage rule). The electric fields are all inside the semiconductor, and thus they cannot be accessed from outside. This will happen regardless of the material chosen to connect the two ends.

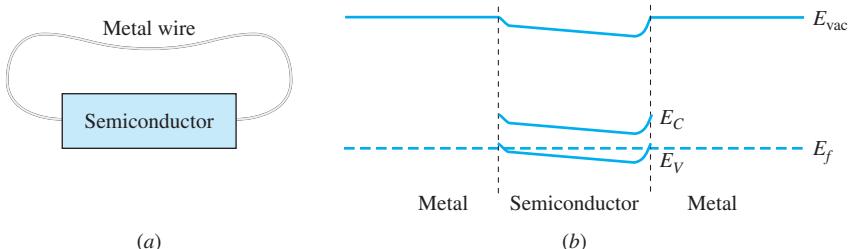


Figure 4.6 (a) A wire is used to connect one end of the graded-doping semiconductor to the other. (b) The energy band diagram. There are built-in fields at the ends of the semiconductor where it is attached to the outside environment (a metal wire in this case), but the voltages add up to zero around the loop.

4.3 NONUNIFORM COMPOSITION

Next, consider a semiconductor whose *composition* is varied gradually from one end to the other. For example, perhaps the crystal is grown such that it starts out as pure Si ($E_g = 1.12$ eV), but Ge ($E_g = 0.67$ eV) is introduced in gradually increasing quantities until the final alloy is $\text{Si}_y\text{Ge}_{1-y}$, where y denotes the fractional concentration of Si. Since Si and Ge each have a valence of four, and have the same (diamond) crystal structure, the resultant SiGe alloy is also a semiconductor with the diamond structure, and it has some intermediate band gap. Let us also assume that the alloy is doped with acceptors such that the Fermi level is equidistant from the valence band. Note that since N_V varies only slightly with position (because of the small variation in electron effective mass as the composition changes [Equation (2.62)]), to keep $E_f - E_V$ constant, the net acceptor concentration is not quite uniform [Equation (4.9)].

To determine the equilibrium energy band diagram of the graded semiconductor alloy, we follow the procedure used in the previous section for graded doping. We first draw the diagram for the case in which electrical neutrality exists in every macroscopic region, as indicated in Figure 4.7, in which the grading in the band gap is exaggerated for clarity. With the vacuum level as a reference, we see that the conduction band edge E_C is not a function of position. This is because for the SiGe alloy, the electron affinity χ is essentially independent of alloy composition. The affinities are 4.05 eV (Si) and 4.0 eV (Ge). The band gap, however, is a noticeable function of alloy composition, since Si has a significantly different band gap (1.12 eV) than Ge (0.67 eV). Therefore as E_g decreases, E_V moves upward. It is known that E_g decreases linearly with the Ge

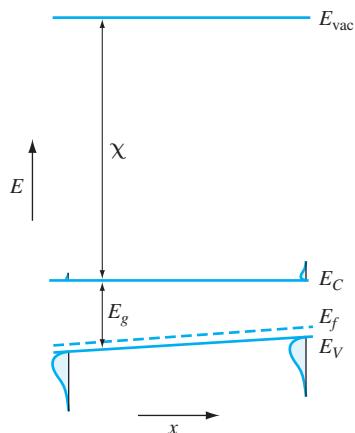


Figure 4.7 The energy band diagram for a SiGe alloy of nonuniform Si composition (exaggerated for clarity), under electrical neutrality.

concentration, by an amount 7.5 meV per each atomic percent of Ge. [1] In this example, the doping level is essentially uniform, so the free hole concentration p_0 is also uniform (approximately N_A). The electron concentration, on the other hand, is not uniform, since the band gap is not constant (recall that n_i increases with decreasing E_g , and that $n_0 p_0 = n_i^2$).

Since the hole concentration p_0 is constant, the hole diffusion is negligible. However, since the valence band edge is tilted, there is a force on the holes. The force exerted on the holes is minus the gradient of the hole's potential energy:

$$F_h = -\frac{dE_{P(\text{holes})}}{dx} \quad (4.15)$$

and for the holes, the potential energy is the valence band edge. Thus

$$\frac{dE_{P(\text{holes})}}{dx} = -\frac{dE_V}{dx} \quad (4.16)$$

The minus sign arises from the fact that we use the energy band diagram for negatively charged electrons and thus the hole energy increases downward, since holes are positively charged. Equation (4.15) then becomes

$$F_h = \frac{dE_V}{dx} \quad (4.17)$$

This force tends to redistribute the holes, until the sample reaches equilibrium, or the Fermi level is constant with position. The case at equilibrium is illustrated in Figure 4.8.

The alert student will observe from Figure 4.8 that a hole in the valence band will not be subjected to any force at equilibrium ($F_h = dE_V/dx \approx 0$), but that

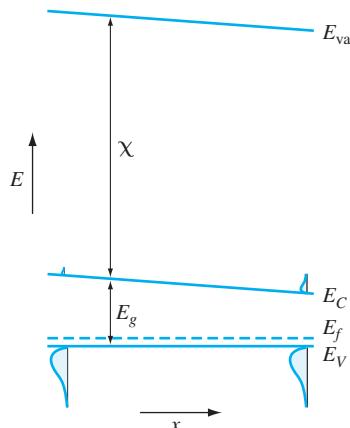


Figure 4.8 The energy band diagram for the semiconductor of nonuniform composition at equilibrium (again, exaggerated).

an electron in the conduction band *will* experience a force accelerating it to the right:

$$F_e = -\frac{dE_{P(\text{electron})}}{dx} = -\frac{dE_C}{dx} \quad (4.18)$$

At equilibrium, there *is* a force on the electrons that causes drift and a gradient in the electron concentration that causes diffusion. These drift and diffusion currents cancel at equilibrium.

Since the forces on electrons and holes are different, we define *effective electric fields* \mathcal{E}_e^* and \mathcal{E}_h^* [2] by the equations:

$$F_e = -q\mathcal{E}_e^* = -\frac{dE_C}{dx} \quad (4.19)$$

for electrons and

$$F_h = q\mathcal{E}_h^* = \frac{dE_V}{dx} \quad (4.20)$$

for holes.

The true electric field is equal to $(1/q)(dE_{\text{vac}}/dx)$ and by definition is equal to the force on a unit positive test charge. This can be understood by imagining a small hole drilled inside the semiconductor from left to right, with a small test charge Q_T on the end of an insulating rod inserted into the hole. Since the test charge is far enough from the semiconductor surface, its potential energy is E_{vac} (its kinetic energy is zero). The force exerted on this test charge is

$$F = Q_T\mathcal{E} = -\frac{dE_P}{dx} = -\frac{dE_{\text{vac}}}{dx} \quad (4.21)$$

EXAMPLE 4.1

Consider a p-type SiGe alloy (in the base region of a bipolar junction transistor) in which the band gap varies linearly by 0.1 eV across the $0.05\text{-}\mu\text{m}$ base width. Assume that the net acceptor concentration is constant. Find the effective electric field and force for holes and electrons, and the true electric field.

Solution

Since ΔE_g varies 7.5 meV per atomic percent Ge, this transistor has a linear grading of the Ge content in its base from zero to 13 percent. We saw earlier that for SiGe alloys, if the acceptor concentration is constant, the valence band level is essentially constant. Therefore, from Equation (4.20) the force (and the effective electric field) for holes is virtually zero:

$$\mathcal{E}_h^* = \frac{F_h}{q} = \frac{1}{q} \frac{dE_V}{dx} = 0$$

Since the valence band edge is flat, then as the band gap decreases, the conduction band edge energy decreases. For a base width of $0.05\text{ }\mu\text{m}$, the resulting effective field for electrons is, from Equation (4.19),

$$\begin{aligned}\mathcal{E}_e^* &= \frac{1}{q} \frac{dE_C}{dx} = \frac{1}{q} \frac{dE_g}{dx} = \frac{1}{1.6 \times 10^{-19} \text{ C}} \frac{(-0.1 \text{ eV})(1.6 \times 10^{-19} \text{ V/eV})}{0.05 \mu\text{m}} \\ &= -\frac{0.1 \text{ V}}{5 \times 10^{-8} \text{ m}} = -2 \times 10^6 \text{ V/m} = -20 \text{ kV/cm}\end{aligned}$$

Such a graded-composition SiGe alloy is sometimes employed in the base region of high-performance bipolar junction transistors. Electrons injected from the emitter into the base are accelerated to the collector by this high field, decreasing the transit time across the base and effectively decreasing the switching time or increasing the operating frequency of the transistor.

The true electric field is given by Equation (4.21). In this case, though, the electron affinity is almost constant across the material, so E_{vac} is nearly parallel to E_C . Since their slopes are the same, in this example the true electric field is the same as the effective electric field for electrons.

$$\mathcal{E} = \frac{1}{q} \frac{dE_{\text{vac}}}{dx} \approx \frac{1}{q} \frac{dE_C}{dx} = -20 \text{ kV/cm}$$

The negative sign on the expressions for both the effective and true fields means the field is in the negative x direction.

4.4 GRADED DOPING AND GRADED COMPOSITION COMBINED

In an actual SiGe transistor the field in the base is further increased by grading the acceptor concentration [3] as described in Section 4.2, which in turn speeds up the device even more, increasing the operating frequency.

EXAMPLE 4.2

A SiGe transistor has a base that is compositionally graded as in Example 4.1, but in addition, the net acceptor density decreases exponentially from 10^{18} cm^{-3} to $3 \times 10^{16} \text{ cm}^{-3}$ over this region. Find the effective electric field for electrons.

Solution

Figure 4.9 shows the energy band diagram in the base region (a) for the case of neutrality and (b) for equilibrium. Since $\Delta E_g = 0.1 \text{ eV}$ and the base width W_B is $0.05 \mu\text{m}$ as in Example 4.1, the effective field due to compositional grading alone is -20 kV/cm . In the previous case, E_V was considered constant with position. Here, however, there is an additional field resulting from the varying doping. Since

$$\begin{aligned}p(x) &= N_A(x) = N_V e^{-(E_f - E_V(x))/kT} \\ E_V(x) - E_f &= kT \ln \frac{N_A(x)}{N_V}\end{aligned}$$

then the additional tilt ΔE_V to E_V due to the varying doping concentration is

$$\Delta E_V = (E_V(W_B) - E_V(0)) = (E_V(W_B) - E_f) - (E_V(0) - E_f)$$

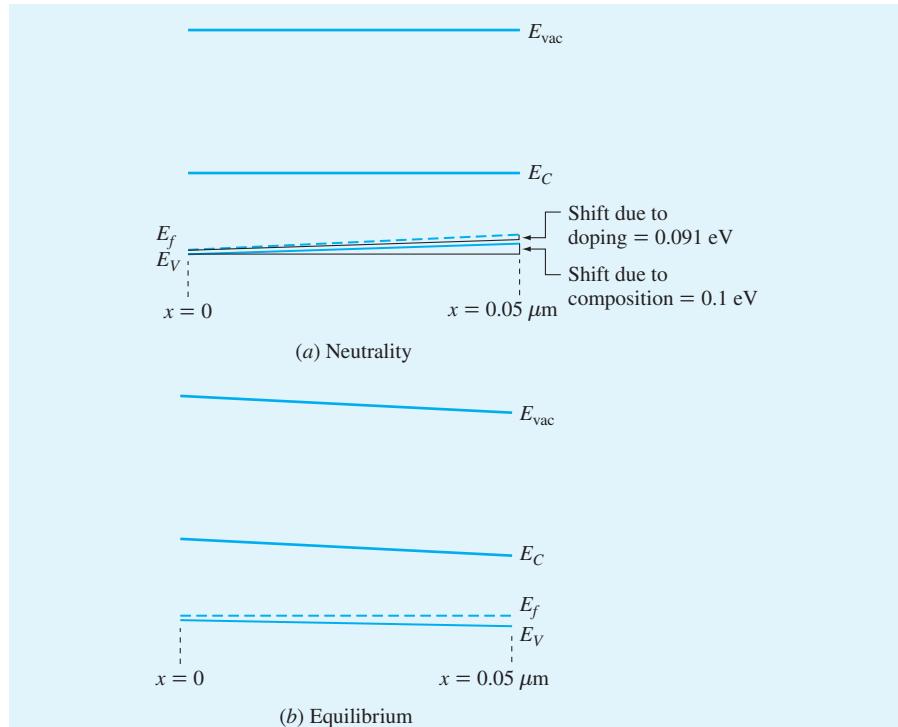


Figure 4.9 The energy band diagram for the sample of Example 4.2, with both graded composition and varying doping. (a) Electrical neutrality; (b) equilibrium.

or

$$\Delta E_V = kT \left(\ln \frac{N_A(W_B)}{N_V} - \ln \frac{N_A(0)}{N_V} \right) = kT \ln \frac{N_A(W_B)}{N_A(0)}$$

$$\Delta E_V = 0.026 \ln \frac{3 \times 10^{16}}{10^{18}} = -0.091 \text{ eV}$$

The negative sign indicates the valence band edge moves down away from the Fermi level (in Figure 4.9a the Fermi level is shown moving up). Combining the field induced by the doping gradient (which moves the conduction band edge down) with the effect of the compositional grading (which moves the valence band and the conduction band down) yields

$$\Delta E_C = 0.1 \text{ eV} + (-\Delta E_V) = 0.191 \text{ eV}$$

Since the linear compositional grading and the exponential doping gradient each contribute to constant fields, the net quasi-electric field for electrons is

$$\mathcal{E}_e^* = \frac{0.191 \text{ V}}{0.05 \mu\text{m}} = 3.8 \text{ V}/\mu\text{m} = 38 \text{ kV/cm}$$

4.5 SUMMARY

In this chapter, we have extended our understanding of the physics of materials to predict what happens in nonuniform semiconductors. The results are interesting indeed. Internal electric fields can be generated in a semiconductor simply by varying the doping concentrations across the sample or by varying the material composition, or both. These fields are utilized in the base regions of bipolar junction transistors to improve their performance.

The built-in electric fields cannot be accessed externally; the sample can't be used as a battery. Furthermore, we saw that at equilibrium the drift currents produced by the built-in electric fields are balanced by opposing diffusion currents; the net electron and hole currents are still zero.

We obtained the useful concept that for a system in intimate contact, at equilibrium the Fermi level is constant throughout the system. While the concept of a Fermi level is valid only at equilibrium, analogous quasi Fermi levels can be used for nonequilibrium cases.

4.6 REFERENCES

1. John D. Cressler and Katsuyoshi Washio, “Bipolar transistors,” in *Guide to State-of-the-Art Electron Devices*, Joachim N. Burghartz, ed., Chap. 1, Wiley-IEEE Press, John Wiley and Sons, Chichester, UK, 2013.
2. H. Kroemer, “Quasielectric and quasimagnetic fields in nonuniform semiconductors,” *RCA Review*, 18, pp. 332–347, 1957.
3. J. D. Cressler (ed.), *Silicon Heterostructure Handbook: Materials, Fabrication, Devices, Circuits and Applications of SiGe and Si Strained-Layer Epitaxy*, CRC Press, Boca Raton, FL, 2006.

4.7 REVIEW QUESTIONS

1. Explain why the Fermi level is constant at equilibrium.
2. How does varying the doping across a sample create a built-in electric field?
3. What is meant by *quasi-neutral region*?
4. What are the steps for drawing the energy band diagram for a new system?
5. When a material has a built-in electric field, what is the current through it at equilibrium?
6. How does grading the composition create a built-in electric field?
7. Why can't a material with a built-in electric field be used as a battery?
8. What is meant by the term *effective electric field* (as for electrons)? How is it different from the true electric field?

4.8 PROBLEMS

- 4.1. Show that if a sample of semiconductor is doped with an exponentially varying doping concentration $N_D(x) = N_D(x=0)e^{-\frac{x}{\lambda}}$ that the electric field is constant and is given by $\mathcal{E} = -\frac{kT}{q}\frac{1}{\lambda}$
- 4.2. A sample of InP is doped with donors with concentration varying exponentially from $N_D = 1.0 \times 10^{16}$ at $x = 0$ to $N_D = 5.0 \times 10^{18}$ at $x = 400$ nm. Find the electric field and indicate the direction.
- 4.3. Consider the graded-composition SiGe alloy discussed in Section 4.3. The hole concentration is assumed constant (N'_A is essentially constant and all dopants are ionized). There is no electric field or diffusion for holes at equilibrium, but there is a field for electrons in the conduction band. Since at equilibrium there is no net current, the drift electron current must be offset by an opposing diffusion current. Identify and explain the source of the varying electron concentration that produces the diffusion current.
- 4.4. Find the time required for an electron to traverse the p region of Example 4.2 due to drift alone.
- 4.5. A graded alloy is manufactured in the AlGaAs system. At $x = 0$, the material is pure GaAs ($\chi = 4.07$ eV, $E_g = 1.43$ eV), and over a distance of $2 \mu\text{m}$ the composition changes to $\text{Ga}_{0.6}\text{Al}_{0.4}\text{As}$ ($\chi = 3.84$ eV and $E_g = 1.92$ eV). The material is intrinsic. Find the effective electric fields for electrons and holes, and find the true electric field.
- 4.6. Consider the equilibrium energy band diagram of Figure P4.1.
 - a. Find the effective electric field for electrons.
 - b. Find the effective electric field for holes.
 - c. Sketch the carrier concentrations.

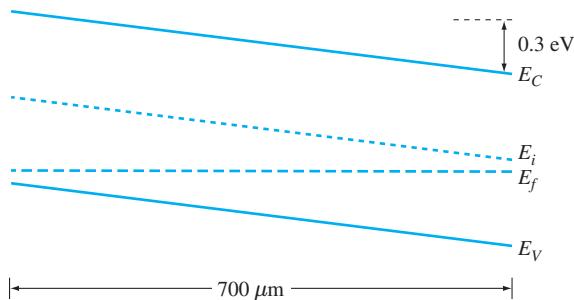


Figure P4.1

- d. Indicate the directions of each of the following:
 - i. Hole diffusion
 - ii. Hole diffusion current
 - iii. Hole drift
 - iv. Hole drift current
 - v. Electron diffusion
 - vi. Electron diffusion current

- vii. Electron drift
 - viii. Electron drift current
 - ix. Electric field for electrons
- 4.7** For the sample whose energy band diagram is shown in Figure P4.2:
- Find the effective electric field for electrons.
 - Find the effective electric field for holes.
 - Sketch the carrier concentrations.
 - Indicate the directions of each of the following:
 - i. Hole diffusion
 - ii. Hole diffusion current
 - iii. Hole drift
 - iv. Hole drift current
 - v. Electron diffusion
 - vi. Electron diffusion current
 - vii. Electron drift
 - viii. Electron drift current
 - ix. Electric field for electrons
 - x. Electric field for holes

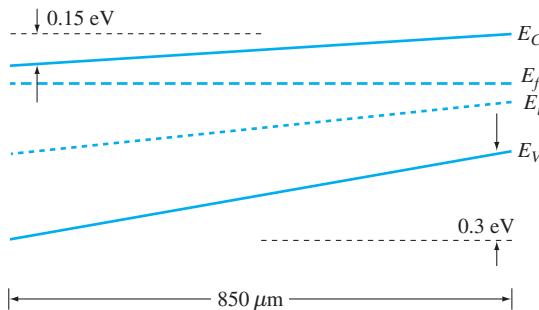


Figure P4.2

- 4.8** Draw the energy band diagram for GaAs that is doped such that $E_C - E_f = 0.2 + (0.8 \text{ eV}/\mu\text{m})x$ for $0 < x < 0.5 \mu\text{m}$.
- Find the effective electric field for electrons.
 - Indicate the directions of $J_{n(\text{diff})}$, $J_{n(\text{drift})}$, $J_{p(\text{diff})}$, and $J_{p(\text{drift})}$.
 - Indicate the direction of the electric field.
- 4.9** A graded-doping n-type Si semiconductor has a phosphorus doping concentration varying linearly from 10^{15} cm^{-3} to 10^{18} cm^{-3} in the region $0 \leq x \leq 1 \mu\text{m}$. Find the electric field as a function of x in this region. What is its value at $x = 0$, 0.5 , and $1 \mu\text{m}$?
- 4.10** Often the net acceptor concentration N'_A in the base region of an npn bipolar junction transistor (BJT) can be approximated as
- $$N'_A = N'_A(x_0)e^{-\frac{(x-x_0)}{\lambda}}.$$
- Determine an expression for the electric field in the base as a function of x .

- 4.11** Find the electric field in the base region of the BJT of Problem 4.10 for a value of $\lambda = 5 \times 10^{-6}$ cm.
- 4.12** A sample of $\text{Al}_y\text{Ga}_{1-y}\text{As}$ has a composition that is graded linearly from pure GaAs to $\text{Al}_{0.35}\text{Ga}_{0.65}\text{As}$ over a length of $8.5 \mu\text{m}$. The sample is not doped. Find the value of the built-in electric field for electrons and holes. The band gap of $\text{Al}_y\text{Ga}_{1-y}\text{As}$ is given by $E_g = 1.43 + 1.425y$ (in eV) for $y < 0.43$. Note the y here is composition fraction, not distance. Assume that the electron affinity is constant and E_i is at midgap.
- 4.13** An optoelectronic device will be made of layers of $\text{Ga}_x\text{In}_{1-x}\text{As}_y\text{P}_{1-y}$ grown on an InP substrate. To make sure the layers will have the same lattice constant as the substrate, the gallium fraction x is kept equal to $0.47y$. The arsenic fraction is varied among the layers from $y = 0$ to $y = 0.5$. The band gap varies as $E_g = 1.35 - 0.72y + 0.12y^2$ eV. Assuming for convenience that the effective masses are constant at m_{dse}^* and m_{dsh}^* , find the ratio of $n_i(y = 0)$ to $n_i(y = 0.5)$ from the pure InP layers to the $y = 0.5$ layer.
- 4.14** Consider the SiGe alloy discussed in Example 4.2. If the acceptor concentration varies exponentially from 10^{18} cm^{-3} to $2 \times 10^{16} \text{ cm}^{-3}$, show that the total effective electric field for electrons is $\mathcal{E}_e^* = \mathcal{E}_{\text{doping}}^* + \mathcal{E}_{\text{composition}}^*$. Assume for simplicity that N_C and N_V are the same for pure silicon and the SiGe alloy. (The effective density of states for electrons N_C actually does not vary much from Si to Ge, but N_V varies by a factor of 10 from one to the other. Since the alloy in question, however, is only 13% Ge, you can roughly take N_V to be constant.)
- 4.15** The energy band diagram of a pn junction, in which the semiconductor changes from n type to p type, is shown in Figure P4.3. Although we have not covered pn junctions, you already have enough knowledge to deduce some things from the energy band diagram.
- In what region is there an electric field?
 - What is the value of the built-in voltage V_{bi} ?
 - Sketch the electron and hole concentrations. Are the directions of the drift and diffusion components of the electron and hole currents correct for equilibrium?

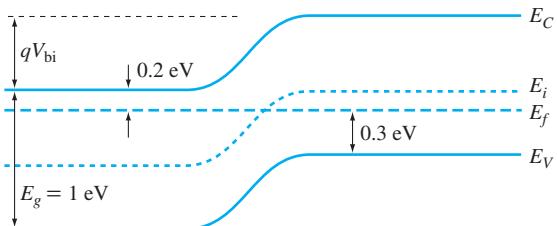


Figure P4.3

Supplement to Part 1: Materials

Throughout this book, supplements are used to expand on the material in each part of the book. In this supplement, several items are discussed.

In Chapter 1, quantum mechanics was used to solve two problems of interest for electrons in semiconductor crystals: the simplified free-electron model and the more realistic quasi-free electron model. In this supplement, quantum mechanics is discussed in slightly more detail, and some specific (though simplified) problems are investigated for electron behavior in semiconductor devices. Some importance is placed on quantum mechanical tunneling of electrons since it has considerable influence on the electrical characteristics of several electronic devices and is a factor in some device failure mechanisms.

The role of phonons on electronic behavior is also discussed since phonons have considerable influence on device operation.

Introduction to Quantum Mechanics

S1.1 INTRODUCTION

A reasonable question in the mind of the student might be, “Why should I know quantum mechanics?”

You saw in Chapter 1 how an understanding of the quantization of the electron states in an atom led eventually to the understanding of energy bands in a crystal. The energy band concept is used repeatedly throughout this book. You also saw how different materials could be expected to have different band structures. The energy band structure can also be changed by the presence of electric fields, crystal defects, impurities, and other factors. You will want to be able to predict the behavior of electrons and holes when they encounter these disturbances because it is precisely these disturbances that make transistors, lasers, and all semiconductor devices work.

The classical (Newtonian) mechanics you learned in freshman physics does not always reliably predict the behavior of the charge carriers (electrons and holes). This is because those carriers have such small sizes, and these nonclassical effects become even more important when the disturbances are also on a scale on the order of nanometers, which is often the case. A qualitative understanding of electron behavior based on quantum mechanics is essential for an understanding of device operation. In Supplement 1, a few principles are presented that will allow you to predict what electrons (and holes) are going to do in the presence of various energy band structures.

S1.2 THE WAVE FUNCTION

All matter can be thought of as consisting of either particles or waves. Since the particle description of electrons is intuitive, we discuss in more detail the other case, in which we consider electrons to be behaving as waves. In quantum mechanics, we say the electron can be described by a wave function $\Psi(x, y, z, t)$.

For simplicity, unless the three-dimensional spatial formulation is required for the concepts being introduced, the one-dimensional formulation will be used here. The concept of wave function is somewhat disturbing the first time one comes across it, partly because it cannot be directly measured. We will see, however, that a knowledge of the wave function along with the *operators* discussed below permit us to predict the behavior of electrons.

Inherent in quantum mechanics is the idea that only the properties of particles (waves) that can be measured, called *observables*, are meaningful. The rule for calculating the average (of several measurements) or expected value for some observable quantity O is:

$$\langle O \rangle = \frac{\int \Psi^*(x, t) O_{\text{op}} \Psi(x, t) dx}{\int \Psi^*(x, t) \Psi(x, t) dx} \quad (\text{S1.1})$$

where Ψ^* is the complex conjugate of the wave function, O_{op} is an operator that is associated with the observable O and operates on the wave function, and the integration is taken over all space. Table S1.1 shows the operators for various

Table S1.1 Quantum mechanical operators

Observable	Operator
x	x
y	y
z	z
r	r
$f(r)$	$f(r)$
v_x	$\frac{\hbar}{jm} \frac{\partial}{\partial x}$
v_y	$\frac{\hbar}{jm} \frac{\partial}{\partial y}$
v_z	$\frac{\hbar}{jm} \frac{\partial}{\partial z}$
$p_x = mv_x$	$\frac{\hbar}{j} \frac{\partial}{\partial x}$
$\vec{p} = \hat{i} p_x + \hat{j} p_y + \hat{k} p_z$	$\vec{p}_{\text{op}} = \hat{i} p_{x_{\text{op}}} + \hat{j} p_{y_{\text{op}}} + \hat{k} p_{z_{\text{op}}}$
v_x^2	$-\frac{\hbar^2}{m^2} \frac{\partial^2}{\partial x^2}$
p_x^2	$-\hbar^2 \frac{\partial^2}{\partial x^2}$
$E_K = \frac{mv^2}{2} = \frac{p^2}{2m} = \frac{\vec{p} \cdot \vec{p}}{2m}$	$-\frac{\hbar^2}{2m} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) = -\frac{\hbar^2}{2m} \nabla^2$
E	$-\frac{\hbar}{j} \frac{\partial}{\partial t}$

observable quantities. Some of the operators, for example, those for position observables, are merely multiplication factors. Others perform an operation such as taking a derivative. We will do an example in the next section, after we have further developed the idea of wave functions.

S1.3 PROBABILITY AND THE WAVE FUNCTION

As indicated in Section 1.5.1, the probability that a particle will be found in region dx near the coordinate x at time t is

$$P(x, t) dx = \Psi^*(x, t)\Psi(x, t) dx \quad (\text{S1.2})$$

where the function $P(x, t)$ is called the *probability density*. Since the particle must be somewhere, the probability of finding it in space at any given time is unity. That is, integrating over all space,

$$\int P(x, t) dx = \int \Psi^*(x, t)\Psi(x, t) dx = 1 \quad (\text{S1.3})$$

in which case the wave function is said to be *normalized*. While Equation (S1.1) is valid for any wave function, for normalized wave functions, the denominator is equal to unity and Equation (S1.1) has the simpler form

$$\langle O \rangle = \int \Psi^* O_{\text{op}} \Psi dx \quad (\text{S1.4})$$

S1.3.1. PARTICLE IN A ONE-DIMENSIONAL POTENTIAL WELL

As an example of applying the operators, we consider an electron of energy E in a potential energy configuration in which the potential energy is some constant $E_p = E_0$ in a region of length L , as shown in Figure S1.1a, and infinitely high everywhere else. Such a structure is called a *potential well*, and the electron inside the well cannot escape. The electron in the well is traveling at some constant and finite energy, so it cannot cross the infinite potential barriers. This is a crude approximation for the potential energy of an electron in an atom (See Figure 1.2), except that here the sides are vertical and go to infinity. From physical intuition, then, we expect that the electron is reflected at each wall and oscillates back and forth. Later, after we have discussed how to find the wave function for a particle, we will show how to find the non-normalized wave functions of the electrons in this potential well. For now, let us accept that the result for the lowest energy state in this problem is¹

¹This results from the complete reflection of the traveling waves at the boundaries of the well where $E_p = \infty$, producing the standing wave as indicated.

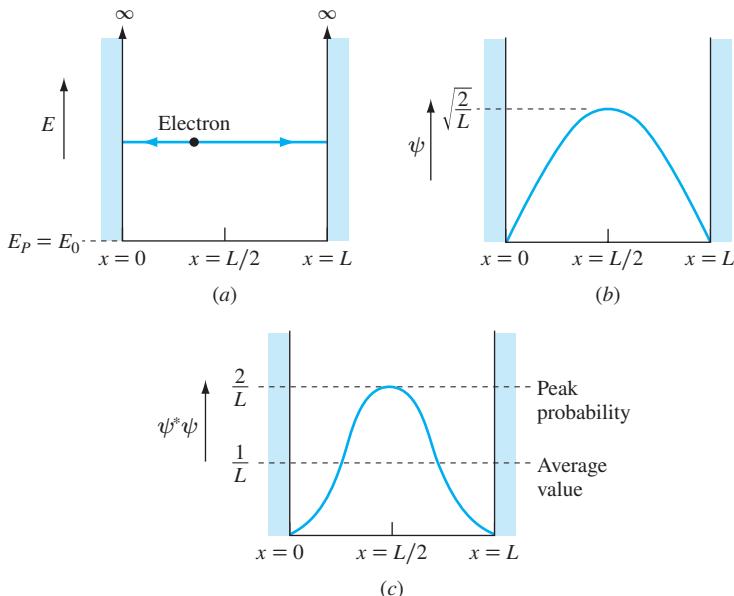


Figure S1.1 Particle in a one-dimensional well. (a) The potential energy diagram. The electron oscillates back and forth between the walls at constant energy. (b) The wave function for the lowest allowed energy state. (c) The probability distribution function.

$$\Psi(x, t) = C \sin\left(\frac{\pi x}{L}\right) e^{-j(E/\hbar)t} \quad 0 < x < L$$

$$\Psi(x, t) = 0 \quad \text{elsewhere}$$

where C is some constant. Since we know that the electron has to be inside the well (it does not have enough energy to escape under any circumstance), the probability is unity that at any time it is somewhere in the well, or

$$\int_0^L \Psi^* \Psi dx = \int_0^L \psi^* \psi dx = 1$$

Here we introduce ψ , the spatial part of wave function Ψ . That is, ψ depends only on x , and the time dependence is treated separately. That is,

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

The time dependence will be derived explicitly later.

We can determine the value of C from the equation by recalling that the probability function integrated over all space must be unity:

$$\int_0^L \Psi^* \Psi dx = \int_0^L \psi^* \psi dx = \int_0^L C^* \sin\left(\frac{\pi x}{L}\right) C \sin\left(\frac{\pi x}{L}\right) dx = 1$$

We simplify our work with a substitution of variables by letting $y = (\pi x/L)$. Now the preceding equation becomes

$$\frac{C^* CL}{\pi} \int_0^\pi \sin^2 y \, dy = 1$$

Since this integral is equal to $\pi/2$, we obtain $C^* C = 2/L$, or

$$C = \sqrt{\frac{2}{L}}$$

Figure S1.1b shows a plot of the wave function $\psi(x)$, and the probability density function $\psi^*\psi$ is shown in Figure S1.1c for this case. Note that the maximum value of $\psi^*\psi$ is $2/L$. The average value of $\psi^*\psi$ is $1/L$ as indicated. This average value, times the well length L , gives the expected unity probability that the particle is in the well.

Now, we find the average position of the electron. Using the normalized wave function and Equation (S1.4) [we have already normalized the wave function, so it is proper to use Equation (S1.4) instead of Equation (S1.1)], we find

$$\begin{aligned}\langle x \rangle &= \int_0^L \Psi^*(x, t) \cdot x \cdot \Psi(x, t) dx = \int_0^L \left(\frac{2}{L}\right)^{1/2} \sin \frac{\pi x}{L} \cdot x \cdot \left(\frac{2}{L}\right)^{1/2} \sin \frac{\pi x}{L} dx \\ &= \frac{2}{L} \int_0^L x \sin^2 \left(\frac{\pi x}{L}\right) dx\end{aligned}$$

This integral is slightly different from the one we used to find C , but we proceed the same way. Again letting $y = \pi x/L$, and using an integral table to find that $\int_0^\pi y \sin^2 y \, dy = \pi^2/4$, we obtain $\langle x \rangle = L/2$ as expected. Thus, the particle's average position of the electron is in the center of the well.

S1.4 SCHRÖDINGER'S EQUATION

We saw that if the wave function for a quantum mechanical particle is known, one can calculate the expected (or average) value of the physical observables, such as position, speed, and momentum. The challenge in quantum mechanics is finding out what the wave functions are for a particular particle in some given situation. Once the wave functions are known, we can determine everything (observables) about the behavior of the particle, with the use of Equation (S1.1) and the appropriate operator for the observable from Table S1.1.

The wave functions are found from Schrödinger's equation, which can be expressed for a one-dimensional problem as

$$-\frac{\hbar^2}{2m} \frac{\partial^2 \Psi(x, t)}{\partial x^2} + E_{P_{op}} \Psi(x, t) = -\frac{\hbar}{j} \frac{\partial \Psi(x, t)}{\partial t} \quad \text{Schrödinger's equation}$$

(S1.5)

Notice that $-(\hbar^2/2m)(\partial^2/\partial x^2)$ is the kinetic energy operator (Table S1.1), that $-(\hbar/j)(\partial/\partial t)$ is the total energy operator, and that $E_{P\text{op}}$ is the potential energy operator. Then Schrödinger's equation can be expressed

$$E_{K\text{op}}\Psi + E_{P\text{op}}\Psi = E_{\text{op}}\Psi \quad (\text{S1.6})$$

This is analogous to the relation in classical mechanics that the kinetic energy plus the potential energy is equal to the total energy:

$$E_K + E_P = E$$

Now, the physical problem to be solved is described through the potential energy, for example the potential energy of a free electron, an electron in an atom, or an electron in a crystal. The potential energy operator depends on the particular problem being solved. The potential energy might vary in space, for example, the way it does in the vicinity of an atom, as seen before in Figure 1.2. For many problems, E_P can be considered to be independent of time to first approximation. In this case

$$-\frac{\hbar^2}{2m}\frac{\partial^2\Psi(x,t)}{\partial x^2} + E_P(x)\Psi(x,t) = -\frac{\hbar}{j}\frac{\partial\Psi(x,t)}{\partial t} \quad (\text{S1.7})$$

where from Table S1.1, $E_{P\text{op}}(x) = E_P(x)$.

S1.5 APPLYING SCHRÖDINGER'S EQUATION TO ELECTRONS

We now have the elements of a procedure for finding the values of observable quantities for electrons. The procedure consists of three steps:

1. Determine the potential energy function. This can be done from knowledge of the forces acting on an electron and considering it to be a point charge.

Since the force is minus the gradient of the potential energy,

$$F = -\nabla E_P \rightarrow -\frac{dE_P}{dx} \quad (\text{S1.8})$$

2. Use the potential energy function in Schrödinger's equation and solve to find the wave function Ψ .
3. Insert the wave function Ψ into Equation (S1.1) along with the appropriate operator to find the value of the corresponding observable.

While in principle this appears straightforward, only a few physical problems have simple analytical solutions in closed form. There is some value, however, in considering simplified cases—they may not be quantitatively realistic, but they can provide qualitative insight.

For the case in which the potential energy is not a function of time, Schrödinger's equation can be solved in part by using the technique of separation

of variables. The result is that the wave function can be expressed as the product of two functions, one time-independent $\psi(x)$ and one space-independent $T(t)$:

$$\Psi(x, t) = \psi(x)T(t) \quad (\text{S1.9})$$

where the lowercase ψ represents the time-independent wave function as before and uppercase Ψ the complete wave function. Equation (S1.9) is inserted into Schrödinger's equation, which is then solved. The result is two separate ordinary differential equations, each with the same constant of separation, E .

The time-independent equation is

$$-\frac{\hbar^2}{2m} \frac{d^2\psi(x)}{dx^2} + E_P(x)\psi(x) = E\psi(x) \quad (\text{S1.10})$$

and the time-dependent equation is

$$-\frac{\hbar}{i} \frac{dT(t)}{dt} = ET(t) \quad (\text{S1.11})$$

We can solve Equation (S1.11) immediately. The solution to the time-dependent equation is

$$T(t) = e^{-j(E/\hbar)t} \quad (\text{S1.12})$$

where E is the constant of separation. This constant E is the total energy of the electron.

The time-independent wave function $\psi(x)$ can be found from the time-independent Schrödinger equation, Equation (S1.10). The solution to this one, however, depends on the particular form of $E_P(x)$, which in turn depends on the particular problem being solved.

Once we have the appropriate expression for $E_P(x)$, Schrödinger's equation can be solved by using the boundary conditions appropriate to the problem. There are two restrictions on the solutions of wave functions $\Psi(x, t)$ and $\psi(x)$:

1. The wave function must be finite, single-valued, and continuous in space.
2. The gradient of the wave function must be continuous.

We will present the results of some interesting (and important) problems in the next section.

S1.6 SOME RESULTS FROM QUANTUM MECHANICS

In Section 1.6, Schrödinger's equation was solved for two cases of interest: the free electron, in which the electron potential energy was considered constant with position (E_0), and the more realistic model for electrons in crystals, the quasi-free electron model in which the potential energy was considered to be

a periodic function of position with the periodicity of the lattice. Some of the major results for these two cases are repeated here.

S1.6.1 THE FREE ELECTRON

The wave function for the free electron can be expressed as [recall Equation (1.42)]

$$\Psi(x, t) = A e^{j[Kx - (E/\hbar)t]} \quad \text{one dimension} \quad (\text{S1.13})$$

$$\Psi(\vec{r}, t) = A e^{j[\vec{K} \cdot \vec{r} - (E/\hbar)t]} \quad \text{three dimensions} \quad (\text{S1.14})$$

where A is a constant and represents the amplitude of the wave function. The total electron energy is represented by E and the wave vector by K , where the magnitude of K is [Equation (1.44)]

$$K = \frac{2\pi}{\lambda} = \sqrt{\frac{2m_0(E - E_0)}{\hbar^2}} = \sqrt{\frac{2m_0E_K}{\hbar^2}} \quad (\text{S1.15})$$

and the direction of K represents the direction of motion. Here, the kinetic energy is $E_K = (E - E_0)$.

The group velocity, $v_g = v$ of the electron wave is the velocity of its center of mass:

$$\begin{aligned} v_g &= \frac{1}{\hbar} \frac{dE}{dK} && \text{one dimension} \\ \vec{v}_g &= \frac{1}{\hbar} \nabla_K E && \text{three dimensions} \end{aligned} \quad (\text{S1.16})$$

where $\nabla_K E$ represents the gradient of E in K space.

The total energy can be expressed as

$$\begin{aligned} E &= E_0 + \frac{\hbar^2 K^2}{2m_0} && \text{one dimension} \\ E &= E_0 + \frac{\hbar^2}{2m_0} (K_x^2 + K_y^2 + K_z^2) = E_0 + \frac{\hbar^2 K^2}{2m_0} && \text{three dimensions} \end{aligned} \quad (\text{S1.17})$$

Equations (S1.17) are parabolas in one dimension (recall Figure 1.13) and paraboloids (three dimensions) respectively. All are with minima at $K = 0$ and $E = E_0$. The mass of the free electron is related to the curvature of the E - K plot:

$$m_0 = \hbar^2 \left(\frac{d^2 E}{dK^2} \right)^{-1} \quad (\text{S1.18})$$

EXAMPLE S1.1

Show that the momentum of the free electron is given by $\hbar K$. We will do this by calculating the average momentum of the free electron using quantum mechanics.

Solution

We use Equation (S1.1) and the momentum operator from Table S1.1. For an electron traveling in the positive x direction, the non-normalized wave function in one dimension is $\Psi(x, t) = Ae^{j[Kx - (E/\hbar)t]}$. The average momentum is then

$$\begin{aligned}\langle p_x \rangle &= \frac{\int [A^* e^{-j[Kx - (E/\hbar)t]}] \frac{\hbar}{j} \frac{\partial}{\partial x} [Ae^{j[Kx - (E/\hbar)t]}] dx}{\int [A^* e^{-j[Kx - (E/\hbar)t]}] [Ae^{j[Kx - (E/\hbar)t]}] dx} \\ &= \frac{\hbar(jK)}{j} \frac{\int [A^* e^{-j[Kx - (E/\hbar)t]}] [Ae^{j[Kx - (E/\hbar)t]}] dx}{\int [A^* e^{-j[Kx - (E/\hbar)t]}] [Ae^{j[Kx - (E/\hbar)t]}] dx} \quad \text{free electron} \\ &= \hbar K\end{aligned}$$

Thus, for the free electron, the quantity $\hbar K$ is the momentum of the particle.

Similarly, for a free electron in three dimensions,

$$\vec{p} = \hbar \vec{K} \quad \text{free electron in three dimensions}$$

since the curvature of the E - K relation is independent of the direction of K .

S1.6.2 THE QUASI-FREE ELECTRON

Recall from Chapter 1 that the solution to Schrödinger's equation for an electron in a periodic potential such as a crystal is a Bloch wave function [Equation (1.57)]:

$$\Psi(x, t) = U_K(x) e^{j[Kx - (E/\hbar)t]} \quad (\text{S1.19})$$

$$\Psi(\vec{r}, t) = U_K(\vec{r}) e^{j[\vec{K} \cdot \vec{r} - (E/\hbar)t]} \quad (\text{S1.20})$$

where U_K is a function with the periodicity of the crystal. Here the E - K relation cannot be determined unless $E_P(x)$ or $E_P(\vec{r})$ is known. However, as for a free electron (recall Section 1.6.4),

$$|\vec{K}| = \frac{2\pi}{\lambda} \quad (\text{S1.21})$$

and the group velocities are again

$v_g = \frac{1}{\hbar} \frac{dE}{dK}$ $\vec{v}_g = \frac{1}{\hbar} \nabla_K E$	one dimension three dimensions
---	-----------------------------------

(S1.16)

The quasi-free electron model predicts multiple allowed energy bands, each with its own E - K relation.

Within any band, the E - K relation is periodic in K space with periodicity

$$\frac{2\pi}{a} \quad \text{one dimension}$$

$$\left(\frac{2\pi}{a}, \frac{2\pi}{b}, \frac{2\pi}{c}\right) \quad \text{three dimensions}$$

S1.6.3 THE POTENTIAL ENERGY WELL

We have so far considered the free electron and the quasi-free electron, which were covered in Chapter 1. In both models, we considered the case of an electron deep within the crystal such that the surface effects were neglected. Here we will consider the effects of the surfaces, which constitute high potential barriers to the electron. The electrons are essentially trapped in the resulting potential well. This situation is of considerable importance in many semiconductor devices.

In Figure S1.2a we show the case of a one-dimensional crystal indicating the periodic potential within the crystal and the potential energy at the surface. Physically, we expect that an electron approaching the surface will be decelerated and repelled, because the force is equal to minus the gradient of the potential energy.

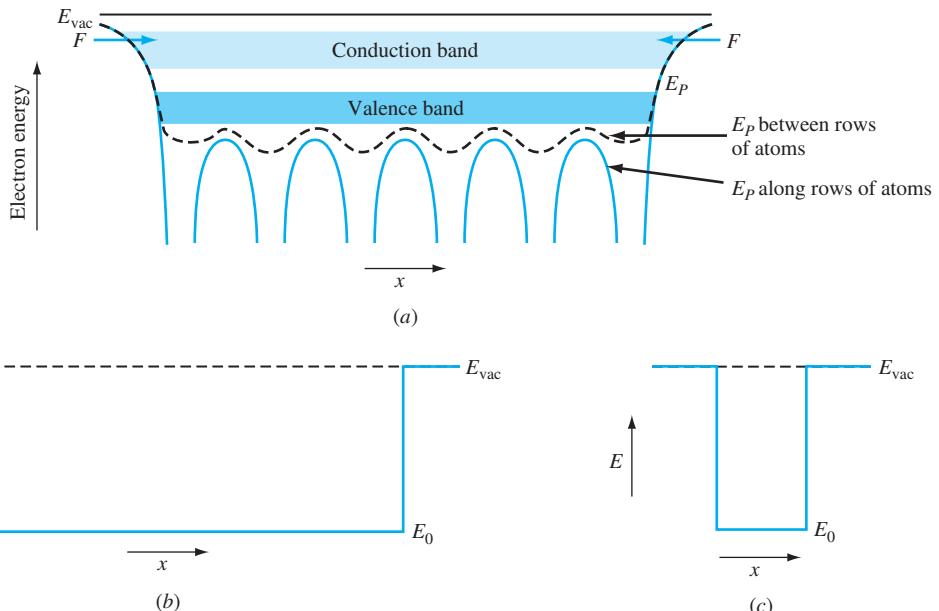


Figure S1.2 The potential energy for an electron in a crystal. (a) The actual potential, including the effects of the surfaces; (b) the idealized potential, which is easier to solve. (c) The potential energy diagram for an electron in an idealized but thin crystal.

In Figure S1.2a, the slope of the left potential energy barrier, for example, is negative, so the force is in the positive x direction, or to the right. An electron traveling toward the left surface will be decelerated and turned around. At the right edge, the slope is positive and therefore the force is in the negative x direction. The electron is trapped in the well. Between collisions, electrons further than about a mean free path from the surface will behave as quasi-free particles.

The potential energy distribution in Figure S1.2a would be hard to describe analytically, but we could simplify the problem as shown in Figure S1.2b. Here we replace the surface potentials with abrupt potential energy barriers, and we replace the periodic crystal potential with a constant potential energy.

When we considered the quasi-free electron, we assumed the material was so thick that the electron never came near enough to the surfaces to be influenced by those barriers. We are now going to complicate the problem by letting the material be so thin that the electron is definitely influenced by the interfaces at the surface; i.e., the electron mean free path is much larger than the width of the well. Consider a very thin sheet of this crystalline material, Figure S1.2c. There are two symmetrical energy barriers here, but now they are very close together (in space), on the order of a few lattice constants. These structures are approximations to quantum wells in light-emitting diodes and lasers, and to some field-effect transistors.

S1.6.4 THE INFINITE POTENTIAL WELL IN ONE DIMENSION

In Example S1.1, we presented the infinite potential well and gave the result for the wave function of the electron in the lowest allowed state. We revisit that problem now, but this time we will solve it. Recall that we assumed the barriers are so high in energy that they appear infinite. This problem is not very realistic but it is easy to solve.

The potential energy diagram for the one-dimensional infinite potential well is shown in Figure S1.3a. We can write the potential energy as

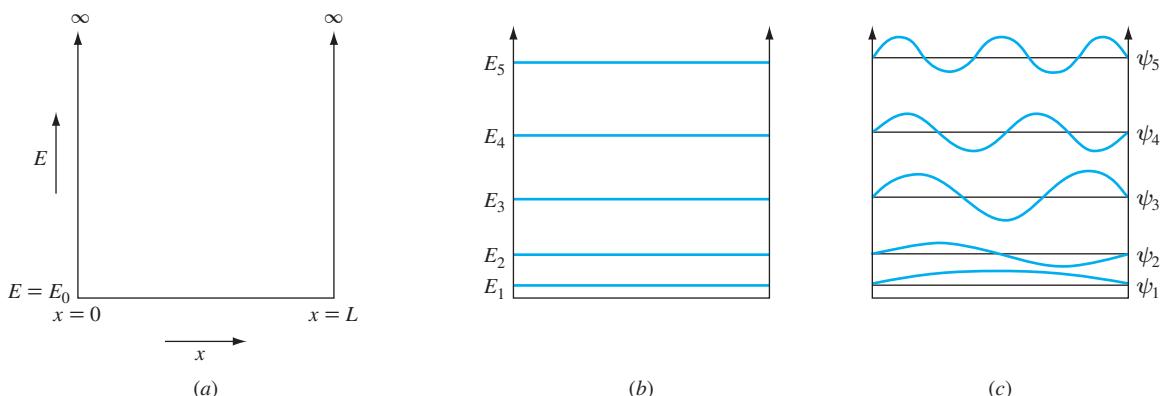


Figure S1.3 The infinite potential well. (a) The potential energy is E_0 everywhere except at $x \leq 0$ and $x \geq L$, where the potential is infinite; (b) the first five energy levels for the electron in the well; (c) the corresponding wave functions.

$$E_P(x) = E_0 \quad 0 \leq x \leq L \quad (\text{S1.22})$$

$$E_P(x) = \infty \quad x \leq 0 \text{ and } x \geq L \quad (\text{S1.23})$$

Now that we have an expression for the potential energy, we can write the time-independent Schrödinger's equation [Equation (S1.10)]:

$$-\frac{\hbar^2}{2m^*} \frac{d^2\psi(x)}{dx^2} + E_0\psi(x) = E\psi(x) \quad 0 \leq x \leq L \quad (\text{S1.24})$$

$$-\frac{\hbar^2}{2m^*} \frac{d^2\psi(x)}{dx^2} + (\infty)\psi(x) = E\psi(x) \quad x \leq 0 \text{ and } x \geq L \quad (\text{S1.25})$$

The time dependence is $e^{-(jE/\hbar)t}$ as usual [Equation (S1.12)], since the potential energy is not changing with time.

The electron cannot possibly be found outside the well, since it would have to have infinite potential energy. Since the probability distribution function for the electron is $\psi^*\psi$, this quantity must be zero outside the well. Therefore, the only solution in these two regions is $\psi = 0$.

Schrödinger's equation for the region inside the well, Equation (S1.24), is the same as the equation for the free electron [Equation (1.36)] because the potential is constant in this region. The only difference is that here we have used the effective mass m^* instead of m_0 since this electron is not strictly free. The solution was found earlier [Equation (1.38)] as:

$$\psi(x) = Ae^{jKx} + Be^{-jKx} \quad (\text{S1.26})$$

where

$$K = \pm \sqrt{\frac{2m^*}{\hbar^2}(E - E_0)} \quad (\text{S1.27})$$

Note that in this case this solution applies only inside the well.

We can expect that the electron will oscillate back and forth inside the well, so we must keep both the positive-traveling and the negative-traveling solutions in Equation (S1.27).

Next, we invoke the boundary conditions. Recall that $\psi(x)$ must be continuous. Since we have established that the wave function $\psi(x)$ is zero outside the well, $\psi(x)$ must also go to zero at $x = 0$ and $x = L$ for the solution inside the well. With these boundary conditions on Equation (S1.26) we find

$$\psi(x = 0) = 0 = A + B \quad (\text{S1.28})$$

$$\psi(x = L) = 0 = Ae^{jKL} + Be^{-jKL} \quad (\text{S1.29})$$

From the first of these,

$$A = -B \quad (\text{S1.30})$$

meaning

$$\psi(x) = A(e^{jKx} - e^{-jKx}) = +2jA \sin(Kx) = C \sin(Kx) \quad (\text{S1.31})$$

where C is a constant ($C = +2jA$), and we have used the Euler relation $\sin \theta = (e^{j\theta} - e^{-j\theta})/2j$. The second condition, (S1.29), can be satisfied only for

$$KL = n\pi \quad n = \pm 1, \pm 2, \pm 3, \dots \quad (\text{S1.32})$$

This means that the wave vector K is quantized (takes on discrete values) and can have only the particular values

$$K = \frac{n\pi}{L} \quad (\text{S1.33})$$

Now, from Equation (S1.27), we know that K is related to the kinetic energy E_K of the electron, since $E_K = E - E_0$, so therefore the kinetic energy is also quantized:

$$E_K = (E - E_0) = \frac{\hbar^2 K^2}{2m^*} = \frac{\hbar^2 \pi^2}{2m^* L^2} n^2 \quad (\text{S1.34})$$

Only discrete energies are allowed, just as we found for the atom. The first few energies are shown in Figure S1.3b. The actual energies are different from the case for the hydrogen atom, since the basic problems are different, but the quantization of the energy levels is a common feature.

Since n can be any integer, there are an infinite number of solutions to the time-independent Schrödinger's equation in the infinite potential well:

$$\psi_n(x) = C \sin(K_n x) = C \sin\left(\frac{n\pi x}{L}\right) \quad (\text{S1.35})$$

each of which has the form of a standing wave. The first five solutions are shown in Figure S1.3c.

The complete solution contains the time dependence as well. From Equations (S1.9) and (S1.12),

$$\Psi_n(x, t) = C \sin(K_n x) e^{-j(E_n/\hbar)t} \quad (\text{S1.36})$$

Note that the wave vector K_n can be positive or negative, depending on the sign of n . From Equation (S1.34) then, we see that a given magnitude of n ($n = |\pm 1|, |\pm 2|, \dots$) results in the same value of E_K .

Also, notice that the wider the well gets (L), the closer together the energy levels get, as indicated in Figure S1.4. For an infinite L [part (c) of the figure], we'd have a continuum of allowed energies. As the well gets narrow enough, the states separate into discrete energy levels [(a) and (b)]. For such a case, the potential well is referred to as a *quantum well*.

The subscript n is a quantum number because it identifies which quantum state is being discussed. Each wave function ψ_n is called an *eigenfunction*, describing a particular state (*eigenstate*); each quantized parameter such as K_n and E_n is called an *eigenvalue*.

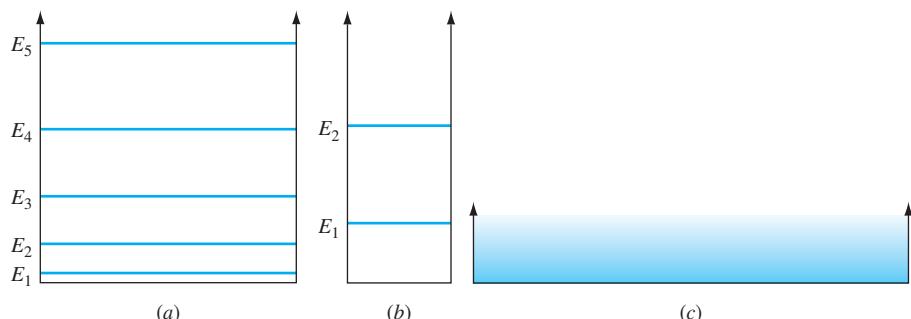


Figure S1.4 The effect of the well width on the energy levels in an infinite potential well. (a) Some discrete states; (b) as the well gets narrowed the energy levels are spaced further apart; (c) for a very wide well the energy levels are so close together they appear to be quasi-continuous (an energy band).

S1.6.5 REFLECTION AND TRANSMISSION AT A FINITE POTENTIAL BARRIER

We saw that the electron wave in the infinite well reflected off the two infinite barriers and oscillated back and forth in the well. We will now consider a case where the electron approaches a potential barrier that is finite. There are two possible configurations for this problem. The electron approaching the barrier can have enough energy to go over the barrier, or the electron's energy can be smaller than the barrier.

We begin by assuming the electron has energy *higher* than the barrier. Consider an electron in a material, traveling toward one of the surface barriers, as in Figure S1.5. We assume an abrupt barrier of finite height. Classical mechanics would predict that the electron would continue past the barrier, although with lower kinetic energy.

As would be suspected, the results will be different when we consider very small particles like electrons, where classical mechanics doesn't apply. In

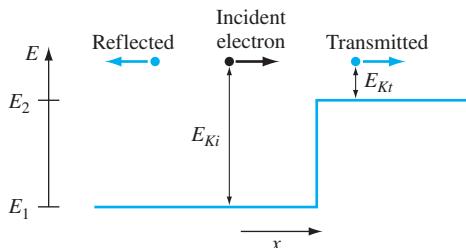


Figure S1.5 Reflection and transmission of an electron by a finite potential barrier.

quantum mechanics, we find that sometimes the electron will go past the barrier, and sometimes it will turn back. For any given approach to the barrier, we cannot say for certain whether the electron will be reflected or continue on. Using Schrödinger's equation, however, we can predict the probability of reflection or transmission for any given attempt at the barrier by the electron.

As always, we need to know the wave function to predict the behavior of the electron. The incident electron is in a region of constant potential energy. Until it gets close enough to be influenced by the barrier, it looks like a free electron. We can write the wave function ψ_i for the incident electron, having positive K since it is traveling to the right, as:

$$\psi_i = A e^{jK_i x} \quad (\text{S1.37})$$

where

$$K_i = \sqrt{\frac{2m}{\hbar^2} E_{Ki}} \quad (\text{S1.38})$$

If the electron wave is reflected, we can write

$$\psi_r = B e^{-jK_r x} \quad (\text{S1.39})$$

where $K_r = K_i$ since $E_{Kr} = E_{Ki}$.

For the transmitted wave, the wave function reflects the changed potential and kinetic energies:

$$\psi_t = C e^{jK_t x} \quad (\text{S1.40})$$

where the subscripts i , r , and t refer to incident, reflected, and transmitted waves respectively.

We have expressions now for the wave functions for the electron whether it is incident, reflected, or transmitted. The relationships between coefficients A , B , and C will tell us the probability of reflection and transmission. These can be found by using the boundary conditions, and the analysis is the subject of a homework problem. We give the results here. The probability of the electron reflecting from the barrier is

$$R = \left[\frac{K_i - K_t}{K_i + K_t} \right]^2 \quad (\text{S1.41})$$

and the probability of its being transmitted across the barrier is

$$T = 1 - R = \frac{4 K_i K_t}{(K_i + K_t)^2} \quad (\text{S1.42})$$

Note that for an electron at energy E crossing the barrier, from Equations (S1.41) and (S1.42), the reflection and transmission probabilities are independent of the direction of the electron. That is, it doesn't matter whether the electron approaches the barrier from the low-energy or high-energy side.

EXAMPLE S1.2

Obtain an expression for the reflection coefficient in terms of total energy and the potential energy barrier ΔE_P .

■ Solution

Since $E_K = E - E_P = \hbar^2 K^2 / 2m^*$,

$$K_i = \sqrt{\frac{2m^*}{\hbar^2} (E - E_{Pi})}$$

$$K_t = \sqrt{\frac{2m^*}{\hbar^2} (E - E_{Pt})}$$

where E_{Pi} and E_{Pt} are respectively the potential energies of the incident and transmitted electrons. Letting $\Delta E_P = (E_{Pt} - E_{Pi})$ gives, from Equation (S1.41),

$$R = \left[\frac{1 - \sqrt{1 - \frac{\Delta E_P}{E}}}{1 + \sqrt{1 - \frac{\Delta E_P}{E}}} \right]^2$$

S1.6.6 TUNNELING

Recall from Section 1.6.5 that, in quantum mechanics, an electron may tunnel through to the other side of a potential barrier if the barrier is thin enough. Here we consider three cases of tunneling. First we consider the case in which the tunneling barrier has a finite width, the case we saw in Chapter 1. We then consider the case for infinite width and lastly the case in which the electron tunnels into the forbidden gap of a semiconductor and then returns, a common occurrence in semiconductor devices.

Case 1: Finite Barrier Width We repeat the problem of the previous section, but this time we take the energy of the electron to be *less* than the height of the barrier, and the barrier to have width L , as shown in Figure S1.6. From a classical point of view, this should turn out to be a pointless exercise because, of course, the particle cannot cross the barrier; it does not have enough energy, so it must be reflected at the barrier. From a quantum mechanical point of view, we will obtain very different results. The result is a concept that has no analog in classical mechanics: the ability of a particle to pass through a region where its total energy is less than its potential energy.

In Figure S1.6, assume the electron starts out in region a , with a total energy E that is less than the barrier height. We will show that the electron can travel through the barrier, even though this region of space is classically forbidden. Keep in mind that the energy of the electron is conserved, so it travels at constant energy.

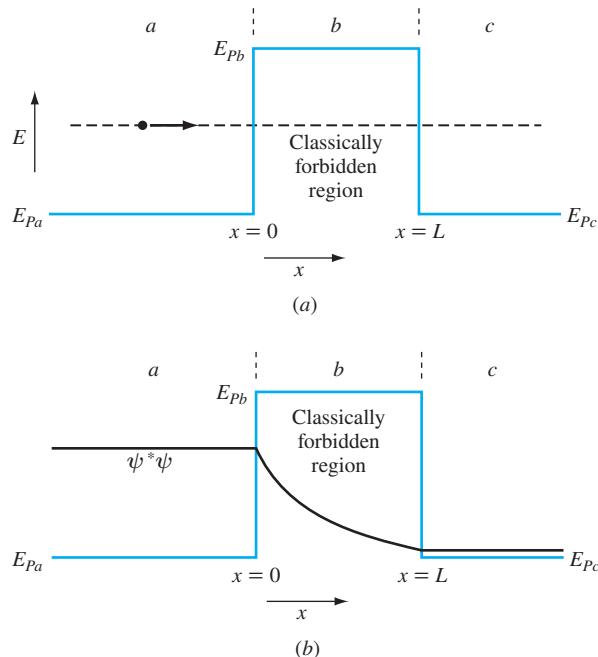


Figure S1.6 Tunneling. (a) An electron approaches a finite potential barrier; (b) the probability density function shows there is a small chance the electron will appear on the far side of the barrier.

We begin by writing the wave function of the incident electron in region *a* as

$$\psi_a = A e^{i K x} = A e^{i \sqrt{(2m^*/\hbar^2)(E - E_{Pa})} x} \quad (\text{S1.43})$$

This is the usual solution for an electron in a region of constant potential energy, and it could have been written as sines and cosines. The point is it has an oscillatory wavelike nature. (The figure shows not the wave function, but $\psi^* \psi$, which is constant in region *a*, as we will see later.) Let us neglect reflection at the *a/b* boundary ($x = 0$). Then the amplitude of the wave function at $x = 0$ is A .

In the forbidden region, since $K = \sqrt{(2m^*/\hbar^2)(E - E_{Pb})}$, and $(E - E_{Pb})$ is negative, K is imaginary in region *b*. We therefore have

$$\psi_b = A e^{-\sqrt{(2m^*/\hbar^2)(E_{Pb}-E)} x} \quad (\text{S1.44})$$

In this case the exponent in the wave function is real, so this function does not oscillate but decays exponentially with x .

Both ψ_a and ψ_b have the same coefficient because ψ must be continuous across the boundary $x = 0$ (since reflection at the boundary is neglected). At $x = L$, just inside the far side of the barrier, we have

$$\psi_b(L) = A e^{-\sqrt{(2m^*/\hbar^2)(E_{Pb} - E)}L} \quad (\text{S1.45})$$

We also neglect the reflection at this boundary. Then in region *c* we have

$$\psi_c(x) = C e^{j\sqrt{(2m^*/\hbar^2)(E - E_{Pc})}x} \quad (\text{S1.46})$$

where

$$C = A e^{-\sqrt{(2m^*/\hbar^2)(E_{Pb} - E)}L} \quad (\text{S1.47})$$

Let us examine these results. In regions *a* and *c*, the wave functions are oscillatory. The probability density function $\psi^* \psi$ is $\psi^* \psi = A^2$ in region *a* and $\psi^* \psi = C^2$ in region *c*. These are both constant, as shown in Figure S1.6b.

In region *b*, however, the wave function is a decaying exponential function. The probability density function $\psi^* \psi$ will also be a decaying exponential, given by

$$\boxed{\psi^* \psi = A^2 e^{-2\sqrt{(2m^*/\hbar^2)(E_{Pb} - E)}x}} \quad (\text{S1.48})$$

We interpret the figure, then, to suggest that there is some reduced, but non-zero, probability of finding the electron on the opposite side of the barrier. The electron will be reflected part of the time, but not always.

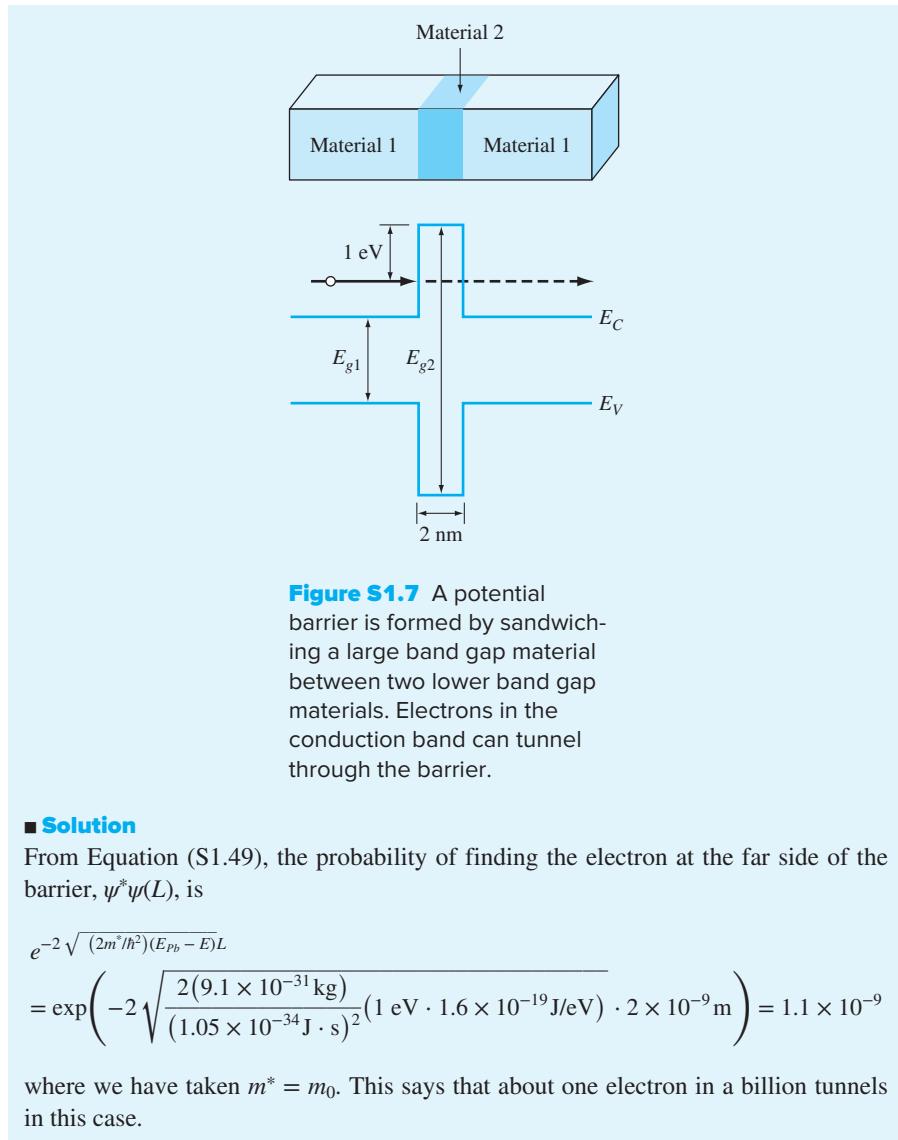
It seems from the figure that the thicker the barrier, the smaller the probability of finding the electron on the other side. We can calculate this probability by using the boundary conditions. The wave function ψ must be continuous across both boundaries. Since we are neglecting reflection at the boundaries, this becomes

$$\boxed{\frac{\psi_b^* \psi_b(L)}{\psi_b^* \psi_b(0)} = e^{-2\sqrt{(2m^*/\hbar^2)(E_{Pb} - E)}L}} \quad (\text{S1.49})$$

EXAMPLE S1.3

Let us consider how to make a tunneling structure in semiconductors. Consider a semiconductor material of band gap E_{g1} . We grow a thin (2 nm) layer of another material whose band gap is larger, E_{g2} . Then we add a thick layer of the first material. The resulting energy band diagram is shown in Figure S1.7.

Now suppose an electron in the conduction band approaches the thin layer at energy 1 eV below the top of the barrier. Neglecting reflection and the influence of the valence band edge in the wide-gap material, what is the probability that the electron will penetrate the barrier?



While the probability of tunneling in this example may seem small, the number of incident electrons in a real material can be large enough that this thin insulating region can, in fact, conduct electrons. Tunneling through barriers similar to these is used extensively in nonvolatile memories (e.g. solid-state drives and flash memories). Electron tunneling is also the basis for tunnel diodes and tunnel field effect transistors.

Let us do another example. While this text is about semiconductor materials and devices, it is instructive to consider a case of tunneling in superconductors.

EXAMPLE S1.4

SUPERCONDUCTOR (GIAEVER) TUNNELING

As another example of tunneling, we consider two superconductors (Sn) separated by a thin semiconductor (SnO_2). This structure, shown schematically in Figure S1.8a, is fabricated by oxidizing the surface of a sample of tin, producing about 2 nm of SnO_2 .

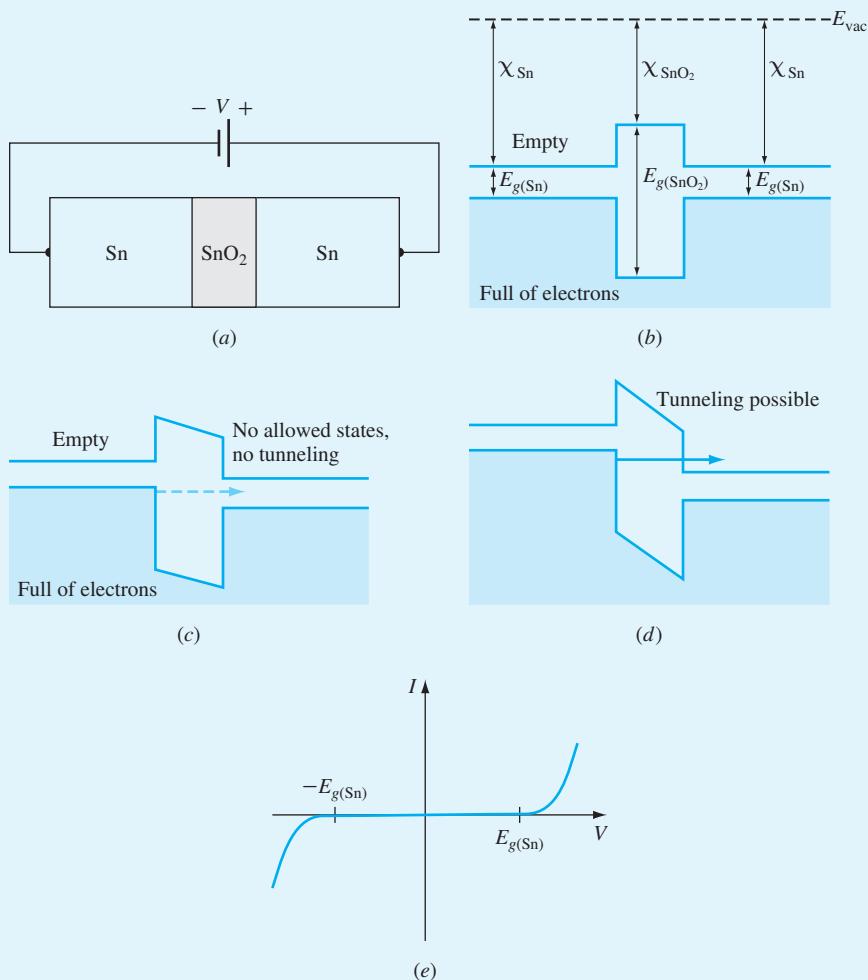


Figure S1.8 Giaever tunneling. (a) The physical structure consists of two layers of superconductor (Sn is a superconductor below 3.7 K) with a thin layer of semiconductor (SnO_2) between them. (b) The energy band diagram at equilibrium. (c) The energy band diagram under small bias. Electrons cannot tunnel because empty states are not available at the same energies as those occupied by electrons. (d) Under larger bias, electrons from the valence band on the left tunnel through the forbidden region to empty states in the conduction band on the right. (e) A plot of current versus voltage for this device.

Then another layer of Sn is deposited, and electrical contacts are made to each end. While Sn is normally considered a metal, below 3.7 K it becomes a superconductor (zero electrical resistance). Furthermore, in a superconductor the normally continuous conduction band splits into two bands separated by a small forbidden energy E_g (1.15 meV for Sn). The equilibrium energy band diagram of this structure is shown schematically (not to scale) in Figure S1.8b. The applied voltage is zero. The lower energy band is completely filled with electrons while the upper band is completely empty because of the low temperature. The thin layer of SnO_2 is a semiconductor with a forbidden band on the order of four electron volts.

At equilibrium, we do not expect current to flow. The bottom band is filled, but there are no empty states at the same energy as a filled state, so the electrons cannot move since they must move at constant energy.

We now apply a small voltage (e.g., 0.5 mV) across the device. The energy band diagram will appear as shown in Figure S1.8c. The band tilts in the SnO_2 because the applied voltage is dropped across the semiconductor (the superconductors, having infinite conductivity, cannot support a voltage), and a change in voltage implies a change in potential energy. Remember that the energy band diagrams are for the negatively charged electrons, and so on these diagrams positive voltage corresponds to a reduced electron potential energy. Tunneling occurs at constant energy, and when tunneling does occur, an electron tunnels from a filled state to an empty state. In this case, as at equilibrium, there are no empty states on the far side of the barrier at the same energies as filled states on the near side, so current still cannot flow.

Figure S1.8d shows the energy band diagrams for an applied voltage just greater than the band gap of Sn. Now electrons can tunnel from the filled states on the left to the empty states on the right, and current flows. The current-voltage characteristic of this device is shown in (e). Note that since the structure is symmetrical, so is the I - V characteristic.

While this structure is not very useful as an electronic device, it does provide a means of measuring the band gap of superconducting materials. Incidentally, this experiment with its interpretation, along with another type of tunneling (Josephson tunneling) in superconductors and tunneling in semiconductors (Esaki tunneling), resulted in a Nobel Prize in 1973, shared by Ivar Giaever, Brian D. Josephson, and Leo Esaki.

Case 2: Infinite Barrier Width Next, suppose the region b in Figure S1.6 is infinitely thick. For $L = \infty$, from Equation (S1.49), $\psi^* \psi$ decays exponentially to zero. This implies, however, that the probability of finding the electron on the other side of the barrier (in the classically forbidden region) is not zero, at least not close to the barrier. But the electron cannot penetrate infinitely far, either, since its probability density function decays to zero. We thus conclude that the electron penetrates some distance past the finite height barrier, but is reflected back to the low potential energy side.

Since the electron has some charge associated with it, we can say that some of that charge is actually within the classically forbidden region.

Case 3: Tunneling into the Forbidden Band of a Semiconductor Next, consider the case of Si, which has a band gap of $E_g = 1.12$ eV. An electron is incident to the Si surface at energy E where $E_V < E < E_C$, as indicated in Figure S1.9a. The electron can tunnel some distance into the silicon, even though classically its energy is forbidden. This situation actually resembles reflection from a barrier because the electron will penetrate some distance into the silicon and then be repelled back. While the electron is in the silicon, however, its charge can affect the potentials near the surface and alter the energy band structure. Here we investigate the distance that an electron can tunnel into the silicon and calculate the associated charge concentration.

For this situation (tunneling into a forbidden region of a material) Equation (S1.48) has the form

$$\psi^* \psi = A^2 e^{-2 \sqrt{(2m^*/\hbar^2)(E_P^* - E)} x} \quad (\text{S1.50})$$

where m^* represents the electron effective mass. We know that an electron in the conduction band has a particular effective mass, and a hole in the valence band has a (different) effective mass, but what about an electron in the forbidden gap? Its effective mass is some combination of electron and hole effective masses (tunneling effective mass). Next, the term E_P^* is the effective potential energy of the tunneling electron. We have the same difficulty here—the potential energy for an electron in the conduction band is E_C and the potential energy for a hole in the valence band is E_V . The electron in the forbidden region sees two potential energies, E_C and E_V . The term $(E_P^* - E)$ in Equation (S1.50) then has the form

$$(E_P^* - E) = \frac{(E_C - E)(E - E_V)}{(E_C - E) + (E - E_V)} = \frac{(E_C - E) + (E - E_V)}{E_g} \quad (\text{S1.51})$$

similar to the effective resistance of two resistors in parallel. Since $(E_P^* - E)$ is a function of E , so is $\psi^* \psi$.

EXAMPLE S1.5

We define the characteristic tunneling distance x_T as the value of x such that $\psi^* \psi(x_T)/\psi^* \psi(0) = e^{-1}$. Estimate the characteristic tunneling distance into the forbidden band of Si as a function of energy. Assume $m^* = m_0/2$ and is independent of energy.

Solution

From the definition of characteristic tunneling distance in the problem statement,

$$\frac{\psi^* \psi(x_T)}{\psi^* \psi(0)} = e^{-2 \sqrt{(2m^*/\hbar^2)(E_P^* - E)} x_T} = e^{-1}$$

or

$$x_T = \frac{1}{2 \sqrt{\frac{2m^*}{\hbar^2}(E_P^* - E)}} \quad (\text{S1.52})$$

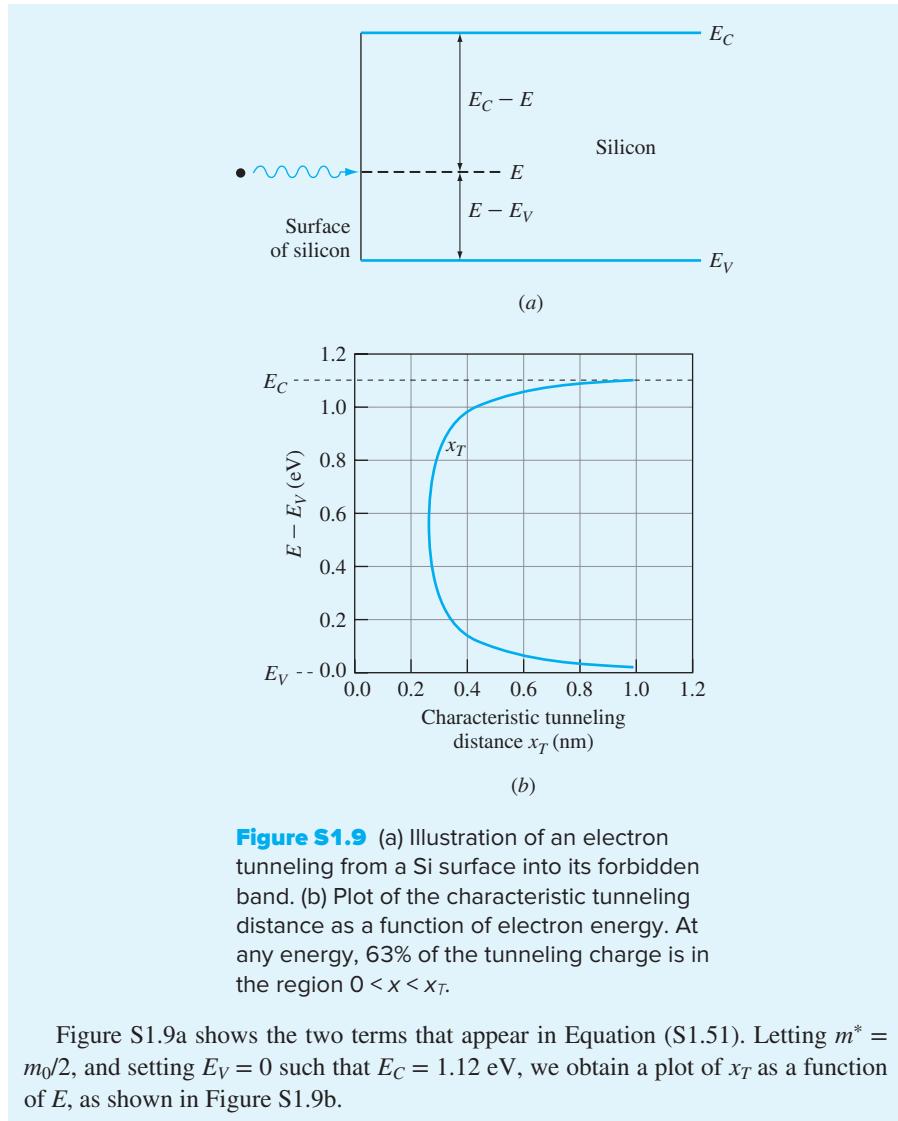


Figure S1.9 (a) Illustration of an electron tunneling from a Si surface into its forbidden band. (b) Plot of the characteristic tunneling distance as a function of electron energy. At any energy, 63% of the tunneling charge is in the region $0 < x < x_T$.

Figure S1.9a shows the two terms that appear in Equation (S1.51). Letting $m^* = m_0/2$, and setting $E_V = 0$ such that $E_C = 1.12$ eV, we obtain a plot of x_T as a function of E , as shown in Figure S1.9b.

Note that the classically forbidden region contains charge. In this region, from Equation (S1.50) and Figure S1.6,

$$\psi^* \psi(x) = A^2 e^{-2 \sqrt{(2m^*/\hbar^2)(E_p^* - E)} x}$$

Since the probability of finding an electron in a region dx is $\psi^* \psi dx$ and since the electron has charge q , the forbidden region will contain electronic charge density $Q_{VT}(x)$ (tunneling charge density) proportional to $\psi^* \psi(x)$; i.e., the charge density will decrease exponentially with distance into the classically forbidden region.

Since for a given electron $Q_{VT} dx = -q\psi^*\psi dx$, the fractional tunneling charge within one tunneling characteristic length x_T is

$$\frac{Q_{VT}(x < x_T)}{Q_{VT}(x < \infty)} = \frac{\int_0^{x_T} \psi^* \psi dx}{\int_0^\infty \psi^* \psi dx} = (1 - e^{-1}) = 0.63 = 63\%$$

We can see that most of the tunneling charge is concentrated within x_T of the surface, a distance on the order of a few tenths of a nanometer, as Figure S1.9b shows.

The preceding result is of considerable importance for metal-semiconductor diodes, in which the tunneling electrons originate from the conduction band in the metal. It is also important for the band lineup in heterojunctions (junctions between two different semiconductors).

S1.6.7 THE FINITE POTENTIAL WELL

In this section we consider what would happen if a layer of narrow band-gap semiconductor were sandwiched between two materials of wide band gap. Figure S1.10 shows an idealized one-dimensional energy band diagram for the resulting structure, called a *quantum well*. Quantum wells made this way are widely used in lasers. We will consider only the conduction band here. It appears to the electron as a finite potential well.

We consider the case in which the electron is at energy lower than the top of the well, as shown in Figure S1.10b. We intuitively expect the particle to be

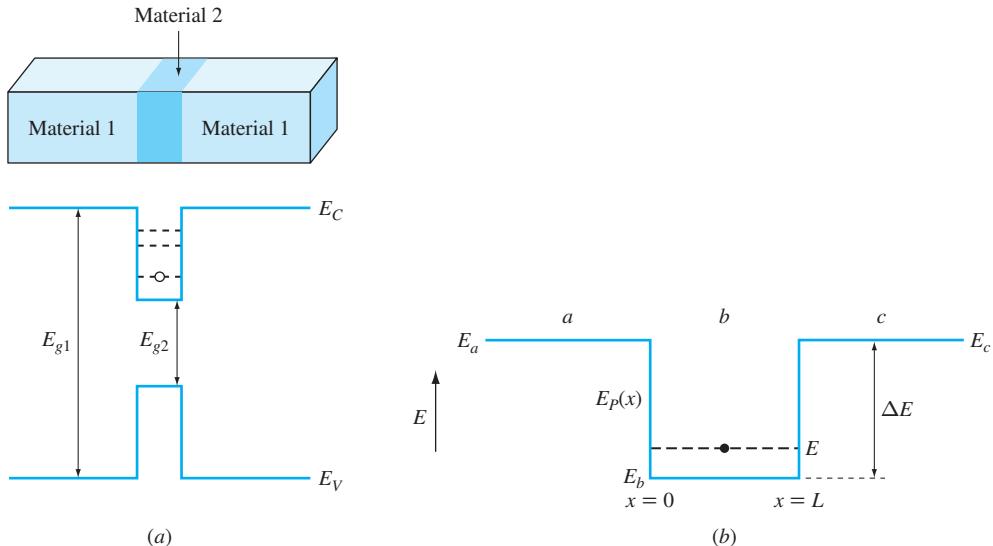


Figure S1.10 A quantum well is formed when a layer of narrow band gap material is sandwiched between two layers of wide band gap material.

trapped in the well and oscillate back and forth. We will not go through the derivation, but instead we will outline the steps and discuss the results. Again, there are three regions of space, each with a constant potential energy. We have solved this problem many times now. The wave functions are

$$\psi_{a,b,c}(x) = A_{a,b,c} e^{jK_{a,b,c}x} + B_{a,b,c} e^{-jK_{a,b,c}x} \quad (\text{S1.53})$$

where

$$|K_{a,b,c}| = \sqrt{\frac{2m^*}{\hbar^2}(E - E_{P_{a,b,c}})} = \sqrt{\frac{2m^*}{\hbar^2}E_{K_{a,b,c}}} \quad (\text{S1.54})$$

and the K 's are different for the different regions because the E_P 's are different.

Inside the well, (region b) K is real, resulting in solutions of the form $\psi(x) = C \sin(Kx)$, as in the case of the infinite potential well. The solutions outside the finite well where K is imaginary are decaying exponentials as in the case for tunneling. The first three solutions are shown in Figure S1.11, where we have plotted both the wave functions and the probability distribution functions.

Looking at the probability distribution functions, we can see that now there is a small, but finite, probability that the electron can be found outside the well. Another way to express that is to say that the electron spends some small amount of time outside the well as it oscillates back and forth between the barriers.

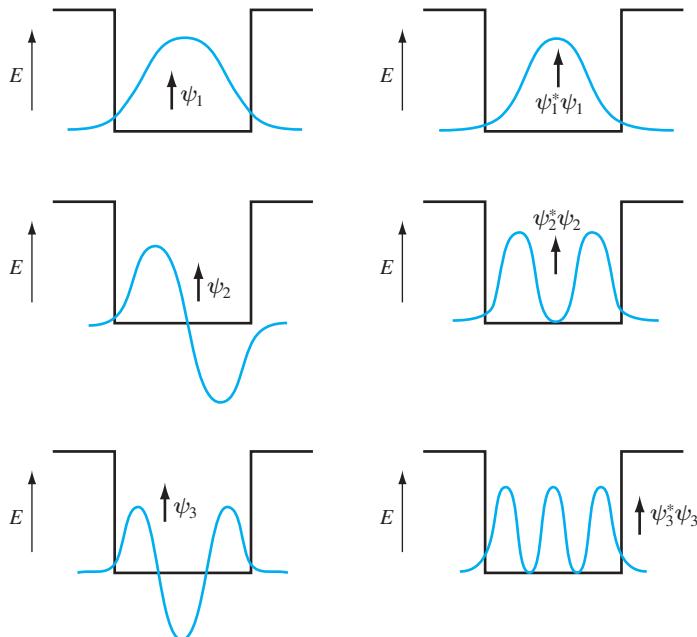


Figure S1.11 The first three wave functions in the finite potential well and the associated probability density functions $\psi^*\psi$.

The electron actually penetrates the barrier on either side for a short distance. The higher the step in the potential well, the shorter this distance is, or the more tightly confined the electron is to the well.

As was the case with the infinite well, the allowed energy states get farther apart as the well becomes narrower. For example, the well shown in Figure S1.11 happened to have three discrete states. Above the top of the well is a continuum of allowed states—the rest of the conduction band. Thus, designers of quantum well devices can control the number and spacing of allowed energy levels by adjusting the thickness of the well and by choice of materials and thus barrier heights. This is called *band-gap engineering*. One example where band-gap engineering is used is in lasers—by controlling the well width and depth, the energies of the discrete states in the potential well for electrons (in the conduction band) and the potential well for holes (in the valence band) can be adjusted to produce emission at a particular wavelength of light.

S1.6.8 THE HYDROGEN ATOM REVISITED

You will recall that in Chapter 1 we used the Bohr model to find the allowed energy levels in the hydrogen atom. Here we briefly outline the steps used to apply quantum mechanics to this problem.

The hydrogen atom is necessarily a three-dimensional object, and it is best described in spherical coordinates. We will need to follow the usual steps:

1. Describe the potential energy as a function of position (in spherical coordinates this time), assuming the electron to be a point charge.
2. Solve Schrödinger's equation (also in spherical coordinates) for the allowed wave functions. As with every problem, once we know the wave function(s), we can use Equation (S1.1) to find the average or expected value of any observable.

We begin with the potential energy of an electron in the presence of a single positive charge, the nucleus of the atom. Solving Schrödinger's equation with $E_P(r)$ given by Equation (1.6) results in quantized energies of values

$$E_n = E_{\text{vac}} - \frac{m_0 q^4}{2(4\pi\epsilon_0)^2 n^2 \hbar^2} \quad (\text{S1.55})$$

This is the same result obtained from the Bohr model.

S1.6.9 THE UNCERTAINTY PRINCIPLE

An important concept in quantum mechanics is the *uncertainty principle*, often referred to as the *Heisenberg uncertainty principle*. “Uncertainty” is a way of saying that a quantity is not exactly known. For example, an electron oscillating back and forth in a potential well is known to be in the well somewhere, but since its wave function is spread out in space we can't say it is located precisely at a given point. The uncertainty principle states that for certain pairs of observables,

the more accurately one of them is known, the less accurately the other is. Examples of the currently accepted expressions of the uncertainty principle are

$$\Delta p \Delta x \sim \frac{\hbar}{2} \quad (\text{S1.56})$$

$$\Delta E \Delta t \sim \frac{\hbar}{2} \quad (\text{S1.57})$$

where Δx is the uncertainty in position and so forth, and the symbol \sim is taken to mean “on the order of.” The observables x and p are said to be conjugate variables.

From Equation (S1.56), then, the more accurately a particle’s momentum is known, the less accurately you can determine its position. Equation (S1.57) says that the more accurately the electron energy is known, the less is known about the amount of time it spends at that energy.

Let us do an example to relate the uncertainty principle to real life.

EXAMPLE S1.6

Find the spread of the photon energy spectrum for electron transitions from an excited state E_2 to a ground state E_1 .

Solution

In a laser, light is emitted when an electron makes the transition from a state of energy E_2 to a state of lower energy E_1 , emitting a photon of angular frequency, $\omega = 2\pi\nu$. An electron in the upper state E_2 has a lifetime—the amount of time that it spends in that state on the average. If Δt is the time the electron spends in the upper state, then from Equation (S1.57),

$$\Delta E \sim \frac{\hbar}{2\Delta t} \quad (\text{S1.58})$$

The spread in the emitted photon energy spectrum can be measured with a spectrometer. The width of the photon energy spread (Figure S1.12) is a measure of Δt .

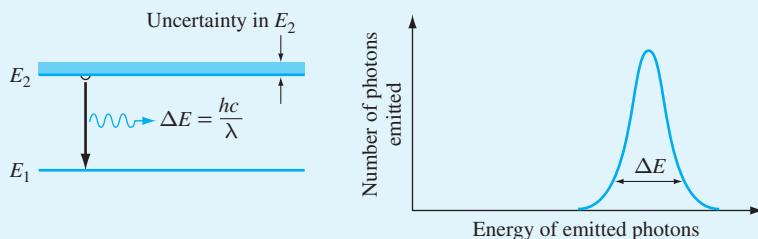


Figure S1.12 The uncertainty in the energy levels of a laser creates some uncertainty in the energy of the emitted photons.

For a lifetime of $\Delta t = 1$ ns,

$$\Delta E = \frac{\hbar}{2\Delta t} = \frac{6.58 \times 10^{-16} \text{ eV} \cdot \text{s}}{2 \times 10^{-9} \text{ s}} = 3.29 \times 10^{-7} \text{ eV}$$

This corresponds to a spectral width of 80 MHz.

S1.7 PHONONS

In Chapter 3, the term *phonon* was used to describe lattice vibrations in crystals. Phonons influence the mean free time between electron or hole collisions and thus the mobilities of the carriers. In Chapter 3, we mentioned that at high electric fields, the drift velocity is limited by carrier interactions with *optical phonons*. At low fields, the *acoustical phonons* influence the mobility. In this section, we discuss these phonons in slightly more detail. We will see that phonons can be treated as particles or waves, and since the phonons travel in a periodic structure, they develop energy bands.

Figure S1.13 shows the longitudinal and transverse pressure waves in a crystal in which the wave is traveling in the x direction. The dashed lines represent planes of atoms at their equilibrium positions, while the solid lines represent the planes displaced by the pressure wave. A longitudinal wave is one in which the displacement is in the direction of motion as indicated in (a). In a transverse wave (b) the displacement is perpendicular to the direction of motion.

We will develop the idea of energy bands for phonons by analogy with electrons. In Chapter 1, electrons traveling in a periodic structure were treated as waves (Bloch waves), with wavelength $\lambda = 2\pi/|K|$, and also as particles of energy E and crystal momentum $\hbar K$. These electrons exist in energy bands, each band

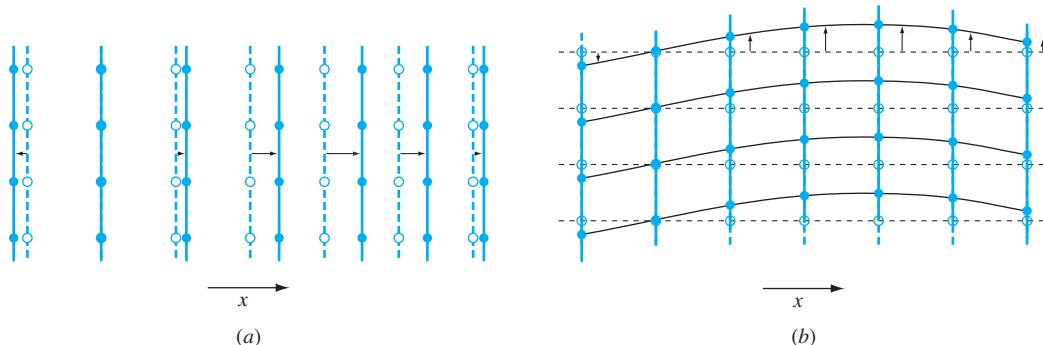


Figure S1.13 Displacement of atomic planes under the influence of a pressure wave. For a longitudinal wave (a), the displacement is in the direction of motion. For a transverse wave (b), the displacement is transverse to the direction of motion. For a three-dimensional crystal, for each longitudinal wave there are two transverse waves. The dashed lines represent the equilibrium positions, and the solid lines indicate the deflected positions at a given time.

with its own E - K relation. All the important information for electrons (effective mass, allowed energies, potential energy, etc.) is contained in the energy band diagrams (E - K relations) of the first Brillouin zone. The first Brillouin zone is in the region $-\pi/a \leq K \leq \pi/a$, where a is the periodicity of the lattice.

Phonons and photons are also waves that can move through the lattice. We are interested in phonons in this section. Phonon waves have some angular frequency ω and some wave vector $|K| = 2\pi/\lambda$.

The energy associated with lattice vibrations of angular frequency ω is given by

$$E = \left(n + \frac{1}{2}\right)\hbar\omega \quad (\text{S1.59})$$

where n is an integer. Defining $\hbar\omega$ as the energy of one phonon, E_{phonon} ,

$$E_{\text{phonon}} = \hbar\omega \quad (\text{S1.60})$$

The integer n in Equation (S1.59) indicates the number of phonons at frequency ω . A phonon thus has the characteristics of a wave and a particle, like electrons and photons. The frequencies of the lattice vibrations are on the order of 10^{13} Hz, while the phonon energies are a fraction of an electron volt. Since we consider the interaction of phonons with electrons and photons whose energies are expressed in electron volts, it is convenient to consider phonon energies rather than their frequencies.

A detailed treatment of these lattice vibrations [1] is beyond the scope of this text, but to summarize the major results:

1. There exist, for phonons, Brillouin zones in K space, periodic with periodicity $2\pi/a$ where a is the periodicity of equivalent planes.
2. All information exists in the reduced zone (first Brillouin zone) extending from $-\pi/a \leq K \leq +\pi/a$.
3. This zone is symmetric around $K = 0$.
4. There are as many allowed energy bands as there are atoms per primitive unit cell. Most semiconductors of interest (e.g., Si, Ge, GaAs) have two atoms per primitive unit cell.² Thus in these materials there are two energy bands for phonons.
5. In each band, there are three branches, one due to the longitudinal wave and two transverse branches corresponding to the two transverse directions.
6. In a cubic crystal, from symmetry the two transverse branches are the same.

Expressions for the longitudinal branches can be easily calculated [1]. We will not do the calculation here, but the procedure is to treat the primitive unit cell as having two atoms of masses M_1 and M_2 (in silicon these would be the same

²The “primitive unit cell” of a crystal is the smallest arrangement of atoms that when repeated and translated will produce the entire lattice. There are four primitive unit cells in one conventional unit cell in a diamond or zinc blende lattice, where the conventional unit cell is a cube.

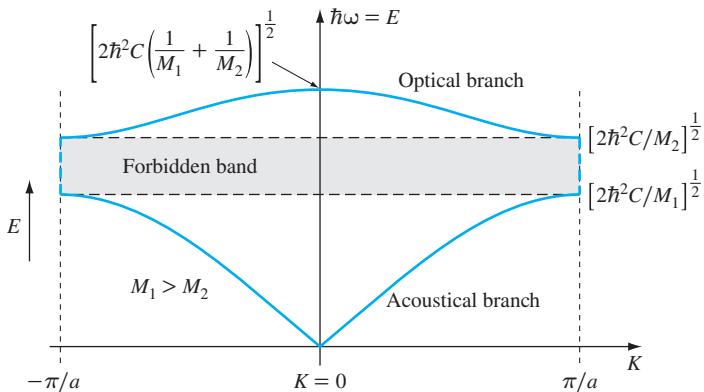


Figure S1.14 Phonon E - K diagram for longitudinal waves in a crystal having two different atoms of masses M_1 and M_2 . The maximum frequency is on the order of 10^{13} Hz. The force constant between adjacent planes is C .

since all atoms are silicon atoms) connected by a “spring” of spring constant C representing the attractive and repulsive forces between successive planes. The resulting E - K diagram is plotted in Figure S1.14 for the case of unequal masses (e.g., GaAs).

The “acoustical phonons” correspond to the lower phonon energy band, and the “optical phonons” occupy the upper energy band. The reasons for these names will emerge shortly.

We make the following observations from the figure:

1. The group velocity of a phonon at a given frequency is proportional to the slope of the E - K plot, since $v = (1/\hbar)dE/dK$.
2. In the lower branch, for small K (large λ) the velocity is constant and equal to the velocity of sound, hence the term *acoustical branch*.
3. In the upper branch at $K = 0$, the slope is zero and thus the velocity $v = 0$. Physically, this corresponds to the two adjacent planes in the crystal vibrating out of phase. Since the center of mass is constant in time, the wave does not propagate in the crystal.
4. The top of the upper branch at $K = 0$ corresponds to a frequency on the order of 10^{13} Hz. This is comparable to the frequencies of photons in the infrared. We will see later that this allows phonons to interact with photons. For now it suffices to say that this is why this branch is referred to as the *optical branch*.
5. The forbidden energy band increases with the difference in mass of the two atoms. For elemental semiconductors (e.g., Si, Ge, diamond) $M_1 = M_2$ and the forbidden band disappears.
6. The energy at $K = 0$ in the optical band is dependent on the sum of the reciprocal masses. The smaller the masses, the larger the energy.

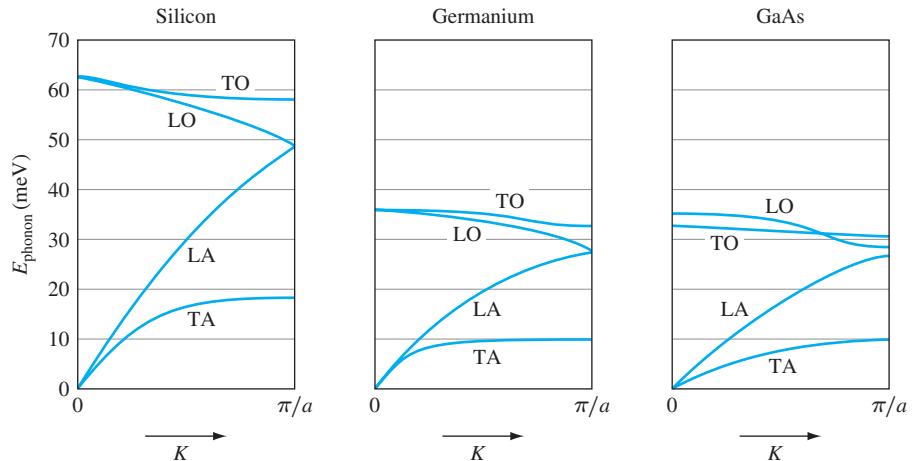


Figure S1.15 Energy-wave vector diagrams for Si, Ge, and GaAs in the $<100>$ directions. Shown are transverse acoustical (TA), longitudinal acoustical (LA), longitudinal optical (LO), and transverse optical (TO) branches.

Figure S1.15 shows the phonon E - K curves for Si, Ge, and GaAs. The transverse as well as the longitudinal branches are shown in the (100) direction for $K \geq 0$. For Si and Ge, where $M_1 = M_2$, the longitudinal branches have the same value at the zone edge. The masses of Ga and As are only slightly different, and so in GaAs a small forbidden energy band exists.

We indicated that the upper band is called the optical band since it can interact directly with optical photons. Consider the situation in which a photon produces a phonon. To transform the optical energy into acoustic energy, both energy and wave vector must be conserved. On the E_{phonon} - K diagram of Figure S1.16, the E - K diagram of the photon is superimposed on the phonon diagram. The two intersect at a point in the optical branch of the phonon curve. Thus, only phonons in the optical branch can interact directly with photons. Since for a wave, $v = (1/\hbar) dE/dK$, and the velocity of light is on the order of four orders of magnitude greater than the velocity of sound, the photon E - K characteristic is almost vertical on the phonon E - K characteristic.

Let us compare the energies and K vectors for the photons and phonons. Since $v = (1/\hbar) dE/dK$, for a photon of small K (large λ), the velocity of light and thus dE/dK are constant. Thus the photon wave vector is

$$K = \frac{E_{\text{photon}}}{\hbar c}$$

where c is the velocity of light in the material. In silicon, at $K = 0$, $E_{\text{phonon}} = 0.063$ eV. Assuming $c = 3 \times 10^8$ m/s, a photon of light of the same energy, 0.063 eV, has a wave vector $K_{\text{photon}} \approx 2 \times 10^5$ m $^{-1}$. Compare this with the K of a phonon at the edge of the zone: $K_{\text{phonon}} = \pi/a \approx 5.6 \times 10^9$ m $^{-1}$. Thus only phonons near $K = 0$ are able to interact directly with light.

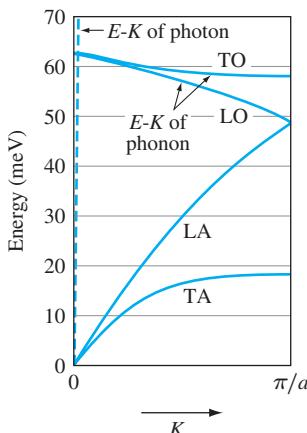


Figure S1.16 Interaction of a photon and a phonon requires conservation of both E and K . Thus a photon can create a phonon only in the optical band of the phonon's E - K curve.

What about phonon interaction with electrons? Can a phonon excite an electron from the valence band to the conduction band, for example? The range of wave vector for phonons is the same as that for electrons, so from a conservation of K point of view such an event is possible. The phonon energies, however, are much smaller than that needed to excite an electron across the forbidden gap.³ The phonons can, however, interact with electrons via scattering, as illustrated next.

S1.7.1 CARRIER SCATTERING BY PHONONS

Recall from Section 3.3 that electrons can be scattered by lattice vibrations (phonon scattering). There the scattering was described in terms of particle-particle collisions. This scattering can also be described by electron wave–phonon wave interaction.

We know that, at any temperature, the vibrating atoms create pressure waves in the crystal. The pressure waves cause periodic compression and dilation of the atomic spacing, as shown in Figure S1.17a. Since the electron band gap is pressure sensitive (it is sensitive to atomic spacing), the regions of compression cause E_g to increase, while regions of dilation tend to decrease E_g . This results in the electron energy band structure shown in Figure S1.17b. Electrons in the conduction band are partially reflected (scattered) off these changes in potential

³However, a phonon and a photon can simultaneously collide with a valence band electron and excite it to the conduction band.

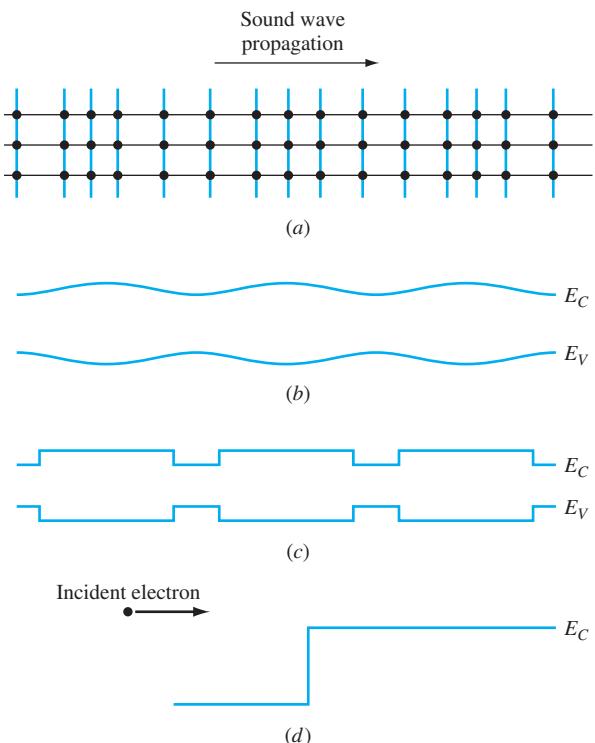


Figure S1.17 Phonon scattering. Phonons are the particle description of acoustic or compressional waves in the crystal. (a) The phonon causes compression and dilation of the crystal. (b) The varying lattice spacing causes periodic electron band-gap narrowing and widening. (c) These band-gap variations can be modeled as abrupt steps, in which case they resemble the reflection-from-a-potential-barrier problem as shown in (d).

energy E_C . This is easier to see for the extreme cases in which E_C is assumed to change abruptly as shown in the idealized model of (c). As the electron travels, it encounters a series of potential barriers, each of which has some probability of reflecting it as indicated in Section S1.2.6. Similarly, holes are reflected (scattered) by the changes in E_V .

The temperature also has an effect. With increasing temperature, the stronger lattice vibrations cause an increased variation in electron band gap and thus increased scattering or reduced mobility.

In addition, high-energy electrons could be scattered by the creation of optical phonons. This last effect limits the electron's kinetic energy and thus its velocity (velocity saturation).

These two interactions (electron with optical phonons and electrons with acoustical phonons) are illustrated in Figure S1.18, in which the kinetic energy–wave

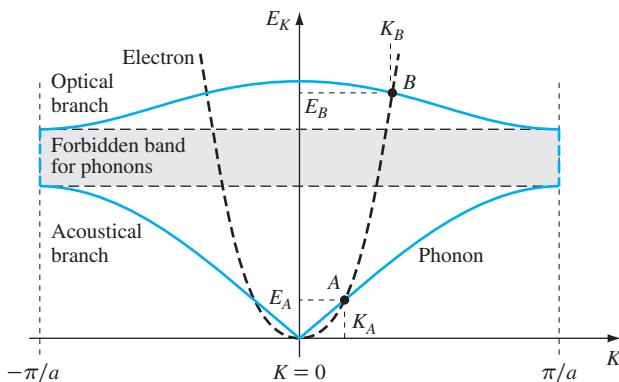


Figure S1.18 Kinetic energy–wave vector relations for phonons and conduction band electrons in a direct-gap semiconductor. An electron can create an optical phonon of energy E_B and wave vector K_B , or an acoustical phonon of energy E_A and wave vector K_A .

vector (E_K - K) diagrams of electrons and phonons are superimposed. An electron at position A has a high probability of creating an acoustical phonon of energy E_A and wave vector K_A . To conserve energy the electron must lose this energy, and to remain on its allowed E - K diagram it must simultaneously lose wave vector.

In the preceding process, an acoustical phonon was created, to carry off the energy and wave vector given off by the electron. In the reverse process, an acoustical phonon can be annihilated, giving its energy and wave vector to an electron, thus increasing its energy and wave vector. In either case the electron is scattered—its energy (and direction) is changed. At equilibrium, the creation and annihilation rates of phonons are equal, and they result in mean free times between scattering events on the order of 10^{-13} seconds.

Assume an electron can gain enough kinetic energy to reach the point B in Figure S1.18. This could happen if the electron is accelerated strongly by a high electric field. An electron at B can lose its energy and wave vector, this time creating an optical phonon of energy E_B and wave vector K_B . Since this interaction is highly probable, it effectively puts a limit on the kinetic energy an electron can achieve, and this is what causes the electron velocity to saturate.

S1.7.2 INDIRECT ELECTRON TRANSITIONS

Next, we examine the role of phonons in semiconductors with indirect energy gaps. In Chapter 3, we indicated that optical absorption in an indirect gap semiconductor requires a three-particle process involving an electron, a photon, and a phonon. This is illustrated in Figure S1.19 for the case of Si, where again the E - K diagrams of electrons, photons, and phonons are superimposed. We saw earlier that a single phonon does not have enough energy to excite an electron from the valence band to the conduction band. Similarly, for energies of interest

in semiconductor electronics, photons have significant energy $h\nu$ but negligible wave vector. An electron can, however, be excited from the valence band to the conduction band by simultaneously absorbing a photon and a phonon. The phonon furnishes the necessary wave vector K_C and energy $\hbar\omega_{\text{phonon}}$, while the energy furnished by the photon is $h\nu$. The photon energy required to excite an electron across the energy gap is at least:

$$h\nu = E_g - \hbar\omega_{\text{phonon}}$$

to conserve energy. To conserve wave vector, we require

$$K_{\text{phonon}} \approx K_C$$

where K_C is the electron wave vector at the bottom of the conduction band.

It is also possible for the band-to-band transition to result from the absorption of a photon accompanied by the *emission* of a phonon (as opposed to absorption of a phonon). The minimum photon energy required in this case is

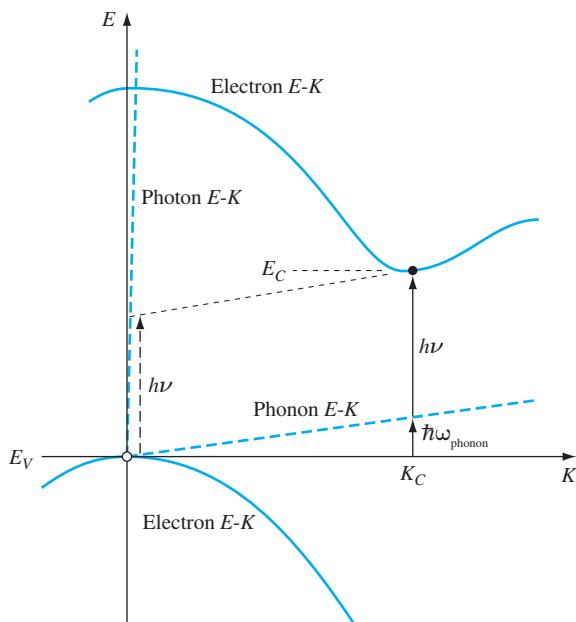


Figure S1.19 Illustration of optical absorption in an indirect semiconductor involving an electron, a phonon, and a photon. The wave vector needed to make the transitions comes almost entirely from the phonon; the phonon contributes a small amount of energy, $\hbar\omega_{\text{phonon}}$, as well, but most of the energy is supplied by the photon.

$$h\nu = E_g + \hbar\omega_{\text{phonon}}$$

and

$$K_{\text{phonon}} \approx -K_C$$

EXAMPLE S1.7

Considerable information about the band gap of a semiconductor can be obtained from the measured photocurrent spectrum. Consider the circuit of Figure S1.20, in which a semiconductor is illuminated at room temperature with monochromatic light of varying photon energy, $h\nu$, and the current is measured as a function of photon energy. The result of such a (hypothetical) semiconductor is indicated in Figure S1.21. Since for most semiconductors the valence band maxima are at $K = 0$, we assume that is the case for the semiconductor under test.

There are four distinct regions.

Region 1. For $h\nu < 0.88$ eV, the current is independent of photon flux. The photon energy is insufficient to excite electrons from the valence band to the conduction band. This current is referred to as “dark current” or the current in absence of illumination.

Region 2. For $0.88 < h\nu < 0.96$ eV the current increases slowly with increasing $h\nu$. This slow increase indicates that the semiconductor has an indirect band gap. Electrons are excited into the conduction band by the simultaneous absorption of photons and optical phonons. The phonons have energies, $\hbar\omega_0$, and wave vectors at the conduction band minimum.

Region 3. For $0.96 < h\nu < 1.08$ eV, the photocurrent rises more rapidly with photon energy than in region 2. In this region in addition to electrons being excited from valence to conduction band by the simultaneous absorption of photons and phonons, they can be excited by absorption of photons and emission of phonons.

Region 4. The sharp increase in current at 1.08 eV indicates band-to-band transitions by photons without the aid of phonons. This indicates that minima exist in the conduction band at $K = 0$ and at 1.08 eV above the valence band maximum.

The minimum photon energy for band-to-band transitions by simultaneous absorption of photons and phonons occurs for $h\nu = 0.88$ eV.

$$0.88 \text{ eV} + \hbar\omega_{\text{phonon}} = E_g \quad (\text{S1.61})$$

The minimum photon energy for simultaneous absorption of photons and emission of phonons is

$$0.96 \text{ eV} - \hbar\omega_{\text{phonon}} = E_g \quad (\text{S1.62})$$

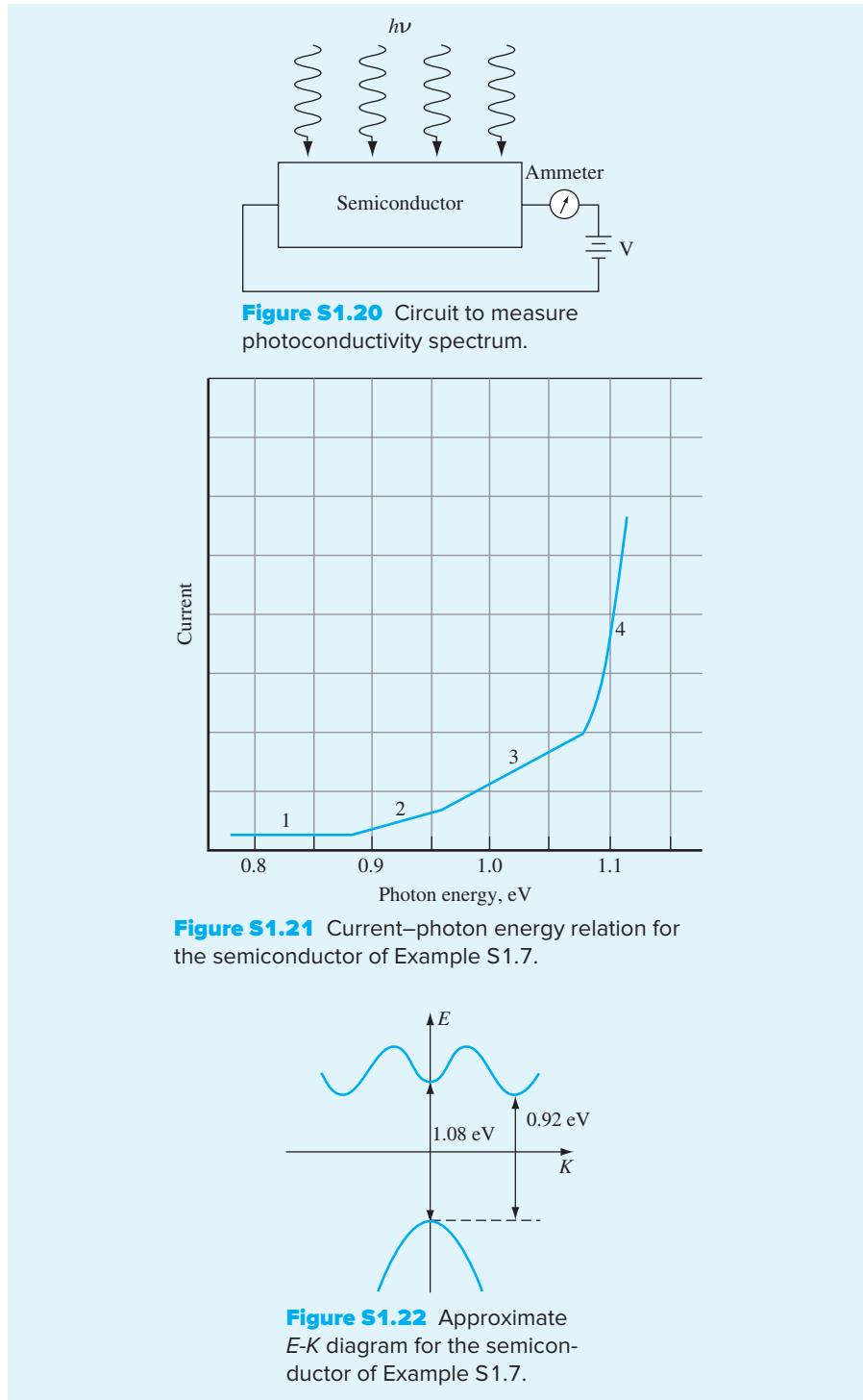
Adding Equation (S1.61) and (S1.62), and solving for E_g ,

$$E_g = 0.92 \text{ eV}$$

Subtracting Equation (S1.61) from Equation (S1.62) and solving for $\hbar\omega_{\text{phonon}}$,

$$\hbar\omega_{\text{phonon}} = 0.04 \text{ eV} = 40 \text{ meV}$$

The resulting energy band diagram for this semiconductor would look something like Figure S1.22.



S1.8 SUMMARY

We have expanded our understanding of quantum mechanics as described by Schrödinger's equation. The solutions to Schrödinger's equation are called wave functions, and we can use them to tell us everything about the observable quantities, or quantities that can, in principle, be measured. The value of an observable is found by using the operator corresponding to the observable, and the wave function representing the particle whose observable needs to be known.

We saw that once we know the wave function for a particle in a given physical situation, then we can calculate all of the observables for that particle. We also found that for most realistic problems, the potential energy cannot be accurately described. We therefore resorted to simplified models, which we used to get some physical intuition about certain kinds of structures. For example, we saw that for an electron in a potential well, the energy levels are quantized, consistent with the electron in the hydrogen atom. The hydrogen nucleus is a potential well for electrons.

We found that in quantum mechanics, particles can tunnel through potential barriers that classical physics predicted would reflect them every time.

We also began to explore the uncertainty principle. This nonintuitive concept predicts (for example) that the more accurately one knows a particle's position, the less accurately one will know its momentum (or its wave vector). The uncertainty principle prevents us from making precise statements. In quantum mechanics, we tend to express everything as a probability—the electron has a probability of tunneling through this barrier, or this electron is, on the average, at this position. We can never say where the electron is, but we can discuss where it's likely to be.

In this section, we began applying the principles of quantum mechanics to some simple but common device structures: potential barriers and potential wells.

Although Equations S1.56 and S1.57 are sometimes referred to as the Heisenberg uncertainty principle, they are Fourier transforms of the p - x and E - t relations. W. Heisenberg obtained the result $\Delta p \Delta x \geq h$ (for free electrons) without a precise definition of uncertainty [2]. E. H. Kennard obtained the expression S1.56 with Δp and Δx as standard deviations of p and x [3].

We spent some time discussing the role of phonons in semiconductors. Phonons can scatter electrons, thus influencing the mean free time between collisions (and therefore mobility), and they can also assist with electron transitions.

S1.9 REFERENCES

1. See for example, C. Kittel, *Introduction to Solid State Physics*, 8th ed., Chapter 4, John Wiley and Sons, New York, 2004.
2. W. Heisenberg, “Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik,” *Zeitschrift für Physik* 43 (3–4): 172–198, 1927.
3. E. H., Kennard, “Zur Quantenmechanik einfacher Bewegungstypen,” *Zeitschrift für Physik* 44 (4–5): 1927.

S1.10 REVIEW QUESTIONS

1. Explain why the wave function is important if it can't be measured.

2. Consider the electron reflecting off a finite potential barrier of finite width L , as shown in Figure S1.6, but suppose this time that electron has a total energy greater than the barrier height. In Section S1.6.5, we saw that when the electron approaches the a/b boundary, it has some probability of being reflected. Suppose it is not reflected, however, but travels across the barrier. What happens when it reaches the b/c boundary of the barrier? What happens after that?
3. Explain the difference between a transverse acoustic wave and a longitudinal acoustic wave.
4. Can an electromagnetic wave (for example, a photon) be longitudinal?
5. Explain how an electron scatters from an acoustic wave.
6. With the help of Figure S1.19, explain how an electron in an indirect material can make a band-to-band transition. What is the probability of this transition compared with that for a direct gap material?

S1.11 PROBLEMS

- S1.1** Assume the wave function Ψ is separable, as shown in Equation (S1.9). Insert that into Schrödinger's equation and show that Equations (S1.10) and (S1.11) result. If the procedure is not obvious, review separation of variables from your differential equations course.
- S1.2** Solve Equation (S1.11) to show that Equation (S1.12) is a solution.
- S1.3** Consider the $n = 3$ state of the infinite potential well, where $\Psi_3 = A \sin(2\pi x/L)e^{-j(E_3/\hbar)t}$ for $0 \leq x \leq L$.
- a. Sketch the time-independent part of the wave function $\psi_3(x)$.
 - b. Find the average value of the position x .
 - c. Sketch $\psi^* \psi$. What is the most probable location for the electron? Explain how you drew your conclusion.
- S1.4** a. Sketch the time-independent wavefunction ψ_2 for the infinite potential well. Where do you expect the average value of the position to be?
- b. Sketch the probability density function for the same energy level. Identify the most probable position(s) for the electron to be found.
- c. What is the probability of finding the electron at the average position at any given moment?
- S1.5** For the electron in the lowest energy state of the infinite potential well, find
- a. The average momentum p_x
 - b. The average velocity v_x
 - c. The average energy E
- S1.6** Consider a plane wave $\Psi(x, t) = e^{j[Kx - (E/\hbar)t]}$. What is the average velocity v_x ?
- S1.7** Consider a free electron with kinetic energy 2.0 eV.
- a. What is its classical velocity?
 - b. What is its classical momentum?

- c. If the total energy of the electron is 3.0 eV, write the plane wave expression for this particle, e.g. put it in the form $\psi(x, t) = Ae^{j(kx)}$.
- d. Using your expression from (c), show that the group velocity obtained using quantum mechanics (i.e. using the observable operator for group velocity, $(v_{op} = \frac{\hbar}{jm} \frac{\partial}{\partial x})$) is the same as the classical result (part a).
- S1.8** Consider the (artificial) wave function shown in Figure PS1.1. The wave function ψ is zero for $x < 0, x > L$.
- Sketch the probability distribution function.
 - Indicate the region(s) in which the electron is most likely to be found.
 - To find the probability of an electron being in a certain region, one integrates $\Psi^*\Psi$ over the region of interest and divides by the integral over all space [see Equation (S1.1)]. Calculate the probability of finding the electron in the region $L/4 < x < L/2$.

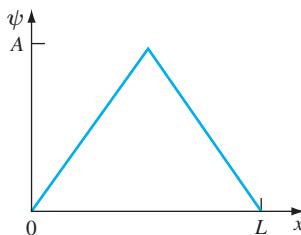


Figure PS1.1

- S1.9** Find the eigenfunctions for the infinite potential well, but let the well begin at $x = -L/2$ and end at $x = L/2$. Since the electron can't read, it doesn't know you have changed coordinates and should end up with exactly the same states as before. Sketch the first three eigenfunctions and compare with Figure S1.3. *Hint:* Make sure that you have a complete set of solutions.
- S1.10** Consider an electron approaching a potential barrier with an energy lower than the barrier, as shown in Figure S1.6. Explain why neglecting reflections at the two sides of the barrier in Figure S1.6 is a reasonable approximation. Take the reflection coefficient at either side of the barrier to be on the order of 20 to 30%, and take the barrier to be thick. *Hint:* Compare the transmission probabilities, neglecting reflection, with the reflection probabilities.
- S1.11** Sketch the electrical characteristics of a tunneling junction in which the superconductor Pb ($E_g = 2.73$ meV) is used instead of Sn (i.e., the structure is Pb/SnO₂/Pb). Label the significant points on your horizontal axis with numeric values.
- S1.12** Consider the potential energy distribution function shown in Figure PS1.2. It consists of two finite wells separated by a finite barrier. Sketch what you expect the wave function to look like. Can the electron move from one well to the other? If so, by what process? Can the electron escape from the wells?

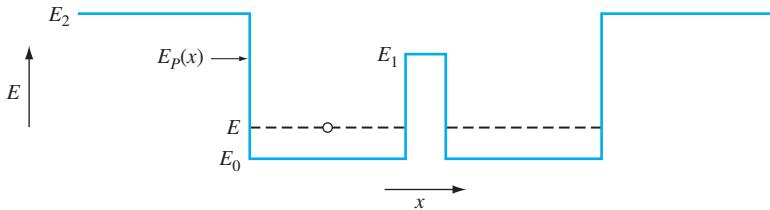


Figure PS1.2

- S1.13** An electron is in an asymmetric potential well as shown in Figure PS1.3. The wave function is indicated for a particular energy state.

- Sketch the probability density.
- On the average, where is the electron most likely to be found?
- What is the approximate uncertainty in the electron's position? Explain your reasoning.
- Can the electron tunnel into the barrier on the left side? Right side?
- Can the electron in energy level E_1 escape the well?
- Sketch a realistic wavefunction for the second energy in this well (if there is one).

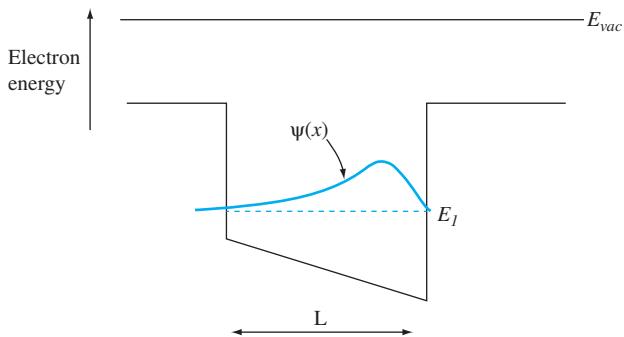


Figure PS1.3

- S1.14.** Consider the potential energy diagram (in real space) shown in Figure PS1.4.

- Indicate the region(s) in which an electric field is present.
- Can an electron in energy level E_1 escape the well?

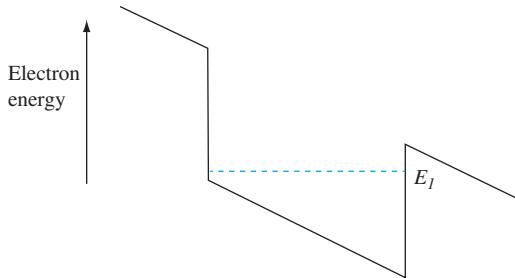


Figure PS1.4

- S1.15** (Lengthy.) In this problem you will solve the problem of the finite potential well, using results presented in this supplement. The geometry of the problem is shown in Figure PS1.5. The energy of the bottom of

the well is E_w , and the energy of the barriers is E_a . The difference is $E_a - E_w = \Delta E$. The well extends from $x = 0$ to $x = L$. We will call the three regions a, b, and c as shown in the figure.

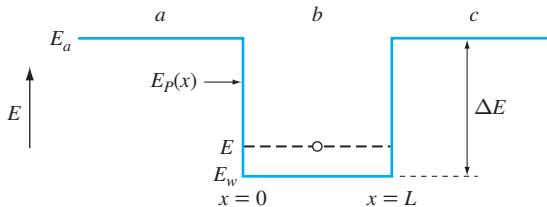


Figure PS1.5

- Write the time-independent Schrödinger's equation for this problem (it will require three equations, one for each region).
- Solve Schrödinger's equation for each of the three regions. The results should be of the form:

$$\begin{aligned}\psi_a(x) &= A e^{ax} \\ \psi_b(x) &= B \cos(Kx) + C \sin(Kx) \\ \psi_c(x) &= D e^{-a(x-L)}\end{aligned}$$

Express the constants a and K in terms of the energies in the regions.

- Use the boundary conditions to express the constants B , C , and D in terms of A . You will need to apply (1) continuity of ψ at both boundaries and (2) continuity of $d\psi/dx$ at $x = 0$. The results are: $A = B$, $C = A(a/K)$, and

$$D = A \left[\cos(KL) + \frac{a}{K} \sin(KL) \right]$$

- We need to find the value(s) of K . Use the final boundary condition, continuity of $d\psi/dx$ at $x = L$, to show that

$$\tan(KL) = \frac{2a}{K \left(1 - \frac{a^2}{K^2} \right)}$$

- The result of (d) is known as the characteristic equation, and it cannot be solved in closed form. You can solve it graphically, however, by plotting the left-hand side versus K and the right-hand side versus K and seeing where they cross.
 - To do this, first find the maximum value that K can have for an electron trapped in the well. (*Hint:* Its kinetic energy must be less than $E_a - E_w$)
 - Given that $a = \sqrt{(2m/\hbar^2)(E_a - E)}$ and $K = \sqrt{(2m/\hbar^2)(E - E_w)}$, show that $a = \sqrt{K_{\max}^2 - K^2}$.
 - Write a program that plots the two sides of the characteristic equation from $K = 0$ to $K = K_{\max}$. As an example, use $L = 2$ nm and $\Delta E = 0.5$ eV.
 - How many states are in this well, and what are their energies?
- Plot the wave function for the lowest energy state, from $x = -0.2L$ to $1.2L$. The wave function should be symmetrical.

- g. Use your program to find the number of allowed states for

$$L = 3 \text{ nm}, \Delta E = 0.5 \text{ eV}$$

$$L = 5 \text{ nm}, \Delta E = 0.5 \text{ eV}$$

$$L = 2 \text{ nm}, \Delta E = 1 \text{ eV}$$

$$L = 3 \text{ nm}, \Delta E = 0.2 \text{ eV}$$

- S1.16** Figure PS1.6 shows the conduction band edge for another tunneling device, the resonant tunnel diode. It consists of two tunneling barriers with a quantum well between them. Assume the quantum well has only one state. Explain what is going on in the four regions of the I - V curve. The region where the slope is negative is called the negative resistance region.

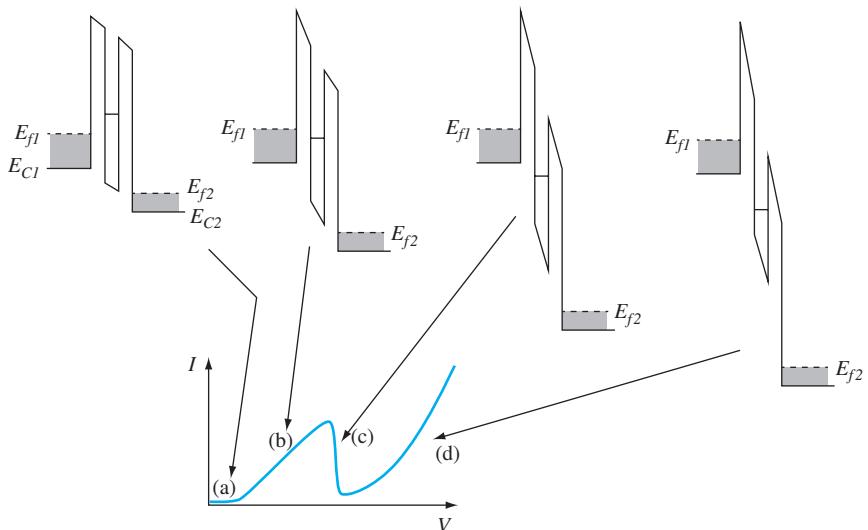


Figure PS1.6

- S1.17** Consider the phonon E - K diagram of Figure S1.14 for a crystal with two types of atoms, of mass M_1 and M_2 in a zinc blende lattice.

- Since the lattice constants are nearly equal, we can assume the force between the atoms to be comparable. For a material consisting of a single atom type (e.g., silicon), the lattice becomes a diamond lattice and $M_1 = M_2$. Explain why the optical phonon energy for diamond (carbon) is greater than that for silicon, which in turn is greater than that for germanium.
- Explain why the saturation velocity of germanium is less than that for silicon.
- Explain why the saturation velocity of electrons in GaAs is less than its peak value.

- S1.18** Acoustic waves travel faster in materials that are stiffer (harder). The group velocity of phonons is proportional to the slope of the E - K diagram for acoustic phonons (as in Figure S1.14). At the edge of the Brillouin zone, this slope is zero, corresponding to standing waves. Traveling waves would be found nearer to the origin on the E - K diagram. Use the phonon E - K curve to predict which should be harder, diamond (carbon) or lead, assuming they have the same force constant C , and comparable lattice constants.

Diodes

We now have the necessary background in the electronics of semiconductor materials to progress to devices. Most electronic devices depend on the electrical characteristics of junctions between different materials. Usually (except for the case of superconductor junctions), at least one of these is a semiconductor. Such junctions are two-terminal devices and are referred to as *diodes*.

In Part 2 of this book, we investigate the electrical characteristics of a variety of junctions. These include:

1. *Homojunctions.* Junctions between two differently doped regions of the same semiconductor material.
2. *Heterojunctions.* Junctions between two different materials. Although technically this includes combinations such as a semiconductor-insulator junction, in common usage *heterojunctions* refers to junctions between two different semiconductors.
3. *Metal-semiconductor junctions.* Junctions between a metal and a semiconductor.

An understanding of semiconductor junctions and their electrical characteristics is essential to the understanding of most semiconductor devices—transistors, lasers, etc.—since these devices are composed of such junctions.

Chapter 5 deals with diodes. A semiconductor diode is a junction. It is also a two-terminal device that acts as a switch: for one polarity of applied voltage it acts as a near open circuit and for the other polarity it acts as a near short circuit. This is illustrated by the current-voltage characteristics of a Si homojunction diode in Figure II.1. For an applied voltage less than about 0.7 V, the current is small enough that the diode can be considered an open circuit. For applied voltages greater than about 0.7 V, the current increases rapidly with voltage, and the diode can be approximated by a voltage source of 0.7 V. An approximate equivalent circuit is indicated in Figure II.2. A generalized symbol for a diode is indicated in Figure II.3. Current flows in the direction indicated for any

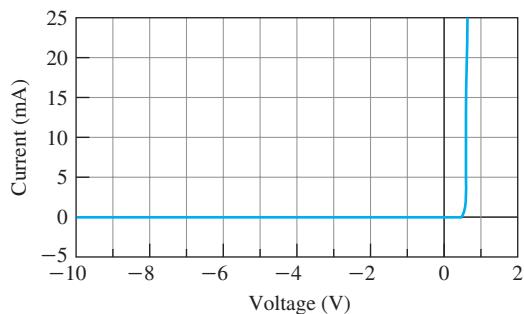


Figure II.1 The current-voltage characteristic of a silicon homojunction diode. For $V < 0.7$ V, the current can be considered negligible. For current flow, the diode voltage is nearly constant and equal to 0.7 V.

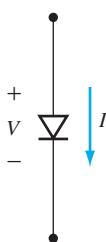


Figure II.3 Symbol commonly used to represent a diode. For forward bias ($V > 0$) current flows in the direction indicated. For reverse bias ($V < 0$) the current is essentially zero.

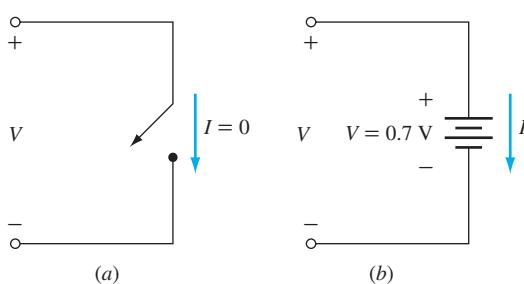


Figure II.2 (a) Approximate equivalent circuit for a silicon homojunction diode. For $V < 0.7$ V, the diode acts as an open circuit. (b) For an applied voltage greater than this, the diode approximates a constant voltage $V = 0.7$ V.

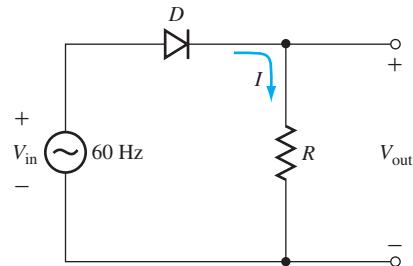


Figure II.4 Circuit diagram for a rectifier. The output voltage is $V_{\text{out}} = IR$. The diode conducts, and thus current flows only for forward bias of $V_{\text{in}} > 0.7$ V.

voltage greater than zero, but until the voltage reaches about 0.7 V, the current is negligibly small.

A common application of a diode is in a rectifier circuit as indicated in Figure II.4. Here an input 60 Hz sine wave is applied to a series combination of a diode D and a resistor R . The output voltage ($V_{\text{out}} = IR$) is taken across the resistor.

EXAMPLE

Consider the rectifier circuit of Figure II.4 in which $V_{\text{in}} = 5 \sin(2\pi ft)$, where $f = 60$ Hz, $R = 2 \text{ k}\Omega$, and the diode is represented by the equivalent circuit of Figure II.2. Plot the input and output waveforms and the current waveform.

Solution

The input waveform is given as $V_{\text{in}} = 5 \sin(2\pi ft) = 5 \sin(377t)$, as indicated in Figure II.5. For $V_{\text{in}} < 0.7$ V the current is zero and thus $V_{\text{out}} = 0$. For $V_{\text{in}} > 0.7$ V the voltage across

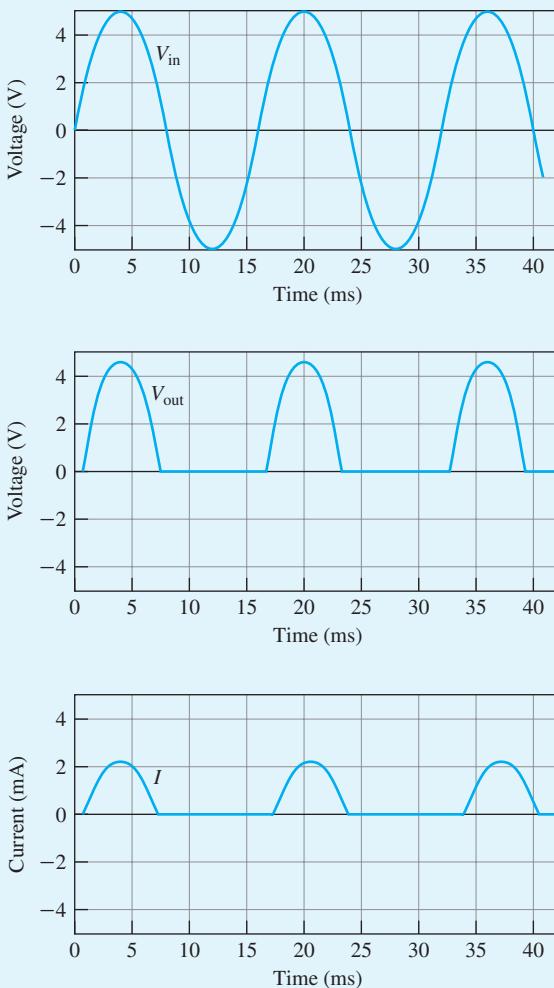


Figure II.5 Input voltage, output voltage, and current for the rectifier circuit of Figure II.4, with $V_{in} = 5 \sin(\omega t)$, $R = 2 \text{ k}\Omega$, and frequency of 60 Hz ($\omega = 2\pi f = 377 \text{ rad/s}$).

the diode is 0.7 V, and thus $V_{out} = V_{in} - 0.7 \text{ V}$. The current then is

$$I = \frac{V_{out}}{R} = \frac{5 \sin 377t - 0.7 \text{ V}}{2 \text{ k}\Omega}$$

This particular circuit is referred to as a half-wave rectifier. It converts an ac voltage, which has zero average value, to a pulsating voltage whose average is now nonzero. With additional circuitry this pulsating voltage can be smoothed out to produce a constant dc voltage.

The diode equivalent circuit of Figure II.2 is reasonably accurate for the rectifier circuit of the illustration. However, most electronic circuits require more accurate equivalent circuits as developed in the next chapter.

The operation of semiconductor junctions is explained in terms of energy band models, from which we can predict the electrical characteristics of the devices. To draw the energy band diagrams for junctions, the following procedure is used:

1. We use the technique of Chapter 4 for drawing the energy band diagram of each material. That is, we begin with the energy band model in which each material is electrically neutral in every macroscopic region. We normally start out by considering the regions to be electrically isolated from each other.
2. Next, the regions are considered to be in intimate contact. Charge will transfer, such that the Fermi levels in the regions are equalized. The energy band diagram at equilibrium is then constructed, taking into account the locations of the transferred charge, the band gaps, and the electron affinities and work functions and any interface states or electric dipoles in the vicinity of the junction.
3. The influence of applied voltage on the energy band diagram is then investigated.
4. The current as a function of voltage is then calculated on the basis of the shape of the energy band diagram and specific models for carrier transport.

In Chapter 5, the basic principles of operation of semiconductor diodes are discussed. As examples, semiconductor homojunctions with step function doping concentrations at the metallurgical junction are analyzed. These structures can be considered *prototype homojunctions*.

In Chapter 6, deviations from the ideal prototype homojunction model are treated along with heterojunctions and metal-semiconductor junctions. ■

Prototype pn Homojunctions

5.1 INTRODUCTION

A pn homojunction (often called simply a pn junction) consists of a single crystal of a given semiconductor in which the doping level changes from p type to n type at some boundary. The term *homojunction* implies that the junction is between two regions of the same material (e.g., silicon), as opposed to the term *heterojunction* in which the junction is between two different semiconductors (for example, Ge and Si).

There are several methods of fabricating pn junctions. A common technique is to implant phosphorus (donor) atoms into a p-type Si substrate. The resulting structure is shown schematically in Figure 5.1a. In Figure 5.1b a cross-sectional view is shown along the cut A-A'. Since the depth of the n-type region is small compared with its lateral dimension, for most purposes the side effects can be ignored and the properties of the junction can be determined by examining the cross section B-B'. This is indicated schematically in Figure 5.1c.

A typical doping profile for ion implantation is shown schematically in Figure 5.1d. The doping in the p-type substrate (N_A) is assumed constant throughout the crystal. When the donors are implanted, they penetrate the substrate with high energy, causing crystal damage. A donor implantation is followed by a short annealing step (heating the crystal up to eliminate defects). Wherever there are more donors than acceptors ($N_D > N_A$), the semiconductor is n type in that region. It is p type wherever $N_A > N_D$. The metallurgical junction is at x_0 , the point at which $N_D = N_A$, or the net doping is zero. It is the net doping profile that determines the energy band diagram, so in Figure 5.1e, $N'_D = N_D - N_A$ and is plotted as a function of position.

To solve for the electrostatic properties of the junction, the N'_D profile must be known, but it is typically not a simple mathematical function. Thus, an

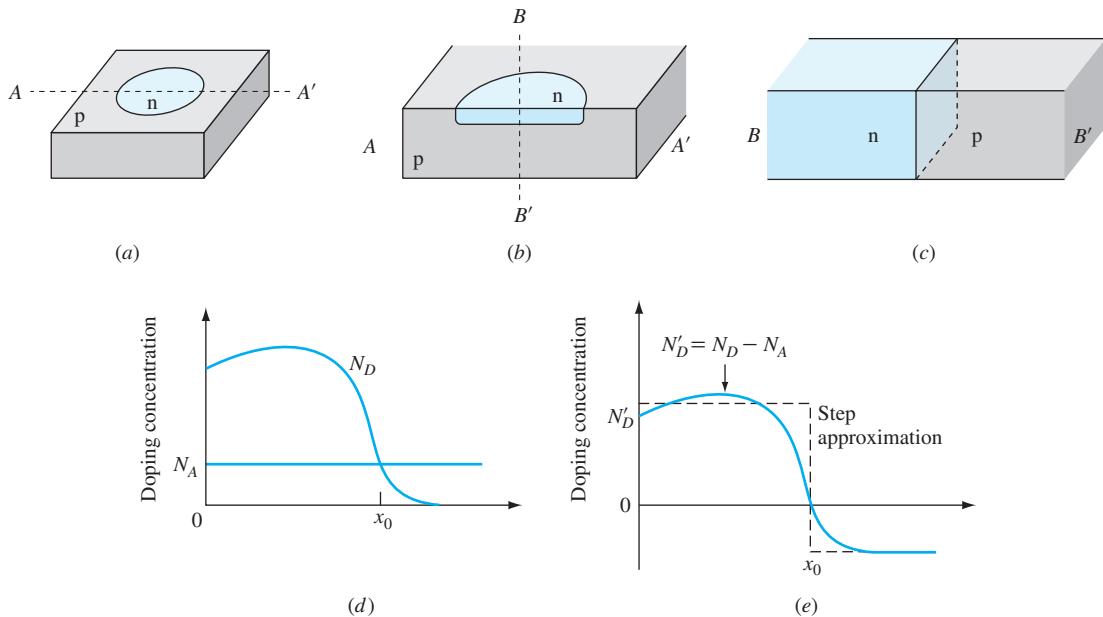


Figure 5.1 (a) The physical picture of a planar pn junction; (b) cross section through $A - A'$; (c) schematic representation of the pn junction; (d) typical doping profile showing a p-type substrate with implanted donors (the junction occurs where $N_D = N_A$); (e) the net doping concentration $N_D - N_A$ for this junction, and the step approximation (dashed line).

approximation to the $N_D - N_A$ profile is often used.¹ The simplest of these is the step approximation, in which the net doping is assumed to be a step function, as shown by the dashed line in Figure 5.1e.

To develop some physical understanding of diodes, we will analyze this simplified model (step junction) of a semiconductor homojunction. We refer to this as the *prototype homojunction*. The purpose is to illustrate the basic principles of operation of semiconductor diodes. In practice, examples of the prototype junction are seldom encountered, but the simplifying approximations permit analytical calculations for many of the device properties. These results, then, can be used as first approximations for more realistic junctions, which are considered in Chapter 6.

The approximations used in the step-junction model are:

1. The doping profile is a step function. On the n-type side, $N'_D = N_D - N_A$ and is constant. On the p side, $N'_A = N_A - N_D$ and is constant.
2. All impurities are ionized. Thus the equilibrium electron concentration on the n side is $n_{n0} = N'_D$. The equilibrium hole concentration on the p side is $p_{p0} = N'_A$.

¹However, in device simulators using numerical solution methods, the more complicated profiles can be used.

3. Impurity-induced band-gap narrowing effects are neglected. Therefore, for the purposes of this simple model, if one side of the junction is degenerate ($N_D, N_A > 10^{18} \text{ cm}^{-3}$ in Si), the Fermi level is assumed to be at E_{C0} , the bottom of the intrinsic conduction band (n type), or at E_{V0} for p type.²

5.2 PROTOTYPE pn JUNCTIONS (QUALITATIVE)

The first step to understanding any junction is to draw its energy band diagram.

5.2.1 ENERGY BAND DIAGRAMS OF PROTOTYPE pn JUNCTIONS

Electrical Neutrality To draw the energy band diagram for the prototype pn junction, we begin by imagining that the n and p regions are physically separated and are electrically neutral in every macroscopic region. Again, by *electrically neutral* we mean that there is no region having more positive charges than negative, a situation that will change when the materials are in contact. The energy band diagram of each of the two isolated semiconductors is given in Figure 5.2.

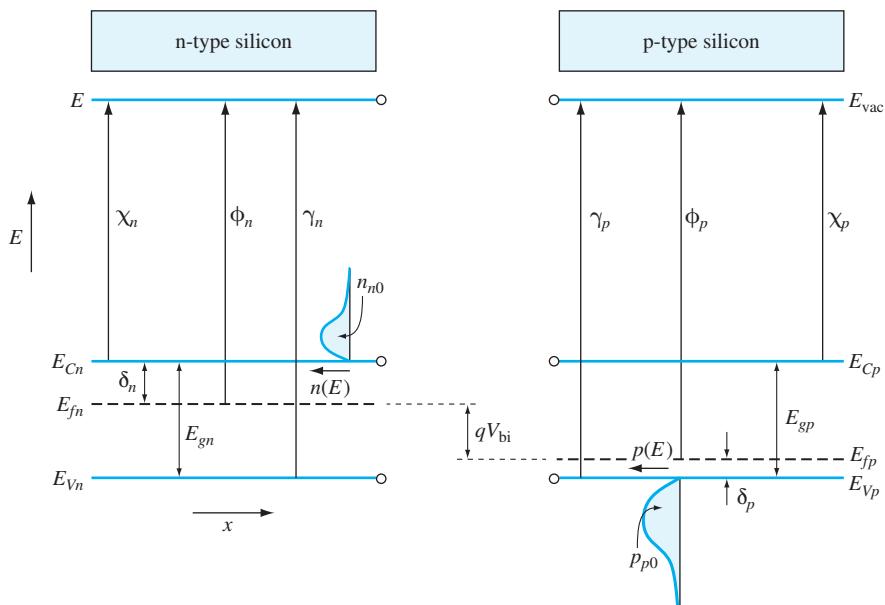


Figure 5.2 Two similar semiconductors, doped differently, before being joined in a junction. The minority carrier concentrations (holes in n type, electrons in p type) are small and not shown.

²Actually, the Fermi level is generally within the (reduced) conduction band in n-type and in the (reduced) valence band for p type degenerately doped materials.

The electron affinity χ , the ionization potential γ , and the energy gap E_g are indicated for each material. The subscript n indicates n-type semiconductor while the subscript p indicates p-type semiconductor. Also shown is an additional parameter, the work function Φ . The work function is equal to the energy difference between the vacuum level and the Fermi level, $\Phi = E_{\text{vac}} - E_f$.

Because of the assumption of space charge neutrality everywhere, the energy required for an electron to escape the material is the same in any region. Therefore, the vacuum level is the same for either material at any position. It is convenient to choose as reference the vacuum level for each material on the edge facing the other material. The circles in the vacuum level represent this.

Since the material is silicon on both sides, $\chi_n = \chi_p$, $E_{gn} = E_{gp}$, and $\gamma_n = \gamma_p$. This implies that the bottom of the conduction band is (for neutrality) at the same energy for both materials, and $E_{Cn} = E_{Cp}$. Similarly $E_{gn} = E_{gp}$ and $E_{Vn} = E_{Vp}$. Since electron affinities and ionization potentials are constant, E_C and E_V at the material edges are secondary references as indicated by the additional circles in Figure 5.2. However, because the doping is different in the two materials, the positions of the Fermi levels are not the same, and thus $\Phi_n \neq \Phi_p$.

Equilibrium Upon contact between the two materials,³ electrons flow (diffuse) from the n-type semiconductor to the p-type semiconductor because there are more quasi-free electrons on the n side than on the p side. As the electrons move toward the p-type region, they leave behind ionized donors (charged positively) that are locked into the crystal lattice. At the same time, holes flow from the p semiconductor to the n semiconductor, leaving behind negatively charged acceptors. This separation of charges sets up an electric field, as shown in Figure 5.3. This is the situation at equilibrium. The presence of the electric field is evident because there is a gradient in the vacuum level and in the band edges. We also observe that the Fermi level is now continuous across the entire sample. We will discuss this figure in more detail shortly, but first let us examine the currents. Because the concentrations of electrons and holes are different on either side of the junction, we expect diffusion current to flow across the junction. At the same time, the presence of an electric field sets up drift current as well. In the transition region between n and p, the electron and hole currents are

$$\boxed{J_n = q\mu_n n\mathcal{E} + qD_n \frac{dn}{dx} = q\mu_n \left[n\mathcal{E} + \frac{kT}{q} \frac{dn}{dx} \right]}$$

$$J_p = q\mu_p p\mathcal{E} - qD_p \frac{dp}{dx} = q\mu_p \left[p\mathcal{E} - \frac{kT}{q} \frac{dp}{dx} \right]$$

At equilibrium there is no net current, so $J_n = J_p = 0$ and the Fermi levels are equalized. A built-in field \mathcal{E} is generated in what is referred to as the *transition region*. In the transition region, the field produces a drift current for electrons

³Here we must assume that in contact the atomic periodicity is continued across the pn interface, or that the entire structure is a single crystal.

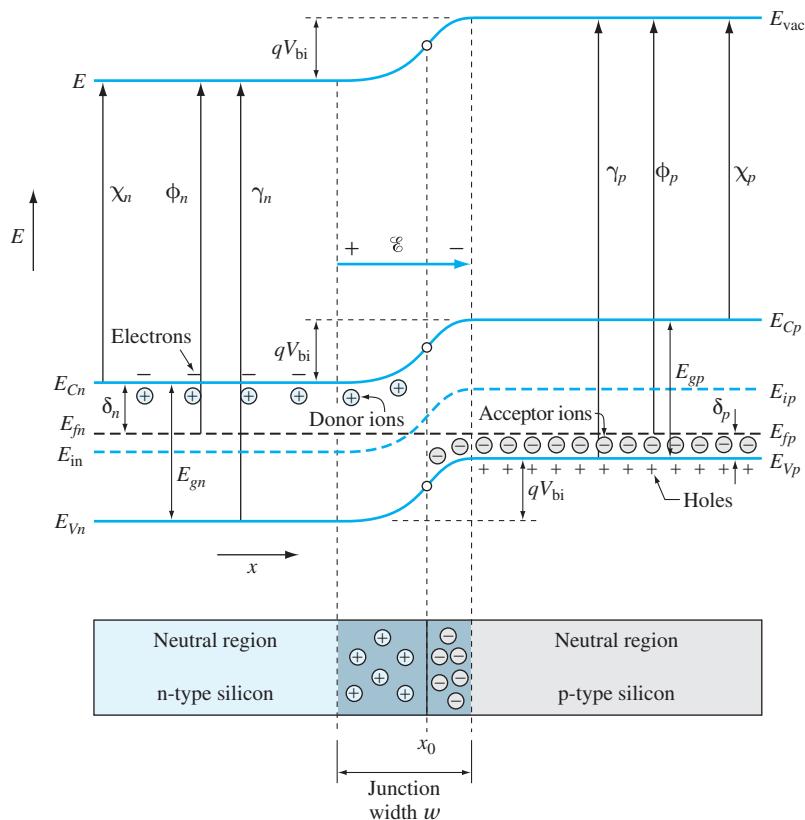


Figure 5.3 The energy band diagram for the pn homojunction at equilibrium.

that at every position exactly compensates for the diffusion current caused by the electron concentration gradient. A similar balance of hole drift and diffusion currents exists.

The built-in field at the junction, however, alters the bands. Since we are using the vacuum levels *at the interface* as a reference, this point will remain unchanged. Also, χ and γ are fundamental properties of silicon, so they will remain unchanged as well. The result is that the vacuum level, conduction band, and valence band all pivot about the positions indicated by the circles at the interface, as Figure 5.3 shows. The details of the shape of this energy band diagram will be discussed later. Notice, however, that:

1. Near the interface, the electrons from the n-type semiconductor fill the holes on the p side resulting in a region of negative charge (non-neutralized acceptors) on the p side. Similarly, electrons annihilate holes that diffused into the n-type region, giving a region of positive charge (non-neutralized donors) on the n side. This creates an electric field at the interface between

non-neutralized (positive) donor ions in the n region and non-neutralized (negative) acceptor ions in the p region. By *non-neutralized*, we mean that there are no corresponding free electrons or free holes in the same region to neutralize the charge of the ions.

2. The energies E_{vac} , E_C , and E_V are everywhere parallel. This is a direct result of the quantities χ , γ , and E_g being constant.
3. The electron affinity is unchanged across the device. However, an electron escaping from the bottom of the conduction band in the n region to the vacuum level on the right side must overcome the electrostatic potential of the built-in field in addition to the electron affinity of the material.
4. On each side of the junction, there is a region of uncompensated charge. This *space charge region*⁴ extends on both sides of the interface, and contains non-neutralized impurity ions. Its width depends on the concentration of impurities on each side and the charge transfer required to align the Fermi levels.
5. The concentration of free carriers in the space charge region is negligible. The carriers are swept out by the electric field. The space charge region is said to be depleted of carriers, and thus is often also referred to as the *depletion region* (and sometimes as the transition region).
6. The total space charge on either side of the junction is the same (but has opposite sign).
7. A built-in potential energy barrier qV_{bi} exists across the junction, where V_{bi} is referred to as the *built-in voltage*. The magnitude of V_{bi} is proportional to the energy difference in the Fermi levels in Figure 5.2 for the case of neutrality:

$$qV_{\text{bi}} = \Phi_p - \Phi_n \quad (5.1)$$

where as indicated earlier, Φ_p and Φ_n are respectively the work functions of the two semiconductors. By convention, V_{bi} is taken as a positive quantity.

8. The potential energy barrier is the same for the conduction band as for the valence band. This implies that the potential energy barrier is the same for electrons as for holes.
9. Because the magnitude of the space charge on either side of the junction is equal, for equal doping levels, the width of the space charge region is the same on each side of the junction. For unequal doping levels most of the space charge region is on the side with the lighter doping.
10. The built-in voltage increases with increased doping level on either side, resulting from the dependence of work function on doping level [Equation (5.1), Figure 5.2].
11. The electric field, which is proportional to the slope of the vacuum level:

$$\mathcal{E} = \frac{1}{q} \frac{dE_{\text{vac}}}{dx} \quad (5.2)$$

⁴Another name for the transition region.

has its maximum value at the metallurgical junction, x_0 . This will be shown in Section 5.3. The field given by Equation (5.2) is the true electric field, as discussed in Chapter 4.

12. The built-in voltage is mostly across the more lightly doped region.

Recall from item 9 that the transition region is primarily on the more lightly doped side. For a junction with one side degenerate and the other side nondegenerate, essentially all of the depletion region will be on the lightly doped side, as indicated in Figure 5.4. This is referred to as a *one-sided step junction*. In this class are n^+p and p^+n junctions. The notation n^+ indicates degenerately or heavily doped n type, and p^+ indicates heavily doped p-type material.

Energy Band Diagram under Bias We have seen how to determine the energy band diagram for a pn homojunction at equilibrium. Now we investigate what happens to the energy band diagram when a voltage is applied. Consider the case of Figure 5.5, in which the p region is made negative with respect to the n region

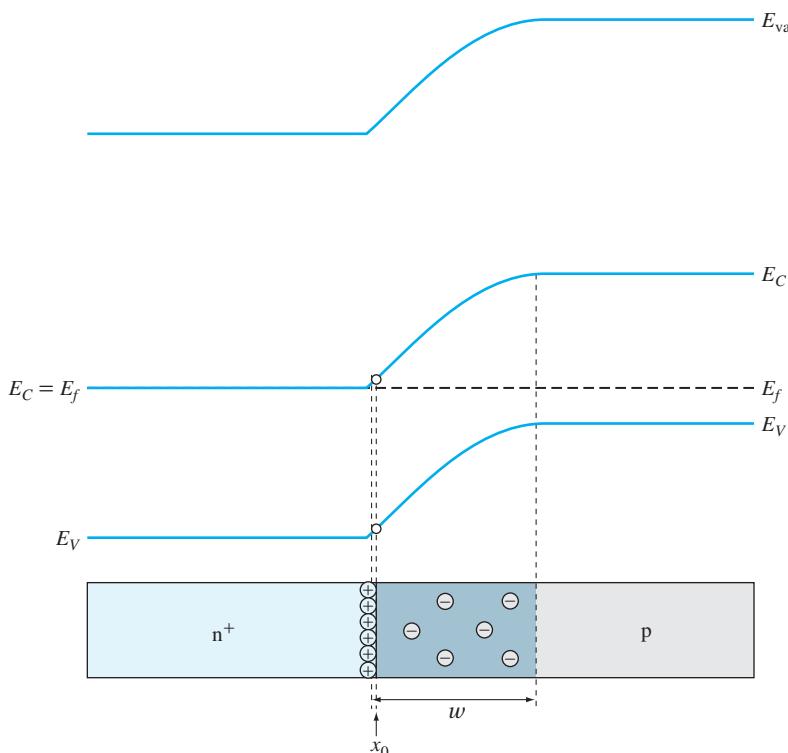


Figure 5.4 The one-sided n^+p junction has one heavily doped side. The designation n^+ indicates degenerately doped n type. On the degenerately doped side, we approximate that the Fermi level is at the conduction band edge.

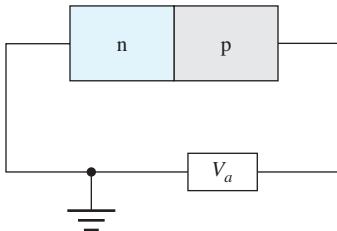


Figure 5.5 A pn homojunction with a bias V_a applied. By convention the applied voltage is measured from p side to n side, or with the n side as a reference. Thus, the diode is forward biased if V_a is positive, and reverse biased if V_a is negative.

by the applied voltage V_a . By convention, V_a is measured from the p side to the n side, i.e., with the n side as reference, and is considered to be negative (reverse bias) if it has the same polarity as the built-in voltage and to be positive (forward bias) if it is of opposite polarity. We will consider the case of reverse bias first, so V_a is negative (for example, $V_a = -1$ V).

But how is V_a distributed across the device? Let us consider the device to consist of three regions:

1. The region from the left terminal to the edge of the junction depletion region on the n side. This is called a *quasi-neutral* region (often called a *neutral region*), since in this area the net number of donors is equal (or nearly equal) to the number of electrons in the conduction band.
2. The depletion region itself, which contains ions but virtually no (or a negligible number of) free carriers.
3. The region from the edge of the depletion region on the p side to the right terminal (also quasi-neutral).

We proceed by considering the pn junction as a series connection of the resistances of these three regions. Recall that resistivity is inversely proportional to the concentration of free carriers at a point, i.e., n_{n0} on the n side and p_{p0} on the p side. In the transition region, however, the carrier concentration is much less than either of these values, since virtually all the free carriers are swept out by the built-in electric field. This implies that the resistance of the transition region is much greater than that of the other two, and virtually all of the applied voltage is dropped across this depletion region.

To adjust the equilibrium energy band diagram to reflect the applied bias, we proceed as we did before in constructing the energy band diagrams at equilibrium. Using the vacuum level at the metallurgical junction as a reference, the energy band diagrams will pivot around the circles, as indicated in Figure 5.6.

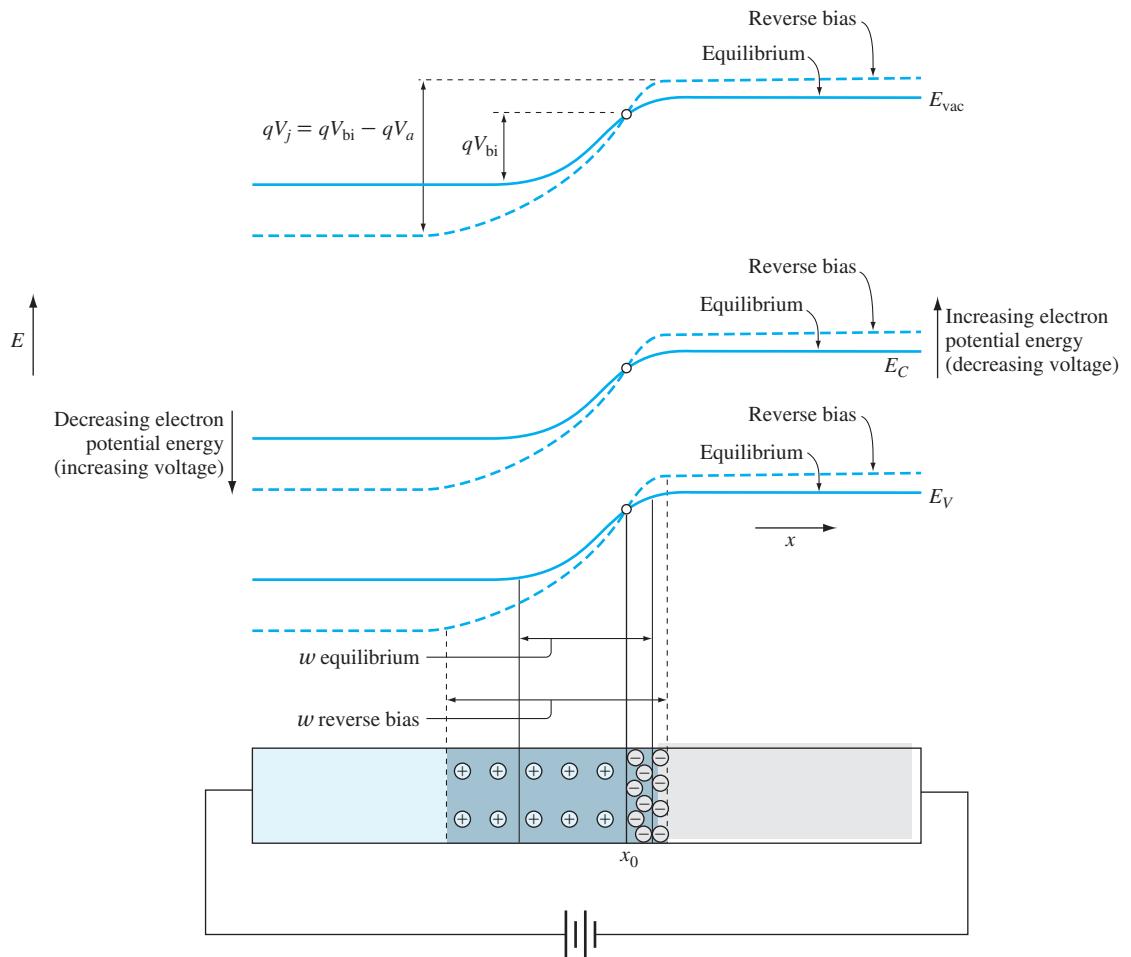


Figure 5.6 The pn homojunction under reverse bias. Solid line: equilibrium energy band diagram; dashed line: energy band diagram under reverse bias. The field increases; this requires more ionized acceptors and donors, so the depletion region gets wider under reverse bias.

Since the p side is made electrically more negative, that represents higher potential energy for electrons and that side of the energy band diagram moves upward. The n side is more positive, representing a lower potential energy for the electrons, and thus the n side moves downward on the energy band diagram.

Figure 5.6 shows the result. The applied negative bias effectively increases the potential barriers for both electrons and holes.

We can see from Figure 5.6 that the junction voltage V_j is increased:

$$V_j = V_{\text{bi}} - V_a \quad (5.3)$$

(with V_a negative because of reverse bias), resulting in a greater field at the metallurgical junction. The internal electric field is generated, however, by the negatively charged acceptor and positively charged donor ions near the junction, so this increased field requires an increased quantity of charge on either side of the junction. Since the charge is a result of ionized impurities that are fixed in the crystal lattice (remember that the mobile carriers are swept away from this region by the field), the transition region must expand on each side. Therefore the width of the transition region increases, as shown in the bottom of Figure 5.6.

Similarly, for a positive or forward bias V_a , the voltage across the junction V_j and the transition width both decrease as shown in Figure 5.7. Here the vacuum level is not indicated since it is parallel to E_C and E_V , and thus holds no new information.

5.2.2 DESCRIPTION OF CURRENT FLOW IN A pn PROTOTYPE HOMOJUNCTION

Now let us qualitatively discuss current flow in a diode. We will start by considering the currents at equilibrium, and then progress to the situations under bias.

Equilibrium The energy band diagram for a diode at equilibrium is shown in Figure 5.8. The electrons are diffusing from their region of high concentration (the n side) to their region of low concentration (the p-type material).

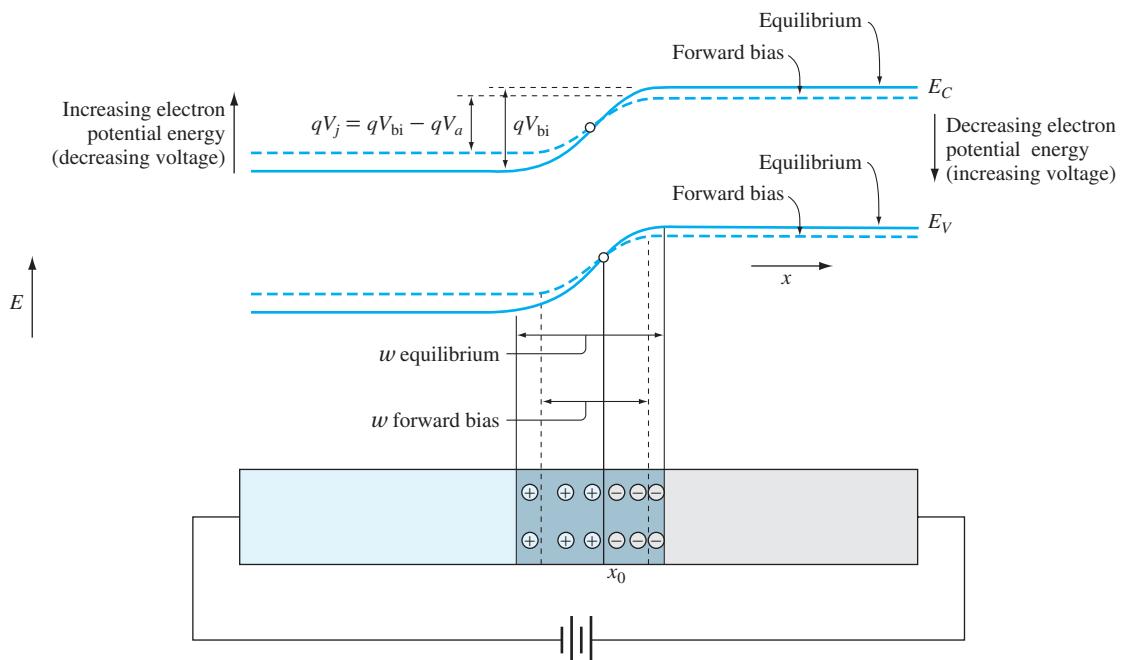


Figure 5.7 The space charge region width under equilibrium and forward bias, and the corresponding energy band diagrams.

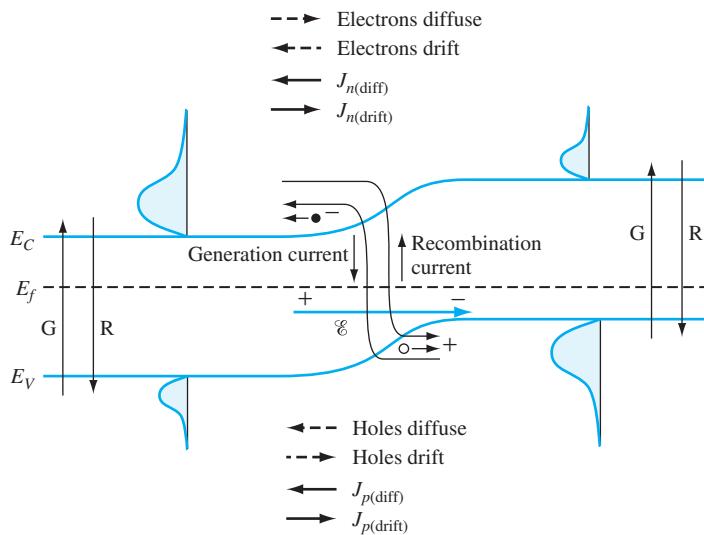


Figure 5.8 The pn homojunction at equilibrium. The built-in field in the transition region causes drift currents that exactly oppose the diffusion current resulting from the different carrier concentrations along the junction. Within the transition region, the generation (G) and recombination (R) currents are equal and opposite at any point.

There is also a built-in electric field at the junction that produces an electron drift that exactly cancels (at equilibrium) the electron diffusion at every point. Similarly, the holes are diffusing from p to n, but the electric field forces them back toward the p region. Neither a net electron nor hole current flows.

Electron generation and recombination occur throughout the device. In the neutral n region, the generation and recombination rates are equal at equilibrium. Thus, since there is no field in this region and no concentration gradients, the generation-recombination processes do not contribute to current. From a similar argument the generation-recombination process in the neutral p region does not produce current.

Within the transition region, however, a field does exist. What is the effect on generated or recombining electron-hole pairs? The electrons generated in the transition region are accelerated toward the n side, producing a current from n to p, as indicated in the figure. Note that the hole produced by the same generation event is accelerated to the p side. This also produces current in the direction n to p. Thus, an electron-hole pair generation event in the transition region produces current across the junction.

Similarly, recombining electrons produce a current in the opposite direction. An electron entering the depletion region from the n side recombines with a hole supplied by the p side. This produces a net current from p to n. At equilibrium, the generation and recombination rates at any position are equal and so the net generation-recombination current is zero.

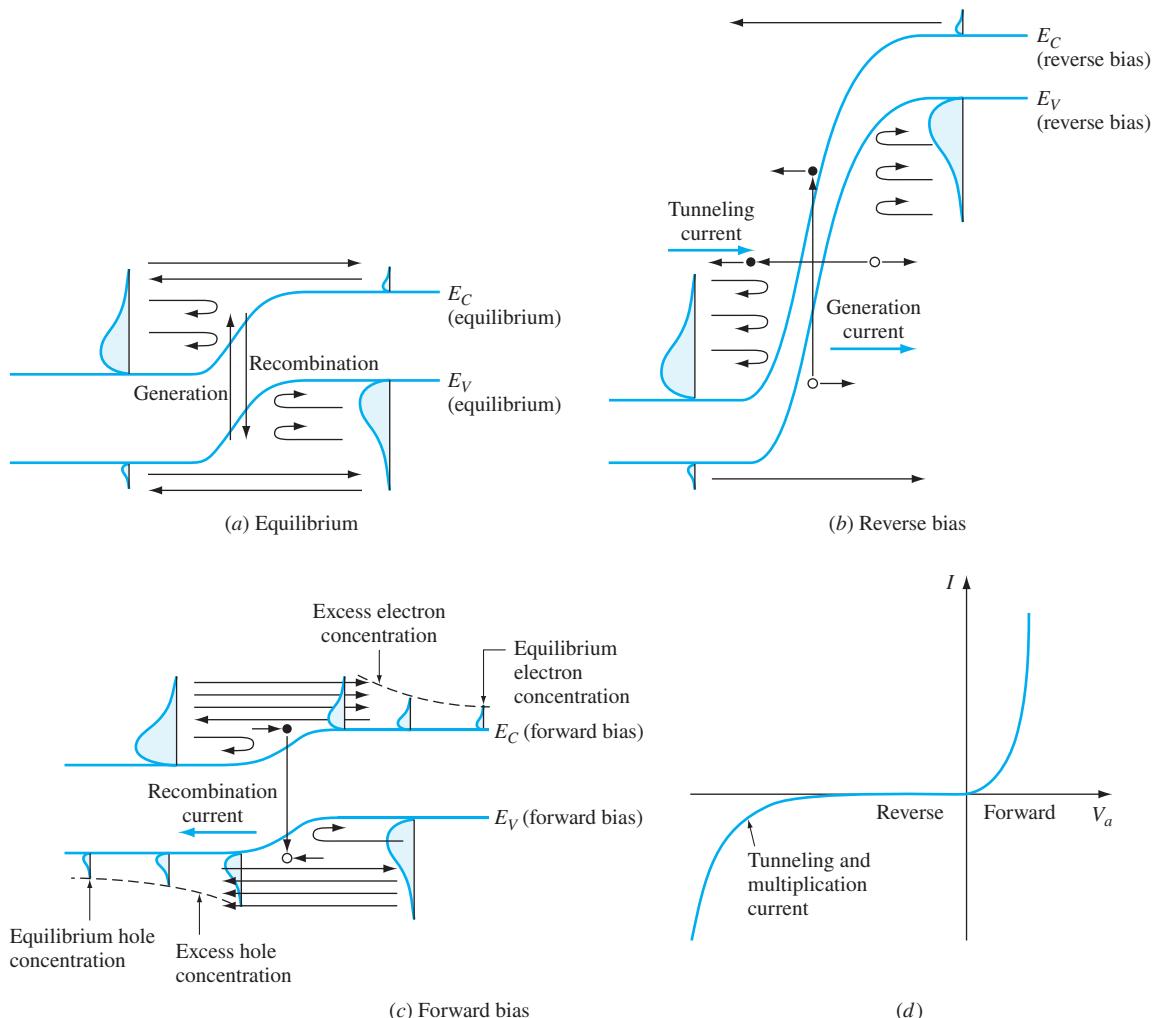


Figure 5.9 Qualitative view of current flow in a pn junction diode. (a) Under equilibrium, both diffusion currents are cancelled by opposing drift currents; (b) under reverse bias, only a small number of carriers are available to diffuse across the junction (once within the junction they are repelled back due to the electric field). With increasing reverse bias the reverse current increases due to tunneling and carrier multiplication. (c) Under forward bias, the drift current is slightly reduced but the diffusion current is greatly increased; (d) the expected current-voltage characteristic.

Reverse Bias Next, we consider the currents that flow under reverse bias. The equilibrium case is shown again in Figure 5.9a, and, for comparison, the reverse-bias case is shown in Figure 5.9b. Under reverse bias, the electrons in the n-type material (majority carriers) attempt to diffuse toward the p-type material (region of lower concentration). However, there is a potential energy barrier

there. A negligible number of electrons have enough kinetic energy to surmount the energy barrier. As for the rest of the electrons, as they travel to the right, their kinetic energy decreases and they eventually stop, whereupon the junction's field accelerates them back to the left. Thus from this process there is negligible net electron current across the junction.

The few conduction band electrons on the right of the junction, however, also have some kinetic energy, and at any given moment, 50 percent of them are traveling to the left. If some of those electrons wander into the transition region, the field accelerates them and they cross into the n-type region. Thus, there is a net electron current, but it is very small since there are few electrons available to participate.

Similarly, for holes, those on the p side tend to diffuse to the n side where their concentration is smaller, but they also are prevented from crossing the junction by the barrier. The very small number of holes on the n side that diffuse to the junction, however, are accelerated to the right. The current resulting from electrons and holes diffusing to the junction and being swept across by the field is referred to as *minority carrier diffusion current*.

There are three additional current mechanisms under reverse bias. A generation current is produced by electrons in the valence band in the transition region being thermally excited into the conduction band then swept to the n side. The hole produced by each generated electron is swept to the p side. Note that recombination in the depletion region under reverse bias is negligible, since both the electron and hole concentrations there are minuscule. The generation current under reverse bias is normally referred to as *leakage current* because of its small size. A second current is a tunneling current. Although electrons on the n side see a high potential barrier, the barrier can be quite thin, as shown in the figure, and some electrons from the valence band on the right can tunnel through the forbidden region into the available states in the conduction band on the left. Tunneling was discussed in the Supplement to Part 1 and is addressed again in Section 5.3 under "Reverse-Bias Tunneling." This tunneling current, then, is caused by electrons in the valence band of the p side that tunnel through the forbidden region into the conduction band on the n side and are swept to the left.

A third current mechanism, *carrier multiplication*, results from electron or hole collisions within the depletion region, which produce electron-hole pairs. These additional carriers also contribute to current. This mechanism is considered in more detail in Section 5.3 under "Reverse-Bias Carrier Multiplication and Avalanche."

Forward Bias For forward bias, we refer to Figure 5.9c. Because the potential energy across the junction is reduced in this case, an appreciable number of electrons now have sufficient energy to cross the junction. Thus, there is a net diffusion of electrons across the barrier from the n side to the p side, producing net electron current from p to n.

Similarly, the applied voltage also reduces the barrier to holes; thus there is a net hole flow and a net hole current from p to n. The hole and electron diffusion currents can add up to a significant current under forward bias.

Once the electrons are injected into the p region, their charge is immediately (within the dielectric relaxation time—on the order of 10^{-12} to 10^{-13} seconds) neutralized by an equal number of oppositely charged holes, which are supplied through the ohmic contact to the p material. The dielectric relaxation time is discussed in more detail in the Supplement to Part 2, but the point here is that except in the transition region, the p region is therefore virtually electrically neutral ($\mathcal{E} \approx 0$). Similarly, outside of the transition region the n side is almost neutral. As indicated earlier, these regions are referred to as being *quasi-neutral*.

Under forward bias, there are excess minority carriers near the junction but outside of the transition region that have been injected across the junction. That is, there are excess electrons near the junction on the p side and excess holes near the junction on the n side. The concentrations of excess electrons Δn and holes Δp decay by recombination as they diffuse away from the junction, as shown in Figure 5.9c. But in the quasi-neutral regions, the electric field $\mathcal{E} \approx 0$, so electron flow into the p region is almost entirely by diffusion. A similar argument applies to holes injected into the n-type region. Thus, the minority carrier currents are often referred to as *diffusion currents*.

What about farther from the junction, where there are no excess carriers? What maintains the current flow? In the quasi-neutral regions, away from the junction, the majority carriers carry the current by drift. Although we said the field is zero in these regions, in reality the field is finite but small. Still, the number of majority carriers available to carry current is so large that the diode current can be maintained by drift even with a small electric field.

We noted earlier that with increasing distance from the junction, the concentration of excess minority carriers decreases because of recombination. This results in decreasing electron diffusion current and a corresponding increase in hole drift current to keep the total current constant.

An additional current under forward bias conditions results from electrons recombining with holes *within* the transition region. While at equilibrium the recombination and generation rates are equal, for forward bias the recombination rate predominates because of the increased concentration of both electrons and holes in the depletion region (as they diffuse across).

For the prototype junction considered here, tunneling cannot occur under forward bias. This is because tunneling occurs at constant energy. There are no allowed states at the same energies across the gap from the valence band, as there were in the reverse bias case. Therefore, electrons cannot tunnel from the valence band on the p side to the conduction band on the n side.

To summarize the current mechanisms under forward bias, when a voltage is applied that reduces the potential barrier, electrons and holes are injected across the depletion region. When they reach the other side, they become minority carriers. As they diffuse away from the junction, they recombine, so the excess minority carrier concentration decreases (exponentially) away from the junction, and thus so does the minority carrier diffusion current. The majority carriers make up the difference between the minority carrier diffusion current and the

total current. The majority carriers travel by drift. The field is small in the quasi-neutral regions, but the majority carrier concentrations are large.

Normally in a pn junction the reverse bias leakage currents are small compared with the currents under forward bias. Thus, to good approximation, the pn junction is a device that permits current to flow in only one direction. The pn junction is often called a *junction diode*.

Current as a function of voltage is discussed quantitatively in the next section. Qualitatively, however, let us consider what to expect. We know that at equilibrium, the number of electrons and holes that have enough energy to diffuse across the barrier is equal to the number that drift back as a result of the built-in field. We also know that electron and hole distribution functions vary exponentially with energy. Therefore, if the barrier is lowered by some amount qV_a , the number of carriers that can diffuse across the barrier increases exponentially. (Note that they diffuse against the electric field.) We expect that under forward bias, then, the current across the junction will increase exponentially with V_a , as shown in Figure 5.9d.

Under reverse bias, the barrier is increased, but the barrier height has no influence on the diffusion current from n to p since the number of minority carriers available doesn't change with barrier height. However, with increasing reverse bias the junction width increases, resulting in increased generation current. In addition, the tunneling distance (at constant energy) decreases, causing increased tunnel current. Further, with increasing reverse bias, the field strength within the transition region increases. Thus, the carriers gain more kinetic energy between collisions. During a collision, if the kinetic energy is large enough, an electron-hole pair can be created, resulting in an increase in multiplication current. Thus, we might qualitatively expect a complete current-voltage ($I-V_a$) characteristic to look something like Figure 5.9d.

Finally, we comment that the diffusion current under forward bias is often referred to as *injection current*. Electrons from the n-type region are injected into the p-type region (over the barrier), and holes from the p region are injected into the n region.

5.2.3 TUNNEL DIODES

Tunnel diodes (also known as Esaki diodes or Esaki tunnel diodes) are pn junctions in which both sides are degenerately doped, such that the Fermi levels are within the (reduced) conduction and valence bands. While tunnel diodes have been largely replaced by other solid-state devices, they do demonstrate quantum mechanic electron tunneling, and they have convinced many skeptics that tunneling actually exists. Further, this tunneling mechanism is the basis for tunnel field effect transistors (TFETs) discussed in Chapter 8.

The current-voltage characteristics of a typical tunnel diode are shown in Figure 5.10. There is a negative slope region to the $I-V$ characteristic, and since the differential resistance is $R = dV/dI$, the diode is characterized by a pronounced negative differential resistance region. The peak and valley currents I_P and I_V are

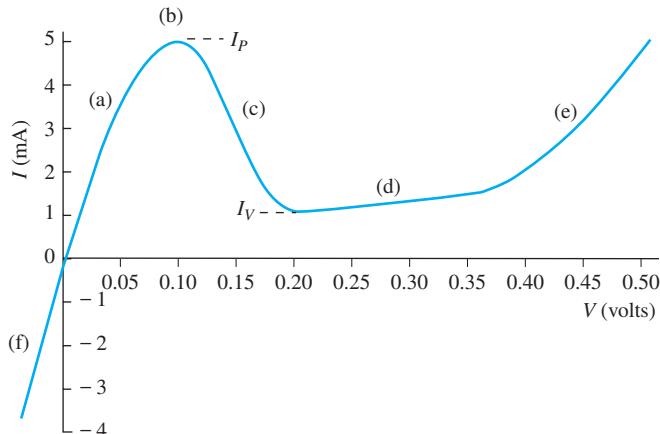


Figure 5.10 I - V characteristics of a germanium tunnel diode. The letters correspond to bias levels shown in parts (a)–(f) of Figure 5.12.

indicated. The negative resistance characteristic of tunnel diodes can be used in applications such as oscillators⁵, amplifiers, and switches.

Figure 5.11 shows an idealized energy band diagram of the p^+n^+ tunnel diode at equilibrium. Due to the high level of doping, the depletion region across the metallurgical junction is narrow. The thickness of this layer is on the order of 10 nm, narrow enough for tunneling of electrons across the junction. Because the built-in voltage is on the order of one volt and the depletion region is on the order of 10 nm, the average electric field in the depletion region is on the order of 1×10^6 V/cm. At such a high field the periodic model of the electron potential energy is a poor approximation, and the band edges, E_C and E_V , are not well-defined, or are fuzzy. To explain qualitatively the I - V characteristics of tunnel diodes, however, the band edges in the depletion region are assumed to be well defined as was done for pn junctions.

For simplicity we assume the following:

1. In the discussion of tunneling, current is primarily from electrons below the Fermi energy on one side of the junction to vacant states (holes) above the Fermi energy on the other side.

⁵Historical note: While Jack Kilby (Texas Instruments) and Robert Noyce (Fairchild Semiconductor) are generally credited with inventing the integrated circuit, in the same time period Dick Rutz (IBM) demonstrated an integrated circuit consisting of a microwave oscillator integrating a germanium tunnel diode and a germanium resistor in the same chip and using the diode capacitance and the inductance inherent to the structure: R. F. Rutz, "A 3000 Mc Lumped-parameter oscillator using an Esaki negative-resistance diode," *IBM Journal of Research and Development*, 3, no. 4, pp. 372–374, October 1959.

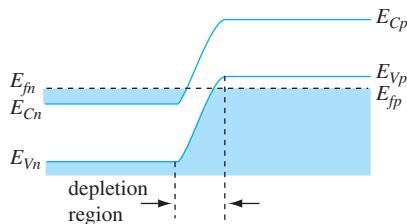


Figure 5.11 Energy band diagram of an n^+ p^+ tunnel diode at equilibrium. The Fermi energies are within the conduction band (n^+) and the valence band (p^+).

2. Electrons cross the “forbidden” region at constant energy and occupy vacant states (holes) of the same energy [conservation of energy (within the uncertainty principle)].
3. Electrons that traverse the forbidden region occupy vacant states with the same wave vector [conservation of wave vector or crystal momentum (within the uncertainty principle)].

Note that assumptions 2 and 3 are for a direct-gap semiconductor where the electrons in the conduction band of the n^+ material are all opposed by vacant states in the p^+ material having the same energies and wave vectors (within the uncertainty principle).

As expected, at equilibrium as indicated in Figure 5.11, the net diode current is zero. Figure 5.12 shows simplified energy band diagrams at various bias levels. For clarity, the Fermi levels are shown to be well inside the conduction and valence bands on the n^+ and p^+ sides respectively.

Figure 5.12a shows the energy band diagram at a small (millivolt) forward bias applied. Electrons can tunnel at constant energy from n^+ to p^+ , and positive current flows from p^+ to n^+ . In (b), at a somewhat higher voltage the current increases. This shows the maximum number of electrons across from equal-energy holes, and the tunnel current reaches a maximum. In (c) the bias is increased such that the concentration of equal-energy electrons and holes is decreased from that of (b), and the current decreases. At higher applied voltages, as in (d), there are no equal-energy electrons and holes, so the tunnel current becomes zero. The mechanism responsible for the observed current is not well understood, but it is thought to result primarily from electron recombination via interband states within the transition region. At still higher voltage as in (e) the current can be fitted to the expression $I = C e^{\frac{qV}{n\tau}}$ where n is in the range $2 > n > 1$ indicating a combination of recombination and injection current. For negative applied voltage, indicated in (f), electrons tunnel from p^+ to n^+ and the negative current increases rapidly with negative voltage.

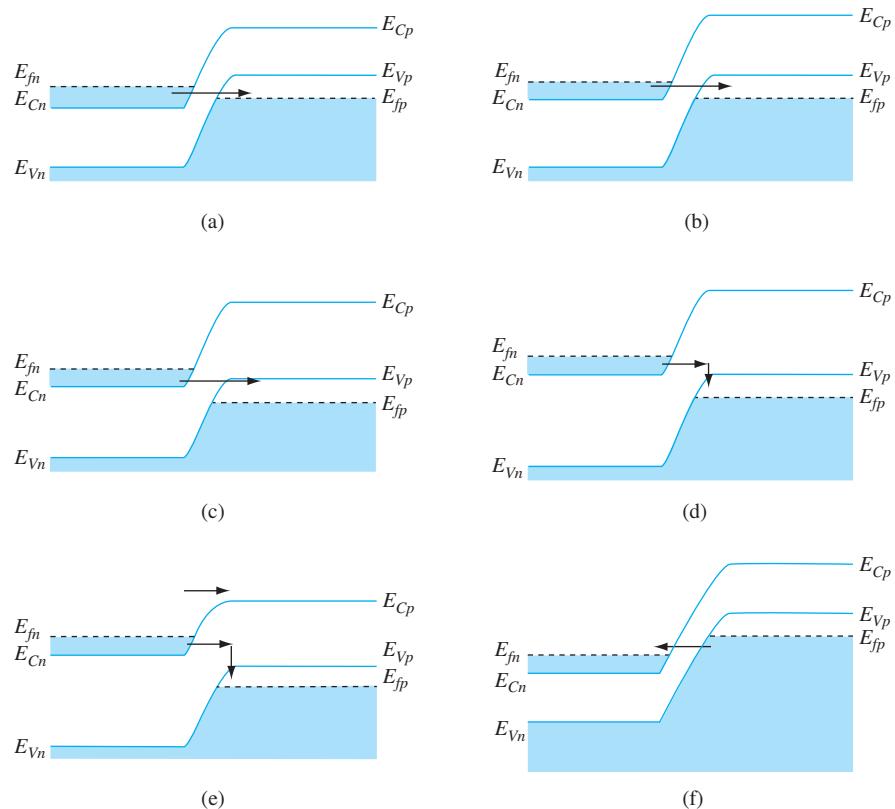


Figure 5.12 Energy band diagrams for an n^+p^+ tunnel diode at several values of forward bias (a) – (e) and negative bias (f) indicating mechanisms of current flow.

A figure of merit for tunnel diodes is the peak to valley current ratio, I_P/I_V . Values of I_P/I_V on the order of 20 have been observed in GaAs tunnel diodes.

When they are operated as oscillators, i.e., in the negative resistance region, the electrical characteristics of tunnel diodes are quite stable with time. When operated like switches, however, such that the electrical characteristics enter the high-voltage, high-current region labeled (e) in Figure 5.10, the valley current increases over time. This degradation is thought to result from the creation of Frenkel defects (displacement of atoms from their equilibrium positions, creating vacancies and interstitial atoms) and thus interband states caused by the recombination of electrons in the depletion region.

The preceding discussion is for p^+n^+ junctions in direct-gap semiconductors, (assumptions 1, 2, and 3). For example, in gallium arsenide (direct-gap semiconductor) the tunneling is primarily between the electrons in the conduction

band minimum at $K \approx 0$ and the valence band maximum at $K \approx 0$. The negative resistance region has been observed in indirect gap semiconductors. In Si, which has conduction band minima in the $\langle 100 \rangle$ directions near the edge of the Brillouin zone and a valence band maximum at $K = 0$, assumption 3 is not satisfied. While negative resistance has been observed, the peak-to-valley current ratio is much reduced, on the order of 3 to 4. It is thought that the law of conservation of wave vector is satisfied by simultaneous generation or absorption of phonons. Ge also is an indirect gap semiconductor having valence band maxima at $K = 0$ and conduction band minima at the edge of the Brillouin zone in the $\langle 111 \rangle$ directions ($K = 3\pi/2a$). It has a relative minimum at $K = 0$, however, slightly above the absolute minima. Tunneling in Ge is primarily between this relative minimum and the valence band. Peak-to-valley current ratios in excess of 16 have been observed.

5.3 PROTOTYPE pn HOMOJUNCTIONS (QUANTITATIVE)

We now have a qualitative understanding of the operation of a prototype pn homojunction in terms of its energy band diagram and transport processes. In this section, we will obtain a quantitative description of the energy band diagram of a pn homojunction. We will also obtain quantitative descriptions of the various current mechanisms. Once these characteristics are known and understood, we can find the junction $I-V_a$ characteristics.

5.3.1 ENERGY BAND DIAGRAM AT EQUILIBRIUM (STEP JUNCTION)

Since the junction characteristics depend on the built-in voltage, we will start by finding an expression for V_{bi} . From Equation (5.1),

$$qV_{bi} = \Phi_p - \Phi_n \quad (5.4)$$

This can be expressed as

$$qV_{bi} = E_g - (\delta_n + \delta_p) \quad (5.5)$$

where the quantity δ_n is the energy difference between conduction band edge and the Fermi level in the neutral region on the n side of the junction, as indicated in Figure 5.2. It is given by

$$\delta_n = E_C - E_f = kT \ln \frac{N_C}{n_{n0}} = kT \ln \frac{N_C}{N'_D} \quad \text{n side, nondegenerate} \quad (5.6)$$

If the material is degenerately doped, we take the Fermi level to be *at* the band edge (remember we are neglecting band-gap narrowing for the prototype pn junction). In this case

$$\delta_n = 0 \quad n \text{ side, degenerate} \quad (5.7)$$

Similarly, δ_p is the energy between the Fermi level and the top of the valence band in the neutral p region:

$$\delta_p = E_f - E_V = kT \ln \frac{N_V}{p_{p0}} = kT \ln \frac{N_V}{N'_A} \quad p \text{ side, nondegenerate} \quad (5.8)$$

and

$$\delta_p = 0 \quad p \text{ side, degenerate} \quad (5.9)$$

As indicated earlier, we usually refer to a junction in which both sides are nondegenerate as a *pn junction*. Often, however, one side is degenerately doped, so the nomenclature then is that an n^+p junction has the n side degenerate and the p side nondegenerate. A p^+n junction is the opposite.

Let us now calculate the built-in voltage of the step pn junction of Figure 5.3. From Equations (5.5), (5.6), and (5.8), we get

$$qV_{bi} = \left[E_g - kT \left(\ln \frac{N_C}{N'_D} + \ln \frac{N_V}{N'_A} \right) \right] \quad \text{pn junction} \quad (5.10)$$

If the junction were n^+p or p^+n , then δ_n or δ_p would be zero, respectively, and Equation (5.10) would be reduced to:

$$\begin{aligned} qV_{bi} &= \left[E_g - kT \ln \frac{N_V}{N'_A} \right] && n^+p \text{ junction} \\ qV_{bi} &= \left[E_g - kT \ln \frac{N_C}{N'_D} \right] && p^+n \text{ junction} \end{aligned} \quad (5.11)$$

We can simplify Equations (5.10) and (5.11) by recognizing that for a non-degenerate semiconductor,

$$\begin{aligned} n_i^2 &= N_C N_V e^{-E_g/kT} \\ E_g &= kT \ln \frac{N_C N_V}{n_i^2} \end{aligned} \quad (5.12)$$

Then from Equations (5.5) through (5.12) we can write

$$\begin{aligned}
 V_{bi} &= \frac{kT}{q} \ln \frac{N'_D N'_A}{n_i^2} && \text{pn junction} \\
 V_{bi} &= \frac{kT}{q} \ln \frac{N_V N'_D}{n_i^2} && \text{p}^+ \text{n junction} \\
 V_{bi} &= \frac{kT}{q} \ln \frac{N_C N'_A}{n_i^2} && \text{n}^+ \text{p junction}
 \end{aligned} \tag{5.13}$$

For a pn junction, we can also express the built-in voltage as

$$V_{bi} = \frac{E_g}{q} - \frac{kT}{q} \ln \frac{N_C N_V}{N'_D N'_A} \quad \text{pn junction} \tag{5.14}$$

This form shows us that the built-in potential approaches the value of the band gap only if both N'_D and N'_A approach N_C and N_V , respectively. Because for Si, $N_C \approx N_V$, the built-in voltages of n⁺p and p⁺n junctions are virtually the same for a given doping on the lightly doped side.

If impurity-induced band-gap narrowing (see approximation 3 in Section 5.1) were included, the built-in voltage would be somewhat reduced from that in Equations (5.13). (In bipolar junction transistors, the band-gap narrowing has an appreciable effect on the transistor characteristics.)

Let us apply our knowledge of how to calculate the built-in voltage of a junction.

EXAMPLE 5.1

Calculate the value of the built-in voltage for a Si pn junction in which the n region is uniformly doped with 10^{16} net donors per cm³ and the p region has a uniform net acceptor concentration of 10^{15} per cm³.

Solution

From Equation (5.13),

$$\begin{aligned}
 V_{bi} &= \frac{kT}{q} \ln \frac{N'_A N'_D}{n_i^2} = \frac{(0.026 \text{ eV})(1.6 \times 10^{-19} \text{ J/eV})}{1.6 \times 10^{-19} \text{ C}} \ln \left[\frac{10^{16} 10^{15}}{(1.08 \times 10^{10})^2} \right] \\
 &= 0.026 \text{ V} \ln [8.57 \times 10^{10}] = 0.655 \text{ V}
 \end{aligned}$$

For a one-sided step junction, the appropriate version of Equation (5.13) would be used.

Because one-sided step junctions are so common, it is worthwhile to plot Equation (5.13) and thus avoid many repetitive calculations. A plot of V_{bi} for a silicon one-sided step junction in which one side is degenerate is shown in Figure 5.13 as a function of net doping on the lightly doped side. The curves for a p⁺n and an n⁺p junction are indistinguishable. If impurity-induced band-gap narrowing is considered, the results of Figure 5.13 will be slightly reduced.

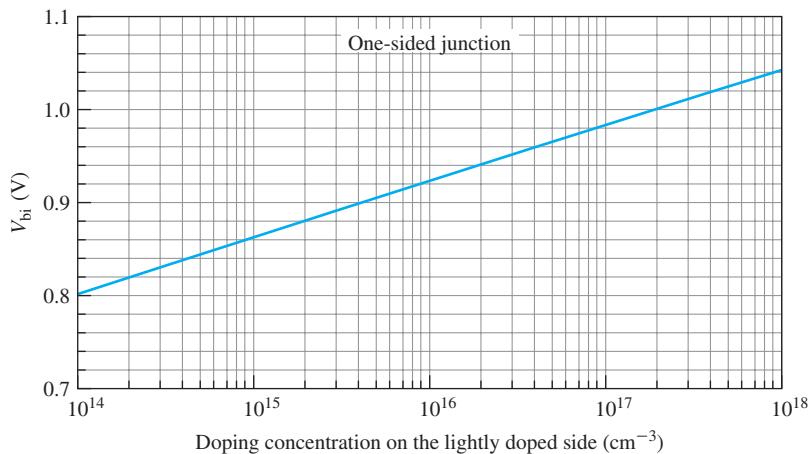


Figure 5.13 The built-in voltage in a one-sided step junction in silicon, as a function of the net doping concentration on the lightly doped side. The curves for the p⁺n and n⁺p junctions are indistinguishable on the plot.

5.3.2 ENERGY BAND DIAGRAM WITH APPLIED VOLTAGE

Next we will quantitatively examine the energy band diagram of a pn junction under applied bias. To determine the shape of the energy band diagram, and see how it varies with applied voltage, we start with the variation of charge density Q_V as a function of position. From that we can find the electric field $\mathcal{E}(x)$, and from the electric field we can determine the variation of the voltage $V(x)$ with position in the junction.

We will use a coordinate system as shown in Figure 5.14a where x_0 is the position of the metallurgical junction. The material is n type for $x < x_0$ and p type for $x > x_0$. The boundary of the depletion region on the n side is x_n and on the p side is x_p . The width of the depletion region on the n-type side is w_n , and the width on the p side is w_p . The overall width is w . The n side is more heavily doped than the p side in this example.

We know the distribution of charge in the device. Under the step approximation, the net doping is constant on the n side and on the p side. In the quasi-neutral regions, the ionized dopants are compensated since there are electrons near the donor ions and holes in the region of acceptor ions. Thus in the quasi-neutral regions there is no uncompensated charge that could give rise to an electric field. In the junction, however, the mobile carriers have been swept out, leaving uncompensated positive donor ions on the n side and uncompensated negative acceptor ions on the p side. These charges, then, will set up the built-in electric field.

Earlier we observed that if a bias is applied, the depletion width changes. Under reverse bias, the junction voltage increases and the depletion width gets wider as more uncompensated charges become “uncovered.” Under forward bias the junction voltage decreases and the transition region narrows. We now show

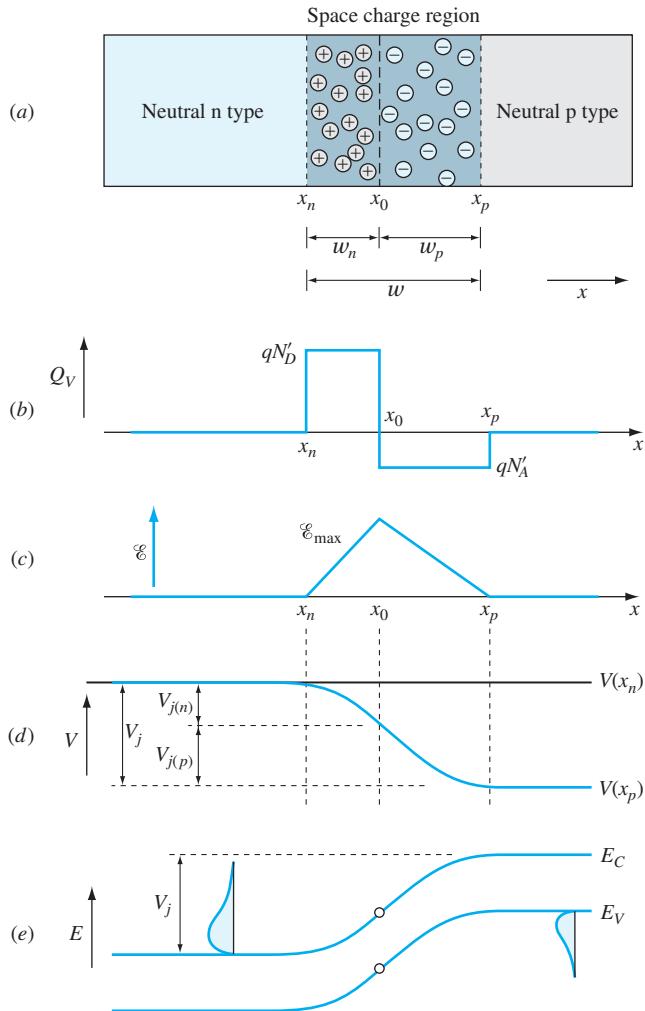


Figure 5.14 A prototype homojunction: (a) the physical diagram; (b) the distribution of charge; (c) the electric field, obtained by integrating the charge; (d) the voltage, obtained by integrating the field; (e) the energy band diagram, with the same shape as the voltage but inverted. The energy band diagram of (e) is for the device reverse-biased.

this mathematically. The following derivation is general and applies under any bias condition.

To find the electric field $\mathcal{E}(x)$, we solve Poisson's equation:

$$\frac{d\mathcal{E}}{dx} = \frac{Q_V(x)}{\epsilon} \quad (5.15)$$

along with the appropriate boundary conditions. The charge density $Q_V(x)$ is the charge per unit volume at a given position x (the V subscript is for volume, to remind us that this is a charge density, not a total charge), and ϵ is the permittivity of the semiconductor.⁶ Note that $\mathcal{E}(x)$ is continuous in x .

To solve Poisson's equation we need to know $Q_V(x)$. The possible charges are the electrons, the holes, and the ionized donors and acceptors. We expect, however, very few electrons or holes in the depletion region itself. We therefore use the *depletion approximation* that, in the transition region, $n = p = 0$.⁷ Since we are assuming that all dopants are ionized,⁸ on the n-type side, where $x_n < x < x_0$,

$$Q_V = qN'_D \quad (5.16)$$

and for $x_0 < x < x_p$

$$Q_V = -qN'_A \quad (5.17)$$

since the ionized acceptors are negatively charged. We further assume that $Q_V(x)$ is zero and thus $\mathcal{E}(x) = 0$ outside the depletion region:

$$\mathcal{E}(x) = 0 \quad x \leq x_n \quad \text{and} \quad x \geq x_p \quad (5.18)$$

The charge density under the depletion approximation is plotted in Figure 5.14b.

Using these expressions in Equation (5.15), on the n side for $x_n \leq x \leq x_0$ we have

$$\int_0^{\mathcal{E}(x)} d\mathcal{E} = \int_{x_n}^x q \frac{N'_D}{\epsilon} dx \quad (5.19)$$

and on the p side for $x_0 < x < x_p$,

$$\int_{\mathcal{E}(x)}^0 d\mathcal{E} = - \int_x^{x_p} q \frac{N'_A}{\epsilon} dx \quad (5.20)$$

On the n side the result is

$$\mathcal{E}(x) = q \frac{N'_D}{\epsilon} (x - x_n) \quad x_n \leq x \leq x_0 \quad (5.21)$$

and on the p side

$$\mathcal{E}(x) = q \frac{N'_A}{\epsilon} (x_p - x) \quad x_0 \leq x \leq x_p \quad (5.22)$$

⁶The permittivity is given by $\epsilon = \epsilon_r \epsilon_0$, and ϵ_r , the relative permittivity (dielectric constant) for silicon, is 11.8.

⁷Actually, there are electrons and holes diffusing and drifting across this region, but their numbers are small.

⁸At low (cryogenic) temperatures, it is possible that not all donors and acceptors are ionized. In this case we must subtract the number of un-ionized donors n_D from the total number of donors N_D , and similarly subtract the un-ionized acceptors n_A from the total acceptor concentration N_A . Including these, plus accounting for possible electrons and holes, yields $Q_V(x) = q[N_D - n_D - n - N_A + p + p_A]$. The un-ionized donor and acceptor concentrations are discussed in the online module OM3.

Since $\mathcal{E}(x)$ is continuous, matching these solutions at $x = x_0$ gives

$$qN'_D(x_0 - x_n) = qN'_A(x_p - x_0) \quad (5.23)$$

The electric field for the step junction increases linearly with x on the n side, and decreases linearly on the p side, as plotted in Figure 5.14c. The maximum of $\mathcal{E}(x)$ occurs at x_0 , the metallurgical junction.

Earlier we had predicted that the junction would extend further into the lightly doped side. We will see now whether we were right. Letting w_n and w_p be the transition region widths on the n and p sides respectively, from Equation (5.23) we can write

$$\frac{w_n}{w_p} = \frac{(x_0 - x_n)}{x_p - x_0} = \frac{N'_A}{N'_D} \quad (5.24)$$

Therefore if N'_D is the larger doping concentration (on the n side), then w_p is the larger depletion width, verifying our prediction.

Now we find the functional form of $V(x)$, the voltage distribution. We use the expression

$$\mathcal{E} = -\frac{dV}{dx} \quad (5.25)$$

On the n side, we integrate V from $x = x_n$ to $x \leq x_0$ and find:

$$\int_{V(x_n)}^{V(x)} dV = - \int_{x_n}^x \mathcal{E}(x) dx = - \int_{x_n}^x \frac{qN'_D}{\epsilon} (x - x_n) dx \quad (5.26)$$

or

$$V(x) - V(x_n) = -\frac{qN'_D}{2\epsilon} (x - x_n)^2 \quad x_n \leq x \leq x_0 \quad (5.27)$$

Likewise for the p side:

$$V(x_p) - V(x) = -\frac{qN'_A}{2\epsilon} (x_p - x)^2 \quad x_0 \leq x \leq x_p \quad (5.28)$$

We can find the voltage drops across the n side of the transition region and across the p side by equating Equations (5.27) and (5.28) at $x = x_0$:

$$V(x_n) - V(x_0) = V_j^n = \frac{qN'_D}{2\epsilon} (x_0 - x_n)^2 = \frac{qN'_D}{2\epsilon} w_n^2 \quad (5.29)$$

and

$$V(x_0) - V(x_p) = V_j^p = \frac{qN'_A}{2\epsilon} (x_p - x_0)^2 = \frac{qN'_A}{2\epsilon} w_p^2 \quad (5.30)$$

where V_j^n and V_j^p are the voltages across the n and p sides of the junction respectively.

Then we can find the total voltage across the junction, V_j :

$$V_j = V_j^n + V_j^p = \frac{q}{2\epsilon} [N'_D (x_0 - x_n)^2 + N'_A (x_p - x_0)^2] \quad (5.31)$$

These quantities are indicated on the plot of $V(x)$ shown in Figure 5.14d. Note that the voltage decreases from x_n to x_p , since the electric field is positive.

Since $V(x)$ is continuous, we match Equations (5.27) and (5.28) at x_0 :

$$V(x_0) = V(x_n) - \frac{qN'_D}{2\epsilon}(x_0 - x_n)^2 = V(x_p) + \frac{qN'_A}{2\epsilon}(x_p - x_0)^2 \quad (5.32)$$

Evaluating Equations (5.29) and (5.30) at x_0 and taking the ratio,

$$\frac{V(x_n) - V(x_0)}{V(x_0) - V(x_p)} = \frac{V_j^n}{V_j^p} = \frac{N'_D w_n^2}{N'_A w_p^2} \quad (5.33)$$

With the aid of Equation (5.24), Equation (5.33) becomes

$$\frac{V_j^n}{V_j^p} = \frac{N'_A}{N'_D} \quad (5.34)$$

indicating that most of the junction voltage is dropped across the more lightly doped region.

Next, let us find expressions for the junction widths. From Equations (5.29), (5.32), and (5.33) we can obtain an expression for w_n :

$$w_n = (x_0 - x_n) = \left[\frac{2\epsilon V_j^n}{qN'_D} \right]^{1/2} = \left[\frac{2\epsilon V_j}{qN'_D \left(1 + \frac{N'_D}{N'_A} \right)} \right]^{1/2} \quad (5.35)$$

Similarly

$$w_p = (x_p - x_0) = \left[\frac{2\epsilon V_j^p}{qN'_A} \right]^{1/2} = \left[\frac{2\epsilon V_j}{qN'_A \left(1 + \frac{N'_A}{N'_D} \right)} \right]^{1/2} \quad (5.36)$$

which results in a total junction width

$$w = w_n + w_p = \left[\frac{2\epsilon V_j (N'_A + N'_D)}{qN'_A N'_D} \right]^{1/2} \quad \text{pn junction} \quad (5.37)$$

Solving for the junction voltage gives

$$V_j = \frac{qN'_D N'_A w^2}{2\epsilon (N'_D + N'_A)} \quad \text{pn junction} \quad (5.38)$$

For a one-sided junction, the junction is almost entirely on the lightly doped side. For an n^+p junction, $w \approx w_p$, since in this case $N'_D \gg N'_A$. From Equation (5.37) we have

$$w = \left[\frac{2\epsilon V_j}{qN'_A} \right]^{1/2} \quad n^+p \quad (5.39)$$

and for a p⁺n junction

$$w = \left[\frac{2\epsilon V_j}{qN_D} \right]^{1/2} \quad \text{p}^+ \text{n} \quad (5.40)$$

We found earlier that the maximum electric field occurs at the junction $x = x_0$; now we can find its value. From Equations (5.21), (5.37), and (5.38),

$$\mathcal{E}_{\max} = \frac{qN'_D}{\epsilon} w_n = \frac{qN'_A}{\epsilon} w_p = \left[\frac{2qV_j N'_D N'_A}{\epsilon(N'_D + N'_A)} \right]^{1/2} = \frac{2V_j}{w} \quad (5.41)$$

Finally, we can obtain the shape of the energy band diagram by recognizing that the potential energy is related to the electric potential:

$$\frac{dE_P}{dx} = -q \frac{dV}{dx} = \frac{dE_C}{dx} = \frac{dE_V}{dx} \quad (5.42)$$

Thus, the conduction band has the same shape as $V(x)$, but inverted, resulting in the energy band diagram of Figure 5.14e.

The width of the transition region is important for determining the various current mechanisms in a pn junction. Figure 5.15 shows the width as a function of junction voltage, $V_j = V_{bi} - V_a$ as calculated from Equation (5.37) for a one-sided step junction. The plot shows this for several values of doping on the lightly doped side. As expected, w increases with reduced doping and increased junction voltage.

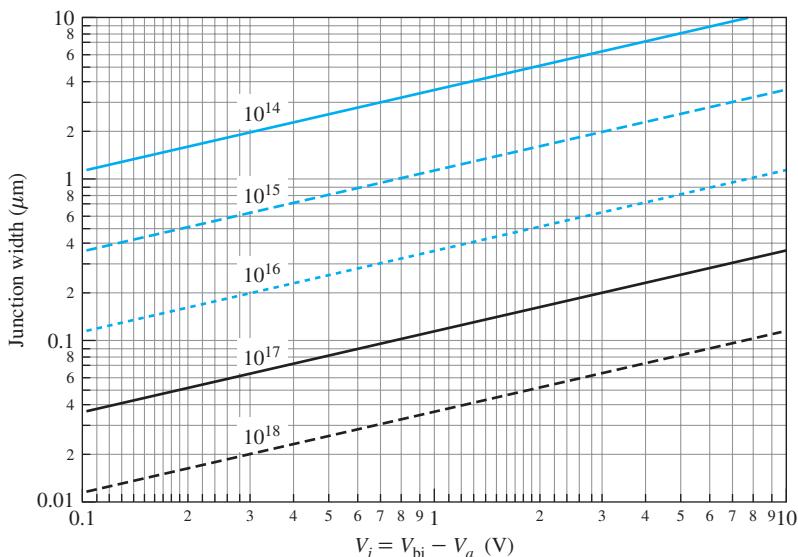


Figure 5.15 The junction width for one-sided Si step junctions in silicon as a function of junction voltage with the doping on the lightly doped side as a parameter.

In Figure 5.16, the junction width for a one-sided step junction is shown as a function of doping for applied voltages of 0.5 V (forward bias), 0 V, and -5 V (reverse bias).

5.3.3 CURRENT-VOLTAGE CHARACTERISTICS OF pn HOMOJUNCTIONS

We now obtain expressions for the steady-state current versus applied voltage ($I-V_a$) characteristics of prototype pn homojunctions. The case of one-sided step junctions, i.e., n⁺p and p⁺n junctions, will be discussed specifically later. In addition to the assumptions already made, we assume that:

1. On either side of the junction the minority carrier concentration is everywhere much less than the majority carrier concentration:

$$\begin{aligned} n_p &\ll p_p \approx N'_A \\ p_n &\ll n_n \approx N'_D \end{aligned} \quad (5.43)$$

This is referred to as the *low-level injection condition*.

2. In the bulk region of a semiconductor the majority carrier concentration is essentially the equilibrium value. This is a result of space-charge neutrality ($\Delta n = \Delta p$) and the low injection condition. This implies that

$$\begin{aligned} n_n &\approx n_{n0} = N'_D \\ p_p &\approx p_{p0} = N'_A \end{aligned} \quad (5.44)$$

3. For minority carriers the drift current can be neglected compared with diffusion current in the quasi-neutral regions. Thus

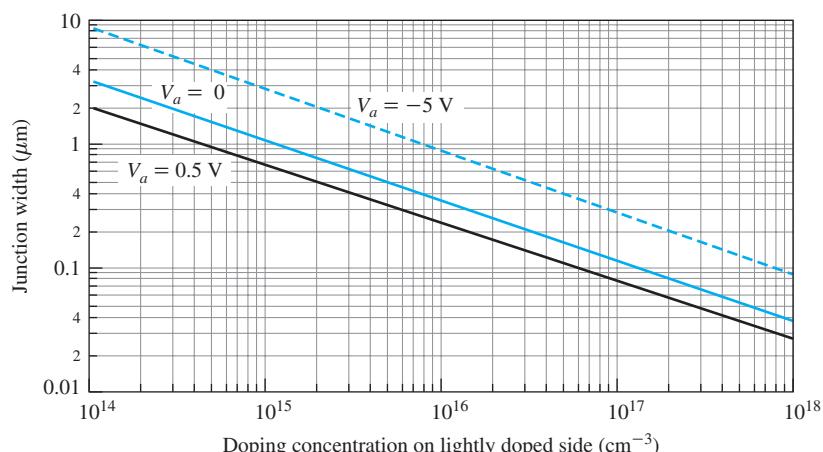


Figure 5.16 Junction width for a one-sided Si junction is plotted as a function of doping on the lightly doped side for three different operating voltages.

$$\begin{aligned} J_{n(p)} &= qD_n \frac{dn_p}{dx} \\ J_{p(n)} &= -qD_p \frac{dp_n}{dx} \end{aligned} \quad (5.45)$$

where D_n and D_p are the minority carrier diffusion coefficients.

4. The semiconductors are nondegenerate. This implies that Boltzmann's statistics can be used, and at equilibrium

$$\begin{aligned} n_0 &= N_C e^{-[(E_C - E_f)/kT]} \\ p_0 &= N_V e^{-[(E_f - E_v)/kT]} \end{aligned} \quad (5.46)$$

5. We define current to be positive for positive V_a and negative for negative V_a . Thus for positive current, electrons flow from n to p and holes flow from p to n.

To find the I - V characteristic, we first consider a *long-base diode* in which both sides of the quasi-neutral regions are much longer than their minority carrier diffusion lengths, L_n or L_p . After that, we consider the case of a *short-base diode* in which one or both sides are much shorter than a diffusion length.

The most important mechanism for current flow across a pn junction is referred to as minority carrier injection-extraction (I-E) current. It is the fundamental quantity of interest in devices such as bipolar junction transistors as well as in optoelectronic devices such as lasers, light-emitting diodes, and photodetectors. For the prototype junctions considered in this chapter, the I-E current is entirely diffusion current.⁹ We will discuss this next. Then we will go on to discuss generation-recombination (G-R) current, tunnel current, and carrier multiplication and avalanche currents.

Diffusion Current Diffusion current consists of minority carriers that diffuse toward or away from the junction. To obtain the I - V characteristic of a diode it is convenient to begin with the continuity equations

$$\frac{\partial n}{\partial t} = \frac{\partial \Delta n}{\partial t} = \frac{1}{q} \left(\frac{\partial J_n}{\partial x} \right) + \left(G_{op} - \frac{\Delta n}{\tau_n} \right) \quad (5.47)$$

$$\frac{\partial p}{\partial t} = \frac{\partial \Delta p}{\partial t} = -\frac{1}{q} \left(\frac{\partial J_p}{\partial x} \right) + \left(G_{op} - \frac{\Delta p}{\tau_p} \right) \quad (5.48)$$

where Δn and Δp are the excess electron and hole concentrations.

⁹In nonprototype junctions, i.e., those with nonuniform doping on either side, drift current is also important, but the injection-extraction current is often referred to as *diffusion current*.

First we consider electrons within the p region of a pn junction, where they are minority carriers. Then from Assumption 3 [Equation (5.45)],

$$J_n = J_{n(\text{diff})} = q D_n \frac{dn}{dx} \quad (5.49)$$

Combining Equations (5.47) and (5.49) gives

$$\frac{\partial n}{\partial t} = \frac{\partial \Delta n}{\partial t} = D_n \left(\frac{\partial^2 n}{\partial x^2} \right) - \frac{\Delta n}{\tau_n} \quad (5.50)$$

For steady state, $\partial n / \partial t = 0$ and Equation (5.50) can be expressed

$$\frac{\partial^2 n}{\partial x^2} = \frac{d^2 n}{dx^2} = \frac{\Delta n}{D_n \tau_n} = \frac{\Delta n}{L_n^2} \quad (5.51)$$

where the minority carrier diffusion length for electrons is, from Equation (3.82),

$$L_n = \sqrt{D_n \tau_n}$$

Since $n_p = n_{p0} + \Delta n_p$ where Δn_p is the excess electron concentration in the p region, and n_{p0} is constant, Equation (5.51) can be expressed

$$\frac{d^2 \Delta n_p}{dx^2} = \frac{\Delta n_p}{L_n^2} \quad (5.52)$$

whose solution is

$$\Delta n_p = A e^{x/L_n} + B e^{-x/L_n} \quad (5.53)$$

The constants A and B are determined from the boundary conditions. For a long-base diode, the p region extends several diffusion lengths and thus the excess carriers will have all recombined by the end of the region. Evaluating Equation (5.53) with $\Delta n = 0$ at $x = \infty$ gives $A = 0$. At $x = x_p$, the boundary condition is that $\Delta n(x_p) = B e^{-(x-x_p)/L_n}$, giving $B = \Delta n(x_p) e^{x_p/L_n}$ and

$$\boxed{\Delta n(x) = \Delta n(x_p) e^{-(x-x_p)/L_n}} \quad (5.54)$$

That is, the excess electron concentration decays exponentially with position from its value at the edge of the junction.

Since

$$J_n = q D_n \frac{dn}{dx} = q D_n \frac{d\Delta n}{dx} \quad (5.55)$$

the electron current is

$$\boxed{J_n = \frac{q D_n}{L_n} \Delta n(x_p) e^{-(x-x_p)/L_n}} \quad (5.56)$$

Note that the electron current density decreases exponentially with x on the p side. But since the total current must remain constant, the hole current increases

by the same amount. This hole current is drift current resulting from the small electric field in this quasi-neutral region.

At the edge of the transition region where $x = x_p$, the electron current density is

$$J_n(x_p) = q \frac{D_n}{L_n} \Delta n(x_p) \quad (5.57)$$

It is useful to obtain an expression for $\Delta n(x_p)$ as a function of applied voltage. We first consider the case of equilibrium where $\Delta n(x_p) = 0$ or $n(x_p) = n_{p0}$.

Equilibrium Consider the equilibrium energy band diagram of Figure 5.17. On the n side, the electron concentration is $n_{n0} = N'_D$ while the hole concentration on that side is $p_{n0} = n_i^2/N'_D$. Both are uniform in the quasi-neutral n region. On the p side the hole concentration is $p_{p0} = N'_A$ and the electron concentration is $n_{p0} = n_i^2/N'_A$. The potential energy barrier for both electrons and holes is qV_{bi} .

We now relate N'_D , n_{p0} , N'_A , p_{n0} , and V_{bi} . In the neutral p region, we have

$$n_{p0} = N_C e^{-[(E_{Cp}-E_f)/kT]} \quad (5.58)$$

Multiplying by $e^{-E_{Cn}/kT} e^{+E_{Cn}/kT}$ allows this to be written

$$n_{p0} = N_C e^{-[(E_{Cn}-E_f)/kT]} e^{-[(E_{Cp}-E_{Cn})/kT]} \quad (5.59)$$

But since the energy step in the conduction band edge is proportional to the built-in voltage:

$$(E_{Cp} - E_{Cn}) = q V_{bi} \quad (5.60)$$

we can rewrite Equation (5.59) as

$$n_{p0} = [N_C e^{-[(E_{Cn}-E_f)/kT]}] e^{-qV_{bi}/kT} \quad (5.61)$$

The part in the square brackets is the electron concentration on the n side:

$$n_{n0} = N'_D = N_C e^{-q(E_{Cn}-E_f)/kT} \quad (5.62)$$

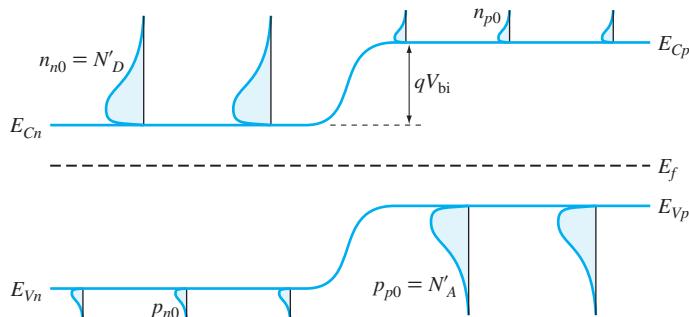


Figure 5.17 Equilibrium energy band diagram for a step junction.

and, therefore, combining Equations (5.61) and (5.62) gives the minority carrier concentration on the p side:

$$n_p(x_p) = n_{p0} = N'_D e^{-qV_{bi}/kT} = n_{n0} e^{-qV_{bi}/kT} \quad (5.63)$$

Similarly

$$p_n(x_n) = p_{n0} = N'_A e^{-qV_{bi}/kT} = p_{p0} e^{-qV_{bi}/kT} \quad (5.64)$$

These may seem counterintuitive because the carrier concentration on one side appears to depend on the doping on the other side, but remember that the qV_{bi} term takes into account the doping on *both* sides.

Diffusion Current: Forward Bias Next, we consider the case of an externally applied voltage. For example, under a forward bias of V_a (Figure 5.18), the energy barrier is changed to $(E_{Cp} - E_{Cn}) = q(V_{bi} - V_a)$. The p side has been raised to a higher electric potential, or lowered to a lower electron potential energy, and the n side has increased in potential energy. Equations (5.63) and (5.64) can be adapted to express the concentrations in terms of the doping on the other side and the barrier height. At the edges of the depletion region, x_n and x_p , the minority carrier concentrations are

$$\begin{aligned} n_p(x_p) &= N'_D e^{-q(V_{bi}-V_a)/kT} \\ p_n(x_n) &= N'_A e^{-q(V_{bi}-V_a)/kT} \end{aligned} \quad (5.65)$$

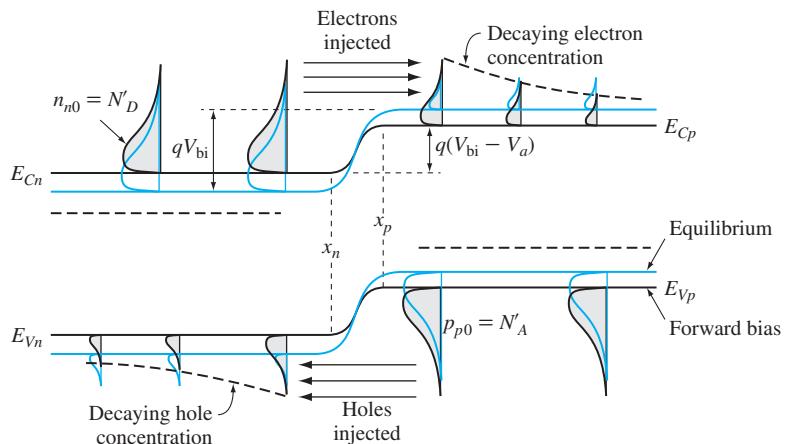


Figure 5.18 Under forward bias, the n side moves up in potential energy and the p side moves down. The colored lines correspond to equilibrium; the black lines represent forward bias. Carriers are injected across the junction by diffusion. They become minority carriers once they get across, and they recombine as they continue to diffuse. The result is a varying concentration of minority carriers with distance from the junction.

or from Equations (5.63) and (5.64)

$$\begin{aligned} n_p(x_p) &= n_{p0} e^{qV_a/kT} \\ p_n(x_n) &= p_{n0} e^{qV_a/kT} \end{aligned} \quad (5.66)$$

When forward bias is applied, the excess electrons that are injected into the p region diffuse to the right and recombine with holes on the p side. Holes are also injected into the n side, where they also recombine. Let Δn_p represent the excess electron concentration on the p side and Δp_n represent the excess hole concentration on the n side:

$$\begin{aligned} \Delta n_p(x_p) &= n_p(x_p) - n_{p0} \\ \Delta p_n(x_n) &= p_n(x_n) - p_{n0} \end{aligned} \quad (5.67)$$

Then from Equations (5.66),

$$\begin{aligned} \Delta n_p(x_p) &= n_{p0}(e^{qV_a/kT} - 1) \\ \Delta p_n(x_n) &= p_{n0}(e^{qV_a/kT} - 1) \end{aligned} \quad (5.68)$$

These are excess carriers injected across the junction. The excess electrons will diffuse, and as they diffuse in the p material they will recombine because of the abundance of holes. The injected excess electron concentration decreases exponentially with x as

$$\Delta n_p(x) = \Delta n_p(x_p)e^{-(x-x_p)/L_n} \quad (5.69)$$

and on the n side the hole concentration will obey

$$\Delta p_n(x) = \Delta p_n(x_n)e^{-(x_n-x)/L_p} \quad (5.70)$$

These injected carrier concentrations are shown in Figure 5.19. Since the concentrations vary with distance, they set up diffusion currents. For the electron diffusion current on the p side,

$$J_n = qD_n \frac{dn}{dx} = qD_n \frac{d\Delta n}{dx} \quad (5.71)$$

where D_n and L_n are the electron minority carrier diffusion coefficient and the minority carrier diffusion length respectively.

Substituting Equations (5.69) and (5.68) into Equation (5.71) gives us

$$J_n = q \frac{D_n}{L_n} n_{p0}(e^{qV_a/kT} - 1) e^{-(x-x_p)/L_n} \quad (5.72)$$

At the edge of the transition region, where $x = x_p$, the electron diffusion current density is

$$J_n(x_p) = q \frac{D_n}{L_n} n_{p0}(e^{qV_a/kT} - 1) \quad (5.73)$$

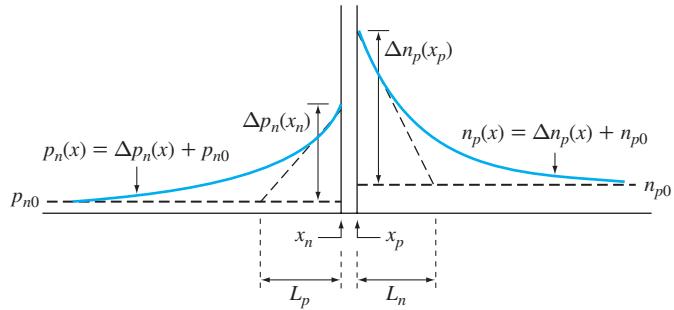


Figure 5.19 The minority carrier concentrations on either side of the junction.

Similarly, the hole diffusion current density in the n region at $x = x_n$ is

$$J_p(x_n) = q \frac{D_p}{L_p} p_{n0} (e^{qV_a/kT} - 1) \quad (5.74)$$

where D_p and L_p are hole minority carrier diffusion coefficient and diffusion length respectively.

Equations (5.73) and (5.74) give the minority carrier diffusion currents at the edge on either side of the depletion region. Since recombination and generation in the transition region are neglected, all of the excess electrons on the p side crossing $x = x_p$ had to come from the n side, and therefore also had to cross $x = x_n$. This is shown in Figure 5.20a. Similarly the hole diffusion current $J_p(x = x_n)$ must equal the hole current crossing x_p (Figure 5.20b). We also agreed by Assumption 3 that the drift of minority carriers was negligible, so on either side of the junction the total current of the minority carriers is due to diffusion. Therefore both electron and hole currents are continuous across the transition region and at any point in the transition region, as shown in Figure 5.20c. The total current anywhere between x_n and x_p is given by the sums of the diffusion currents from either side of the junction:

$$J = J_n(x_p) + J_p(x_n) \quad (5.75)$$

The figure also shows the decrease in diffusion current and corresponding increase in majority carrier drift currents in the bulk regions needed to keep the total current constant.

Substituting in Equations (5.73) and (5.74) gives the total current density:

$$J = q \left(\frac{D_n n_{p0}}{L_n} + \frac{D_p p_{n0}}{L_p} \right) \left(e^{qV_a/kT} - 1 \right) \quad (5.76)$$

This is normally written in the form

$$J = J_0 (e^{qV_a/kT} - 1) \quad (5.77)$$

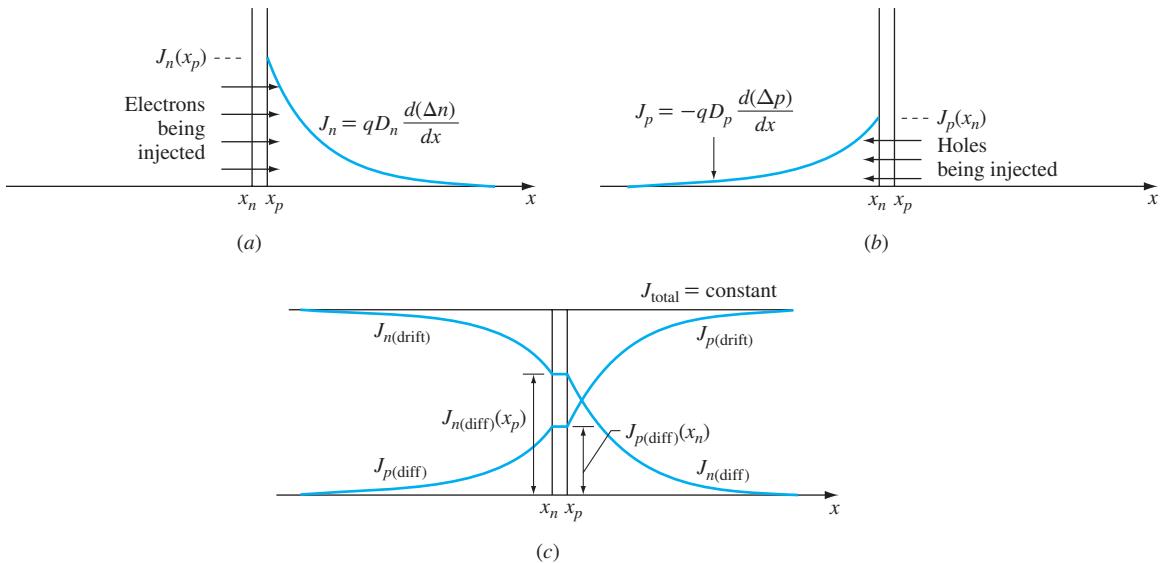


Figure 5.20 (a) The electrons injected into the p region under forward bias must cross the plane $x = x_n$. (b) Similarly, the holes injected into the n region must also cross the plane $x = x_p$. (c) The total current in the junction (neglecting recombination current within the depletion region) is the sum of the two. Outside the junction, the minority diffusion currents decrease and the majority carrier drift current increases to keep the total current constant.

Note that at x_n and x_p the current is entirely due to diffusion, and

$$J_0 = q \left(\frac{D_n n_{p0}}{L_n} + \frac{D_p p_{n0}}{L_p} \right) \quad (5.78)$$

From the relations $L_n = \sqrt{D_n \tau_n}$, $L_p = \sqrt{D_p \tau_p}$, $n_{p0} = n_i^2/N'_A$, and $p_{n0} = n_i^2/N'_D$, Equation (5.78) can be expressed

$$J_0 = q n_i^2 \left(\sqrt{\frac{D_n}{\tau_n}} \cdot \frac{1}{N'_A} + \sqrt{\frac{D_p}{\tau_p}} \cdot \frac{1}{N'_D} \right) \quad (5.79)$$

We can see that J_0 is proportional to n_i^2 .

Equation (5.77) is the familiar diode equation from circuits courses. Now we see that the exponential nature of the diode current is due to lowering the potential barriers.

The ratio of electron to hole current at x_p (or x_n) is

$$\frac{J_n(x_p)}{J_p(x_n)} = \frac{D_n n_{p0}}{L_n} \frac{L_p}{D_p p_{n0}} = \frac{D_n L_p N'_D}{D_p L_n N'_A} \quad (5.80)$$

We can see that this ratio depends on N'_D/N'_A . This result is used in the analysis of bipolar junction transistors.

From Equation (5.77), for forward voltage ($V_a \gg kT/q$), the diode current can be expressed

$$I = I_0 e^{qV_a/kT} \quad (5.81)$$

or

$$V_a = \frac{kT}{q} \ln \frac{I}{I_0} = \frac{kT}{q} (\ln I - \ln I_0)$$

For a decade change in current, $\ln(I) = \ln(10) = 2.3$. Thus, for $\Delta V_a = 2.3kT/q = 60$ mV, the current I changes by a factor of 10 at room temperature.

Diffusion Current: Reverse Bias The energy band diagram under reverse bias is shown in Figure 5.21. Here the diffusion current results from minority carriers diffusing to the transition region and being swept across by the junction field. At the edge of the depletion region, the minority carrier concentration is essentially zero, and the excess carrier concentration is $\Delta n_p(x_p) = n_p(x_p) - n_{p0} = 0 - n_{p0} = -n_{p0}$. On the n side the minority hole concentration is also zero, and $\Delta p_n(x_n) = p_n(x_n) - p_{n0} = -p_{n0}$. We say that minority carriers are *extracted* (as opposed to injected). Since there are few minority carriers near the junction

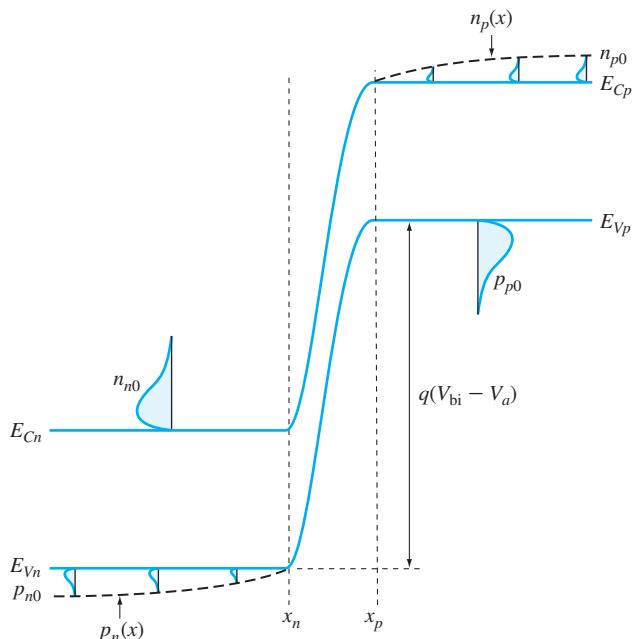


Figure 5.21 Reverse-biased prototype junction. The minority carriers generated within the quasi-neutral regions diffuse to the junction and contribute to the reverse current.

under reverse bias, the minority carriers diffuse toward the junction. In effect, all of the minority carriers generated within a diffusion length of the transition region contribute to current. There are not many of them, so the resulting current is small, as indicated earlier.

Under reverse bias, Equations (5.76) and (5.77) are still valid. However, for $V_a < -3kT$, the exponent becomes large and negative, and the exponential term approaches zero. Therefore, to good approximation the reverse diffusion current density is $-J_0$.

Diffusion Current under Reverse Bias: Short-Base Diode Equation (5.78) for J_0 was derived on the assumption that the thickness on each side of the junction was much larger than a minority current diffusion length. In other words, the neutral region was long enough that the excess minority carrier concentrations could decay by recombination. Often, however, one side has length much less than a diffusion length. This structure is called the *short-base diode*, because this kind of junction occurs in the base region of bipolar transistors. We will prepare for the discussion of that topic now by considering the pn junction of Figure 5.22. Here the thickness W_B of the “base,” in this case the p region thickness excluding the transition region, is much less than L_n , the diffusion length for electrons. In this diode, the n side is long and behaves normally as discussed in the previous section. The p side, however, is different.

Under forward bias, the electron carrier injection process is the same. There is negligible recombination in the p region, however; because it is so thin the electrons are not in it long enough to recombine naturally. There is an assumed contact (not shown) on the far end of the region, and that contact supplies the equilibrium electron and hole concentrations, n_{p0} and p_{p0} respectively. That is, instead of decaying exponentially, the carrier concentration is forced to the

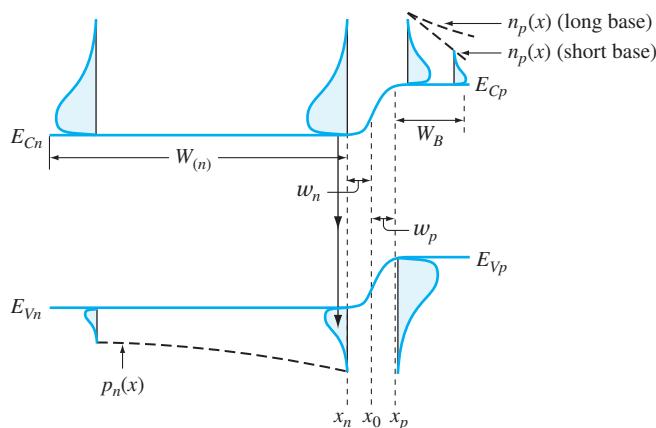


Figure 5.22 Illustration of the diffusion current in a pn junction in which the quasi-neutral region W_B is much shorter than an electron diffusion length.

equilibrium concentration in a distance W_B , the distance between the transition region edge and the end of the quasi-neutral region. Since negligible recombination occurs, the change is very nearly linear. The expression for $d\Delta n/dx$ in Equation (5.71) becomes

$$\frac{d\Delta n}{dx} = \frac{\Delta n_p(x_p)}{W_B} \quad (5.82)$$

and the diffusion current on the p side of the transition region becomes, by Equation (5.73)

$$J_n(x_p) = q \frac{D_n}{W_B} n_{p0} (e^{qV_a/kT} - 1) \quad (5.83)$$

On the n side the diffusion is the same as for the long-base diode, so Equation (5.78) becomes

$$J_0 = q \left(\frac{D_n n_{p0}}{W_B} + \frac{D_p p_{n0}}{L_p} \right) \quad \text{short p-side diode} \quad (5.84)$$

If both sides are short, e.g., on the p side $W_{B(p)} \ll L_n$ and on the n side $W_{(n)} \ll L_p$, as is often the case for the emitter-base junction of a transistor,

$$J_0 = q \left(\frac{D_n n_{p0}}{W_{B(p)}} + \frac{D_p p_{n0}}{W_{(n)}} \right) \quad \text{both sides short} \quad (5.85)$$

and the ratio of electron to hole current at $x = x_p$ or x_n is

$$\frac{J_n}{J_p} = \frac{D_n}{D_p} \cdot \frac{W_{(n)}}{W_{B(p)}} \cdot \frac{N'_D}{N'_A} \quad \text{both sides short} \quad (5.86)$$

We emphasize that the preceding results are for the case of nondegenerate semiconductors on both sides of the homojunction.

Generation and Recombination Current in pn Homojunctions In the analysis of diffusion current in the previous section, we assumed that all carrier generation and recombination occurred in the quasi-neutral regions, and we ignored generation and recombination in the transition region. However, carrier generation and recombination do occur in the transition region and also contribute to current [generation-recombination (G-R) current].

In many semiconductors (including Si), carrier generation and recombination are primarily via interband (trap) states with energy near the intrinsic level, $E_T \approx E_i$, where E_T is the trap energy. Let us consider such a case. For simplicity we assume $\tau_n = \tau_p = \tau_0$ and $E_T = E_i$. Under these conditions, the net recombination rate is given by

$$R - G = \frac{np - n_i^2}{\tau_0(n + n_i + p + n_i)} \quad (5.87)$$

At equilibrium, $np = n_i^2$ everywhere and the net recombination reduces to zero as expected.

The net recombination rate depends on the numbers of available electrons and holes as indicated in Equation (5.87). At position x within the transition region, the carrier concentrations are given by

$$\begin{aligned} n(x) &= n_{n0} e^{-[E_C(x) - E_C(x_n)]/kT} \\ p(x) &= p_{p0} e^{[E_V(x) - E_V(x_p)]/kT} \end{aligned} \quad (5.88)$$

or the electron concentration decreases rapidly with increasing x and E_C . Likewise, $p(x)$ increases rapidly as x and thus E_V increase. We will use these results in the following sections to examine recombination and generation under reverse bias and under forward bias.

Generation and Recombination: Reverse Bias To consider recombination and generation under reverse bias, we refer to Figure 5.23 in which the energy band diagram is shown. Recall that under reverse bias the transition region widens. When the width is large, there can be significant opportunities for both generation and recombination. Under reverse bias, however, n and p are both small in most of the transition region. To simplify the mathematics, the depletion approximation is used, which assumes n and p can be neglected in the transition region for $x_n \leq x \leq x_p$. In this case the recombination can be ignored, since there

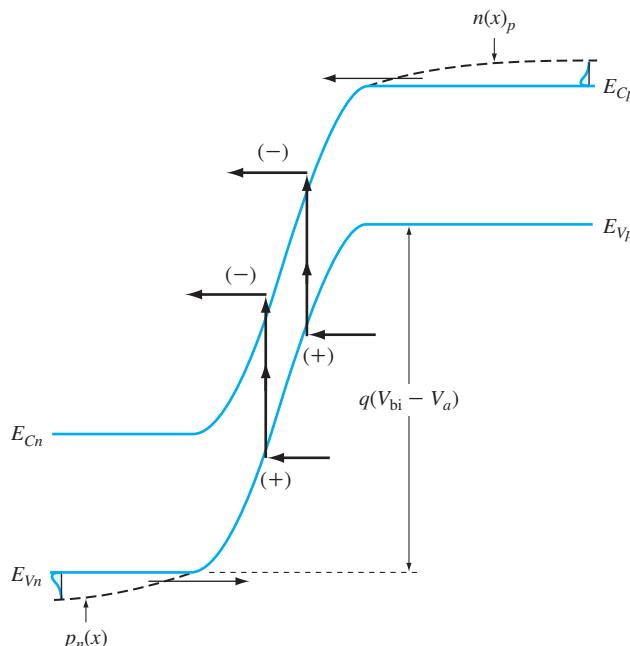


Figure 5.23 Illustration of the diffusion current and generation current in a reverse-biased pn junction. The generation current is normally much larger than the diffusion current.

are few electrons to recombine and few holes with which to recombine. Then from Equation (5.87), we have

$$R = 0 \quad G = \frac{n_i}{2\tau_0} \quad (5.89)$$

This implies that the generation rate is constant in position within the transition region. The generation current density J_G is

$$J_G = -qGw = -\frac{qn_i w}{2\tau_0} \quad (5.90)$$

However, since the depletion width w for a step junction is given by

$$w = \left[\frac{2e(N'_D + N'_A)(V_{bi} - V_a)}{qN'_D N'_A} \right]^{1/2} \quad (5.91)$$

substituting Equation (5.91) into (5.90) results in

$$J_G = -\frac{n_i}{\tau_0} \left[\frac{qe(N'_D + N'_A)(V_{bi} - V_a)}{2N'_D N'_A} \right]^{1/2} \text{ reverse bias} \quad (5.92)$$

EXAMPLE 5.2

Estimate the value of generation current relative to the diffusion current for a typical Si pn junction under reverse-bias conditions.

Solution

Let us take for a prototype junction in silicon:

$$N'_A = 10^{17} \text{ cm}^{-3}$$

$$N'_D = 10^{17} \text{ cm}^{-3}$$

$$\tau_n \approx \tau_p = \tau_0 \approx 6 \times 10^{-6} \text{ s}$$

$$V = (V_{bi} - V_a) = 5 \text{ V}$$

where τ_0 is taken as the average minority carrier lifetime from Figure 3.21.

From the graph of Figure 3.11, at doping concentrations of $N'_A = N'_D = 10^{17} \text{ cm}^{-3}$, the diffusion constants for minority carriers are

$$D_n = 20 \text{ cm}^2/\text{s}$$

$$D_p = 11 \text{ cm}^2/\text{s}$$

The minority carrier diffusion lengths are found from Figure 3.23 to be $L_n = 73 \mu\text{m}$ and $L_p = 102 \mu\text{m}$. The minority carrier concentrations are $n_{p0} = p_{n0} = n_i^2/10^{17} = 1.16 \times 10^3 \text{ cm}^{-3}$, and the built-in voltage is, from Equation (5.13)

$$V_{bi} = \frac{kT}{q} \ln \frac{N'_D N'_A}{n_i^2} = 0.026 \text{ V} \ln \left[\frac{10^{17} \times 10^{17}}{(1.08 \times 10^{10})^2} \right] = 0.83 \text{ V}$$

We are given the junction voltage, so we can find the applied voltage:

$$V_j = 5 \text{ V} = V_{bi} - V_a = 0.83 - V_a \quad \text{or} \quad V_a = -4.17 \text{ V}$$

For the diffusion current, we find the leakage current from Equation (5.78),

$$\begin{aligned} J_0 &= q \left(\frac{D_n n_{p0}}{L_n} + \frac{D_p p_{n0}}{L_p} \right) \\ &= 1.6 \times 10^{-19} \text{ C} \left[\frac{(20 \text{ cm}^2/\text{s})(1.16 \times 10^3 \text{ cm}^{-3})}{73 \times 10^{-4} \text{ cm}} + \frac{(11 \text{ cm}^2/\text{s})(1.16 \times 10^3 \text{ cm}^{-3})}{102 \times 10^{-4} \text{ cm}} \right] \\ &= 7.1 \times 10^{-13} \text{ A/cm}^2 \end{aligned}$$

Inserting this into Equation (5.77), we have for the diffusion current

$$\begin{aligned} J_{\text{diff}} &= J_0 \left(e^{qV_a/kT} - 1 \right) \\ &= 7.1 \times 10^{-13} \text{ A/cm}^2 (e^{[(1.6 \times 10^{-19} \text{ C})(-4.17 \text{ V})]/[(0.026)(1.6 \times 10^{-19}) \text{ J/eV}]} - 1) \approx -J_0 \\ &= -7.1 \times 10^{-13} \text{ A/cm}^2 = -7.1 \times 10^{-21} \text{ A}/\mu\text{m}^2 \end{aligned}$$

For the generation current, we calculate the junction width from Equation (5.91) to be $w = 0.36 \mu\text{m}$. Then, using this value in Equation (5.90), we find

$$\begin{aligned} J_G &= -\frac{qn_i w}{2\tau_0} = -\frac{(1.6 \times 10^{-19} \text{ C})(1.08 \times 10^{10} \text{ cm}^{-3})(0.36 \times 10^{-4} \text{ cm})}{2(6 \times 10^{-6} \text{ s})} \\ &= -5.2 \times 10^{-9} \text{ A/cm}^2 = 5.2 \times 10^{-17} \text{ A}/\mu\text{m}^2 \end{aligned}$$

or

$$\frac{J_G}{J_{\text{diff}}} \approx \frac{5.2 \times 10^{-17}}{7.1 \times 10^{-21}} = 7.3 \times 10^3$$

The generation current under reverse bias is appreciably larger than the diffusion current. Therefore, under reverse bias the diffusion current can normally be neglected.

Generation and Recombination: Forward Bias Next, let us consider the generation and recombination current under forward bias. The energy band diagram for forward bias is shown in Figure 5.24. In the transition region, the electrons are in thermal equilibrium with the n side of the junction and the holes are in thermal equilibrium with the p side. Therefore, from Equation (5.46), one can show that in the transition region,

$$\begin{aligned} n(x) &= n_{n0} e^{-[E_C(x) - E_{Cn}]/kT} \\ p(x) &= p_{p0} e^{-[E_{Vp} - E_{V(x)}]/kT} \end{aligned} \tag{5.93}$$

where x is position within the transition region.

Therefore, within the transition region the np product is, from Equations (5.93) and (5.13),

$$np = n_i^2 e^{qV_a/kT} \tag{5.94}$$

For reasonable forward bias levels, such that $V_a \geq 3kT/q$, the recombination current is large compared with the generation current, because there are now

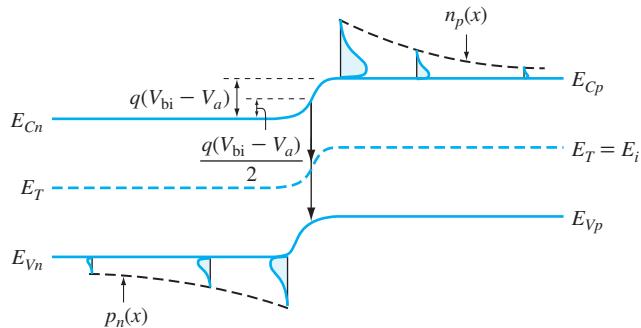


Figure 5.24 Diffusion current and recombination current in a forward-biased pn junction. The barrier for diffusion current is $q(V_{bi} - V_a)$. The barrier for recombination current is about half that value.

many electrons and many holes in the same region. Since from Equation (5.94) the np product is much greater than n_i^2 under forward bias, Equation (5.87) then becomes

$$R = \frac{np}{\tau_0(n+p)} = \frac{n_i^2 e^{qV_a/kT}}{\tau_0(n+p)} \quad (5.95)$$

Where in the junction is the recombination rate the greatest? The maximum recombination rate is obtained by setting $dR/dn = 0$ [or more conveniently, $d(1/R)/dn = 0$, since n is in the denominator of Equation (5.95)]. Using $p = (n_i^2/n)e^{qV_a/kT}$ from Equation (5.94) gives the maximum R (or the minimum $1/R$) which is found at the value of x where

$$n = p = n_i e^{qV_a/2kT} \quad (5.96)$$

or

$$R_{\max} = \frac{n_i e^{qV_a/2kT}}{2 \tau_0} \quad (5.97)$$

and decreases rapidly on either side.

Since most of the recombination current occurs near where $R \approx R_{\max}$, the barrier for electrons and holes for recombination current is $(V_{bi} - V_a)/2$. This is half the barrier for diffusion current of $(V_{bi} - V_a)$.

The G-R current density is often approximated as

$$J_{GR} = J_{GR0} (e^{qV_a/2kT} - 1) \quad (5.98)$$

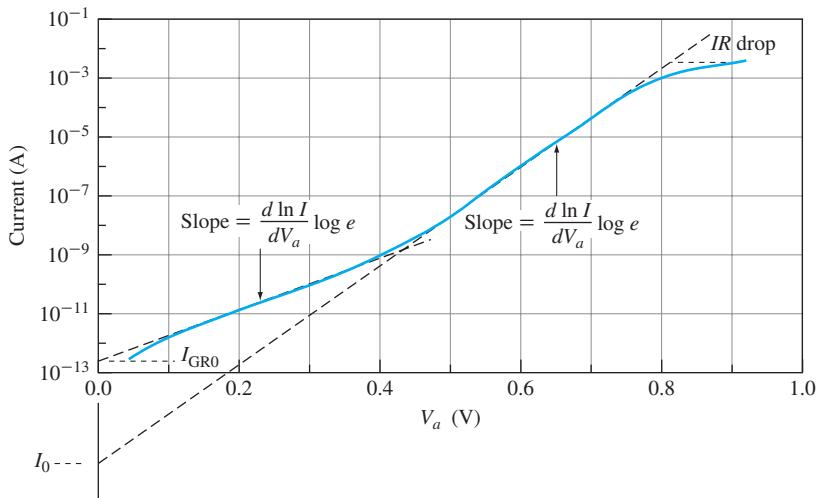


Figure 5.25 The $I-V_a$ characteristic for a forward-biased junction showing the recombination current and diffusion current. At high currents the plot deviates from a straight line by the IR drop in the bulk.

where the term J_{GR0} is slightly voltage dependent.¹⁰ Comparing Equation (5.92) for generation-recombination under reverse bias with (5.79) for diffusion under reverse bias, we can conclude that J_{GR} is proportional to n_i while J_{diff} is proportional to n_i^2 .

The total current density is then

$$J = J_{GR} + J_{diff} = J_{GR0}(e^{qV_a/2kT} - 1) + J_0(e^{qV_a/kT} - 1) \quad (5.99)$$

where usually $J_{GR0} \gg J_0$ as expected from the analysis for reverse bias.

For forward bias with $V_a > 3kT/q$,

$$J = J_{GR0}e^{qV_a/2kT} + J_0e^{qV_a/kT} \quad (5.100)$$

At small V_a , recombination current predominates (because $J_{GR0} \gg J_0$), but at larger V_a , diffusion current predominates. This is indicated in Figure 5.25¹¹ for a junction area of $100 \mu\text{m}^2$. Values for J_{GR0} and J_0 can be extrapolated from the straight-line regions as indicated. At high currents, the line deviates from

¹⁰Its value is obtained by integrating over the width of the junction, and the junction width changes with applied voltage.

¹¹Because the current is plotted on a common log scale, the slope in Figure 5.25 is $d \log I/dV_a = (d \ln I/dV_a) \log e$.

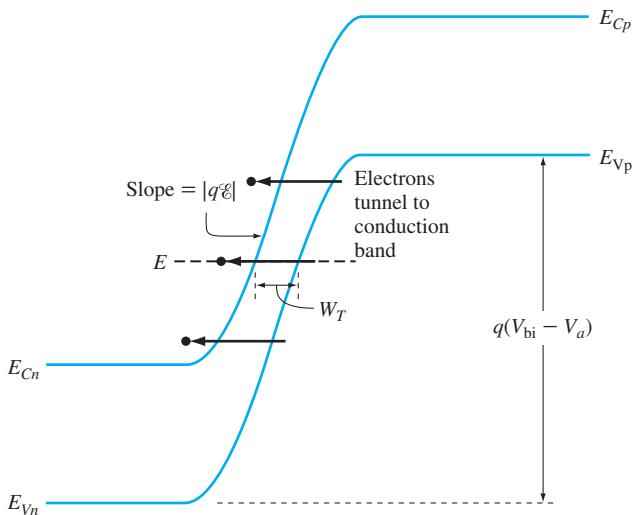


Figure 5.26 Reverse-biased pn junction illustrating the one-step tunneling process.

linearity because part of the applied voltage (IR) is dropped across the series resistance of the material. Over a range of current, J is often approximated as

$$J = J_s (e^{qV_d/nkT} - 1) \quad (5.101)$$

where J_s is a function of J_{GR0} and J_0 , and the *diode quality factor* n (sometimes called the ideality factor) is intermediate between 1 and 2.¹²

Reverse-Bias Tunneling Two other current mechanisms must be considered under reverse bias. These are *tunneling* and carrier *multiplication*. We consider the reverse-biased junction of Figure 5.26. Tunneling can occur when an electron at a given energy (for example energy E in the figure) sees an empty state *at the same energy* on the other side of the potential barrier. In this case, the barrier is the forbidden gap, and the electrons in the valence band are opposite empty states in the conduction band under reverse bias. This is termed a *one-step tunneling process* since the electron goes directly from the valence band to the conduction band without passing through any intermediate states.

As indicated in the Supplement to Part 1, an electron has a finite probability of penetrating the classically forbidden band gap. This process is called *tunneling* or sometimes *Zener tunneling* after the person who predicted the effect.

¹²The diode quality factor of 2 for recombination current is a result of the assumption that the trap level is at the intrinsic level ($E_T = E_i$), and that $\tau_n = \tau_p = \tau_0$. If these conditions are not met, the diode quality factor for recombination current will be less than 2.

The tunneling probability T for an electron at energy E normally incident to the forbidden region is

$$T = e^{-2/\alpha} \quad (5.102)$$

where

$$\alpha = \left[\frac{2m^*}{\hbar^2} (E_P^*(x) - E) \right]^{1/2} \quad (5.103)$$

and the integration is across the forbidden region along the tunneling path. Since the electron effective mass is different in the two bands, m^* in this case represents some average (tunneling) effective mass.

The quantity E_P^* in Equation (5.103) is the “effective potential energy” for the electron. Recall that the potential energy for an electron in the valence band is E_V and the potential energy for an electron in the conduction band is E_C . Both of these are varying with position in this case, as can be seen with the aid of Figure 5.27a. An electron in the forbidden region (i.e., during tunneling) of energy E at position x is affected by two potential energies, $E_C(x)$ and $E_V(x)$. The effective potential energy $E_P^*(x)$ seen by the tunneling electron is analogous to the effective resistance of two parallel resistances:

$$(E_P^*(x) - E) = \frac{(E_C(x) - E)(E - E_V(x))}{(E_C(x) - E) + (E - E_V(x))} = \frac{(E_C(x) - E)(E - E_V(x))}{E_g}$$

The normalized effective potential energy for tunneling as a function of position is indicated in Figure 5.27b. At $x = W_T$ and $x = 0$, $(E_C(0) - E)$ and $[E - E_V(W_T)]$ are equal to zero and thus $(E_P^* - E)$ is equal to zero. Therefore an electron at $x = 0$ (the valence band edge) sees no immediate barrier and enters the forbidden region. As it penetrates deeper into the forbidden region, because $(E_P^* - E)$ increases, the probability of the electron being turned back (reflected) increases. If it tunnels as far as $x = W_T/2$, $(E_P^* - E)$ reaches its maximum value of $E_g/4$. If the electron makes it all the way to the other side, $E_P^* - E$ becomes zero again and the electron enters the conduction band.

After some tedious algebra the tunneling probability becomes

$$T = e^{-\pi W_T \sqrt{m^* E_g} / 2^{3/2} \hbar} \quad (5.104)$$

where W_T is the tunneling distance. Tunneling distance depends on the band gap and the slope of the band edges, as shown in Figure 5.27c, where $W_T = E_g/q\mathcal{E}$. The slope increases with increasing applied voltage, reducing W_T and thus increasing the tunneling probability. Since W_T depends on applied voltage and doping level, from Equation (5.104) we can see that the tunneling probability and thus the tunnel current depend strongly on effective mass, band gap, doping level, and applied voltage.

The $I-V_a$ characteristics resulting from tunneling in a reverse-biased junction are shown in Figure 5.28, along with that for multiplication and avalanche, discussed next.

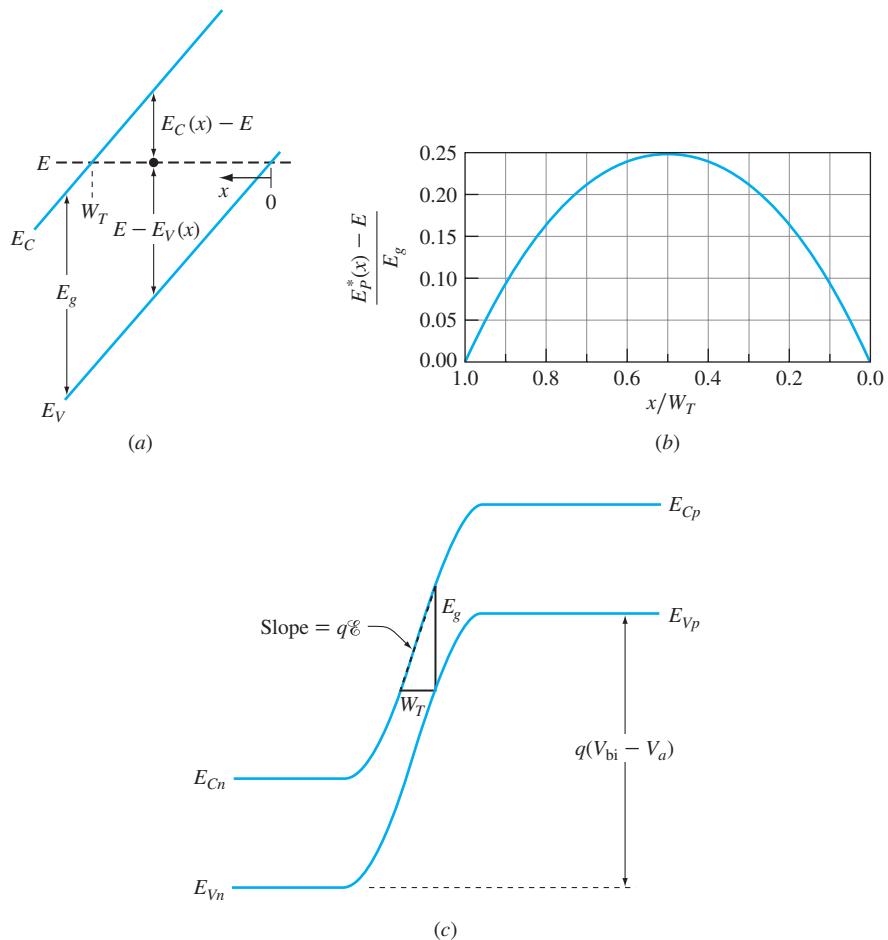


Figure 5.27 (a) The effective potential energy during tunneling. (b) The difference between the total energy and the effective energy, normalized to the band gap, as a function of normalized position of the tunneling electron. At either end of the trip, the electron's energy is entirely potential energy. (c) The geometry used to find the tunneling distance W_T .

Reverse-Bias Carrier Multiplication and Avalanche The last current mechanism to be discussed is *multiplication and avalanche current*, also a reverse-bias phenomenon. Consider an electron that is thermally excited to the conduction band on the p side of a reverse-biased junction as shown in Figure 5.29. Let the thermal generation event be called process 1. We recall that electrons (and holes) travel at constant energy between collisions. The thermally generated electron diffuses to the edge of the transition region, where it is accelerated toward the n side until it makes a collision (process 2). It then loses energy (process 3). It can give this

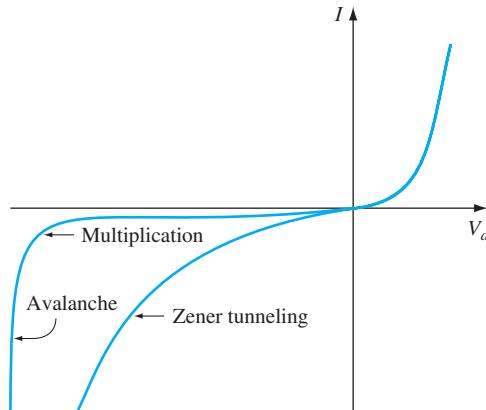


Figure 5.28 The $I-V_a$ characteristics illustrating Zener tunneling, carrier multiplication, and avalanche. The reverse and forward characteristics are not plotted on the same scale.

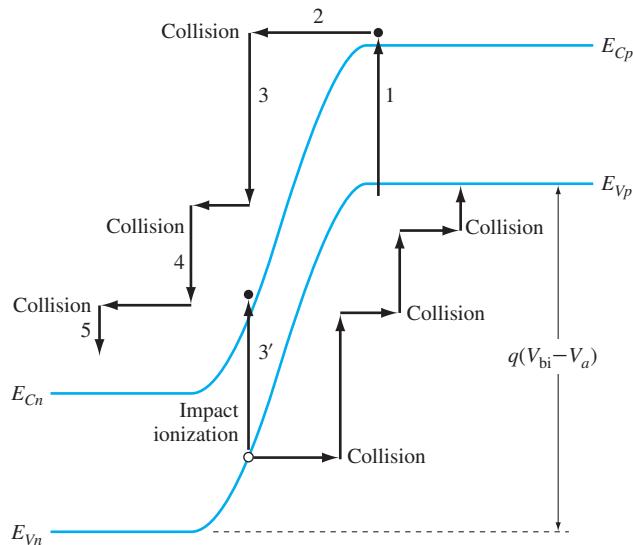


Figure 5.29 The carrier multiplication process under reverse bias. An electron is generated thermally (1), and is accelerated to the left by the field (2). Its kinetic energy increases by more than the band gap, so when the electron collides with another electron in the valence band (3), the second electron can be excited up to the conduction band (3'), a process called *impact ionization*. The ionization results in an electron-hole pair, so one original electron produced two current carriers, effectively multiplying the current.

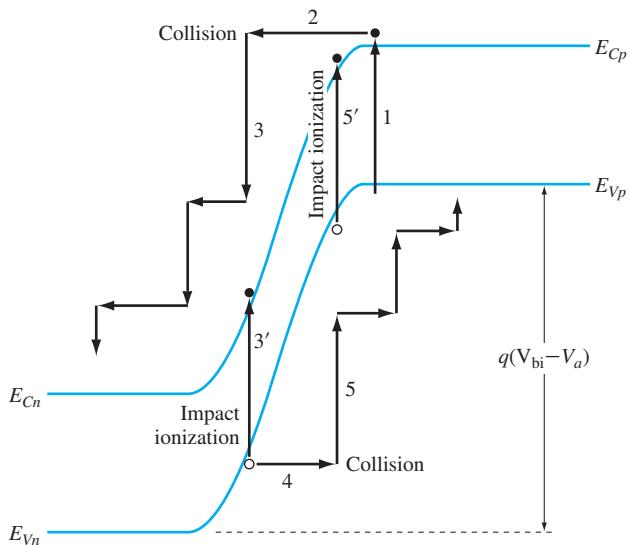


Figure 5.30 Avalanche occurs when the field is large enough to cause the newly excited carriers to in turn create more carriers.

energy to an electron in the valence band, exciting it to the conduction band (3'). This process is called *impact ionization*. There are now three carriers contributing to current—two electrons and one hole. These are swept to their respective sides of the junction. Although one electron started to cross the junction, two carriers finished,¹³ meaning the original current was multiplied by a factor of 2. This multiplication process relies on an electron gaining kinetic energy between collisions (in one mean free path), by an amount greater than the band gap, to generate another electron. Collisions 4 and 5 in Figure 5.29 do not meet this requirement and thus do not cause more multiplication. The multiplication factor depends strongly on the electric field in the junction.

In carrier multiplication, it is also possible for holes to gain enough kinetic energy between collisions to create hole-electron pairs. This is shown in Figure 5.30 as process 5'. If no more carriers were generated, the multiplication factor would be 3. However, above a critical field this generated electron can generate another hole-electron pair. The hole thus generated can do likewise, etc., and the multiplication factor can become infinite. This process is called *avalanche*.

Let us examine the multiplication. Let P be the probability that either a hole or an electron creates an electron-hole pair while it traverses the junction.¹⁴ Let n_{in} be the number of electrons entering the transition region from the p side. Then there will be Pn_{in} ionizing collisions giving $n_{in}(1 + P)$ electrons reaching the n

¹³It might appear that the current is multiplied by a factor of 3. Note, however, that the impact-generated electron and hole each traverse only a portion of the junction, contributing to one carrier crossing the junction.

¹⁴For simplicity, we assume that the probabilities for holes and electrons are equal and the field in the transition region is constant.

side. But Pn_{in} holes are also generated, which generate $P(Pn_{\text{in}}) = P^2n_{\text{in}}$ pairs, etc., or the number of total carriers crossing the junction is

$$n_{\text{in}}(1 + P + P^2 + P^3 + \dots)$$

This can be expressed as

$$\frac{n_{\text{in}}}{(1 - P)}$$

The multiplication factor then, is

$$M = \frac{1}{(1 - P)} \quad (5.105)$$

For $P = 1$, $M = \infty$ and avalanche occurs. The I - V_a characteristic for multiplication and avalanche is *sharp* or *hard*, as was shown in Figure 5.28. If a diode reverse current exceeds a certain value (set by convention, e.g., $10 \mu\text{A}$), the diode is said to have *broken down*. This refers to the I - V_a curve turning sharply downward, but it does not imply the device is damaged, since this breakdown is nondestructive. For Si pn junctions with breakdown voltages greater than about 8 V, the breakdown mechanism is primarily avalanche; if it is less than about 6 V, it is Zener tunneling.¹⁵

5.3.4 REVERSE-BIAS BREAKDOWN

As indicated, with increasing reverse bias the current in a diode increases as a result of tunneling as well as multiplication and avalanche.

The breakdown voltage of one-sided step junctions in each of five common semiconductors is shown in Figure 5.31.¹⁶ The breakdown voltage is plotted as a function of the doping on the lightly doped side. It is seen that for a given semiconductor, the breakdown voltage decreases with increasing doping. For a given doping level, the breakdown voltage increases with increasing band-gap energy of the semiconductor.

These observations can be explained with the aid of Figures 5.30 and 5.26 for avalanche breakdown and tunneling breakdown respectively. For avalanche to occur (Figure 5.30) the field in the junction must be large enough that between collisions, a carrier gains sufficient kinetic energy to create hole-electron pairs. With increased doping on the lightly doped side at a given reverse voltage, the junction width decreases and thus the field increases. With increasing band gap, the energy required to create electron-hole pairs increases, and thus the field and reverse voltage required for avalanche breakdown increase.

For tunneling breakdown (Figure 5.26), the electron tunneling probability depends on tunneling distance (determined by the doping and bias) and band gap as indicated in Equation (5.104). Since W_T is inversely proportional to \mathcal{E} , maximum tunneling occurs at the energy associated with \mathcal{E}_{max} .

¹⁵*Historical note:* Diodes that break down by the avalanche process are often referred to as *Zener diodes* because it was originally thought that the breakdown was a result of tunneling.

¹⁶The wide bandgap semiconductors, 4H-SiC and GaN, are included since they are used in power semiconductor devices as discussed in Chapter 12.

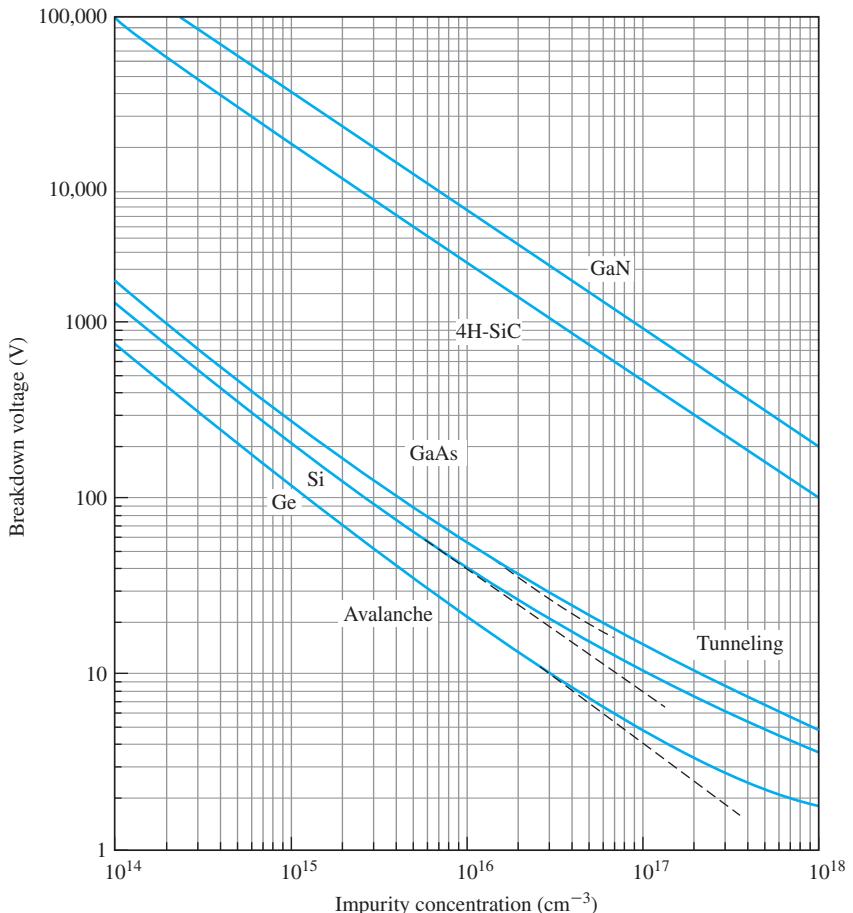


Figure 5.31 Breakdown voltage as a function of impurity concentration for one-sided n^+p or p^+n junctions in Ge, Si, GaAs, 4H-SiC, and GaN. The dashed line indicates the doping level that separates avalanche from tunneling as the dominant breakdown mechanism.

For most one-sided semiconductor diodes of interest, the breakdown due to avalanche can be expressed as

$$V_{br}(AV) = CN^{-\frac{3}{4}} \quad (5.106)$$

where N is the doping level on the lightly doped side and C is a constant for a given semiconductor. For the semiconductors of Figure 5.31, the coefficient C is given in Table 5.1.

Over most of the doping range of Figure 5.31, avalanche breakdown predominates, and the breakdown voltage is given by Equation (5.106). On the log-log plot of Figure 5.31 the curves are straight lines with slopes of negative $\frac{3}{4}$ decade of V_{br} per decade of N . At high doping levels, the plots deviate from straight lines due to an additional tunneling component as seen for Ge, Si, and GaAs.

Table 5.1 Bandgap and doping-dependent avalanche breakdown coefficient, C , as a function of doping concentration on the lightly doped side of a one-sided pn junction for the junctions of Figure 5.31

Semiconductor	Bandgap (eV)	C
Ge	0.67	2.4×10^{13}
Si	1.12	5.3×10^{13}
GaAs	1.43	7.0×10^{13}
4H-SiC	3.26	3.0×10^{15}
GaN	3.44	6.1×10^{15}

EXAMPLE 5.3

Estimate the tunneling distance for appreciable tunnel current. Consider a p⁺n Si junction with $N_D' = 8.0 \times 10^{17} \text{ cm}^{-3} = 8.0 \times 10^{23} \text{ m}^{-3}$.

Solution

Appreciable tunnel current begins to occur around breakdown. We therefore begin by finding the junction voltage at breakdown, which is the difference between the applied breakdown voltage and the built-in voltage.

From Figure 5.13, the built-in voltage is 1.04 V.

From Figure 5.31, the breakdown voltage is 4 V ($V_a = -4 \text{ V}$). The junction voltage is then $V_j = (V_{bi} - V_a) = 5.04 \text{ V}$.

From Equation (5.41), the maximum field is $\mathcal{E}_{\max} = 2 V_j / w$, where from Equation (5.40),

$$w = \left[\frac{2eV_j}{qN_D'} \right]^{1/2} = \left[\frac{2 \times 11.8 \times (8.85 \times 10^{-12} \text{ F/m}) \times (5.04 \text{ V})}{(1.6 \times 10^{-19} \text{ C}) \times (8 \times 10^{23} \text{ m}^{-3})} \right]^{1/2} \\ = 9.1 \times 10^{-8} \text{ m} = 91 \text{ nm}$$

Then

$$\mathcal{E}_{\max} = \frac{2 \times 5.04 \text{ V}}{9.1 \times 10^{-8} \text{ m}} = 1.1 \times 10^8 \text{ V/m}$$

Finally, from Figure 5.27c,

$$W_T = \frac{E_g(\text{eV})}{q\mathcal{E}_{\max}} = \frac{1.12 \text{ eV}}{(1.6 \times 10^{-19} \text{ C})(1.1 \times 10^8 \text{ V/m}) \left(\frac{1 \text{ eV}}{1.6 \times 10^{-19} \text{ V}} \right)} \\ = 1.0 \times 10^{-8} \text{ m} = 10 \text{ nm}$$

We conclude that for Si, an appreciable tunneling current flows for a tunneling distance on the order of 10 nm or less.

5.4 SMALL-SIGNAL IMPEDANCE OF PROTOTYPE HOMOJUNCTIONS

The dc $I-V_a$ characteristics of prototype homojunctions were discussed in Section 5.3. In many applications the small-signal ac response of a diode is important. Consider the circuit of Figure 5.32a, in which a dc voltage V_a and an ac voltage

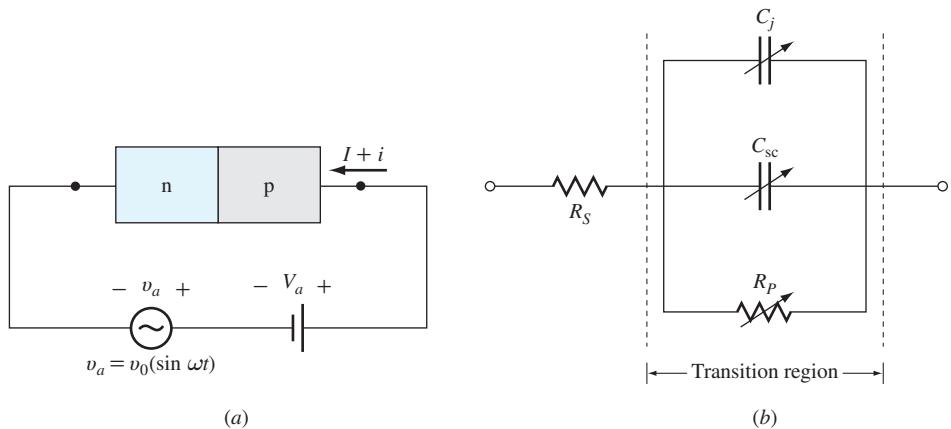


Figure 5.32 (a) A pn junction with dc voltage V_a and ac voltage v_a applied. The resulting dc and ac currents are I and i . (b) Small-signal equivalent circuit for a pn junction indicating the series resistance R_S (of the contacts and quasi-neutral regions, here lumped together into one resistance), the junction resistance R_P , the junction capacitance C_j , and the stored-charge capacitance C_{sc} . The arrows indicate that the parameters change with applied voltage.

v_a are applied to a diode, producing a dc current I and an ac current i , respectively. The small-signal equivalent circuit of the diode is shown in Figure 5.32b. The series resistance R_S represents the contact resistance plus the resistance of the quasi-neutral regions of the diode. The small-signal (differential) resistance of the transition region is designated as R_p , while C_j and C_{sc} represent the junction and stored charge capacitances, respectively, associated with the junction. We will discuss each of these in turn.

5.4.1 JUNCTION (DIFFERENTIAL) RESISTANCE

The small signal (differential) conductance G_p of the transition region is

$$G_p = \frac{dI}{dV_a}$$

This is equal to the slope of the dc $I-V_a$ characteristic at any given point.¹⁷ Figure 5.33 shows how the small-signal conductance can be used to determine the output current for a small, slowly varying input voltage where capacitance current is negligible. The (small-signal) junction resistance is the reciprocal of the slope:

$$R_p = \frac{1}{G_p} = \left[\frac{dI}{dV_a} \right]^{-1} \quad (5.107)$$

¹⁷Here, the voltage across R_S is neglected.

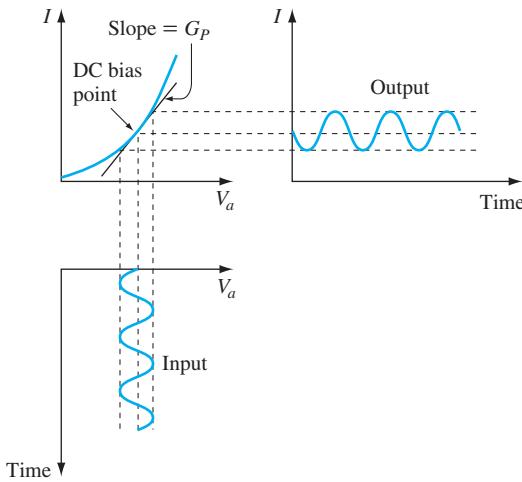


Figure 5.33 The small-signal resistance $R_P = 1/G_P$ is used to find the output current variation for an input voltage variation.

EXAMPLE 5.4

Find the junction resistance of a diode at forward currents of 1 mA and 1 μ A. Assume the ideality factor is unity and $R_S = 0$.

Solution

From Equation (5.107),

$$R_P = \left[\frac{dI}{dV_a} \right]^{-1}$$

The diode current can be expressed

$$I = I_0(e^{qV_a/kT} - 1) = I_0 e^{qV_a/kT} - I_0 \approx I_0 e^{qV_a/kT}$$

where the term $-I_0$ (in the parenthesis) is neglected since it is typically on the order of 10^{-14} A, and so $I \gg I_0$. Then

$$\frac{dI}{dV_a} = \frac{q}{kT} I_0 e^{qV_a/kT} = \frac{qI}{kT}$$

and

$$R_P = \frac{kT}{qI}$$

Since $kT/q = 0.026$ eV,

$$R_P = 26 \Omega \quad (I = 1 \text{ mA})$$

$$R_P = 26 \text{ k}\Omega \quad (I = 1 \mu\text{A})$$

5.4.2 JUNCTION (DIFFERENTIAL) CAPACITANCE

There are two sources of capacitance in pn junctions, the *junction capacitance* and the *stored-charge capacitance*. Both of these limit the speed at which the

diode can respond to changes of the input voltage. We will discuss the junction capacitance first.

Recall from Section 5.3.2 that the charge on either side of the metallurgical junction consists of ionized impurities. The amount of charge on each side of the junction is dependent on junction width, which in turn is a function of applied voltage. Thus, a change in applied voltage produces a change in the number of charges on each side of the junction as indicated in Figure 5.34. If the applied voltage is changed by an amount dV_a , the space charge on one side of the junction changes by an amount dQ and the space charge on the other side of the junction changes by $-dQ$. As the applied voltage changes, the electrons and holes must move out of the junction (if the junction width is increased) or into the junction (if the junction width is decreased) to change the number of non-neutralized ions on each side. The electron current resulting from this movement of electrons and holes flows through the external circuit rather than through the junction. Because current must be continuous, an equal displacement current flows across the junction, or the junction acts as a capacitor.

The small-signal or differential junction capacitance C_j is

$$C_j = \left| \frac{dQ}{dV_a} \right| \quad (5.108)$$

At equilibrium, there is already some charge on either side of the junction—the ionized dopants in the space charge region. When a voltage is applied, this charge is either increased (reverse bias) or decreased (forward bias). Figure 5.34 emphasizes this point by showing that the space charge changes at the edges of the space charge region. Thus the junction resembles a parallel-plate capacitor, and we can write for the junction

$$C_j = \frac{\epsilon A}{w} \quad (5.109)$$

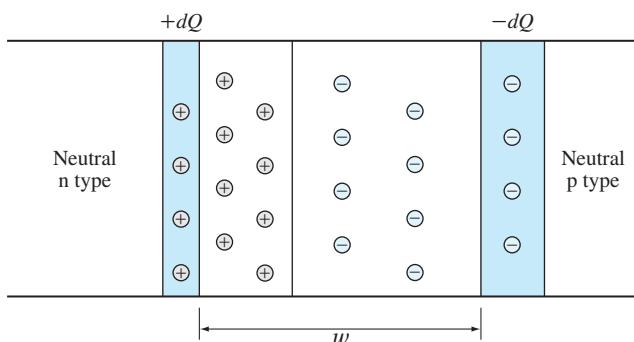


Figure 5.34 Illustration of the junction capacitance C_j . A change in applied voltage V_a produces a change in the number of uncompensated ions on either side of the junction. This produces a change in charge dQ on either side of the junction, making the junction look like a parallel-plate capacitor of width w .

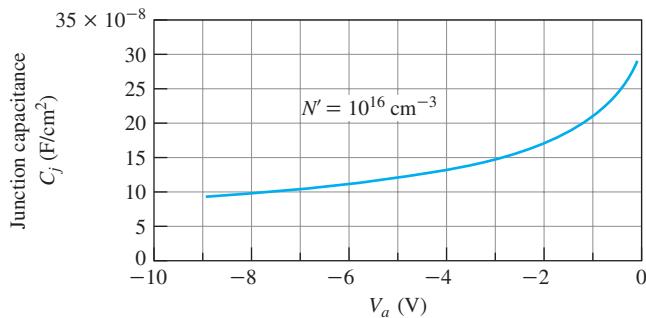


Figure 5.35 Plot of junction capacitance per unit area as a function of reverse bias for a one-sided junction with $N' = 10^{16} \text{ cm}^{-3}$.

where A is junction area and w is the space charge region width. It is understood that C_j is the differential capacitance defined by Equation (5.108). This capacitance is often called *depletion capacitance*. Since w is dependent on the square root of the voltage, C is nonlinear. With substitution of w from Equation (5.37), the junction capacitance becomes, for the step junction,

$$C_j = A \left[\frac{q\epsilon N'_D N'_A}{2(N'_D + N'_A)(V_{bi} - V_a)} \right]^{1/2} = A \left[\frac{q\epsilon N'_D N'_A}{2(N'_D + N'_A)V_j} \right]^{1/2} \quad \text{pn step junction} \quad (5.110)$$

For a one-sided step junction,

$$C_j = A \left[\frac{q\epsilon N'}{2(V_{bi} - V_a)} \right]^{1/2} = A \left[\frac{q\epsilon N'}{2V_j} \right]^{1/2} \quad \text{one-sided step junction} \quad (5.111)$$

where N' is the net doping concentration on the lightly doped side. Junction capacitance per unit junction area is plotted in Figure 5.35 as a function of applied voltage V_a for $N' = 10^{16} \text{ cm}^{-3}$. One can see that C_j increases with V_a . From Equation (5.110) or (5.111), it would appear that for $V_a = V_{bi}$, the junction capacitance would be infinite. For large V_a , however, the current and thus the $I R_S$ drop become large. The junction voltage

$$V_j = V_{bi} - (V_a - I R_s)$$

therefore is always greater than zero.

5.4.3 STORED-CHARGE CAPACITANCE

The other type of capacitance in junctions is called the *stored-charge capacitance*. In the case of the previously discussed junction capacitance, the change in charge with a change in voltage was due to a change in the number of non-neutralized ions on each side of the junction. In stored-charge capacitance, the

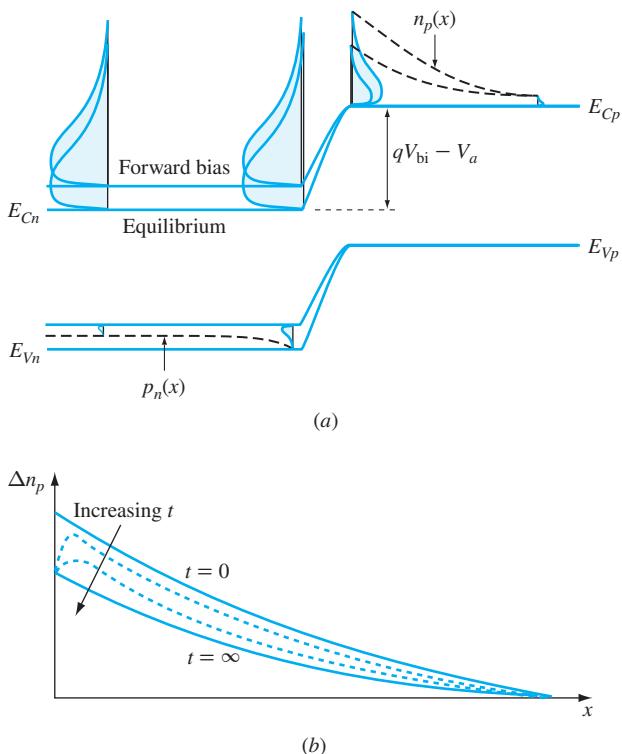


Figure 5.36 Illustration of the stored-charge capacitance in an n^+p junction. (a) As the forward bias changes, the number of injected electrons (minority carriers) changes. In a one-sided junction the hole injection is negligible. (b) An abrupt decrease in the forward voltage ΔV_a causes some of the injected (“stored”) electrons to return to the n side of the junction and contribute to capacitance.

pertinent changing charge on either side of the junction is the change in minority carrier density, as the minority carriers are injected or extracted at the junction edges with changing bias.

We consider the case of an n^+p junction under forward bias, Figure 5.36a. We consider the injection of electrons only, since it is large compared with the hole injection in an n^+p diode. In steady state, assuming the p -type region is much longer than the electron diffusion length (long-base diode), we have from Equation (5.69), which is repeated here,

$$\Delta n_p(x) = \Delta n_p(x_p) e^{-(x-x_p)/L_n} \quad (5.69)$$

In steady state, this distribution of carriers is maintained constant. We say there is a charge “stored” in this distribution, even though the individual stored electrons or holes are changing from moment to moment. The total stored minority carrier (electron) charge in the p region is the integral of the distribution (times the area of the junction):

$$Q_s = -qA \int_{x_p}^{\infty} \Delta n_p(x) dx = qA \Delta n_p(x_p) L_n \quad (5.112)$$

This can be related to the current. We recall that the current across the plane x_p is carried entirely by the diffusion of minority carriers, and that the total current anywhere is equal to the current crossing this plane. The diffusion current is, from Equation (5.69) (again neglecting the small contribution due to holes),

$$I|_{x_p} = qAD_n \frac{d\Delta n_p(x)}{dx} \Big|_{x_p} = -qAD_n \frac{\Delta n_p(x_p)}{L_n} \quad (5.113)$$

Combining Equations (5.113) and (5.112) gives

$$Q_s = \frac{IL_n^2}{D_n} = I\tau_n \quad (5.114)$$

where we have used $L_n = \sqrt{D_n\tau_n}$. This result shows that the steady-state stored minority carrier charge is proportional to current.

If the applied voltage is changed by dV_a , the steady-state stored charge is changed by dQ_s . The rate of change of Q_s with V_a is

$$\frac{dQ_s}{dV_a} = \frac{dI}{dV_a} \tau_n = \frac{\tau_n}{R_P} \quad (5.115)$$

where R_P is the diode differential resistance given by Equation (5.107).

The expression dQ_s/dV_a has the dimensions of capacitance, but it is not the capacitance as seen from the junction terminals. This can be understood with the aid of Figure 5.36b for the case of V_a being abruptly reduced by dV_a . Since from Equation (5.68), $\Delta n(x_p)$ is proportional to $e^{qV_a/kT}$, the excess carrier concentration at the plane $x = x_p$ also abruptly decreases. The rest of the distribution doesn't change instantaneously, however. Thus the peak concentration of excess electrons is no longer at the plane $x = x_p$ but somewhere to the right of it. Then since $J_n = qD_n(dn/dx)$, the stored electrons will diffuse in both directions to regions of lower concentration. Only those that diffuse to the left into the n^+ region will flow in the external circuit and thus contribute to the capacitance. This charge is referred to as the *reclaimable stored charge* Q_{sr} , and the associated stored-charge capacitance C_{sc} is

$$C_{sc} = \frac{dQ_{sr}}{dV_a} = \delta \frac{dQ_s}{dV_a} = \delta I \tau_n \frac{q}{kT} \quad (5.116)$$

where δ is the fraction of the stored charge that is reclaimable.

While this quantity can be calculated, the procedure is tedious. For the case treated here, half of the stored charge is reclaimable, or $\delta = 1/2$. For diodes in which the doping is nonuniform, δ depends on the doping profile.

The stored-charge capacitance in prototype diodes is proportional to diffusion current, which in turn varies exponentially with V_a while the junction capacitance varies as $\sqrt{V_a}$. As a result, for reverse bias and small forward bias, junction capacitance predominates. For large forward bias, stored-charge capacitance predominates.

EXAMPLE 5.5

Compare the junction capacitance and stored-charge capacitance under reverse bias ($V_a = -5$ V) and forward bias ($V_a = +0.75$ V).

Solution

We consider a prototype silicon n⁺p junction with $N'_A = N_A = 10^{17}$ cm⁻³. Let the junction area be 100 μm² and the fraction of reclaimable charge $\delta = 0.5$.

Junction Capacitance

For an n⁺p junction, from Equation (5.111),

$$C_j = A \left[\frac{q\epsilon N_A}{2(V_{bi} - V_a)} \right]$$

The built-in voltage for this junction is, from Figure 5.13,

$$V_{bi} = 0.98 \text{ V}$$

For $V_a = -5$ V,

$$C_j(-5) = \left(100 \mu \text{m}^2 \frac{10^{-8} \text{ cm}^2}{1 \mu \text{m}^2} \right) \times \left[\frac{(1.6 \times 10^{-19} \text{ C})(11.8)(8.85 \times 10^{-14} \text{ F/cm})(10^{17} \text{ cm}^{-3})}{2(0.98 + 5) \text{ V}} \right]^{1/2}$$

$$C_j(-5) = 0.053 \text{ pF}$$

Similarly, for $V_a = 0.75$ V,

$$C_j(0.75) = 0.27 \text{ pF}$$

Stored-Charge Capacitance

To find the stored-charge capacitance, we will need to find I . This will be the diffusion current from Equation (5.77),¹⁸ and to find that we need $I_0 = AJ_0$. Since $J_n \gg J_p$, holes injected into the n⁺ material can be ignored and

$$I = I_0(e^{qV_a/kT} - 1) = qA \left(\frac{D_n n_{p0}}{L_n} \right) (e^{qV_a/kT} - 1)$$

where D_n and L_n are the minority carrier diffusion constant and diffusion length respectively.

For the p-type material, we find

$$n_{p0} = \frac{n_i^2}{N'_A} = \frac{(1.08 \times 10^{10} \text{ cm}^{-3})^2}{10^{17} \text{ cm}^{-3}} = 1.17 \times 10^3 \text{ cm}^{-3}$$

From Figure 3.11, we look up D_n (the minority carriers on the p side are the electrons) using the doping level on the p side. At $N'_A = 10^{17}$ cm⁻³, we find $D_n = 20 \text{ cm}^2/\text{s}$. From Figure 3.23, $L_n = 70 \mu\text{m}$.

Then I_0 becomes

¹⁸The generation-recombination current does not contribute to stored minority carrier charge.

$$\begin{aligned}
 I_0 &= qA \left(\frac{D_n n_{p0}}{L_n} \right) \\
 &= (1.6 \times 10^{-19} \text{ C})(100 \times 10^{-8} \text{ cm}^2) \left[\frac{(20 \text{ cm}^2/\text{s})(1.17 \times 10^3 \text{ cm}^{-3})}{7 \times 10^{-3} \text{ cm}} \right] \\
 &= 5.3 \times 10^{-19} \text{ A}
 \end{aligned}$$

The diffusion current at $V_a = -5 \text{ V}$ is $I = -I_0 = -5.3 \times 10^{-19} \text{ A}$. The current under a forward bias of $+0.75 \text{ V}$ is $I = I_0(e^{qV_a/kT} - 1) = 5.3 \times 10^{-19} \text{ A}(e^{0.75/0.026} - 1) = 1.8 \mu\text{A}$.

The stored-charge capacitances are:

At $V_a = -5 \text{ V}$:

$$C_{sc}(-5) = \frac{q}{kT} \delta I \tau_n = \left(\frac{1}{0.026 \text{ V}} \right)(0.5)(5.3 \times 10^{-19} \text{ A})(3 \times 10^{-6} \text{ s}) = 3.1 \times 10^{-23} \text{ F} \approx 0$$

where we obtained the minority carrier lifetime from Figure 3.21.

At $V_a = +0.75 \text{ V}$:

$$C_{sc}(0.75) = \frac{q}{kT} \delta I \tau_n = \left(\frac{1}{0.026 \text{ V}} \right)(0.5)(1.8 \times 10^{-6} \text{ A})(3 \times 10^{-6} \text{ s}) = 100 \text{ pF}$$

Let us now compare the two capacitances:

$V_a = -5 \text{ V}$	$C_j = 0.053 \text{ pF}$	$C_{sc} \approx 0$
$V_a = 0.75 \text{ V}$	$C_j = 0.27 \text{ pF}$	$C_{sc} = 100 \text{ pF}$

As expected, the junction capacitance dominates under reverse bias and the stored-charge capacitance dominates under forward bias.

5.5 TRANSIENT EFFECTS

A pn junction is often used as a switch. A voltage or current pulse is applied to change the operating state between forward bias (“on”) and reverse bias (“off”). Since a pn junction has capacitance associated with it, it would be expected that some time would be required to make the transition from **off** to **on** (turn-on time) and from **on** to **off** (turn-off time). These transients are discussed qualitatively with the aid of Figure 5.37. The applied voltage is switched between V_F (forward bias) and V_R (reverse bias). For mathematical simplicity, we assume that both V_F and V_R are much larger in magnitude than V_{bi} . We also assume that $R_1 \gg R_S$ so that the diode series resistance R_S can be neglected (i.e., $R_S = 0$).

Usually the turn-off time is much larger than the turn-on time, so this case is considered first.

5.5.1 TURN-OFF TRANSIENT

For simplicity, we discuss the case of an n^+p junction, so that we can ignore the stored hole charge in the n region. The stored excess electron charge is much larger and dominates in an n^+p junction.

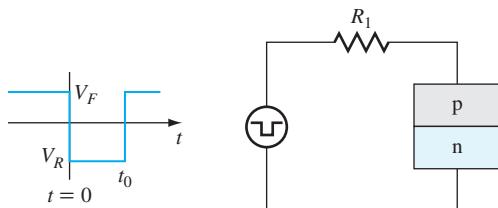


Figure 5.37 The circuit used to illustrate switching turn-off and turn-on transients in a pn junction.

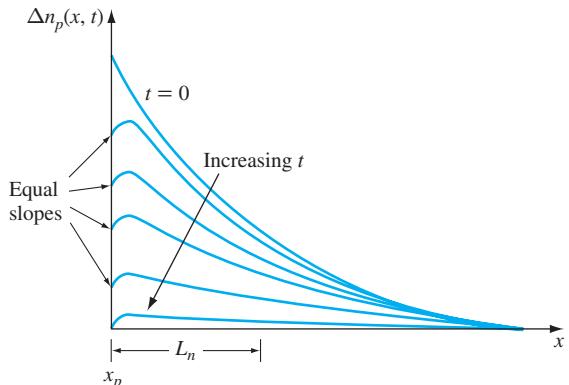


Figure 5.38 Decay of injected electron concentration with time after switching from V_F to V_R .

At $t < 0$, the diode of Figure 5.37 is **on**, with an applied voltage of V_F . The junction is forward biased, and excess electrons are being injected into the quasi-neutral p region. At $t = 0$ the excess electron concentration for $x \geq x_p$ varies with position as given by Equation (5.69):

$$\Delta n_p(x, 0) = \Delta n_p(x_p, 0) e^{-(x-x_p)/L_n} \quad (5.117)$$

The junction voltage V_j (**on**) will be less than V_{bi} (forward bias reduces the barrier). Since $V_F \gg V_{bi} > V_{on}$, the forward current up until $t = 0$ is

$$I_F(0) \approx \frac{V_F}{R_1} \quad (5.118)$$

At $t = 0$ the applied voltage is switched from V_F to V_R . The minority carrier (electron) distribution in the quasi-neutral p region is shown in Figure 5.38 as a function of x and t for $t > 0$. Excess carriers are no longer being injected, and as the carriers diffuse away and recombine, the excess carrier concentration decreases. We also observe that some of the carriers will diffuse back into the junction as discussed earlier.

Notice that the slope of Δn_p at x_p is constant for some time as the carrier concentration decays. We have argued before that the current crossing the plane x_p is the same current that flows through the device, but we can evaluate the current easily at this plane because it is entirely due to diffusion here. Since the slope is constant, the current is constant at

$$I_R \approx \frac{V_R}{R_1} \quad (5.119)$$

until $t = t_s$, where t_s is the *storage time*.¹⁹ The current is plotted as a function of time in Figure 5.39. For $t > t_s$ the slope at $x = x_p$ decreases with t and the magnitude of the current decays toward the steady-state reverse current (≈ 0).

¹⁹Note that since V_R is negative, so is I_R .

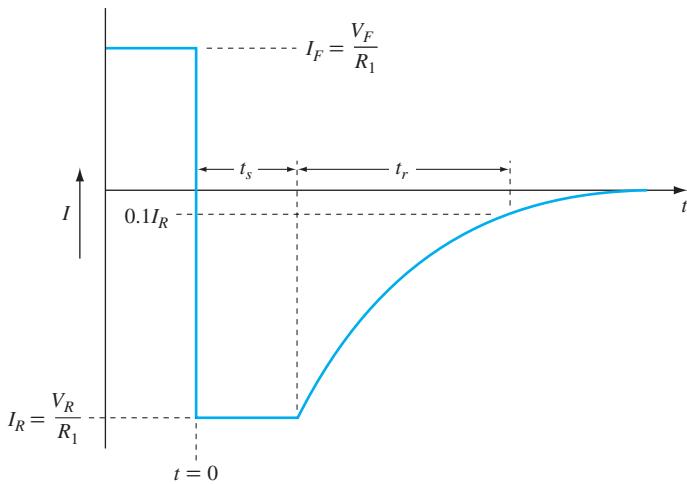


Figure 5.39 Current transient during turn-off of diode, showing the storage time t_s and the rise time t_r .

The storage time can, in principle, be calculated by solving the continuity equation, but this is somewhat involved. An approximate solution²⁰ gives

$$t_s \approx \tau_n \ln \left(1 + \frac{|I_F|}{|I_R|} \right) \quad (5.120)$$

For $V_F = V_R$, $|I_F| \approx |I_R|$, and $t_s = \tau_n \ln 2 = 0.68\tau_n$.

The turn-off time will be shortened if the charge in the p region is reduced, which can be achieved by reducing the electron lifetime or decreasing the $|I_F/I_R|$ ratio. Reducing the carrier lifetime also reduces L_n . This can be accomplished by introducing recombination centers (traps) such as Au or Cu. The traps, however, increase the off current because they provide for more carrier generation in the transition region in the off state. That results in increased power consumption.

The turn-off time can also be reduced by reducing the thickness of the more lightly doped side or by appropriately grading the doping in the more lightly doped side.

5.5.2 TURN-ON TRANSIENT

There is also a transient time associated with switching from **off** back to **on** (V_R to V_F). Some time is required to reach the steady-state forward voltage. This delay time is a result of the time required to discharge the junction capacitance

²⁰The resulting, more accurate relation,

$$t_s \approx \tau_n \left[\operatorname{erf}^{-1} \left(\frac{1}{1 + \frac{|I_R|}{|I_F|}} \right) \right]^2$$

is used in SPICE simulations.

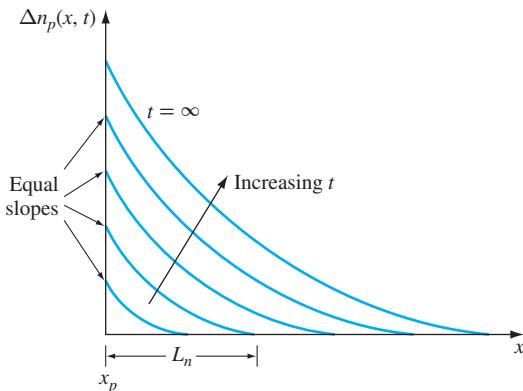


Figure 5.40 The buildup of stored electrons after switching from off to on.

(decrease the junction width) and then to inject carriers to set up the steady-state electron distribution. Switching to V_F at $t = t_0$ results in a current

$$I_F \approx \frac{V_F}{R_1} \quad (5.121)$$

Figure 5.40 shows the variation of stored electron concentration as a function of position and time for the turn-on transient. Since I_F is constant, so are the slopes of the $n_p(x_p, t)$ curves. Steady state is reached when the recombination rate is equal to the rate of electron injection.

The amount of charge required to discharge the junction capacitance is normally much smaller than the steady-state minority carrier stored charge. This, along with the fact that the stored charge is established by diffusion (a slow process), means that the turn-on time is to good approximation equal to that required to set up the minority carrier steady-state distribution.

Figure 5.41 indicates the diode current and voltage waveforms resulting from a rectangular input voltage waveform. Upon switching of the input voltage from V_F to V_R , the diode current switches from $I_F = V_F/R_1$ to $I_R = V_R/R_1$. During the storage time, while the diode current remains constant, the diode voltage remains at V_{on} . In a real diode, the series resistance R_S is finite. Since the diode terminal voltage is that across the junction plus that across the diode series resistance R_S , at $t = 0$, the terminal voltage reduces from $[V_j(\text{on}) + I_F R_S]$ to $[V_j(\text{on}) + I_R R_S]$, where I_F is positive and I_R is negative. This is indicated in Figure 5.41c. After a time $t = t_s$ the diode current reduces in magnitude and the diode voltage decays toward its steady-state value.

Note that the turn-on time is much less than the turn-off time, so the maximum switching frequency is limited by the turn-off time.

In the preceding qualitative discussion of switching transients in a pn junction, approximations were used to illustrate the switching phenomenon resulting from the effect of stored charge. Both forward-bias and reverse-bias voltages

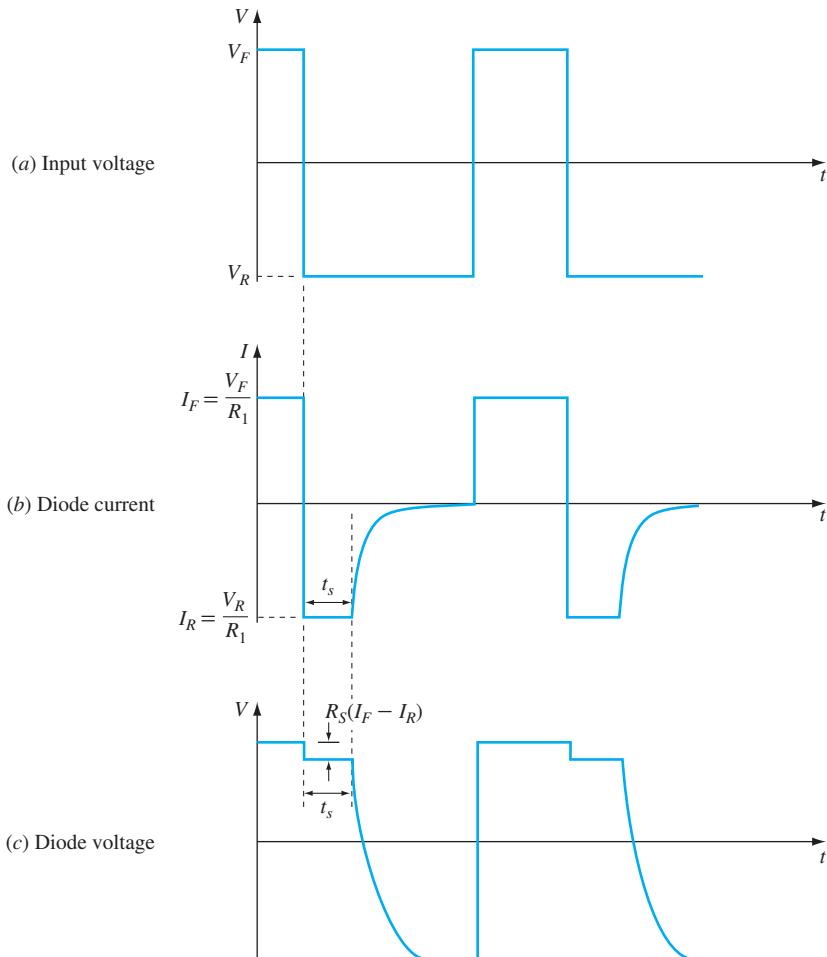


Figure 5.41 The waveform resulting from a diode switched on and off. (a) The input waveform; (b) the diode current; (c) the diode voltage. The turn-off time exceeds the turn-on time significantly.

were assumed large compared with the voltage V_D across the diode, such that V_D could be neglected in determining I_F and I_R . In many circuits it is convenient to have $V_R = 0$ such that the input voltage switches between V_F and zero as shown in the circuit of Figure 5.42a. As indicated, the diode is represented by its equivalent circuit where R_p , C_j and C_{sc} are dependent on voltage and current and the effect of R_S is considered.

It is of interest to determine the diode voltage (V_D) and current (I_D) transients during switching for such a case. This is a complicated problem because of the nonlinearity of R_p , C_j , and C_{sc} . These switching waveforms can be determined, however, with the aid of SPICE (Simulation Program

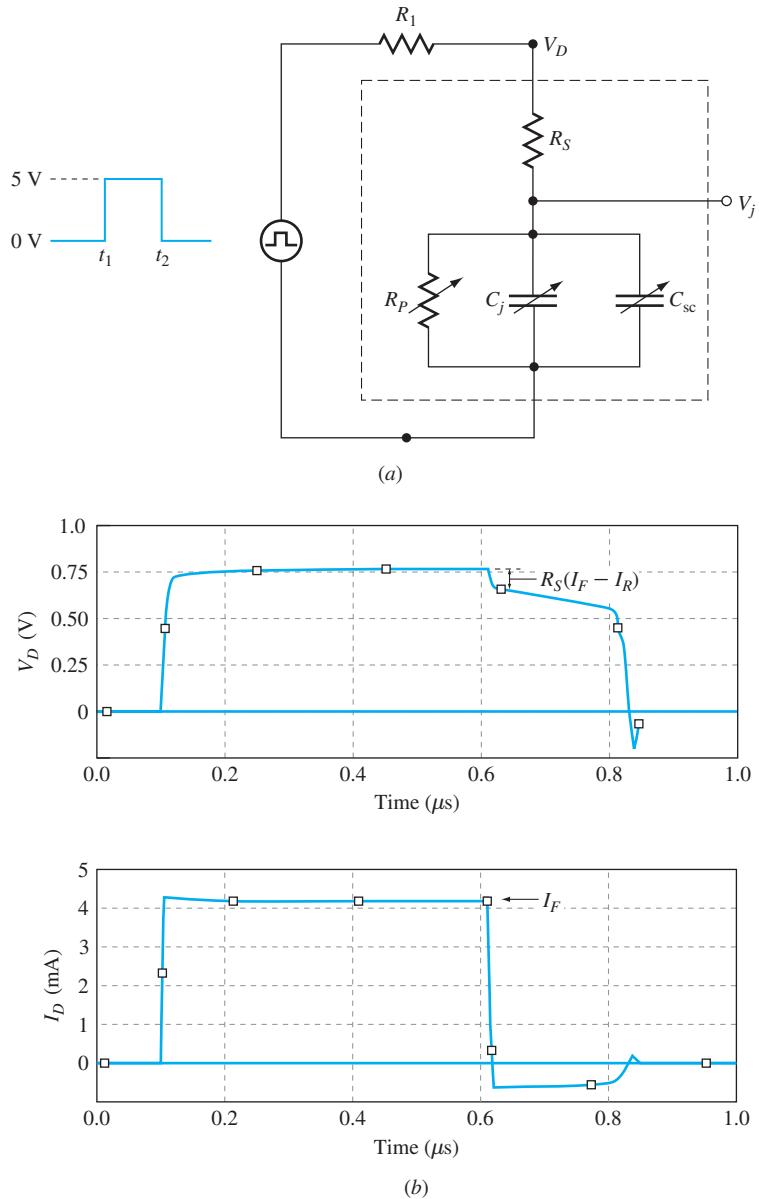


Figure 5.42 (a) The switching circuit model for a pn homojunction diode. (b) The voltage and current response, as determined by SPICE, to an input pulse beginning at $t_1 = 100\text{ ns}$ and ending at $t_2 = 600\text{ ns}$.

with Integrated Circuit Emphasis). This problem is treated in the online module OM6. Plots of V_D and I_D as functions of time as simulated by SPICE are given in Figure 5.42b.

The previous discussion was for a long-base diode. For a short-base n⁺p diode ($W_B \ll L_n$) there is negligible electron recombination in the quasi-neutral p region and $n_p(x)$ can be written

$$\Delta n_p(x) = \Delta n_p(0) \left(1 - \frac{x}{W_B} \right) \quad (5.122)$$

and

$$I_n = qA D_n \frac{d\Delta n}{dx} = -\frac{qA D_n \Delta n(0)}{W_B} \quad (5.123)$$

The total excess electron charge stored in the short base is

$$Q_s = -qA \Delta n(0) \cdot \frac{W_B}{2} = I_n t_T \quad (5.124)$$

where t_T is referred to as the *base-transit time* and is given by

$$t_T = \frac{(W_B)^2}{2 D_n} \quad \text{short-base diode} \quad (5.125)$$

This is the average time it takes for the minority carriers (electrons) to traverse the base.

EXAMPLE 5.6

Compare the amount of minority carrier stored charge in a forward-biased short-base diode with that in a long-base diode.

Solution

From Equations (5.124), (5.125), and (5.114), and using $L_n^2 = D_n \tau_n$,

$$\frac{Q_s(\text{short base})}{Q_s(\text{long base})} = \frac{1}{2} \left(\frac{W_B}{L_n} \right)^2$$

for W_B on the order of 0.1 μm, as in a typical bipolar transistor, and $L_n \approx 31 \mu\text{m}$, we have

$$\frac{Q_s(\text{short base})}{Q_s(\text{long base})} \approx \frac{1}{2} \left(\frac{0.1}{31} \right)^2 \approx 5 \times 10^{-6}$$

It is clear that the stored charge and thus the turn-off recovery time is greatly reduced by the use of a short-base diode.

5.6 EFFECTS OF TEMPERATURE

At a given voltage the current in a diode is quite temperature sensitive. Consider an n⁺p diode. From Equation (5.79),

$$\begin{aligned} J_0 &= \frac{q}{N'_A} \sqrt{\frac{D_n}{\tau_n}} n_i^2 \\ &= \frac{q}{N'_A} \sqrt{\frac{D_n}{\tau_n}} N_C N_V e^{-E_g/kT} \end{aligned} \quad (5.126)$$

Neglecting the temperature dependences of D_n , τ_n , N_C , N_V , and E_g compared with the exponential dependence on $1/T$, the fractional variation of J_0 with temperature can be expressed

$$\frac{dJ_0}{dT} = \frac{\frac{de^{-E_g/kT}}{dT}}{e^{-E_g/kT}} = \left(\frac{E_g}{kT^2} \right) = \frac{E_g}{kT} \cdot \frac{1}{T} \quad (5.127)$$

For Si at room temperature this becomes

$$\frac{1.12}{0.026} \cdot \frac{1}{300} = 0.14$$

or J_0 varies approximately 14 percent per degree Celsius.

Similarly since J_{GR0} varies as n_i rather than n_i^2 , its fractional variation with temperature is about half this value.

An analogous analysis shows that at constant forward current in the typical range of operation (0.1 to 1 mA) the diode forward voltage decreases about 2 mV for an increase of 1°C.

5.7 SUMMARY

In most pn junctions, the doping concentrations are complex functions of position and their electrical characteristics must be calculated numerically. While there are software programs available to do this, their use provides little insight into the physical processes involved. In this chapter, we used a greatly simplified model of a pn junction, the prototype pn homojunction. In this model we assumed that the net doping level on each side is constant with position, or the doping is a step function at the metallurgical junction.

We treated three classes of junctions:

1. pn junctions in which the semiconductor on each side is nondegenerate
2. p⁺n junctions in which the p side is degenerate and the n side is nondegenerate
3. n⁺p junctions with the n side degenerate and the p side nondegenerate

For the n and p regions, the Fermi energy can be expressed

$E_f = E_C - kT \ln \frac{N_C}{N'_B} = E_i + kT \ln \frac{N'_D}{n_i} \quad \text{n region}$	$E_f = E_V + kT \ln \frac{N_V}{N'_A} = E_i - kT \ln \frac{N'_A}{n_i} \quad \text{p region}$
---	---

For a degenerate semiconductor, the approximation is often made that

$E_f \approx E_C \quad \text{n+ region}$	$E_f \approx E_V \quad \text{p+ region}$
--	--

5.7.1 Built-In Voltage

An important parameter required to determine the electrical characteristics of any junction is its built-in voltage. In general

$$V_{bi} = \frac{1}{q} |\Phi_p - \Phi_n|$$

where the work functions Φ_p and Φ_n are evaluated at the edges of the transition region at equilibrium. In general, V_{bi} must be solved numerically. However, for a prototype homojunction,

$V_{bi} = \frac{1}{q} \left[E_g - kT \ln \frac{N_C N_V}{N'_D N'_A} \right] = \frac{kT}{q} \ln \frac{N'_D N'_A}{n_i^2}$	prototype pn junction
$V_{bi} = \frac{1}{q} \left[E_g - kT \ln \frac{N_V}{N'_A} \right] = \frac{kT}{q} \ln \frac{N_C N'_A}{n_i^2}$	prototype n ⁺ p junction
$V_{bi} = \frac{1}{q} \left[E_g - kT \ln \frac{N_C}{N'_D} \right] = \frac{kT}{q} \ln \frac{N_V N'_D}{n_i^2}$	prototype p ⁺ n junction

5.7.2 Junction Width

The junction width of a prototype homojunction can be expressed

$$w = w_n + w_p = \sqrt{\frac{2\epsilon V_j (N'_A + N'_D)}{q N'_A N'_D}} \quad \text{pn}$$

$$w = \sqrt{\frac{2\epsilon V_j}{q N'_A}} \quad \text{n}^+ \text{p}$$

$$w = \sqrt{\frac{2\epsilon V_j}{q N'_D}} \quad \text{p}^+ \text{n}$$

where the junction voltage $V_j = V_{bi} - V_a$ and V_a is the applied voltage.

The ratio of the junction width on the n side to that on the p side is

$$\frac{w_n}{w_p} = \frac{(x_0 - x_n)}{x_p - x_0} = \frac{N'_A}{N'_D}$$

and most of the junction width is on the more lightly doped side.

Likewise, the ratio of the junction voltage dropped across the n side to that across the p side is

$$\frac{V_j^n}{V_j^p} = \frac{N'_A}{N'_D}$$

so most of the junction voltage appears across the more lightly doped side.

5.7.3 Junction Current

The three major current mechanisms for both forward and reverse bias are diffusion current density J_{diff} , drift current J_{drift} , and generation-recombination current density J_{GR} . The total current density can be expressed as $J = J_{\text{GR}} + J_{\text{diff}}$ or

$$J = J_{\text{GR}} + J_{\text{diff}} \approx J_{\text{GR}0}(e^{qV_a/2kT} - 1) + J_0(e^{qV_a/kT} - 1)$$

where J_{diff} is evaluated at the transition region edges, where drift current is negligible.

$$J_{\text{diff}} = J_0(e^{qV_a/kT} - 1)$$

where J_0 increases about 14 percent per degree Celsius. At constant current the diode voltage decreases about 2 mV per degree increase. For forward bias at room temperature, the current increases by a factor of 10 for a voltage change of 60 mV.

The coefficient for the generation-recombination current is

$$J_{\text{GR}0} \approx \frac{qn_iw}{2\tau_0}$$

The transition width w is only slightly voltage dependent (proportional to $\sqrt{V_j}$).

For a device in which the thickness of either side is much longer than a minority carrier diffusion length:

$$J_0 = q \left(\frac{D_n n_{p0}}{L_n} + \frac{D_p p_{n0}}{L_p} \right) \quad \text{long-base diode}$$

If the p-side thickness W_B is much less than a diffusion length,

$$J_0 = q \left(\frac{D_n n_{p0}}{W_B} + \frac{D_p p_{n0}}{L_p} \right) \quad \text{short p-region diode}$$

and if both sides are very short,

$$J_0 = q \left(\frac{D_n n_{p0}}{W_{B(p)}} + \frac{D_p p_{n0}}{W_{(n)}} \right) \quad \text{both sides short}$$

Generally $J_{\text{GR}0} \gg J_0$ and thus $J_{\text{GR}0}$ predominates for reverse bias and small forward bias. However, since the diffusion current increases much more rapidly with forward bias, it predominates at higher V_a and thus higher currents.

5.7.4 Junction Breakdown

There are two additional current mechanisms, which can lead to large reverse currents (breakdown) for reverse-biased junctions.

In Zener tunneling, electrons tunnel from the valence band on the p side to the conduction band on the n side. The tunneling current density can be expressed as

$$J \approx e^{-\pi W_T \sqrt{m^* E_g} / 2^{3/2} h} \quad \text{tunneling}$$

Multiplication current results from carrier impact ionization in the high-field depletion region. The carrier multiplication factor M is

$$M = \frac{1}{(1 - P)}$$

where P is the probability that one carrier crossing the junction will create an additional hole-electron pair. For unity P , M becomes infinite and avalanche occurs.

While Zener tunneling results in a soft breakdown and predominates in heavily doped junctions (small tunneling distance), avalanche gives a sharp breakdown and predominates for more lightly doped junctions.

5.7.5 Capacitance

There are two effects that contribute to capacitance in a pn junction. The junction capacitance C_j results from the variation of ionized charge in the transition region with applied voltage.

$$C_j = A \left[\frac{q e N'_D N'_A}{2(N'_D + N'_A)(V_{bi} - V_a)} \right]^{1/2}$$

Stored-charge capacitance results from the injected minority carrier charge stored in the quasi-neutral regions for forward bias. For an n⁺p junction in which the p region is much greater than an electron diffusion length,

$$C_{sc} = (q/kT) \delta I \tau_n \quad \text{long-base diode}$$

where δ is the reclaimable charge fraction and is equal to 1/2 for step (prototype) junctions.

For the p region much shorter than L_n , the capacitance is proportional to the p-region thickness, or

$$C_{sc} = (q/kT) \delta I t_T = (q/kT) \delta I \left(\frac{W_B^2}{2 D_n} \right) \quad \text{short-base diode with p side short}$$

In this case, $\delta = 2/3$.

For reverse bias and for small forward currents, C_j predominates. For larger forward currents, C_{sc} predominates.

5.7.6 Transient Effects

When operating as a switch, the switching speed is controlled by the time required to charge and discharge the diode capacitances. The major contribution to the time is the time associated with charging and discharging the stored-charge capacitance. Normally, the time required to remove the stored charge in

switching from forward bias to reverse bias (turn-off transient) is appreciably greater than in switching from reverse to forward bias.

We emphasize that, while the prototype model is oversimplified for most real devices, it does provide insight into the physical processes involved. In the next chapter, we will refine the model and discuss some more realistic devices.

5.8 REVIEW QUESTIONS

1. Referring to Figure 5.8, how can electrons enter the transition region from the left if there is an opposing electric field in the transition region itself?
2. Explain qualitatively why the current in a diode is small under reverse bias and large under forward bias.
3. Why does generation current dominate over recombination current under reverse bias in a pn junction, but recombination current dominates over generation current under forward bias?
4. In tunnel diodes, for a given peak current why is the area of a Si diode greater than for a GaAs diode?
5. Sketch the energy band diagrams for a pn junction at equilibrium, at reverse bias, and for forward bias.
6. Sketch the steady-state minority carrier concentrations in a pn junction for forward bias.
7. The built-in voltage of a Si pn junction is on the order of 0.7 V. Why can't this be measured with a voltmeter?
8. What is meant by a "short-base diode"?
9. Why is the drift current negligible in the p-region of an n⁺p junction?
10. Explain the origin of generation and recombination currents in a pn junction.
11. In the text, the band gap narrowing in the n⁺ region of an n⁺p junction is ignored. If this narrowing is considered, how would this affect the built-in voltage? How would it affect the $I-V_a$ characteristics?

5.9 PROBLEMS

- 5.1 A silicon pn junction is formed between n-type silicon doped with $N_D = 5.0 \times 10^{15} \text{ cm}^{-3}$ and p-type silicon doped with $N_A = 2.0 \times 10^{17} \text{ cm}^{-3}$.
 - a. Sketch the energy band diagram. Label both axes and all important energy levels.
 - b. Find n_{n0} , p_{n0} , n_{p0} , and p_{p0} . Sketch the carrier concentrations.
 - c. What is the built-in voltage?
- 5.2 Recall that the circuit symbol for a diode is as shown in Figure II.3. Which is the anode (the end labeled +) of a pn junction diode, the p side or the n side? Explain your reasoning.

- 5.3** A p⁺n junction is formed in silicon. The Fermi level on the n side is at the (intrinsic) conduction band edge $E_f = E_{C0}$. The p side is doped with $N_A = 2.0 \times 10^{17} \text{ cm}^{-3}$.
- Sketch the energy band diagram.
 - Sketch the carrier concentrations.
 - What is the built-in voltage?
- 5.4** Fill in the missing steps to derive Equation (5.35) for the junction width on the n side of the junction.
- 5.5** A step pn junction diode is made in silicon with the n side having $N'_D = 2 \times 10^{16} \text{ cm}^{-3}$ and the p side having a net doping of $N'_A = 5 \times 10^{15} \text{ cm}^{-3}$.
- Draw, to scale, the energy band diagram of the junction at equilibrium.
 - Find the built-in voltage, and compare with the value measured off your drawing in part (a).
 - Find the junction width.
 - Find the widths of the n side of the depletion region and the p side of the depletion region, and the voltage dropped across each side of the transition region.
 - Plot the electric field. What is its maximum value?
 - Plot the voltage distribution.
 - Plot the potential energy for electrons (E_C).
 - Draw the energy band diagram for $V_a = 0.5 \text{ V}$.
 - Draw the energy band diagram for $V_a = -5 \text{ V}$.
- 5.6** Consider the equilibrium energy band diagram for the pn junction diode shown in Figure P5.1a.
- Indicate the region(s) where there exists an electric field.
 - What is the value of the built-in voltage?
 - Sketch the electron and hole concentrations. Indicate the directions of the drift and diffusion components of the electron and hole fluxes and currents.
 - The same device is shown in Figure P5.1b, but now a voltage is applied across it. If the total junction voltage $V_j = V_{bi} - V_a$, what is the value of the applied voltage?
- 5.7** A pn junction is formed in silicon between n-type ($N'_D = 10^{18} \text{ cm}^{-3}$) and p-type ($N'_A = 8.0 \times 10^{16} \text{ cm}^{-3}$) materials. Find, for equilibrium,
- w_n and w_p and w .
 - The built-in voltage.
 - The maximum electric field.
 - How much of V_{bi} is dropped on the n side? On the p side?
 - Sketch an energy band diagram carefully reflecting your calculations above.

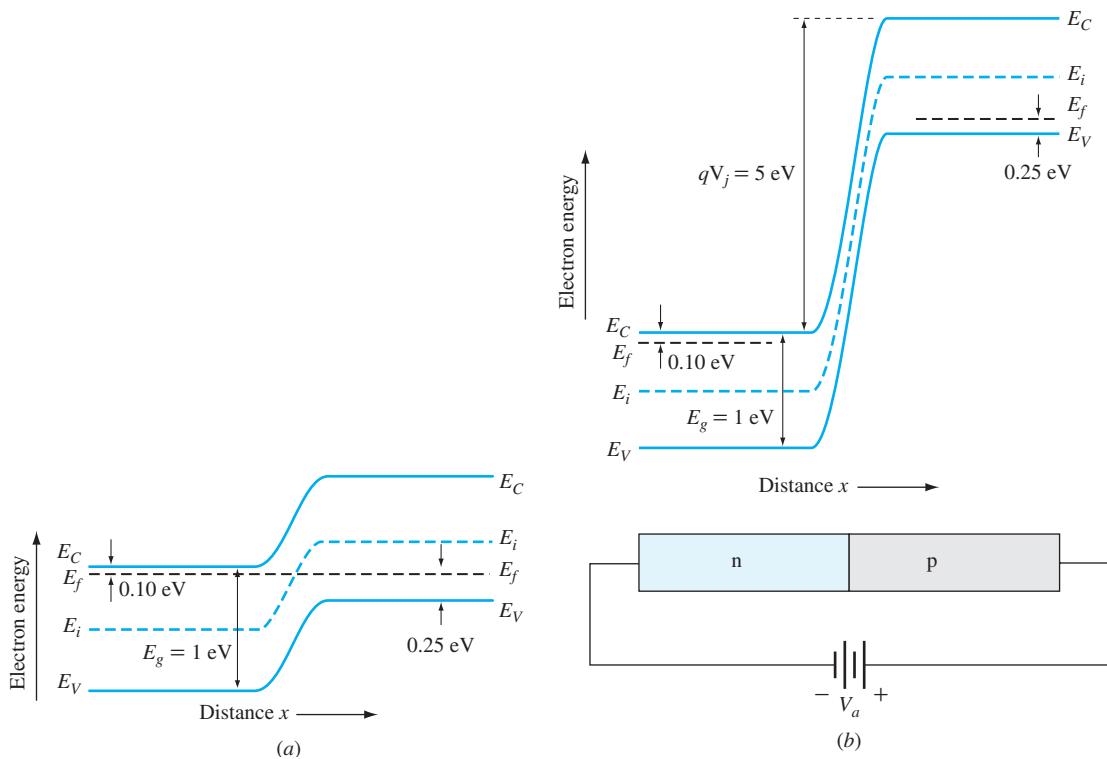


Figure P5.1

- 5.8** Find the equilibrium junction width for a pn junction that is degenerately doped on the n-side such that $E_C = E_f$. The doping on the p side is $N_A = 7.0 \times 10^{16} \text{ cm}^{-3}$.
- 5.9** A silicon diode has $N_D' = 5.0 \times 10^{17}$ on the n side and $N_A' = 5.0 \times 10^{16}$ on the p side. It is forward biased at $V_a = 0.4$ V.
- What is the diffusion current density due to minority carriers at the plane $x = x_p$?
 - What is the minority carrier diffusion density at the plane $x = x_n$?
 - What is the total current density in the junction neglecting recombination and generation?
 - What is the maximum recombination current density in the forward-biased junction? You may approximate the generation current as $J_R = +qwR_{\max}$, where the plus sign indicates that the current flows from p to n. This assumes that the recombination is at its maximum rate throughout the depletion region, an assumption made for simplicity. Compare this result with the injection (diffusion) current.
 - Which is larger, the injection flux density into the lightly doped side or the injection flux density into the heavily doped side?

- f. Repeat part (a) for a reverse bias of $V_a = -5$ V.
 - g. Repeat part (b) for $V_a = -5$ V.
 - h. Estimate the generation current density under reverse bias.
 - i. Compare the generation current density to the diffusion current density in the reverse-biased junction.
- 5.10** a. Calculate the minority excess carrier concentrations at each edge of the transition region for a silicon diode with $N'_D = 5 \times 10^{17} \text{ cm}^{-3}$ and $N'_A = 10^{17} \text{ cm}^{-3}$. The diode is forward biased with $V_a = 0.5$ V.
- b. Sketch the diffusion current as a function of distance on the p side.
- c. Keeping in mind that the total current is constant and the difference between the total current and the minority carrier diffusion current is due to drift of majority carriers, sketch the majority carrier drift current as a function of distance on the p side.
- 5.11** a. Show that Equation (5.93) follows from Equation (5.46).
- b. Derive Equation (5.94).
- 5.12** Consider an n⁺p junction under reverse bias of 5 V. Let $N'_A = 3.0 \times 10^{17} \text{ cm}^{-3}$ and the junction area be $85 \mu\text{m}^2$.
- a. Find the reverse current due to diffusion.
 - b. Find the reverse current due to generation.
- 5.13** Note that diffusion coefficient, diffusion length, and lifetime depend on doping [Equations (3.18), (3.42), (3.82), and (3.76)]. Consider a one-sided n⁺p junction in silicon.
- a. Plot the reverse diffusion current as a function of doping for $V_a = -5$ V.
 - b. Plot the reverse generation current as a function of doping for $V_a = -5$ V.
- 5.14** A Si junction has $N'_D = 5.0 \times 10^{15} \text{ cm}^{-3}$ and $N'_A = 2.0 \times 10^{17} \text{ cm}^{-3}$. The junction area is $125 \mu\text{m}^2$.
- a. Find V_{bi} .
 - b. Find I_0 .
 - c. Find the current at $V_a = -5$ V, 0 V, and +0.5 V. Remember to consider both diffusion and recombination-generation current. Also bear in mind that most of the junction appears on the lightly doped side (thus lifetime should be chosen for that material).
- 5.15** Plot the small signal resistance R_P as a function of applied voltage for V_a positive. Let the reverse leakage current be $I_0 = 10^{-15} \text{ A}$. Use a logarithmic scale for the resistance axis. Comment on your graph in view of your expectations of the resistance of a diode.
- 5.16** Find the reverse-bias breakdown at $N' = 10^{15} \text{ cm}^{-3}$ and 10^{17} cm^{-3} for GaN and Si. Which material is more suited to high-voltage operation? For a device to be able to withstand large voltages without breaking down, is low or high doping preferable?

- 5.17.** Find the electric field required in a Ge diode to create a tunneling probability of 10%.
- 5.18** Consider a symmetrical junction in GaAs in which $N'_D = N'_A = 10^{15} \text{ cm}^{-3}$.
- Find V_{bi} .
 - Calculate the series resistance R_S of the bulk regions if the cross-sectional area of the junction is $75 \mu\text{m}^2$ and the lengths of the bulk regions are each $2 \mu\text{m}$. Note that the series resistance R_S is due to the finite resistivity of the semiconductor materials; it is not the same as the differential resistance R_P .
- 5.19** Consider a forward-biased pn junction:

$$I = I_0(e^{q(V_a - IR_S)/kT} - 1)$$

or

$$V_a = \frac{kT}{q} \ln \left(\frac{I}{I_0} + 1 \right) + IR_S$$

Let $I_0 = 10^{-16} \text{ A}$ and $R_S = 20 \Omega$. Plot I versus V_a for $0.3\text{V} \leq V_a \leq 1\text{V}$.
(Hint: Choose values for I and solve for corresponding values of V_a .)

- 5.20** A Si junction has $N'_D = 5.0 \times 10^{17} \text{ cm}^{-3}$ and $N'_A = 1.0 \times 10^{18} \text{ cm}^{-3}$. The junction area is $100 \mu\text{m}^2$. What is the junction capacitance at $V_a = -5 \text{ V}$?
- 5.21** A junction has a degenerately doped n side and a p side with $N'_A = 10^{16} \text{ cm}^{-3}$. Both sides are long, and the fraction of reclaimable charge is 0.5. Compare the magnitudes of the junction capacitance and the stored charge capacitance at $V_a = -5 \text{ V}$, 0 V , and $+0.5 \text{ V}$. The junction area is $100 \mu\text{m}^2$. Note that generation-recombination current does not contribute to stored charge, only diffusion current contributes.
- 5.22** A bipolar transistor consists of two pn junctions back to back. In the forward active mode, the emitter-base junction is forward biased and the base-collector junction is reverse biased. The base is generally very thin—much shorter than a diffusion length. Figure P5.2 shows the energy band diagram.

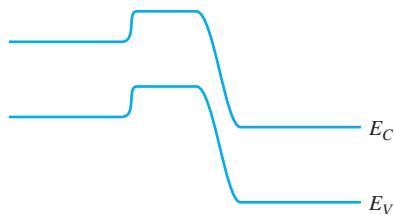


Figure P5.2

- a. Identify the emitter, base, and collector on the diagram.
 - b. Is this an npn or a pnp transistor?
 - c. At the E-B junction, are electrons injected or extracted?
 - d. At the C-B junction, are electrons injected or extracted?
 - e. Sketch the electron concentration distribution you expect across the base.
 - f. This thin base region is the source of the term *short base diode*. The goal in designing a transistor is to ensure that every electron injected into the base from the emitter eventually ends up in the collector (so that $I_E \approx I_C$). Explain how this short-base structure helps. (*Hint:* What else could happen to the electrons?)
- 5.23** In Figure 5.42, explain how the diode current can be negative when the applied voltage is zero.
- 5.24** From Equation (5.120), the storage time is given by

$$t_s = \tau_n \ln \left(1 + \left| \frac{I_F}{I_R} \right| \right)$$

From Figure 5.42 (obtained from a SPICE plot), $I_F = 4.2$ mA and $I_R = -0.6$ mA. Estimate the electron lifetime in the p-type region.

6

CHAPTER

Additional Considerations for Diodes

6.1 INTRODUCTION

In Chapter 5, we treated the case of prototype homojunctions, in which the doping is a step function at the metallurgical junction. While such junctions are seldom encountered in practice, the model is amenable to mathematical analysis and the results are at least indicative of those for a real device.

In this chapter, other, more realistic, junction devices are considered. We will first discuss homojunctions in which the doping profile is not a step function. Then we briefly discuss heterojunctions and metal-semiconductor junctions.

6.2 NONSTEP HOMOJUNCTIONS

As an example of a nonstep homojunction, we consider the case of a silicon bipolar junction transistor (BJT) that consists of two pn junctions back to back. The structure of this npn BJT is shown in Figure 6.1.

The active transistor is the region under the emitter contact (E), shown by the dotted box. It consists of the heavily doped n-type emitter (labeled n^+), the p-type base (p), and the n-type collector (n well). The buried n^+ layer, being heavily doped, is highly conductive and electrically connects the n well to the collector contact.

The doping profile as obtained experimentally for a BJT from a specific process is shown in Figure 6.2a. Here the concentrations of various dopants are plotted as functions of depth (position) in the crystal. Figure 6.2b shows the smoothed doping profile along the A-A' cut on an expanded horizontal scale. We see that the base region is $0.14\ \mu\text{m}$ in width with p-type doping varying from about $6 \times 10^{17}\ \text{cm}^{-3}$ at the emitter edge ($x = 0.13\ \mu\text{m}$, where $N_D(\text{As}) = N_A(\text{B})$), to $4 \times 10^{16}\ \text{cm}^{-3}$ at the collector edge ($x = 0.27\ \mu\text{m}$, where $N_A(\text{B}) = N_D(\text{P})$).

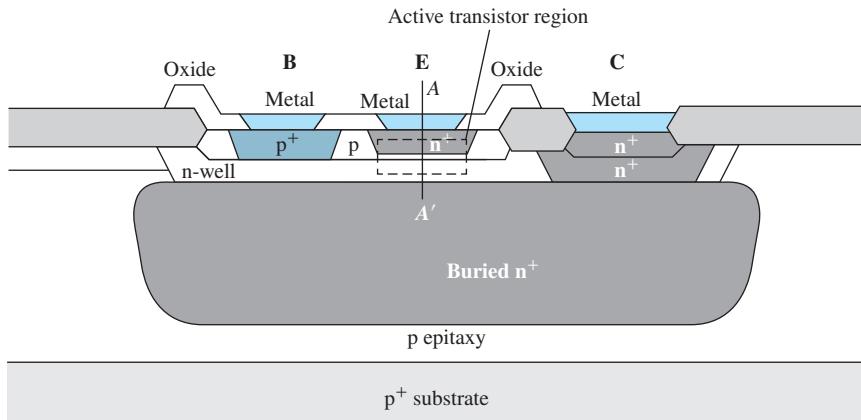


Figure 6.1 Schematic diagram (not to scale) of an npn bipolar transistor.

The doping in each region is nonuniform. The metallurgical junctions (emitter-base and base-collector junctions) where $N_D = N_A$ are shown. Because the two junctions are very close together, the current in one junction affects the current in the other. In this chapter, we investigate the energy band diagrams near the two junctions and discuss some of the consequences of nonuniform doping.

Notice that the doping level in each region varies appreciably with position, so neither junction can be considered a step junction. Most of the emitter region is degenerate, however, and in this region, to a reasonable approximation, $E_f = E_C$. In the base region the doping decreases with distance from the emitter-base (E-B) junction. This is referred to as a *hyperabrupt doping profile*. At the base-collector (B-C) junction, the net doping on each side of the junction increases with distance from the junction. The doping gradient causes an electric field in the base and collector regions.

The equilibrium energy band diagram of the structure is shown schematically in Figure 6.3. The emitter and n⁺ collector are degenerately doped, so there we make the approximation that $E_f = E_C$. The effect of impurity-induced band-gap narrowing is neglected here. We also observe that in the base and collector regions the doping gradients induce electric fields.¹

While the E-B junction is hyperabrupt, the magnitude of the doping increases with distance on either side of the B-C junction. To obtain the doping profile ($N_D - N_A$ versus distance) near the B-C junction, we observe that, on the semilog plot of Figure 6.2b, both acceptor and donor concentrations can be approximated by straight lines. This implies exponential impurity profiles. To describe these profiles on the base side, we use the expression for N_A

$$N_A = N_A(0)e^{-x/\lambda_B} \quad (6.1)$$

¹The doping gradient in the emitter also causes a built-in field to exist there.

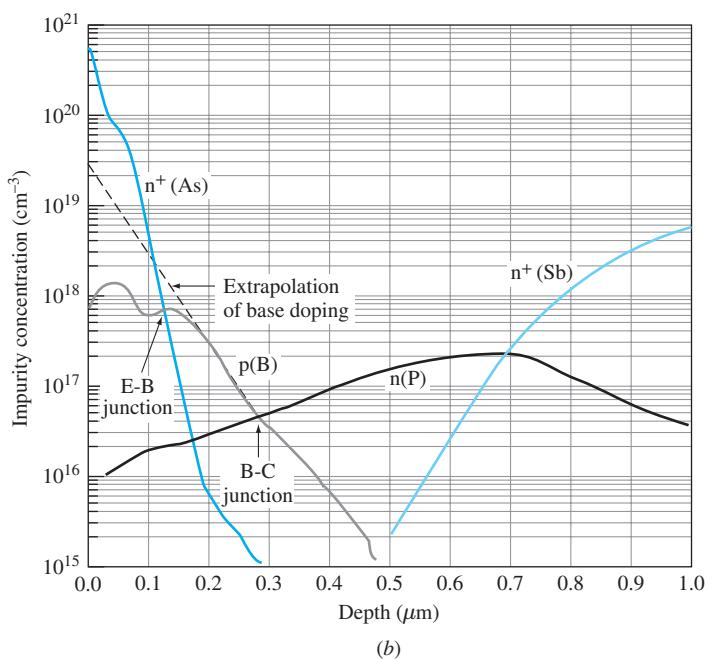
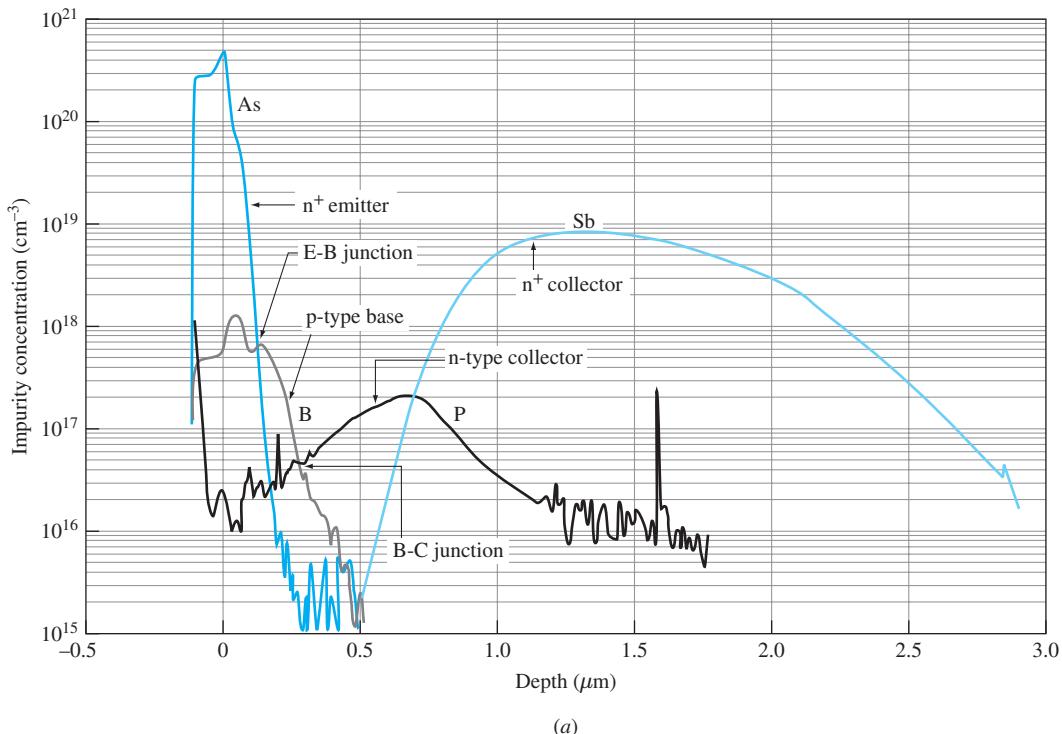


Figure 6.2 (a) Concentration of impurities along the cut A-A' of the device in Figure 6.1. (b) Impurity concentration (smoothed) on an expanded scale.

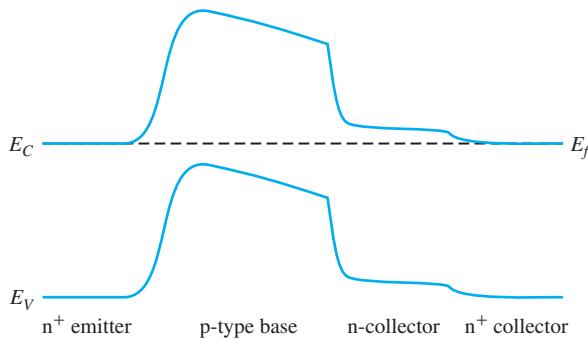


Figure 6.3 Equilibrium energy band diagram corresponding to the impurity profile of Figure 6.2.

where $N_A(0)$ is the (extrapolated) value of N_A at $x = 0$ and λ_B is a constant representing the distance in the base for a change of N_A by a factor of e (2.718).

Performing a similar analysis for N_D on the collector side gives, since over most of the n-type collector $N_D \gg N_A$,

$$N'_D \approx N_D = N_D(0) e^{x/\lambda_C} \quad (6.2)$$

where $\lambda_C = 0.195 \mu\text{m}$ and $\mathcal{E} = -133 \text{ V/cm}$ in the collector.

Figure 6.4 shows a plot of $N_D(x)$, $N_A(x)$, and $N_A(x) - N_D(x)$ near the base-collector junction. We approximate this junction as being linearly graded (along the slope indicated in the figure). The linearly graded junction is discussed in the next section.

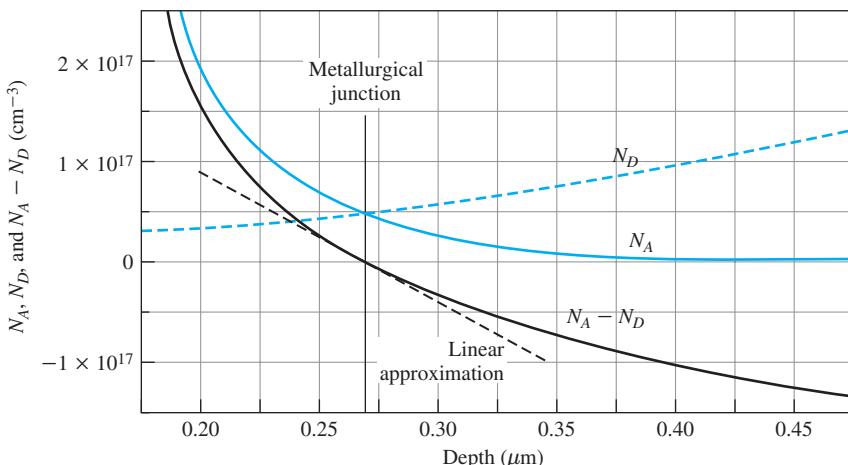


Figure 6.4 Plot of N_A , N_D , $N_A - N_D$ in the vicinity of the base-collector junction. The metallurgical junction is at $x = 0.27 \mu\text{m}$, where $N_A - N_D = 0$.

EXAMPLE 6.1

Find the built-in electric field at equilibrium in the base region of the BJT whose doping profile is shown in Figure 6.2.

Solution

We use Equation 6.1, except that in the p-type base we replace N_D with N_A . Rearranging, we obtain

$$\frac{x}{\lambda_B} = \ln \frac{N_A(0)}{N_A(x)}$$

Evaluating this expression at $x = 0.27 \mu\text{m}$ where $N_A = 4 \times 10^{18} \text{ cm}^{-3}$, and since the (extrapolated) value of N_A at $x = 0$ is $2.5 \times 10^{19} \text{ cm}^{-3}$,

$$\lambda_B = \frac{0.27 \mu\text{m}}{\ln \left(\frac{2.5 \times 10^{19} \text{ cm}^{-3}}{4.5 \times 10^{18} \text{ cm}^{-3}} \right)} = 0.042 \mu\text{m}$$

Assuming complete ionization, we can write for the base region ($0.13 \leq x \leq 0.27 \mu\text{m}$),

$$p(x) = N_A(x) = N_A(0) e^{\frac{-x}{\lambda_B}} = N_V e^{-\frac{(E_f - E_V(x))}{kT}}$$

or

$$e^{\left(\frac{x}{\lambda_B} + \frac{E_f - E_V}{kT}\right)} = \frac{N_V}{N_A(0)}$$

Solving for $[E_f - E_V(x)]$ gives

$$E_f - E_V(x) = kT \left[\frac{x}{\lambda_B} + \ln \frac{N_V}{N_A(0)} \right]$$

But since E_f is constant in x , and recalling that the electric field is proportional to the slope in the band edges, we can write

$$\mathcal{E} = \frac{1}{q} \frac{dE_V(x)}{dx} = -\frac{kT}{q\lambda_B}$$

For this (exponential) doping profile then, the electric field in the base region is

$$\mathcal{E} = \frac{kT}{q\lambda_B} = -0.026 \text{ V} \left(\frac{1}{4.2 \times 10^{-6} \text{ cm}} \right) = -6.1 \times 10^3 \text{ V/cm}$$

or 6.1 kV/cm . This built-in field is used to decrease the electron transit time across the base and thus increase the switching speed of bipolar transistors.

6.2.1 LINEARLY GRADED JUNCTIONS

To a first approximation, $N_A - N_D$ of Figure 6.4 can be taken as the tangent to the $N_A - N_D$ function at the base-collector metallurgical junction x_0 :

$$N_A - N_D = -a(x - x_0) \quad (6.3)$$

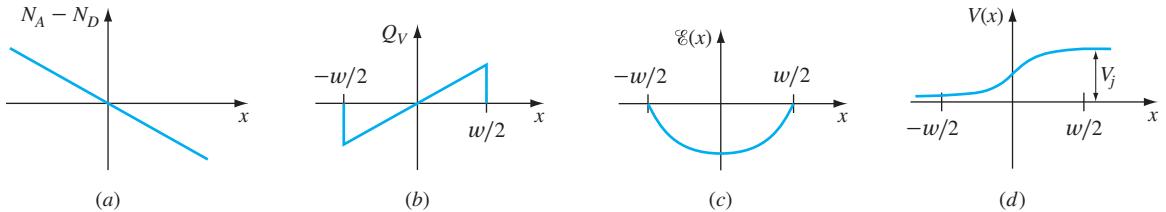


Figure 6.5 In a linearly graded junction the net doping concentration (a) is linear in x . The depletion approximation results in the charge distribution (b). In (c) the electric field is shown, and (d) shows the voltage profile.

where x_0 is the position of the base-collector junction, and a is the magnitude of the slope. This is referred to as the linearly graded approximation. From Figure 6.4 we find $a = 1.2 \times 10^{18} \text{ cm}^{-3}/\mu\text{m}$.

Figure 6.5a indicates the net doping level as a function of position for the linearly graded approximation with the metallurgical junction at $x = x_0$. For simplicity, we let $x_0 = 0$. Using the depletion approximation (that there are virtually no free carriers in the transition region, so the remaining charges are the ionized impurities on either side), we write the charge volume density Q_V as:

$$Q_V = \begin{cases} qax & -\frac{w}{2} \leq x \leq \frac{w}{2} \\ 0 & \text{elsewhere} \end{cases} \quad (6.4)$$

where w is the transition region width. Because of symmetry, the transition region width extends from $-w/2$ to $w/2$ as indicated in Figure 6.5b.

We can find the field in the junction by following the procedure used in Chapter 5 for the step junction. That is, we integrate the charge density to obtain $\mathcal{E}(x)$:

$$\mathcal{E}(x) = \int_{-w/2}^x \frac{Q_V}{\epsilon} dx = \frac{qa}{2\epsilon} \left[x^2 - \left(\frac{w}{2} \right)^2 \right] \quad -\frac{w}{2} \leq x \leq \frac{w}{2} \quad (6.5)$$

which is quadratic, as seen in Figure 6.5c.

Since $\mathcal{E} = -dV/dx$, and $V = 0$ at $x = -w/2$

$$V(x) = \frac{qa}{6\epsilon} \left(\frac{w^3}{4} + \frac{3xw^2}{4} - x^3 \right) \quad -\frac{w}{2} \leq x \leq \frac{w}{2} \quad (6.6)$$

as indicated in part (d) of the figure.

Solving for the junction width w yields

$w = \left[\frac{12\epsilon V_j}{qa} \right]^{1/3}$

(6.7)

where again $V_j = (V_{bi} - V_a)$.

Let us now investigate the validity of using the linearly graded approximation for the base-collector junction.

From Equation (6.7) with $a = 1.2 \times 10^{18} \text{ cm}^{-3}/\mu\text{m}$ and for $V_j = 1 \text{ V}$, we find $w = 0.088 \mu\text{m}$, or the transition region extends $0.044 \mu\text{m}$ on either side of the metallurgical junction. At this distance, the linear approximation in Figure 6.4 is close to the actual doping, so the approximation is reasonable. For reverse bias such that $V_j = V_{bi} - V_a = 5 \text{ V}$, however, the depletion region as calculated from Equation (6.7) extends $0.075 \mu\text{m}$ on each side of the metallurgical junction. At these distances from the junction, the linearly graded approximation is less valid, and the calculated junction width is less credible.

As would be expected, the larger the grading coefficient a , the narrower is the transition region.

The calculation of the built-in voltage V_{bi} is not as straightforward as for the step junction. In Figure 6.6a, the energy band diagram is shown for the case of electrical neutrality. The case for equilibrium is shown in (b). The edges of the transition region are indicated. The built-in voltage is

$$V_{bi} = \frac{1}{q} \left[\phi_p \left(\frac{-w}{2} \right) - \phi_n \left(\frac{w}{2} \right) \right] \quad (6.8)$$

where the work functions are evaluated at the edges of the transition region.

Alternatively, V_{bi} can be expressed in terms of δ_n and δ_p :

$$qV_{bi} = E_g - \left[\delta_n \left(\frac{w}{2} \right) + \delta_p \left(-\frac{w}{2} \right) \right] \quad (6.9)$$

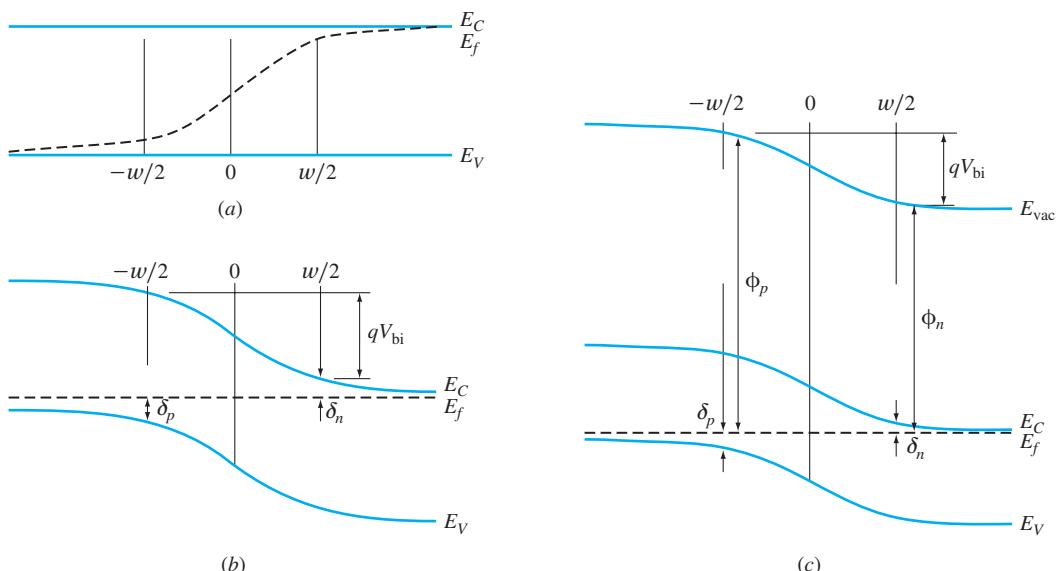


Figure 6.6 Energy band diagrams for the linearly graded junction under (a) neutrality, (b) equilibrium, and (c) equilibrium with vacuum level indicated. The built-in voltage can be found from $qV_{bi} = E_g - \delta_n - \delta_p$ or from $qV_{bi} = \phi_p - \phi_n$.

Following the procedure for a step junction, we write $\delta_p = kT \ln(N_V/p_{p0})$ and $\delta_n = kT \ln(N_C/n_{n0})$, where n_{n0} and p_{p0} are evaluated at the edges of the depletion region ($-w/2$ and $+w/2$), and

$$V_{bi} = \frac{kT}{q} \ln \left[\frac{\left(\frac{aw}{2} \right)^2}{n_i^2} \right] = \frac{2kT}{q} \ln \left(\frac{aw}{2n_i} \right) \quad (6.10)$$

Substituting w from Equation (6.7), and with $V_j = V_{bi}$, we obtain

$$V_{bi} = \frac{2kT}{q} \ln \left[\frac{a}{2n_i} \left(\frac{12\epsilon V_{bi}}{qa} \right)^{1/3} \right] \quad (6.11)$$

which can be solved iteratively.

Note that, for an arbitrary doping profile,

$$V_{bi} = \frac{1}{q} [\Phi_p(x_p) - \Phi_n(x_n)] \quad (6.12)$$

where Φ_p and Φ_n are the work functions evaluated at the edges of the transition region at equilibrium. This is shown in Figure 6.6c.

6.2.2 HYPERABRUPT JUNCTIONS

We treated the base-collector junction as a linearly graded junction, but as indicated earlier, the other junction, the emitter-base n^+p junction, is hyperabrupt. This is because the base doping decreases with increasing distance from the metallurgical junction. If the base doping profile in the vicinity of the junction is known, the junction properties can be calculated numerically. There are software packages for this. Here we simply mention that:

1. The built-in voltage is

$$V_{bi} = \frac{1}{q} [\Phi_p - \Phi_n]$$

where the work functions are evaluated at the edges of the transition region at equilibrium.

2. A field exists in the base that accelerates injected minority carriers away from the E-B junction. The effect of this field is to increase the current for a given forward bias, reduce the stored-charge capacitance, and increase the switching speed of the junction. This is particularly important in bipolar transistors.
3. Hyperabrupt junctions exhibit a large fractional variation of junction capacitance with applied voltage. This property is often used in a class of devices called *varactors* (variable-reactance devices).

6.3 SEMICONDUCTOR HETEROJUNCTIONS

A heterojunction, as indicated earlier, is a junction between two dissimilar materials. Such junctions can be between a semiconductor and a metal, for example, or between two different semiconductor materials. In this section we examine semiconductor–semiconductor junctions.

6.3.1 THE ENERGY BAND DIAGRAMS OF SEMICONDUCTOR-SEMICONDUCTOR HETEROJUNCTIONS

We consider junctions formed between two different semiconductor materials of different band gaps, electron affinities, ionization potentials, and work functions. The electrical properties of the junction depend strongly on these parameters. [1–3]

There are several classes of semiconductor heterojunctions, depending on the relative values of χ and E_g . Three cases are shown in Figure 6.7. These are *straddling* or *Type I*, *staggered* or *Type II*, and *broken-gap* or *Type III*. Because Type I heterojunctions are the most technologically important, we will discuss two examples, in which the forbidden band of the wide-band-gap semiconductor (Figure 6.7) overlaps (in energy) that of the narrow-band-gap semiconductor and the doping levels are constant on each side of the junction.

We will first consider the *electron affinity model* (EAM) in which the bulk semiconductor parameters are assumed invariant up to the metallurgical junction and the lattice constants of the two semiconductors are the same. We will then discuss corrections to this model to include charge dipoles near the interface resulting from the difference in valence band energies of the two semiconductors. Also considered are the effects of states within the forbidden band in the vicinity of the metallurgical junction (*interface states*).

Electron Affinity Model (EAM) The first example is a heterojunction between two different semiconductors, as shown in Figure 6.8a. One is an n-type wide-gap semiconductor (subscript 1) and the other is a p-type narrow-gap semiconductor (subscript 2). To obtain the energy band diagram as predicted by the electron affinity model (EAM), we proceed as for homojunctions, assuming electrical neutrality in every macroscopic region and using the vacuum level at the interface as a reference.

Using the convention in which a capital letter is used to indicate the conductivity type of the wide-band-gap material, this would be labeled an Np junction. Notice that because the electron affinities χ are different, there is a difference in the energies of the conduction band edges, ΔE_C . Since the band gaps (and therefore the ionization potentials γ) are also different, the two valence band edges do not line up either, with a difference of ΔE_V . The discontinuities are:

$$\begin{aligned}\Delta E_C &= |\chi_2 - \chi_1| \\ \Delta E_V &= |\gamma_2 - \gamma_1|\end{aligned}\quad (6.13)$$

At the instant the two materials are brought into contact, electrons in the conduction band of the N-type material flow into empty states that exist at lower energies in the conduction band in the p-type material and then recombine with holes

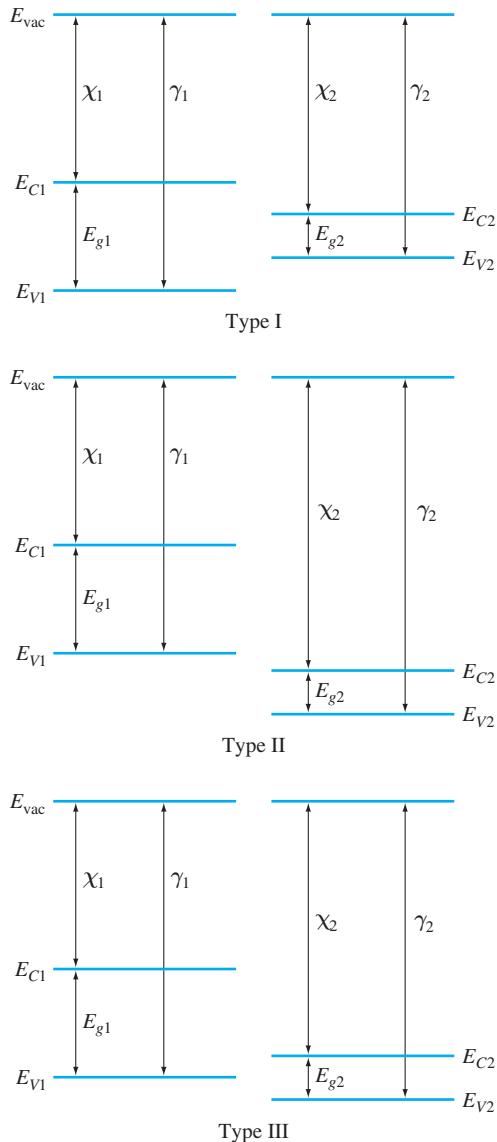


Figure 6.7 Three types of heterojunctions:
Type I (straddling), in which the wide-gap
energy overlaps that of the narrow gap, is
the most important; Type II (broken gap); and
Type III (staggered).

in the valence band. This transfer of carriers is by diffusion, as for a homojunction. Similarly, holes diffuse from p to N. As charges migrate, leaving behind the ionized acceptors and donors near the junction, an electric field builds up, again the same as in a homojunction.

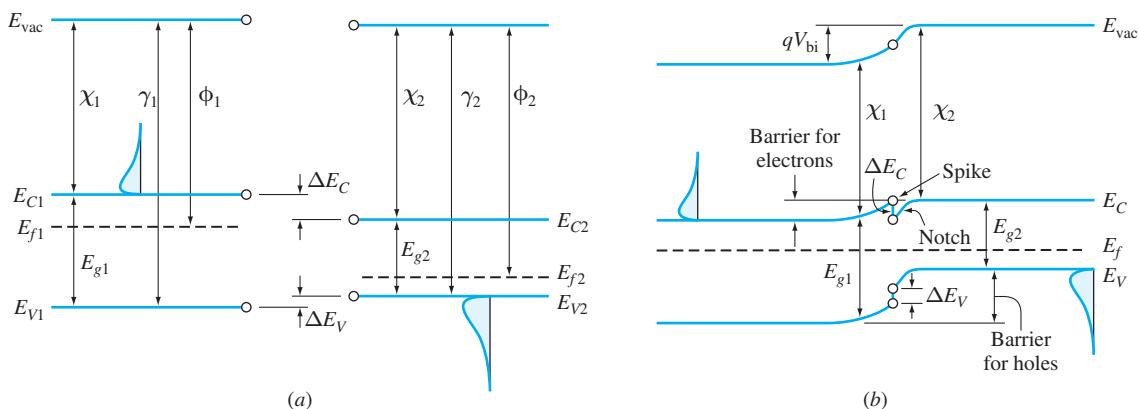


Figure 6.8 Constructing the energy band diagram for an Np Type I heterojunction as predicted by the electron affinity model: (a) electrical neutrality; (b) equilibrium.

The resultant energy band diagram at equilibrium is shown in Figure 6.8b. The energy bands pivot around the reference points E_{vac} , E_{C2} , E_{C1} , E_{V2} , and E_{V1} , all at the interface. When drawing the energy band diagrams, bear in mind that E_{vac} is continuous, and that the electron affinity χ , the band gap E_g , and the ionization potential γ are constants for a given material. Thus, when E_{vac} bends, the conduction band edges must remain parallel to it within each material to maintain constant electron affinity. Similarly, as the conduction band bends, the valence band must remain parallel to it within that material to maintain constant E_g . A potential energy spike and notch (potential energy well) exist in the conduction band. This “glitch” can appear in the valence band instead, depending on the particular semiconductors and their doping levels. The notch or potential well can trap carriers, an effect that can be exploited to improve device performance. For example, in a laser the well is used to trap carriers in the conduction band and increase the probability of stimulated emission.

As was the case for a homojunction, an applied voltage (p side positive, n side negative) tends to decrease the existing barriers and is referred to as a positive bias (Figure 6.9a), while the opposite bias polarity (reverse bias) increases the barriers (Figure 6.9b).

Notice, however, that there are two different barriers—the barrier for electrons in this case is smaller than the barrier for holes. Thus, when the junction is forward biased, electrons will be injected more easily than holes. If we define the *electron injection efficiency* as the electron injection current divided by the total current, this efficiency will be high when the barrier to electrons is less than that for holes. Improving the injection efficiency increases the performance of bipolar transistors.

It is interesting to notice that, for semiconductor heterojunctions, the wide-gap material is transparent to photons of energy less than its band gap and thus serves as an “optical window” to the narrow-gap material. This

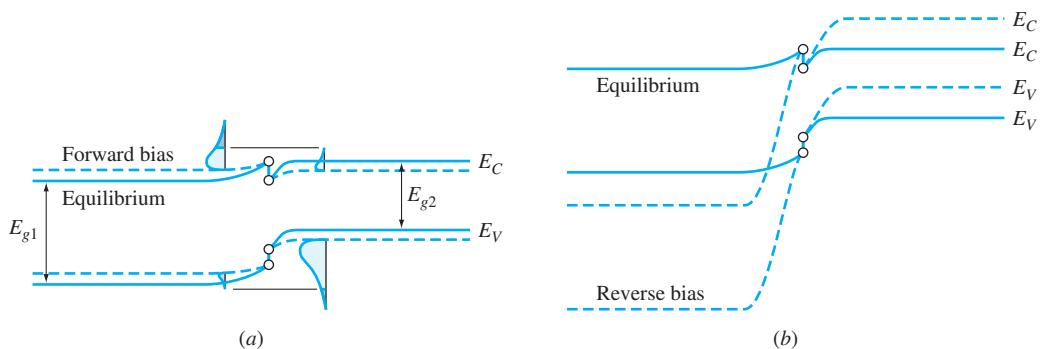


Figure 6.9 Energy band diagram for an Np heterojunction at equilibrium as predicted by the EAM (solid lines) and under bias (dashed lines): (a) forward bias; (b) reverse bias.

window effect can be used in solar cells and in light-emitting diodes, where the photons of interest are transmitted through the wide-gap window with little absorption.

The second type of heterojunction we will analyze using the EAM is an Nn heterojunction. We start as usual with the neutrality diagram as shown in Figure 6.10a. To achieve equilibrium in this case, at the instant the materials are joined, electrons flow across the junction from one semiconductor to the other until equilibrium is reached. The equilibrium case is shown in Figure 6.10b. Here, electrons from the conduction band of the semiconductor with the smaller work function (N) move into the empty states in the conduction band of the material with the larger work function (n). There are a negligible number of holes on either side.

Since electrons from the left transferred to the right, there is a nonuniform distribution of charge across the junction. The result is a built-in electric field. In this case, however, electrons are depleted from the left-hand side, but they accumulate in the right-hand side. That is, there is a depletion region on the left (carriers are swept out of the region by the field) with charge density qN_D' and an accumulation region on the right (carriers swept into the region by the local field). Since the density of states in the conduction band is large, so is the (negative) charge density. Because the total charge in the transition region is zero, the transition region width and junction voltage exist predominantly on the left side.

The EAM outlined above for drawing the energy band diagrams of heterojunctions provides a first estimate of the shape and nature of the bands. It also assumes, however, that there is an exact match in the lattice constants of the two materials, such that there are no dangling bonds at the interface. It also ignores an electric dipole, which can be set up at the interface as a result of the difference in valence band edge of the two materials. These effects, and their influence on the band lineup at the interface as predicted by the EAM, will be discussed qualitatively in the following sections.

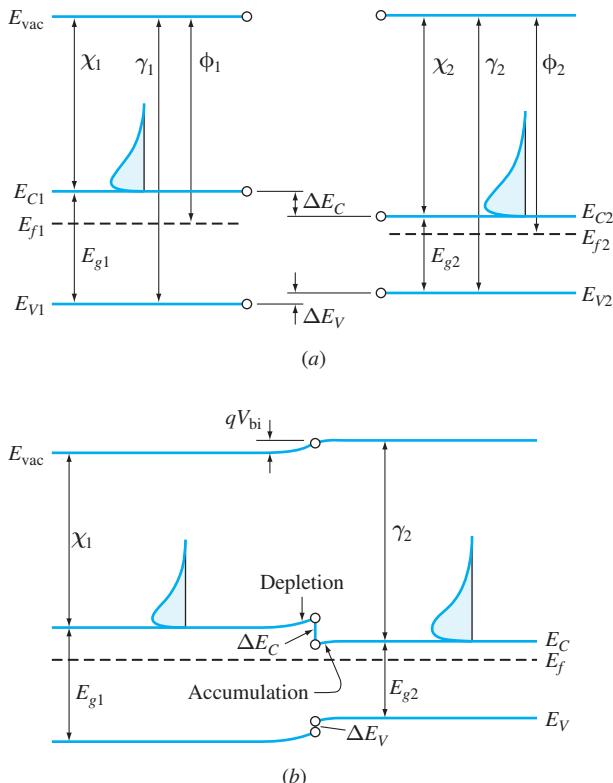


Figure 6.10 The energy band diagram for an Nn Type I heterojunction as predicted by the electron affinity model. (a) Neutrality; to reach equilibrium, electrons flow from the material on the left into the material on the right. (b) Equilibrium.

6.3.2 TUNNELING-INDUCED DIPOLES

Although the electron affinity model is a simple intuitive model for predicting the band lineup at the interface, the measured values of the induced barriers are not generally in agreement with this model. [1–5] To illustrate this, we take as an example a GaAs:Ge heterojunction. The lattice constants of GaAs and Ge are 0.5653 and 0.5646 nm respectively, a reasonably close match (0.124 percent mismatch). The electron affinity of GaAs is 4.07 eV and that of Ge is 4.0 eV. The electron affinity model then predicts $\Delta E_C = 0.07$ eV and $\Delta E_V = 0.83$ eV, as indicated in Figure 6.11a for the neutrality case.² The measured values of ΔE_C and ΔE_V , however, are 0.27 eV and 0.49 eV respectively [1] as indicated in Figure 6.11b.

²This appears to be a Type II or staggered gap heterojunction, but what is experimentally observed is a Type I junction, as we are about to explain.

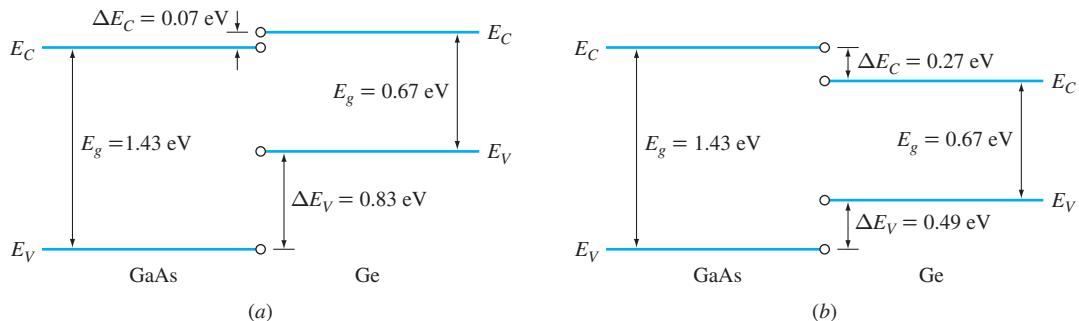


Figure 6.11 Energy band offset for a GaAs:Ge heterojunction (a) as predicted by the electron affinity model and (b) experimentally measured.

This discrepancy is explained (at least in part) by the existence of an electronic dipole at the interface resulting from the discontinuity in the valence band edge.³ To illustrate this, we consider the equilibrium energy band diagram of an arbitrary lattice matched Np heterojunction of Figure 6.12, as predicted by the EAM.

There is some uncertainty in the measured values of electron affinity. A common method is to shine monochromatic light on an n-type semiconductor in a vacuum enclosure, with a positive voltage on a plate (anode) with respect to the semiconductor. The electron affinity is taken as the minimum photon energy which excites electrons into the vacuum, producing a current at the anode (conservation of energy). The conservation of wave vector, however, must be satisfied also. Since the photons have negligible wave vector, electrons must be excited to states above the vacuum level with the same wave vector as at the bottom of the conduction band. Alternatively, a phonon can provide the required wave vector. This can overestimate the value of electron affinity.

We can see that there are electrons in the valence band of semiconductor B at the same energy as the forbidden band of A. These electrons can thus tunnel a short distance (on the order of a nanometer) into the forbidden region before being reflected back by the barrier. This penetration of the electrons into material A has the effect of placing some negative charge on the A side of the junction, creating a negative space charge in this region. An equal positive space charge must then exist in B, also within about a nanometer of the interface. The positive charge is due to the nuclei of the atoms of semiconductor B remaining fixed. This separation of charge results in a dipole layer being established at the interface.

³Much of the early literature reconciled the differences between the EAM and experiment by the assumption of localized states in the vicinity of the interface. More recently, it has been reported that the discrepancy can be primarily attributed to a tunneling-induced dipole layer [2–4], although there exists controversy concerning the relative importance of interface states and the dipole layer.

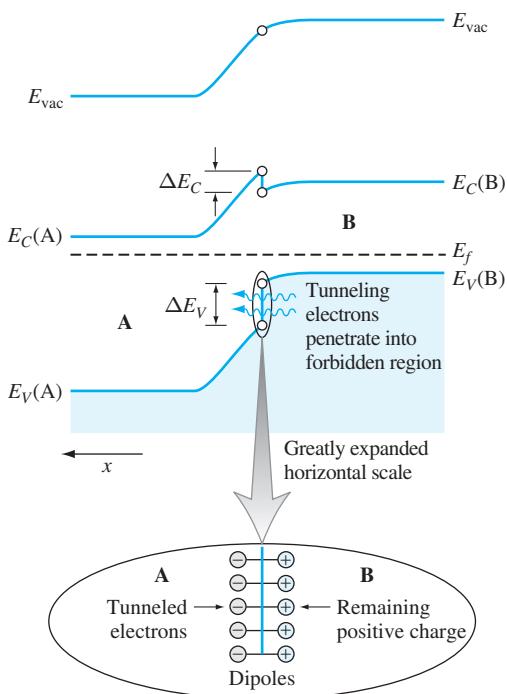


Figure 6.12 Equilibrium energy band diagram of an arbitrary Type I Np heterojunction as predicted by the electron affinity model. Electrons from the valence band of semiconductor B can tunnel a short distance into the forbidden gap of A, thus creating an interfacial dipole.

These interface dipoles influence the energy band diagram. Figure 6.13 shows the junction region of Figure 6.12 on a vastly expanded horizontal scale, to zoom in on the dipole layer. On this scale the bands predicted by the electron affinity model are essentially flat. The circles indicate the band edge location predicted by the EAM. Since negative is “up” on the electron energy band diagram, the tunneled electrons tend to raise the band edges near the surface of material A, pushing the bands upward. Similarly, the positive charge on the surface of B tends to push the bands downward. The squares in the figure indicate the positions of the band edges accounting for the tunneling-induced dipoles.

EXAMPLE 6.2

Estimate the tunneling distance of electrons of semiconductor B into the forbidden band of semiconductor A. Assume $m^* = m_0/2$.

Solution

Consider an electron that tunnels at an energy $E = E_V(A) + 0.1 \text{ eV}$. Then the electron probability penetration into semiconductor A is

$$\frac{\psi^*\psi(x)}{\psi^*\psi(0)} = e^{-2\left[\sqrt{(2|m^*|\hbar^2)(E-E_V(A))}\right]x}$$

where E_V is the potential energy of an electron near the valence band edge in semiconductor A, the influence of E_C is neglected, and positive x is taken in the tunneling direction. Defining the penetration distance x_T as the tunneling distance such that $\psi^*\psi(x_T)/\psi^*\psi(0) = e^{-1}$, we have

$$2\sqrt{\frac{2|m^*|[E - E_V(A)]}{\hbar^2}}x_T = 1$$

or

$$\begin{aligned} x_T &= \frac{\hbar}{2\sqrt{2|m^*|[E - E_V(A)]}} \\ &= \frac{1.05 \times 10^{-34} \text{ J} \cdot \text{s}}{2\sqrt{2(0.5)(9.11 \times 10^{-31} \text{ kg})(1.6 \times 10^{-19} \text{ J/eV})(0.1 \text{ eV})}} \\ &= 0.44 \times 10^{-9} \text{ m} = 0.44 \text{ nm} \end{aligned}$$

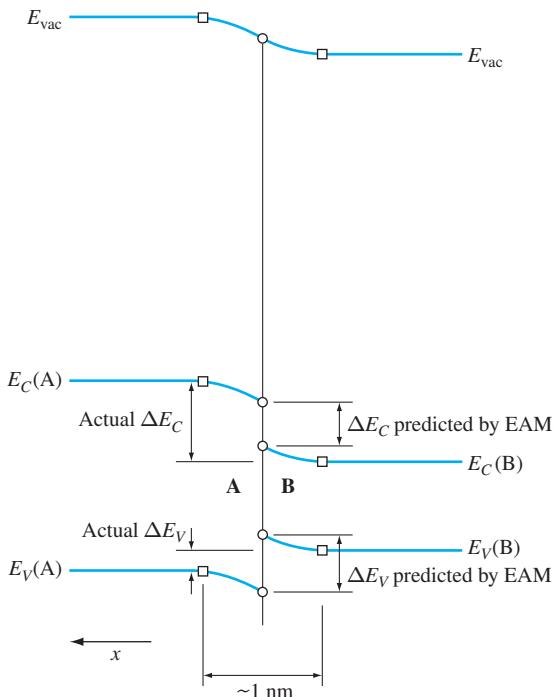


Figure 6.13 Equilibrium energy band diagram within a few nanometers of the interface of the heterojunction of Figure 6.12. The circles indicate the band discontinuities predicted by the electron affinity model; the squares indicate the influence of the tunneling-induced dipoles.

Thus the electrons penetrate approximately one-half nanometer into the wide-gap material. The total extent of the dipole interface (about 1 nm) is small compared with the thickness of the transition region, which is generally on the order of a micrometer.

EXAMPLE 6.3

Estimate the electric field caused by the tunneling-induced band bending at the interface of a GaAs:Ge heterojunction.

■ Solution

To find the true electric field, we examine the slope of the vacuum level E_{vac} . We can determine the change in potential by recognizing that the measured value of the discontinuity in the conduction band edge ΔE_C for this junction is 0.27 eV. From Figure 6.11a, the offset in E_C predicted by the electron affinity model is 0.07 in the opposite direction. Thus, the tunneling-induced dipole effect bends the conduction bands a total of $0.07 + 0.27 = 0.34$ eV. The curvature is repeated in the vacuum level.

This band bending occurs over a distance equal to the length of the dipoles, which we will take to be about 1 nm. Since $\mathcal{E} = (1/q)(dE_{\text{vac}}/dx)$, the effective electric field is⁴

$$\mathcal{E} \cong \frac{0.34 \text{ eV} \left(\frac{1.6 \times 10^{-19} \text{ J}}{1 \text{ eV}} \right)}{(1.6 \times 10^{-19} \text{ C})(10^{-9} \text{ m})} = 3.4 \times 10^8 \text{ V/m} = 3.4 \times 10^6 \text{ V/cm}$$

We emphasize again that the tunneling-induced dipole effect occurs over a very short distance. Figure 6.14 compares the equilibrium energy band diagram predicted by the electron affinity model (circles, black lines) with that considering the interface dipole effect (squares, colored lines). Since the dipole region is on the order of a nanometer and the transition region is on the order of a micrometer, on this diagram the dipole region is neglected, although it is indicated as a discontinuity in the vacuum level. The conduction and valence bands also have discontinuities in addition to those predicted by the EAM.

The magnitude of the tunneling-induced dipole effect depends on the size of the discontinuity in the valence band. The larger the discontinuity in E_V at the interface, the larger the number of electrons that can penetrate the barrier into the forbidden region. Thus, the larger the dipole, the greater the error in the EAM. Note that for a homojunction, $\Delta E_V = 0$ and no tunneling-induced dipole exists.

6.3.3 EFFECTS OF INTERFACE STATES

In the preceding discussion, the presence of any interface states, i.e., localized states near the interface and within the forbidden band, was neglected. The presence of such states can affect the resultant energy band diagram. To examine

⁴Note that at such a high field, the electron potential energy departs substantially from being periodic, and in this region the band gap is not well-defined.

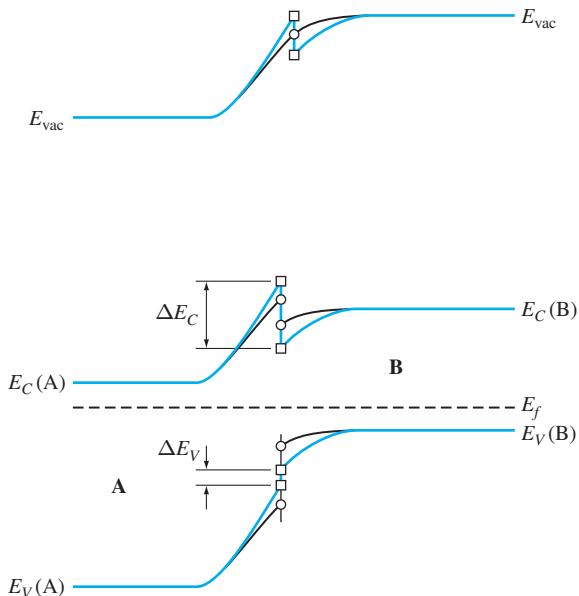


Figure 6.14 Equilibrium diagram of the Np Type I heterojunction considered in Figures 6.12 and 6.13. The lateral scale is reduced by a factor of 10^2 to 10^3 so that the indicated discontinuities appear in E_{vac} , E_C , and E_V . The circles and black lines are for the electron affinity model. The squares and colored lines present the result including the tunneling-induced dipoles.

these effects, we will first discuss surface states, and then extend the discussion to interface states.

The existence of well-defined allowed and forbidden energy bands resulted directly from the periodic nature of the semiconductor crystal. When the periodicity is interrupted, for example at or near a surface or defect, the band structure is modified. For defects, the result is localized allowed states in the forbidden band that can trap electrons or holes. At a surface, the potential energy is no longer periodic and the resultant band structure is no longer applicable. There are two effects: one due to the surface itself and the other due to the nonperiodicity near the surface. These are normally treated by considering the band structure of the bulk to be valid up to the surface, and then surface effects are described by localized *surface states*.

Consider, for example, the Si crystal shown schematically in Figure 6.15a in two dimensions. In the bulk, each Si atom is bound to each of its four neighbors by two electrons (covalent bonding). Surface atoms, however, have only three neighbors. Each surface atom, then, under the condition of neutrality, has one nonbonding electron and one vacant (surface) state. These are referred to as

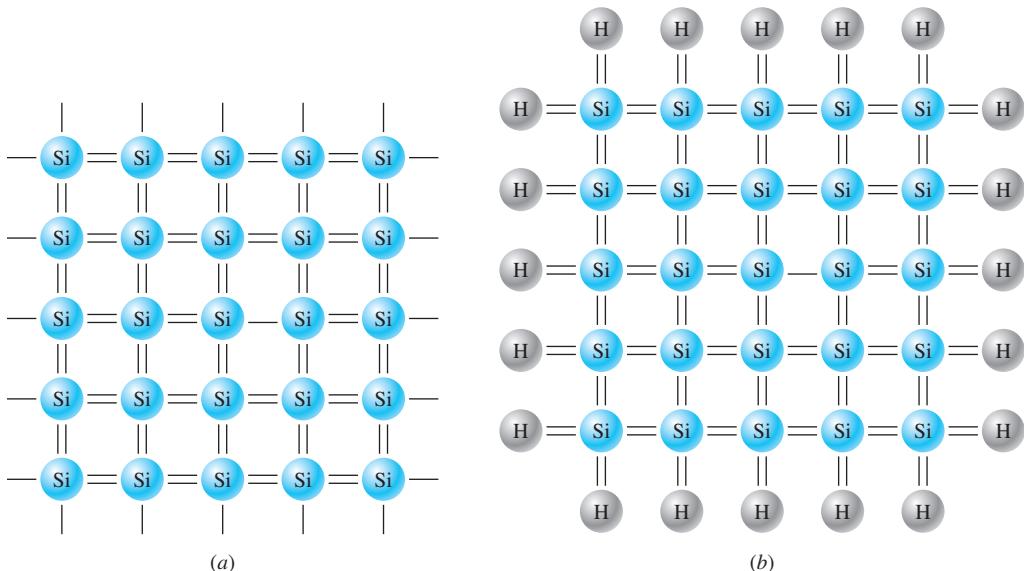


Figure 6.15 Schematics of a crystal showing (a) dangling bonds at the surfaces and (b) passivation of the bonds by atomic hydrogen.

dangling bonds. The surface density of surface states (number per unit area) is then comparable to the surface density of the surface atoms. For N atoms/cm³ in the bulk, there are on the order of $N^{2/3}$ atoms/cm² on a surface. For Si, $N \approx 5 \times 10^{22}$ cm⁻³, giving on the order of 2.7×10^{15} surface atoms/cm². The atoms are more closely spaced in the {111} faces than in the {110} and {100} faces, with consequently more surface states. In a three-dimensional crystal, the actual concentration of surface states is appreciably reduced from this value because the surface atoms are displaced from their bulk positions such that they can bond with their neighbors, reducing the number of dangling bands.

The preceding example is for a “clean” surface.⁵ On a real surface, foreign atoms are adsorbed, making chemical bonds with surface Si atoms. Consider, for example, a clean Si surface exposed to atomic hydrogen. Each H atom has one electron and one vacant state in its outer shell, and thus can bond to Si as shown in Figure 6.15b. In this case, all of the dangling bonds are saturated (filled) and the surface states have disappeared. Unfortunately, atomic hydrogen is not easily obtainable and is quite mobile on the Si surface. The H atoms are also small and diffuse very easily, so surface coverage varies with time.

However, silicon dioxide (SiO₂) can also be used to reduce the surface state density. Silicon dioxide is called a *native oxide* for silicon because it forms

⁵It has been experimentally found, however, that within about 1 to 3 nm of the surface, the atomic periodicity is disturbed, which causes some of the *surface states* to lie within this region.

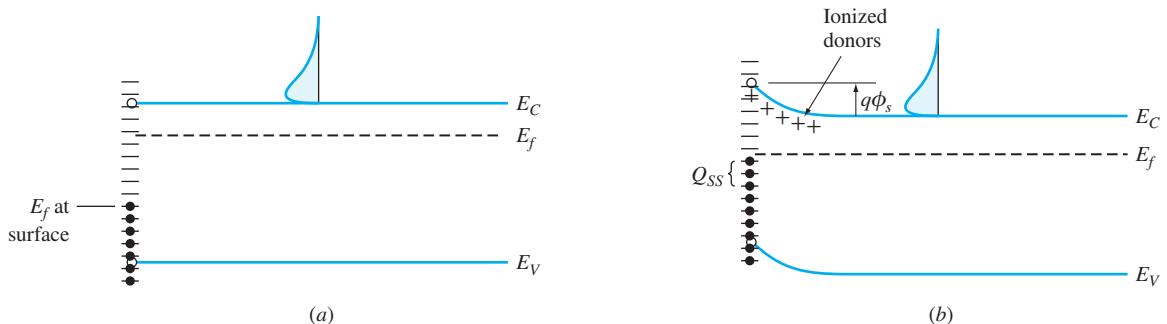


Figure 6.16 Effect of surface states on the energy band diagram. (a) Under the neutrality condition there are empty states at the surface at lower energies than electrons in the n-type semiconductor. (b) At equilibrium the transfer of electrons into the surface states results in a surface potential ϕ_s . The resultant net surface state charge per unit area is designated Q_{ss} .

spontaneously when silicon is exposed to oxygen. It is useful in integrated circuit fabrication because it is insulating, and can be easily grown in places where electrical isolation is needed. It is also more stable than hydrogen for rendering the surface states chemically inactive, a process called *passivating* the surface. Nitrogen, however, is often incorporated in the SiO_2 to more effectively passivate the surface.⁶

What are the energies associated with these surface bonds? The surface states are spread over a range of energies, and many of the states are at energies corresponding to the forbidden band in the bulk. Figure 6.16a shows how these states would appear in the energy band diagram for the case of neutrality. Note that for neutrality, the Fermi level of the surface is not the same as that of the bulk. To achieve equilibrium, electrons from the conduction band of the bulk move in to occupy the empty states at lower energies at the surface. If the surface is not passivated and electrons occupy some of the surface states, a net surface charge builds up (Q_{ss} in Figure 6.16b). A net surface charge requires an equal and opposite charge in the bulk, thus building up an electrostatic field normal to the surface. In this case, the positive charges are ionized donors in the lattice. The bands bend in the semiconductor near the surface to reflect this built-in field, as shown in Figure 6.16b.

We see then that an electrostatic potential difference, or surface potential ϕ_s , exists at the surface of the silicon with respect to the bulk. The value of ϕ_s depends on the density and energies of the surface states and on the doping level of the semiconductor. It is worth pointing out that the polarity of ϕ_s is usually such that minority carriers can be trapped at the surface. In the case of the n-type material in Figure 6.16, holes can be trapped in the valence band at the surface.

⁶Currently nitrogen replaces some of the oxygen to form SiO_xN_y (silicon oxynitride, or SiON) to better passivate the bulk Si surface. Since nitrogen has one less electron than oxygen, it can bond with the Si surface dangling bonds, thus reducing the oxide-silicon interface states.

Note that the band bending is caused by electron capture by the surface states, but the resulting upward band bending creates a potential pocket for holes. The surface state density and surface potential energy induced by the surface states strongly influence many device characteristics. They are controlled by appropriate surface treatments, many of which seem to involve witchcraft.

*6.3.4 EFFECTS OF LATTICE MISMATCH ON HETEROJUNCTIONS

Just as dangling bonds at the surface of a semiconductor affect the energy band profile near the surface, dangling bonds at the interface of two semiconductors with different lattice constants affect the band structure near the interface. Consider the case of a Si:Ge Nn heterojunction, which has a lattice constant mismatch of 4.1 percent. This results in a surface density of dangling bonds on the order of 10^{14} cm^{-2} . [4] The neutrality energy band diagram is indicated in Figure 6.17a. The Fermi levels for the Si, Ge, and interface are indicated. In this case, the Fermi level for the interface states is at a lower energy than in either semiconductor. To achieve equilibrium, electrons from both materials become trapped in the interface states, which causes a potential energy spike at the interface (Figure 6.17b). (Note that, for simplicity, in this diagram the tunneling-induced dipole effect is neglected.) To

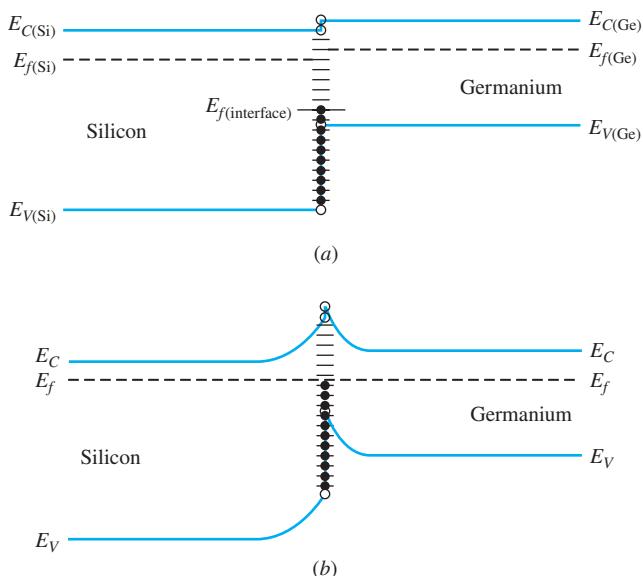


Figure 6.17 Energy band diagram for an Nn heterojunction between silicon and germanium, based on the electron affinity model. The case of neutrality is shown in (a), where the Fermi levels for the Si, Ge, and interface are indicated. The equilibrium case is shown in (b), where the Fermi levels have been aligned.

cross the interface, electrons must have sufficient energy to surmount this spike. Since few have this much energy, a large interface resistance results.⁷

The interface states above were considered to arise from dangling bands due to the lattice mismatch. This mismatch will also create a mechanical strain in the region near the interface. The strain accumulates with distance and eventually leads to crystal defects. These defects can also produce states within the forbidden bands because of the interruption of the lattice periodicity.

Another source of interface states results from the nonperiodicity of the electron wave function across the junction. Even if the two semiconductor lattices are perfectly matched, the Bloch wave function will change somewhat across the junction, since the atomic sizes on each side are different. This effect, however, is thought to have a relatively small effect on the band lineup at the interface.

The interface states associated with the dangling bonds resulting from the lattice mismatch in heterojunctions is a serious problem. As a result, the semiconductors involved in heterojunctions are lattice-matched as closely as possible. In compound semiconductor heterojunctions, this is usually accomplished by using ternary or even quaternary semiconductor alloys.

6.4 METAL-SEMICONDUCTOR JUNCTIONS

Junctions can also occur between a semiconductor and a metal. When a metal is joined to a semiconductor, for example, there are two types of junctions that can result: rectifying, in which current flows in one direction but not the other, and ohmic, or low-resistance, in which current flows easily in both directions. It is important to be able to control which type of junction will occur. Ohmic (low resistance) contacts are important since every transistor must ultimately make contact to the outside circuitry via a conductor. On the other hand, rectifying metal-semiconductor junctions, called Schottky diodes, also have many applications. In this section, the energy band diagrams and the electrical characteristics of these contacts are considered.

6.4.1 IDEAL METAL-SEMICONDUCTOR JUNCTIONS (ELECTRON AFFINITY MODEL)

We consider ideal metal-semiconductor junctions first, using the electron affinity model. In this ideal model, the effects of tunneling-induced interface dipoles and states in the forbidden region at or near the metal-semiconductor interface are neglected. We will come back and treat the more realistic case in the next section. First, we consider a Schottky contact with a metal and a semiconductor. To be specific, we consider the case of a metal semiconductor between Al and n-type Si. The neutrality energy band diagram is indicated in Figure 6.18a.

Aluminum has a work function of 4.1 eV, and silicon has an electron affinity of 4.05 eV. We choose a donor concentration such that the Fermi level in

⁷Note that the tunneling-induced dipole discussed earlier would increase the potential spike in Figure 6.17.

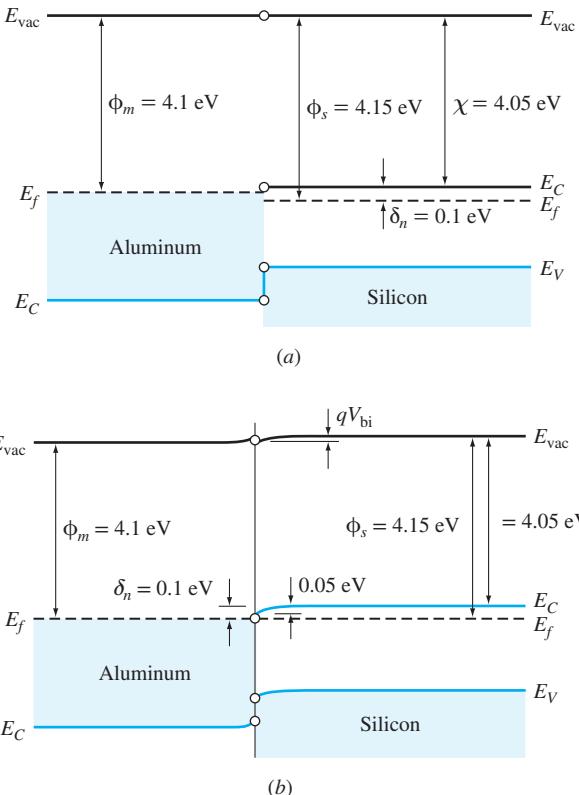


Figure 6.18 Energy band diagram as predicted by the electron affinity model for an Al:n-Si metal semiconductor junction: (a) Neutrality (b) equilibrium. The predicted barrier of 0.10 eV from metal to semiconductor is much less than the experimental value of about 0.7 eV. A more refined model is required.

the silicon is 0.1 eV below E_C . Then we can draw the equilibrium energy band diagram, using the EAM as shown in Figure 6.18b. It can be seen that, for this combination of materials, electrons in the semiconductor do not have a barrier going into the metal. Electrons in the metal see a small barrier of 0.10 eV (measured from the Fermi level in the metal to the conduction band edge of the bulk silicon). The built-in voltage is 0.05 V. Thus we would expect current to flow easily in both directions, and this Al-Si contact would be expected to be low resistance and ohmic. Experimentally, however, this junction is found to be rectifying, with a measured barrier on the order of 0.7 eV. Clearly, the ideal model is inadequate. In the next section, we improve the model by considering the effects of tunneling-induced interface dipoles and interface states.

6.4.2 INFLUENCE OF INTERFACE-INDUCED Dipoles

The discrepancy between the predicted and measured barrier in the aluminum-silicon junction is largely attributed to the existence of an electric dipole at the interface. [2]

As in heterojunctions, electrons in the Al conduction band can tunnel into the forbidden band of the Si, thus creating an electric dipole at the interface. As can be seen from Figure 6.18b, it is possible for electrons to tunnel into virtually the entire forbidden band. This represents a large number of electrons or a large amount of charge, and thus the dipole strength is expected to be larger than for the heterojunction case previously considered. Again, the thickness of the dipole region is on the order of a nanometer.

Figure 6.19a shows a close-up of the EAM neutrality case of Figure 6.18, with the effect of the tunneling-induced dipoles added. As the electrons from the metal penetrate a short distance into the semiconductor forbidden band before being reflected back, they carry some negative charge into the silicon, again only for a very short distance (on the order of a nanometer). This bends the bands upward in the Si adjacent to the Al. Note that there can be no electric field in the interior of a metal since it is a conductor. There can, however, be charge in the metal at the semiconductor interface, in this case the positive charges of the surface dipoles. The equilibrium energy band diagram, taking into account the effect of the dipole layer, is indicated in Figure 6.19b. Here the horizontal scale is adjusted back to the scale of the depletion region, so the tunneling-induced band

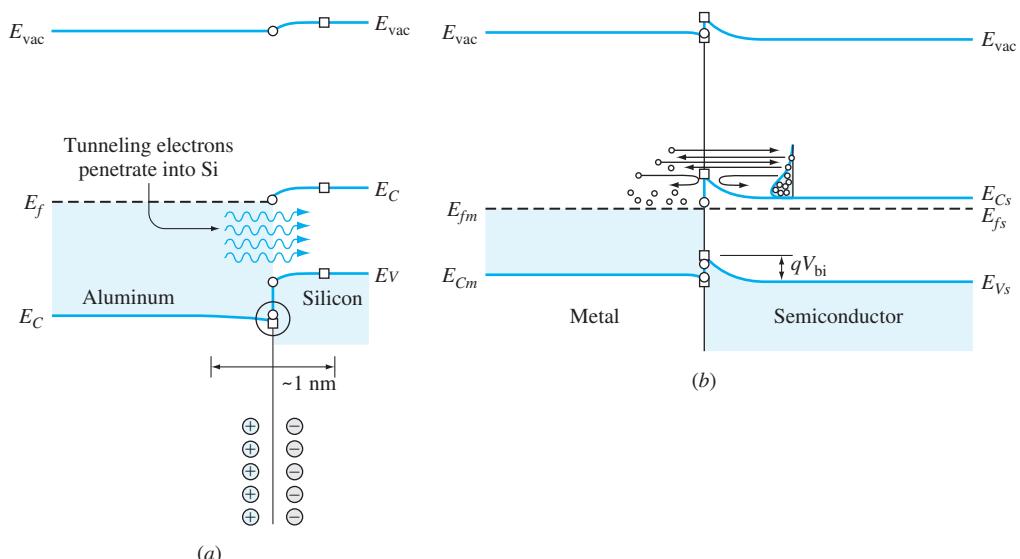


Figure 6.19 (a) The neutrality diagram for the Al:n-Si Schottky barrier diode including the tunneling-induced dipole effect. (b) The equilibrium energy band diagram for an Al:n-Si Schottky barrier diode.

bending appears as abrupt discontinuities. To attain equilibrium in this case, electrons move from the conduction band of the semiconductor to the lower available states in the metal. As the electrons leave the semiconductor, they leave behind ionized donors, which shift the semiconductor energies downward until the Fermi levels line up, as shown. The charge region of the device then consists of the tunneling-induced dipole layer, a depletion region (on the order of a micrometer) in the semiconductor, and a thin region of charge (in this case positive charge) at the surface of the metal. The accumulation region extends into the metal only a short distance—on the order of a fraction of a nanometer, as shown in Figure 6.19a. Because of the relative charge densities in the two materials, virtually the entire voltage drop and space-charge region is within the semiconductor, and usually E_C in the metal is assumed constant as indicated in Figure 6.19b. Also, observe that the barrier for electrons going from the semiconductor to the metal is different from the barrier for electrons going from the metal to the semiconductor. The interfacial dipole creates a near-discontinuity in the vacuum level at the interface.

Whereas for lattice-matched heterojunctions the barrier heights for a given semiconductor pair are quite reproducible, there is considerable scatter in the experimentally obtained barrier heights in metal-semiconductor junctions. This is thought to result from interface states caused by adsorbed foreign atoms on the Si surface before the metal is deposited, or from the structure of the Si semiconductor surface. [2]

6.4.3 THE CURRENT-VOLTAGE CHARACTERISTICS OF METAL-SEMICONDUCTOR JUNCTIONS

The electrical characteristics of a metal-semiconductor junction are a strong function of the doping concentration in the semiconductor. Here we consider two limiting cases. First we consider the case of a lightly doped semiconductor. Such a junction is referred to as a *Schottky barrier diode* or a *Schottky diode*. Next we consider the case of a heavily (degenerately) doped semiconductor, which results in a low-resistance contact.

Schottky Barrier Diodes (First-Order Model) We have referred to the junction of Figure 6.19 as a *Schottky diode*. In this section we will verify that the electrical behavior is diode-like. At equilibrium the net electron current is zero because the number of electrons having enough energy to surmount the barrier going from the metal to the semiconductor is equal to that going from the semiconductor to the metal, since the Fermi levels are equal. This is indicated in Figure 6.19.

With applied bias, the energy bands are altered. Continuing with the convention that forward bias means reducing the barrier, we apply a forward bias V_a to the Al:n-Si diode as indicated in Figure 6.20a. The potential energy barrier from semiconductor to metal is reduced by qV_a , and so the number of electrons in the semiconductor conduction band that have enough energy to surmount the potential barrier has increased (exponentially) with applied voltage. Thus, we expect a large number of electrons to be injected into the metal and a correspondingly

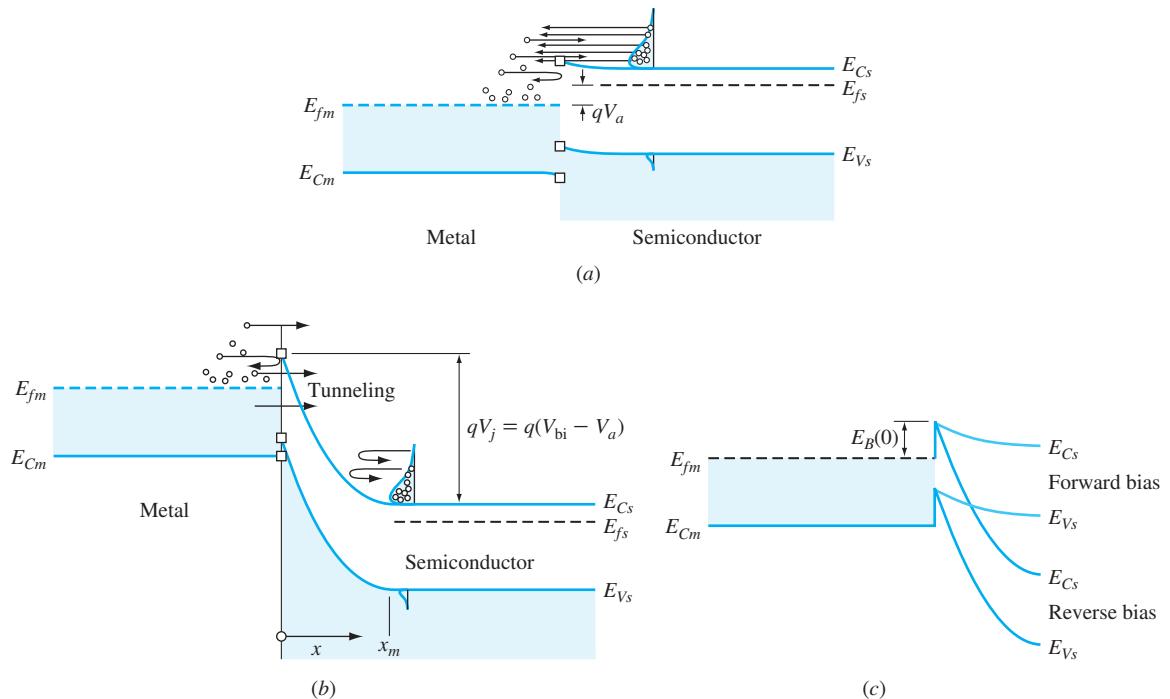


Figure 6.20 Energy band diagrams for a metal:n-semiconductor Schottky barrier. (a) For forward bias, electrons flow from semiconductor to metal. (b) For reverse bias, only a small leakage current flows. (c) For the first-order model, the metal-semiconductor barrier ($E_B(0) = E_C(x = 0) - E_{fm}$) is independent of applied voltage.

high current to flow. It is also possible for electrons in the metal to flow into the holes in the valence band of the semiconductor, but there are few holes, since this is n-type material.

Under reverse bias, the semiconductor-to-metal barrier is increased, as shown in Figure 6.20b. Because of this large barrier, a negligible number of electrons in the semiconductor can flow over the barrier into the metal. There are still a few electrons in the metal that can go over the barrier into the semiconductor (the barrier is smaller in this case than for electrons going in the other direction), contributing a current density of $-J_0$. Since the barrier for electron flow from metal to semiconductor is independent of bias⁸ and the barrier to electron flow from semiconductor to metal varies exponentially with bias, the net current density can be expressed as⁹

⁸The barrier actually decreases slightly with increasing reverse bias because of *image force effects* as discussed in the Supplement to Part 2.

⁹A more accurate expression is $J = J_0 [e^{qV_d/nkT} - 1]$, where the diode quality factor n is greater than unity, as discussed in the Supplement to Part 2.

$$J = J_0 [e^{qV_a/kT} - 1] \quad (6.14)$$

The coefficient J_0 has the form

$$J_0 = \frac{q m^* (kT)^2}{2\pi^2 \hbar^3} e^{-E_B(0)/kT} \quad (6.15)$$

where m^* is the electron conductivity effective mass and $E_B(0)$ is the barrier height for electrons at the Fermi level in the metal to the conduction band in the semiconductor; i.e., $E_B(0) = (E_{Cs} - E_{fm})$ at the interface.

We see from Equations (6.14) and (6.15) that the current has the same bias dependence as does a pn homojunction [compare with Equation (5.77)]. That is, the current I is exponentially related to the potential barrier $E_B(0)$ from metal to semiconductor for electrons at equilibrium. Equation (6.15) is referred to as the *Richardson-Dushman equation*. It was originally developed for thermionic emission of free electrons (mass m_0) from a metal into a vacuum. Schottky barrier current is often referred to as *thermionic emission current*. Equation (6.15) is often expressed [7]

$$J_0 = \frac{m^*}{m_0} A T^2 e^{-E_B(0)/kT}$$

where

$$A = \frac{qm_0 k^2}{2\pi^2 \hbar^3} = 120 \text{ A/cm}^2 \cdot \text{K}^2$$

Since electrons are majority carriers on both sides of the interface, they do not contribute to diffusion current. The current is thus limited by the number of electrons that have sufficient x -directed energy to surmount the barrier.

There is also the possibility of tunneling in a Schottky diode under reverse bias. Figure 6.20c shows that for reverse bias the tunneling distance between metal and semiconductor is less than that for a pn junction, and thus the reverse tunnel current is larger in the Schottky barrier diode.

We have so far considered a metal-semiconductor junction between Al and n-type silicon. It is possible to make a junction with p-type silicon as well. The equilibrium energy band diagram for an Al:p-Si device is shown in Figure 6.21a. The experimentally determined barrier is about 0.6 eV (compared with about 0.9 eV predicted by the EAM). We are now interested in the activity of the holes since they are the majority carriers. The holes in the valence band see a potential barrier. Under forward bias (Figure 6.21b), the barrier for holes is reduced and the hole flow from semiconductor to metal is increased.

Under reverse bias, Figure 6.21c, the barrier to holes is increased, and a negligible number flow into the metal. The electrons in the conduction band of the semiconductor can drift over into the metal, but this is p-type material, so there are few electrons and the current is very small. As for an n-type Schottky diode, tunneling contributes to the reverse current.

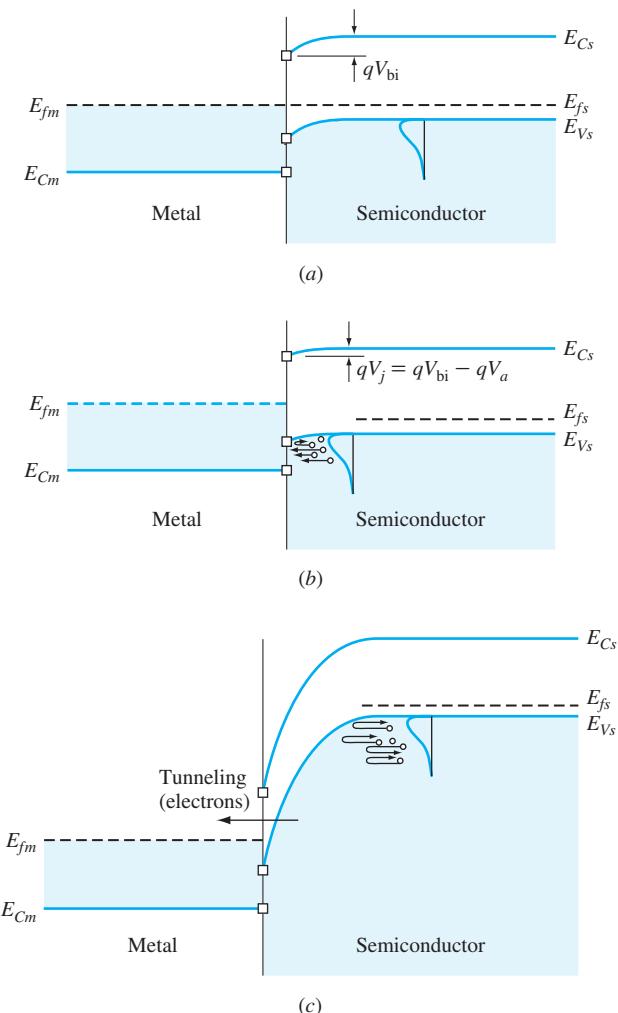


Figure 6.21 A Schottky barrier diode made with a p-type semiconductor. (a) Equilibrium; (b) forward bias; (c) reverse bias.

Rectifying metal-semiconductor contacts are diodes, and they behave qualitatively like pn junctions. The forward current increases exponentially with applied bias, and under reverse bias, there is a small leakage current.

Figure 6.22 compares the $I-V_a$ characteristics of a Schottky diode and a pn junction. The built-in voltage of the Schottky diode is normally less than that of a pn junction, resulting in a larger current at a given forward bias. Under reverse bias, the carrier tunneling discussed previously yields a softer reverse characteristic for the Schottky barrier diode.

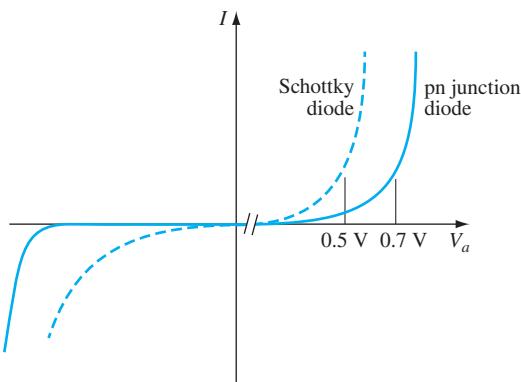


Figure 6.22 Comparison of the $I-V_a$ characteristics of a Schottky diode and a pn junction diode. The scale for the reverse characteristic is compressed compared with the scale for forward bias.

6.4.4 OHMIC (LOW-RESISTANCE) CONTACTS

We saw in the previous section that metal-semiconductor junctions with low doping in the semiconductor act as diodes. Electrical contacts need to be made to devices, however, and these contacts should be of low resistance, not rectifying.

To make these Schottky barriers low resistance rather than rectifying, the semiconductor surface is degenerately doped before the metal is deposited. Using the example of a metallic contact to an n-type semiconductor, this results in a metal-n⁺n junction, as shown in Figure 6.23a. Because of the degenerate doping of the semiconductor at the metal interface, the depletion region at the metal-n⁺ junction is so thin that the electrons tunnel easily through the barrier, resulting in a low resistance. The low resistance is seen for both directions of current flow. Further, the semiconductor-semiconductor n⁺n junction is also of low resistance, and thus the resistance between the metal and the n semiconductor is low. In this case, the contact is ohmic. The p-type case is shown in Figure 6.23b. This tunneling current is discussed further in Supplement to Part 2.

Virtually all low-resistance contacts to semiconductors are made in this manner. This is illustrated for the npn bipolar transistor of Figure 6.24, repeated from Figure 6.1. The active transistor is the vertical n⁺pn (n⁺:p:n well) region under the emitter contact (E). The metal-n⁺ emitter contact is a low-resistance contact as discussed. The collector contact (C) is metal-n⁺n (metal:n⁺:n⁺:buried n⁺:n well), as indicated by the dashed line. The base contact (B) is the metal-p⁺p structure as indicated by the dotted line.

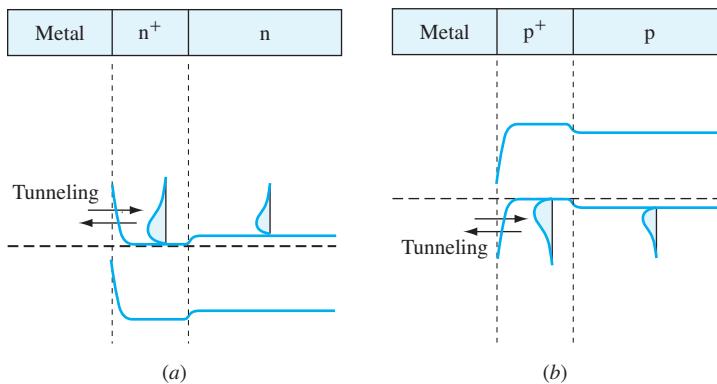


Figure 6.23 Low-resistance metal-semiconductor contacts using degenerate surface layers. Metal-n⁺n contact (a) and metal-p⁺p contact (b). The Schottky barrier is thin enough to permit tunneling.

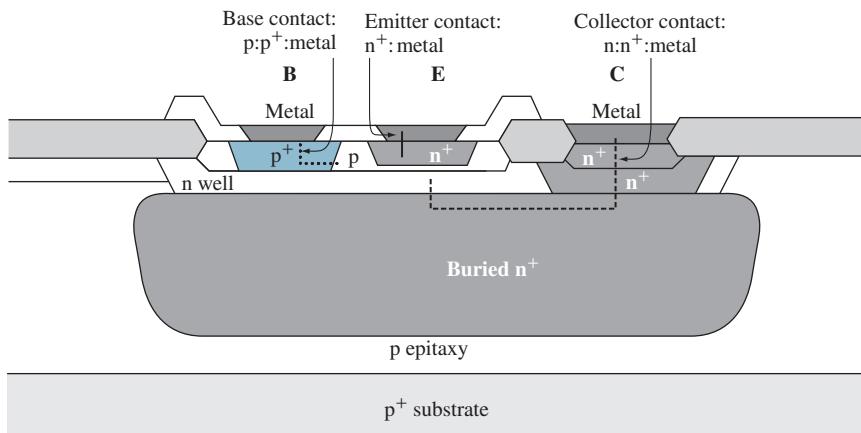


Figure 6.24 Schematic of an npn homojunction transistor indicating the low-resistance contacts. They are the base p:p⁺:metal contact, the emitter n⁺:metal contact, and the collector n:n⁺:metal contact.

6.4.5 $I-V_a$ CHARACTERISTICS OF HETEROJUNCTION DIODES

The $I-V_a$ characteristics of heterojunctions depend largely on the band lineup at the interface and the doping type on each side. Further, interface states at the material interfaces cause appreciable generation-recombination current. Because of these complexities, it is difficult to accurately predict the $I-V_a$ characteristics of heterojunctions.

For the lattice-matched Type I heterojunctions considered in this chapter, current across the interface of Np and Pn heterojunctions is primarily by minority carrier diffusion. For Nn and Pp heterojunctions, it results from majority carrier thermionic emission. As in other diodes, the current increases exponentially with applied voltage in the forward direction, and under reverse bias there is a small leakage current.

*6.5 CAPACITANCE IN NONIDEAL JUNCTIONS AND HETEROJUNCTIONS

Capacitance exists in all diodes, but it is generally more difficult to calculate for real diodes than it is for pn prototype homojunctions. In general, however, for a homojunction, the junction capacitance can be expressed as

$$C_j = \frac{\epsilon A}{w} \quad \text{homojunctions} \quad (6.16)$$

where the junction width w depends on the doping profile and junction voltage. The same equation applies to a Schottky barrier diode since the junction width w is entirely within the semiconductor.

For Np or Pn Type I heterojunctions, the junction capacitance can be considered as two capacitances in series:

$$C_j = \frac{A \left[\frac{\epsilon_n \epsilon_p}{w_n w_p} \right]}{\left[\frac{\epsilon_n}{w_n} + \frac{\epsilon_p}{w_p} \right]} \quad \text{heterojunction} \quad (6.17)$$

where ϵ_n and ϵ_p are the permittivities of the p and n regions respectively, while w_n and w_p are the depletion region widths. For N⁺n or P⁺p heterojunctions and for Schottky barrier junctions, the depletion region is predominantly on one side. In that case, Equation (6.16) is a good approximation.

The stored-charge capacitance is more difficult to calculate. For a Schottky barrier device, however, it is interesting to observe that $C_{sc} \approx 0$ because a negligible number of minority carriers are injected.

A final remark about capacitance in heterojunctions: For a heterojunction, the presence of a potential energy notch at the interface can interfere with the reclaiming of injected minority carriers, thus reducing the stored charge capacitance.

6.6 SUMMARY

In this chapter, we extended the results of the prototype homojunction of Chapter 5 to nonprototype homojunctions, heterojunctions, and metal-semiconductor junctions.

The equilibrium energy band diagram at the metallurgical junction was first predicted by the electron affinity model (EAM). This model is adequate

for homojunctions where the lattice periodicity is constant across the interface and $\Delta E_V = 0$. A nonzero value of ΔE_V in heterojunctions, however, results in a tunneling-induced dipole within about a nanometer of the interface, which alters the band lineup from that predicted by the electron affinity model. A similar tunneling-induced interface dipole for metal-semiconductor junctions along with the presence of interface states also affects their band lineup.

Although in pn junctions current is limited by the rate at which injected minority carriers can diffuse away from the junction, in metal-semiconductor (Schottky barrier) junctions, minority carrier injection is negligible compared with majority carriers thermionically injected over the barrier. In metal-semiconductor junctions in which the semiconductor is degenerately doped, the depletion region is thin enough that tunneling through the barrier results in low-resistance contacts.

The approach presented here is largely qualitative because a quantitative analysis requires a knowledge of the doping profiles and band lineups and is not amenable to closed-form solutions. These analyses are better handled with the use of device simulators, which solve the pertinent equations numerically.

6.7 REFERENCES

1. S. Tiwari and D. J. Franck, “Empirical Fit to Band Discontinuities and Barrier Heights in III-V Alloy Systems,” *Applied Physics Letters*, vol. 60, 1992, pp. 630–632.
2. Winfried Mönch, *Semiconductor Surfaces and Interfaces*, 4th ed., Springer-Verlag Berlin and Heidelberg GmbH & Co. K, 2010.
3. Federico Capasso and Georgio Margaritano, eds., *Heterojunction Band Discontinuities: Physics and Device Applications*, North Holland, Amsterdam, 1987.
4. G. Margaritondo, *Electronic Structure of Semiconductor Heterojunctions*, Kluwer, Dordrecht, 1988.
5. A. G. Milnes and D. L. Feucht, *Heterojunctions and Metal-Semiconductor Junctions*, Chap. 4, Academic Press, New York, 1972.
6. See, for example, A. van der Ziel, *Solid State Physical Electronics*, 2nd ed., Chap. 7, Prentice Hall, Englewood Cliffs, NJ, 1968.
7. C. R. Crowell and V. L. Rideout, “Normalized Thermionic Field Emission in Metal-Semiconductor (Schottky) Barriers,” *Solid State Electronics*, vol. 12, 1969, pp. 89–105.

6.8 REVIEW QUESTIONS

1. Outline the steps used to find the charge distribution, electric field, built-in voltage, and shape of the energy band diagram for a given doping profile.
2. Why is it expected, as stated in the text, that a larger grading coefficient produces a narrower depletion region in the linearly graded junction?

3. What is meant by a *hyperabrupt* junction?
4. In a pn junction, the built-in electric field is generated by the ionized donors and acceptors. In the Nn heterojunction of Figure 6.10, what is the source of the electric field on the N side? What is the source on the n side?
5. What is the electron affinity model, and how is it used? Summarize the steps to finding the energy band diagram for a junction using this model.
6. Explain how tunneling dipoles can produce near-discontinuities in the conduction and valence band edges.
7. For a Schottky barrier between aluminum and p-type semiconductor, the hole flow from semiconductor to metal is increased under forward bias and decreased under reverse bias. Since holes are just an artificial concept, explain these same phenomena in terms of the actual electrons in the valence band of the semiconductor.
8. Explain why, for low-resistance contacts to a semiconductor, the contact region of the semiconductor is made degenerate.

6.9 PROBLEMS

- 6.1 Consider a base-collector junction of a silicon BJT (bipolar junction transistor) like that in Figure 6.1. Assuming a linearly graded junction with $a = 1.2 \times 10^{18} \text{ cm}^{-3}/\mu\text{m}$, find V_{bi} .
- 6.2 A pn homojunction has a doping profile as indicated in Figure P6.1.
 - a. Find the value of the electric field in the bulk on the p side.

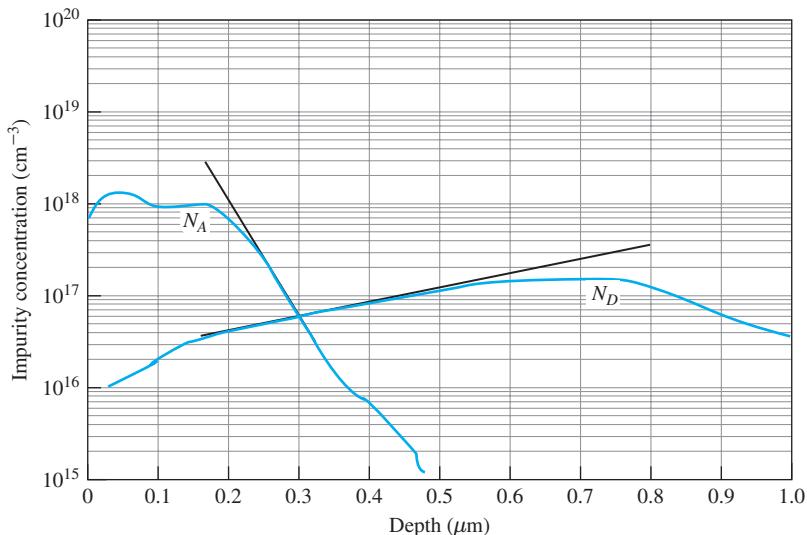


Figure P6.1

- b. Find the electric field in the bulk on the n side.
 - c. Plot $N_D - N_A$ as a function of position and find the slope a .
 - d. Find the built-in voltage.
 - e. Find the junction width at equilibrium.
 - f. For the width you found in (e), comment on the validity of the linear approximation you used over this distance.
- 6.3** In Section 6.2.2, it was claimed that hyperabrupt junctions exhibit a large fractional change in junction capacitance with applied voltage. Explain physically why we should expect this to be the case.
- 6.4** Figure P6.2 shows the equilibrium energy band diagram for a heterojunction between n-type semiconductor A (band gap 1.5 eV) and p-type semiconductor B (1.0 eV).
- Indicate the directions of:
 Electron diffusion
 Electron drift
 Electron diffusion current $J_{n(\text{diff})}$
 Electron drift current $J_{n(\text{drift})}$
 Hole diffusion
 Hole diffusion current $J_{p(\text{diff})}$
 Hole drift
 Hole drift current $J_{p(\text{drift})}$
 Effective electric field for holes
 True electric field

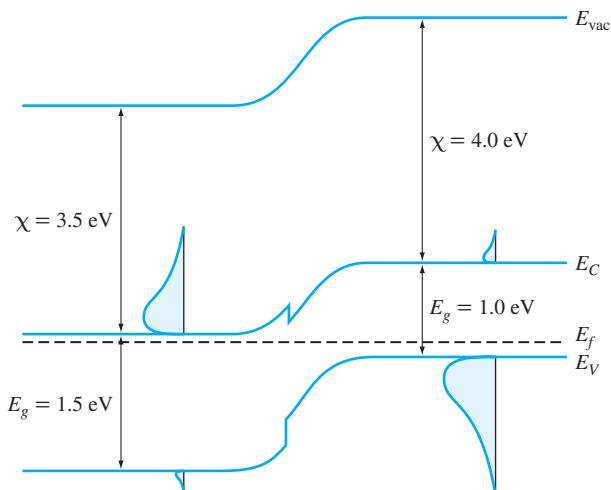
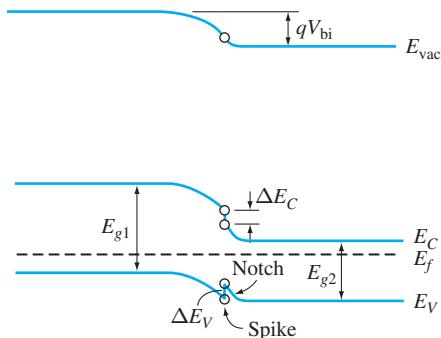


Figure P6.2

- b. In the conduction band edge, there is a notch that looks like a quantum well. The quantum well may have one or two discrete states.
- Indicate on the drawing where the potential well is for the electrons.
 - If an electron is trapped inside the well, list the mechanisms by which it can get out.
- 6.5** Using the electron affinity model, draw (to scale) the equilibrium energy band diagram for a heterojunction between p-GaAs ($E_g = 1.43$ eV, $\chi = 4.07$ eV), whose $\delta_p = 0.1$ eV, and N-Al_{0.3}Ga_{0.7}As ($E_g = 1.8$ eV, $\chi = 3.74$ eV), whose $\delta_n = 0.15$ eV. Neglect interface states.
- 6.6** Repeat the previous problem, only now let the GaAs be n type with $\delta_n = 0.1$ eV and let the AlGaAs be p type with $\delta_p = 0.15$ eV.
- 6.7** Consider the Type I Np heterojunction of Figures 6.8 and 6.9, in which net doping N'_D and N'_A are uniform on the N and p sides respectively. Let ϵ_n be the permittivity on the n side and ϵ_p that on the p side. Solve Poisson's equation to find the depletion widths w_n and w_p on the N and p sides and the total depletion width. (*Hint:* At the interface, displacement ($\epsilon\mathcal{E}$) is continuous.)
- 6.8** Show that the junction capacitance per unit area for Problem 6.7 can be written as
- $$C_j = \left| \frac{dQ_v}{dV_a} \right| = \left| A \sqrt{\frac{q\epsilon_n\epsilon_p N'_A N'_D}{2(\epsilon_n N'_D + \epsilon_p N'_A)(V_{bi} - V_a)}} \right|$$
- 6.9** For the SiGe Nn junction of Figure 6.17, sketch the energy band diagram that you would expect, using the simple electron affinity model (i.e., ignoring tunneling-induced dipoles and interface states). Discuss the difference in current flow that would result from the two energy band diagrams (electron affinity model and the model that includes the effects of the presence of dipoles). Let $E_C - E_f = 0.1$ eV for silicon and 0.15 eV for Ge.
- 6.10** We saw that in a heterojunction in which the spike and notch occurred in the conduction band (as in Figure 6.12), it was possible for electrons in the valence band of the narrow-gap material to tunnel a short distance into the forbidden band of the wide-band-gap material. This tunneling induced a dipole that produced a discontinuity in the bands at the junction. Consider a Pn heterojunction such as that in Figure P6.3. Comment on the possibility of tunneling-induced dipoles in this case.
- 6.11** Consider an n-GaAs Schottky barrier diode of area $10 \mu\text{m}^2$ and potential barrier $E_B(0) = 1$ eV.
- Find the value of I_0 from Equation (6.15).
 - Plot the forward I - V characteristic for a diode quality factor of $n = 1.3$.

**Figure P6.3**

6.12 Consider the junction of Figure P6.4.

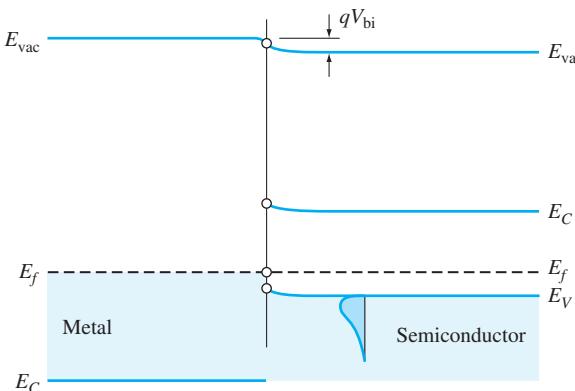
- Draw its energy band diagram under forward and reverse bias.
- Will this junction be ohmic or rectifying? Why?

6.13 Consider an n-Si:Al Schottky barrier diode for which $N'_D = 10^{17} \text{ cm}^{-3}$ and the measured built-in voltage is 0.70 V. Let the junction area be $10 \mu\text{m}^2$.

- Find the depletion width w . Note that this junction can be treated as a one-sided junction.
- Find the junction capacitance at $V_a = -5 \text{ V}$.

6.14 A heterojunction is formed between n-type GaAs of $N'_D = 10^{16} \text{ cm}^{-3}$ and p-type germanium of $N'_A = 10^{17}$. The measured discontinuities in the band edges are $\Delta E_C = 0.27 \text{ eV}$ and $\Delta E_V = 0.49 \text{ eV}$.

- Find the barrier to electrons in electron volts.
- Find the barrier to holes in electron volts.
- Find the built-in voltage (measured at E_{vac}) in electron volts.
- To what is the physical source of the discontinuity in the vacuum level attributed?

**Figure P6.4**

Supplement to Part 2: Diodes

S2.1 INTRODUCTION

In this supplement, we discuss some additional points about diodes. First we will look at the physics of the dielectric relaxation time, the time needed to restore charge neutrality to a region when excess carriers of one polarity are suddenly introduced, for example by injection.

Then we explore the capacitance in diodes in somewhat more detail than done earlier. We will show how measurement of the C - V (capacitance-voltage) characteristics of a pn junction or a Schottky diode can be used to find doping profiles in junctions. In addition to prototype (step) junctions, nonuniformly doped junctions are considered.

The tunneling current in Schottky diodes and low-resistance metal-semiconductor pn junctions is briefly discussed.

S2.2 DIELECTRIC RELAXATION TIME

When a pn junction is turned on, and excess carriers are injected across the junction, at the instant of injection there will be an excess charge. For example, if electrons are injected into the p side, there will be extra negative charge there. In this case extra positive charges must be summoned from the external circuit to reestablish charge neutrality in the p-type material. This process takes a small but finite time, called the dielectric relaxation time τ_D . Let us estimate this time, first for majority carriers and then for minority carriers.

S2.2.1 CASE 1: DIELECTRIC RELAXATION TIME FOR MAJORITY CARRIERS

We take first the case of excess majority carriers. Assume, for example, that Δn excess electrons are injected into n-type material, as might be the case for a forward-biased Schottky barrier junction. At time $t = 0$ there are $\Delta n(0)$ excess

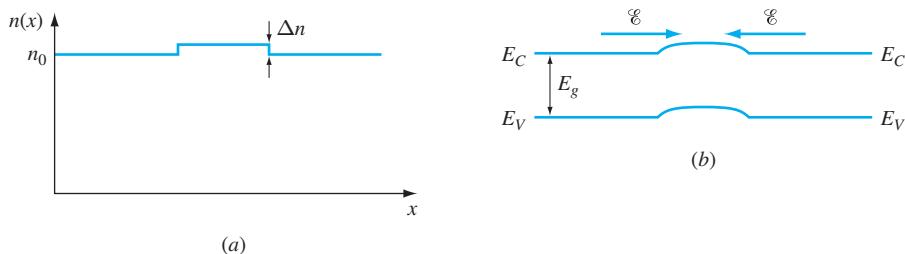


Figure S2.1 Dielectric relaxation time for excess majority carriers. (a) A localized group of excess majority carriers (in this case electrons) is injected into the sample; (b) the charge temporarily distorts the bands. The charge is neutralized with time constant τ_D by electrons that come in from the surrounding area (ultimately through the contacts), and the bands return to flat.

electrons in a small region as indicated in Figure S2.1a. This figure shows the carrier concentration as a function of position. This excess negative charge changes the potential locally, causing the band edges to bend as shown in Figure S2.1b. This local band bending creates an electric field \mathcal{E} , which tends to conduct the excess electrons away from the region.

We can find this current. For low injection such that $\Delta n \ll n_0$, we can ignore the electron diffusion term compared to the conduction term. The electron current is then

$$J_n = \sigma_n \mathcal{E} = q\mu_n n \mathcal{E} \quad (\text{S2.1})$$

The continuity equation for electrons is

$$\frac{\partial n}{\partial t} = \frac{1}{q} \frac{\partial J_n}{\partial x} \quad (\text{S2.2})$$

We neglect the recombination and generation terms because the excess carriers in this case are majority carriers, and recombination and generation will have a negligible effect on such large concentrations. Taking the derivative of Equation (S2.1), we have

$$\frac{\partial J_n}{\partial x} = q\mu_n \left[n \frac{\partial \mathcal{E}}{\partial x} + \mathcal{E} \frac{\partial n}{\partial x} \right] \quad (\text{S2.3})$$

$$\frac{\partial J_n}{\partial x} = q\mu_n \left[(n_0 + \Delta n) \frac{\partial \mathcal{E}}{\partial x} + \mathcal{E} \frac{\partial(n_0 + \Delta n)}{\partial x} \right] \quad (\text{S2.4})$$

The injected excess carriers are small in number because of the assumption of low injection, so we can neglect Δn in comparison with n_0 . We also recognize that n_0 is a constant and thus $\partial n_0 / \partial x = 0$. This results in

$$\frac{\partial J_n}{\partial x} = q\mu_n n_0 \frac{\partial \mathcal{E}}{\partial x} \quad (\text{S2.5})$$

Since the electric field is given by $\mathcal{E} = -dV/dx$,

$$\frac{\partial \mathcal{E}}{\partial x} = -\frac{\partial^2 V}{\partial x^2} \quad (\text{S2.6})$$

and from Poisson's equation

$$\frac{\partial^2 V}{\partial x^2} = -\frac{q\Delta n}{\epsilon} \quad (\text{S2.7})$$

Then by combining Equations (S2.5) to (S2.7), we obtain

$$\frac{\partial J_n}{\partial x} = q\mu_n \left[-n_0 q \frac{\Delta n}{\epsilon} \right] \quad (\text{S2.8})$$

and the continuity equation for electrons becomes

$$\frac{\partial n}{\partial t} = \frac{\partial \Delta n}{\partial t} = -\frac{q\mu_n n_0}{\epsilon} \Delta n = -\frac{\sigma}{\epsilon} \Delta n \quad (\text{S2.9})$$

This is a standard differential equation, whose solution is

$$\Delta n = \Delta n(0)e^{-t/\tau_D} \quad (\text{S2.10})$$

where $\Delta n(0)$ is the excess concentration at a point at $t = 0$ and

$$\tau_D = \frac{\epsilon}{\sigma} = \frac{\epsilon}{q n_0 \mu_n} \quad (\text{S2.11})$$

where μ_n is the majority carrier electron mobility. In *p* material, a similar equation applies involving the majority carrier hole mobility.

The quantity τ_D is called the *dielectric relaxation time*. It is a measure of the time required to neutralize excess carriers. Referring back to Figure S2.1, the net charge density, and with it the energy bands, decays to the normal (flat) condition with a time constant τ_D .

EXAMPLE S2.1

Find the dielectric relaxation time for Si of 1 $\Omega\text{-cm}$ (0.01 $\Omega\text{-m}$) resistivity.

Solution

For this material $\sigma = 100 \text{ S/m}$. From Equation (S2.11),

$$\tau_D = \frac{\epsilon}{\sigma} = \frac{8.85 \times 10^{-12} \text{ F/m} \times 11.8}{100 \text{ S/m}} = 1.04 \times 10^{-12} \text{ s}$$

The dielectric relaxation time in this example is about an order of magnitude greater than the mean free time \bar{t} between collisions. It is also shorter than the switching speed of semiconductor devices, and for most practical cases the relaxation of majority carriers can be considered instantaneous.

S2.2.2 CASE 2: DIELECTRIC RELAXATION TIME FOR MINORITY CARRIERS

The situation is different if excess minority carriers are injected (e.g., in a forward-biased pn junction). In this case the change in minority carrier concentration is locally significant, so the excess carriers will be dispersed by diffusion rather than drift. Diffusion, however, is a relatively slow process.

Suppose holes are injected into an n-type semiconductor. The field built up by the excess holes acts on the electrons. At time $t = 0$, there are excess holes but no excess electrons (Figure S2.2a). The electrons are attracted into this region (Figure S2.2b) by drift because of the field resulting from the excess positive charges, and the negative electrons tend to neutralize those positive charges. This neutralization occurs in a dielectric relaxation time. While electrical neutrality is quickly established, the excess holes (and excess electrons to ensure neutrality) diffuse slowly while recombining, so there are still excess holes and electrons for some time (the carrier lifetime). The bands are flat during this recombination period because neutrality has been established.

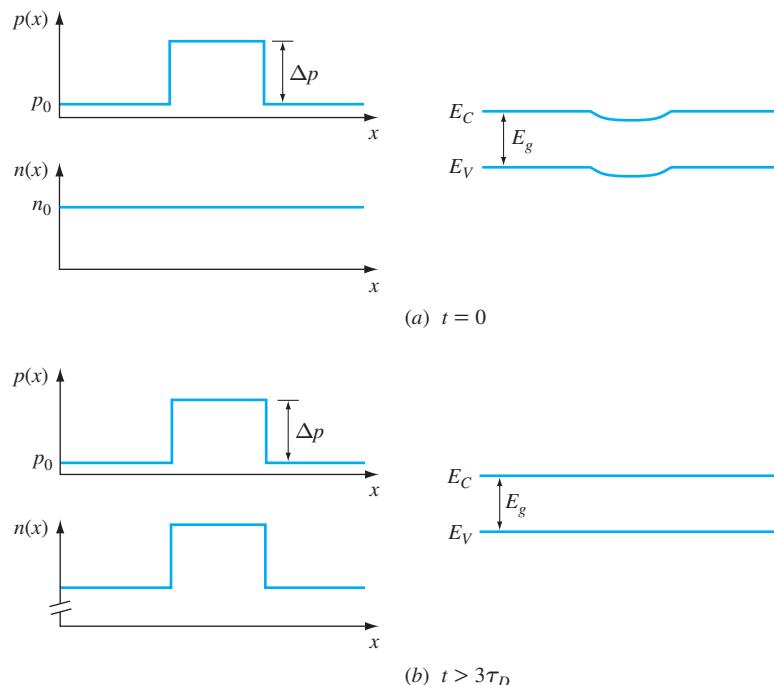


Figure S2.2 Dielectric relaxation for excess minority carriers (in this case holes).

(a) Excess holes are injected at time $t = 0$. At that instant the bands bend downward because of the local excess positive charge. (b) During a period amounting to a few dielectric relaxation time constants, excess electrons are drawn in to neutralize the charge and flatten the bands. Note the carrier concentration plots are not to scale.

To summarize:

1. If excess charge is injected into a semiconductor, it is electrically neutralized within about a dielectric relaxation time τ_D .
2. If the excess charge consists of majority carriers, the field set up by these carriers conducts them out via the contacts.
3. If the excess charge consists of minority carriers, majority carriers flow in from the contacts to neutralize it. The excess electrons and holes then diffuse and recombine.

S2.3 JUNCTION CAPACITANCE

In Chapter 5 we briefly discussed the capacitance of diodes, both the junction capacitance and the stored-charge capacitance. Here we investigate these two in more detail. We will treat junction capacitance, and show how a measurement of the capacitance versus voltage can be used to determine the doping concentration, the built-in voltage of a junction, and the doping profiles of various structures. After that, we will revisit the stored-charge capacitance, looking at the important case of the short-base diode.

S2.3.1 JUNCTION CAPACITANCE IN A PROTOTYPE (STEP) JUNCTION

Recall that the junction (differential) capacitance is defined as

$$C_j \equiv \frac{dQ}{dV_a} \quad (\text{S2.12})$$

Since dQ is the change in charge at the edges of the depletion region for a change in voltage,¹ the junction capacitance has the form of a parallel-plate capacitor

$$C_j = \frac{\epsilon A}{w} \quad (\text{S2.13})$$

While Equations (S2.12) and (S2.13) are general, for the prototype (step) pn junction the result is

$$C_j = A \left[\frac{q\epsilon N'_A N'_D}{2(N'_D + N'_A)(V_{bi} - V_a)} \right]^{1/2} \quad \text{prototype junction} \quad (\text{S2.14})$$

¹More accurately, $C_j = dQ/dV_j$, where $dV_j = dV_a - R_s dI$. However, the capacitance is normally measured for reverse bias or small forward bias such that I is small and $dV_j \approx dV_a$.

For a one-sided step junction, the equation becomes

$$C_j = A \left[\frac{q\epsilon N'}{2(V_{bi} - V_a)} \right]^{1/2} \quad \text{one-sided step junction} \quad (\text{S2.15})$$

where N' is the (uniform) net doping on the lightly doped side. Equation (S2.15) is also valid for a Schottky diode in which the semiconductor is uniformly doped.

Apart from its influence on device response time, experimentally the capacitance yields useful information about the device. The capacitance-voltage characteristic can be used to measure the built-in voltage of a junction. To see this, we square and invert Equation (S2.14) to obtain

$$\frac{1}{C^2} = \frac{2(N'_D + N'_A)(V_{bi} - V_a)}{A^2 q\epsilon N'_D N'_A} \quad (\text{S2.16})$$

Everything in Equation (S2.16) is constant except V_a . A plot of $1/C^2$ as a function of applied voltage gives a straight line as indicated in Figure S2.3. Extrapolating this line to $(1/C)^2 = 0$ gives $V_a = V_{bi}$, so the built-in voltage may be determined directly from the experimental plot. Further, the slope of the C - V curve is

$$\frac{d \frac{1}{C^2}}{d V_a} = \frac{-2(N'_D + N'_A)}{A^2 q\epsilon N'_D N'_A} \quad (\text{S2.17})$$

from which

$$\frac{N'_D + N'_A}{N'_D N'_A}$$

can be found, assuming the area of the junction is known. Furthermore, if it is known that $N'_D \gg N'_A$, the slope determines N'_A . If $N'_A \gg N'_D$, it determines N'_D .

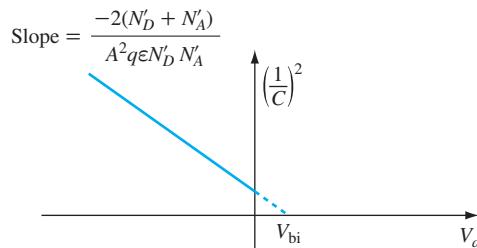


Figure S2.3 The capacitance-voltage characteristic of a prototype (step) pn junction can be used to measure V_{bi} experimentally.

S2.3.2 JUNCTION CAPACITANCE IN A NONUNIFORMLY DOPED JUNCTION

The preceding analysis is valid for a step junction with constant doping on each side, and for a one-sided step junction or a Schottky diode with uniform doping in the semiconductor. The C_j - V_a characteristics are also useful, however, for determining the doping profile for nonuniform doping. Consider the case of Figure S2.4, where a Schottky diode is fabricated on the semiconductor surface, but in which the doping level of the n-type Si varies with distance from the surface. In this case the $1/C^2$ - V_a curve will not be a straight line.

Let $N'_D(w)$ be the net donor doping level at the edge of the transition region (of width w) for an applied voltage V_a across the Schottky diode. Since $C = |dQ/dV_a|$ and

$$dQ = qN'_D(w)Adw \quad (\text{S2.18})$$

we can write

$$C = \left| qN'_D(w)A \frac{dw}{dV_a} \right| \quad (\text{S2.19})$$

But, since $C = \epsilon A/w$, we can express $N'_D(w)$ as

$$\frac{1}{N'_D(w)} = \left| q\epsilon A^2 \frac{1}{C} \frac{d(1/C)}{dV_a} \right| \quad (\text{S2.20})$$

Finally, since

$$\frac{1}{C} \frac{d(1/C)}{dV_a} = \frac{1}{2} \frac{d(1/C^2)}{dV_a} \quad (\text{S2.21})$$

Equation (S2.20) can be expressed as

$$\frac{1}{N'_D(w)} = \left| \frac{q\epsilon A^2}{2} \frac{d(1/C^2)}{dV_a} \right| \quad (\text{S2.22})$$

The doping concentration profile as a function of distance, $N'_D(w)$, can then be found from the C - V_a characteristics and the slope of the experimental $1/C^2$ - V_a

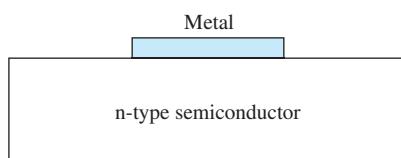


Figure S2.4 A Schottky diode with an n-type semiconductor.

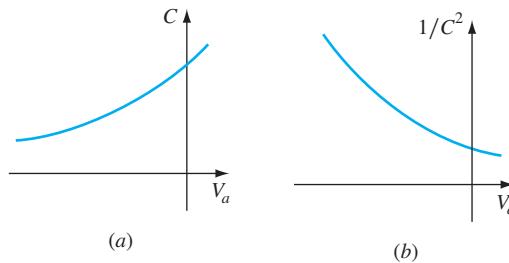


Figure S2.5 The C-V characteristics of a Schottky diode. (a) C versus V_a ; (b) a plot of $(1/C)^2$ as a function of applied voltage.

plot by using Equation (S2.22). This is illustrated in Figure S2.5. In (a) the capacitance is plotted as a function of voltage, and in (b), $1/C^2-V_a$ is plotted. To find w at a given voltage, the distance that the depletion region extends into the n-type material can be determined from the capacitance and Equation (S2.13). The doping concentration at that position is determined from the slope of the $1/C^2-V_a$ characteristic and Equation (S2.22). The built-in voltage cannot be obtained for this case. This method is still quite useful for measuring the uniformity of semiconductor material, however, and automatic systems are available (or easy to make) to plot $1/C^2$ as a function of applied voltage and do the necessary calculation to determine the $N_D'-x$ relation.

S2.3.3 VARACTORS

The variation of junction capacitance with voltage is the basis of a class of devices called *variable capacitance diodes* or *varactors*. They are used for tuning circuits, frequency modulating radio signals, frequency conversion, and parametric amplification. [1]

Consider a one-sided n⁺p junction in which the depletion region is predominantly on the p side. Let the doping be represented by some arbitrary functional form

$$N_A' = B x^m \quad (\text{S2.23})$$

where at the metallurgical junction $x = 0$. Solving Poisson's equation

$$\frac{d^2 V}{dx^2} = \frac{-q N_A'}{\epsilon}$$

with the boundary conditions $V(x = 0) = 0$ and $V(x = w) = V_{bi} - V_a$, where $x = w$ is at the edge of the transition region, gives an expression for the junction width [1, 2]

$$w = \left[\frac{\epsilon(m+2)(V_{bi} - V_a)}{qB} \right]^{1/(m+2)} \quad (\text{S2.24})$$

The capacitance is then found to be

$$C = \left[\frac{qB\epsilon^{m+1}}{(m+2)(V_{bi} - V_a)} \right]^{1/(m+2)} \\ = B^* (V_{bi} - V_a)^{-1/(m+2)} = B^* (V_{bi} - V_a)^{-s} \quad (\text{S2.25})$$

where we have combined the constants into a single constant B^* . We define

$$s = \frac{1}{m+2} \quad (\text{S2.26})$$

where s is called the *sensitivity*. The larger s (the smaller m), the larger the capacitance variation with applied (reverse) voltage. For linearly graded junctions, $m = 1$ and $s = \frac{1}{3}$. For step (prototype) junctions, $m = 0$ and $s = \frac{1}{2}$.

An interesting case is that for $m = -\frac{3}{2}$. In this situation, $s = 2$ and

$$C \propto \frac{1}{(V_{bi} - V_a)^2} \quad (\text{S2.27})$$

If such a capacitance is used with an inductor L in a tuned circuit, the resonant frequency f_r is

$$f_r = \frac{1}{2\pi\sqrt{LC}} \propto (V_{bi} - V_a) \quad (\text{S2.28})$$

and varies linearly with applied voltage. This considerably simplifies the design of tunable circuits.

S2.3.4 STORED-CHARGE CAPACITANCE OF SHORT-BASE DIODES

The junction capacitance of short-base diodes is the same as for their long-base counterparts. The stored-charge capacitance, however, is much reduced from that of a long-base diode because the stored minority carrier charge in the short-base region is appreciably less than in the long-base diode.

Consider the n^+p short-base diode of Figure S2.6 where the p region is uniformly doped and $W_B \ll L_n$. The stored charge consists of the total excess charges in the short base. The electron concentration at the junction edge is $n_p(x_p)$ and at the contact end is n_{p0} . Thus, since recombination is negligible, the total stored excess charge under this distribution is

$$Q_s = \frac{qA[n_p(x_p) - n_{p0}]W_B}{2} \quad (\text{S2.29})$$

and the current is

$$I = \frac{qAD_n[n_p(x_p) - n_{p0}]}{W_B} \quad (\text{S2.30})$$

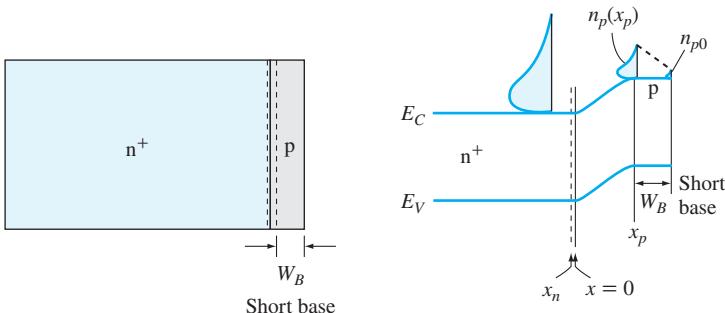


Figure S2.6 A short-base n^+ p junction showing the charge distribution that creates the stored-charge capacitance.

Since this is minority carrier diffusion current, D_n is the minority carrier diffusion coefficient.

From Equations (S2.29) and (S2.30), we have

$$\frac{Q_s}{I} = \frac{(W_B)^2}{2D_n} \quad (\text{S2.31})$$

or the stored charge is

$$Q_s = I \frac{(W_B)^2}{2D_n} = I\tau_T \quad (\text{S2.32})$$

Here we have defined the quantity

$$\tau_T \equiv \frac{Q_s}{I} = \frac{(W_B)^2}{2D_n} \quad (\text{S2.33})$$

which is called the *minority carrier transit time* or simply *transit time*.

The reclaimable stored charge Q_{sr} is, as in Chapter 5, that fraction of the stored charge that contributes to the actual capacitance:

$$Q_{sr} = \delta Q_s = \delta I\tau_T \quad (\text{S2.34})$$

Therefore the stored-charge capacitance C_{sc} is [3]

$$C_{sc} = \frac{dQ_{sr}}{dV_a} = \delta \frac{dQ_s}{dV_a} = \delta I\tau_T \frac{q}{kT} \quad (\text{S2.35})$$

EXAMPLE S2.2

Compare the stored-charge capacitance for long-base and short-base n^+p diodes with $N_A = N_A = 10^{17} \text{ cm}^{-3}$, $I = 10 \text{ mA}$, and a minority carrier lifetime $\tau_n = 2.9 \times 10^{-6} \text{ s}$ (Figure 3.21). Let $W_B = 0.3 \mu\text{m}$ for the short-base diode. The factor δ is $\frac{1}{2}$ for the long-base diode and $\frac{2}{3}$ for the short-base diode. [3]

■ Solution

For the long-base diode, from Equation (5.116), we have

$$C_{sc} = \delta I \tau_n \frac{q}{kT} = \frac{1}{2} \times 10^{-2} \times 2.9 \times 10^{-6} \times \frac{1.6 \times 10^{-19}}{1.38 \times 10^{-23} \times 300} = 5.6 \times 10^{-7} \text{ F}$$

For the short-base diode, we use Equations (S2.35) and (S2.33) to write

$$C_{sc} = \delta I \tau_T \frac{q}{kT} = \delta I \frac{(W_B)^2}{2D_n} (q/kT) = \frac{2}{3} \times 10^{-2} \frac{(0.3 \times 10^{-4} \text{ cm})^2}{2 \times 20 \text{ cm}^2/\text{s}} \times (1/0.026) = 5.8 \times 10^{-12} \text{ F}$$

where we found the minority carrier diffusion length $D_n = 20 \text{ cm}^2/\text{s}$ from Figure 3.11.

For this example, the stored-charge capacitance of the long-base diode is about 10^5 times as large as for the short-base diode, because of its much larger stored charge.

S2.4 SECOND-ORDER EFFECTS IN SCHOTTKY DIODES

In Chapter 6, a first-order model for a Schottky diode was discussed. In this model it was assumed that diode current was thermionic in nature. All carriers with sufficient x -directed energy to surmount the barrier $E_B(0)$ at the metal-semiconductor interface do so. This is illustrated again in Figure S2.7. Part (a) shows the energy band diagram at equilibrium for an n-channel Schottky diode. At equilibrium the net current flow is zero, and the current density J_0 from metal to semiconductor is equal to that flowing from semiconductor to metal as indicated in (a). For forward bias (b), with V_a positive, the barrier from metal to semiconductor, $E_B(0)$, is unchanged while the barrier from semiconductor to metal is reduced by the factor qV_a (V_a positive). In (c) the energy band diagram is shown for reverse bias (V_a negative). Here too, the barrier from metal to semiconductor is unchanged but the barrier from semiconductor to metal is increased. This result is the current-voltage relation given in Equation (6.14), which is repeated here for the first-order model.

$$J = J_0(e^{qV_a/kT} - 1) \quad (\text{S2.36})$$

where J_0 is given by

$$J_0 = \frac{qm^*(kT)^2}{2\pi^2\hbar^3} e^{-E_B(0)/kT} \quad (\text{S2.37})$$

There are two important second-order effects, however, that need to be considered in practical Schottky diodes. These are tunneling through the barrier and barrier lowering due to image effects.

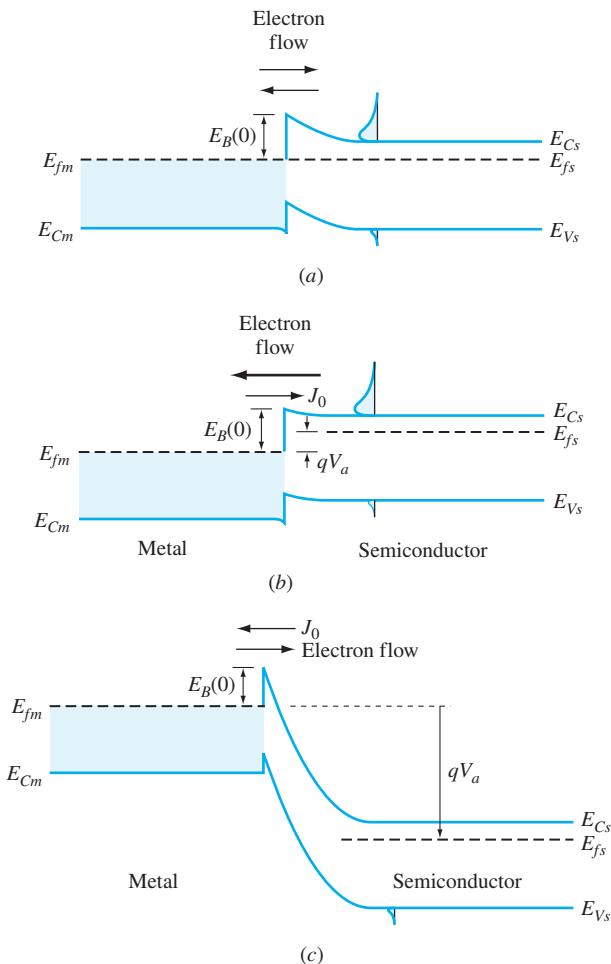


Figure S2.7 First-order current flow in an n-Schottky barrier under (a) equilibrium, where no net current flows; (b) forward bias, where the barrier for electrons is reduced, going from the semiconductor to the metal, but not from the metal to the semiconductor, resulting in a net electron flow from semiconductor to metal; (c) reverse bias. To first order, electrons are assumed to cross the barrier thermionically.

S2.4.1 TUNNELING THROUGH SCHOTTKY BARRIERS

As indicated in Chapter 6, there is a probability that electrons will tunnel through the depletion region in a Schottky barrier. We first treat the case of a rectifier in which the semiconductor is lightly doped (nondegenerate) and then the case of low-resistance metal-semiconductor contacts in which the semiconductor is degenerate.

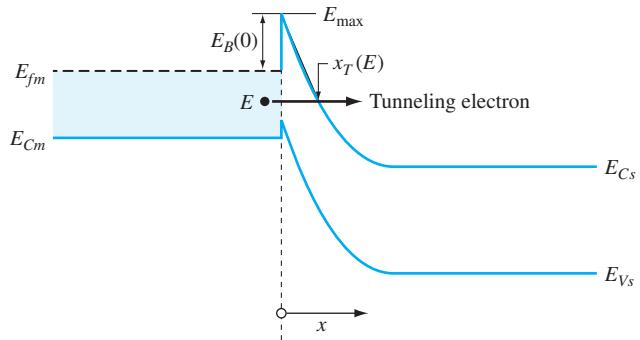


Figure S2.8 In a reverse-biased Schottky diode, electrons can tunnel through the metal-semiconductor barrier. The Schottky barrier is assumed to be close to triangular. The tunneling distance is x_T .

For either case, the probability that an electron at energy E with x -directed velocity tunnels into the semiconductor is

$$T = e^{-2 \int_0^{x_T} [(2m^*/\hbar^2)(E_C(x)-E)]^{1/2} dx} \quad 0 \leq x \leq x_T \quad (\text{S2.38})$$

where x_T is the tunneling distance at energy E and E_C is the conduction band edge in the semiconductor.

Tunneling in Schottky Diodes We first consider the case of a Schottky rectifier. Consider the reverse-biased Schottky diode of Figure S2.8, where x_T is the tunneling distance at energy E , and where the Schottky barrier is assumed triangular. For such a triangular barrier, $x_T = (E_{\max} - E)/q|\mathcal{E}|$ and the tunneling probability becomes

$$T = e^{-4\sqrt{2m^*(E_{\max}-E)^{3/2}}/3q\hbar|\mathcal{E}|} \quad (\text{S2.39})$$

The tunnel current is proportional to the electron concentration in the metal at energy E . For $E > E_{fm}$, the electron concentration decreases exponentially with E but the tunneling probability decreases exponentially with $(E_{\max} - E/q|\mathcal{E}|)^{3/2}$, or it increases as $n(E)$ decreases. The total tunnel current is dependent on \mathcal{E} , which depends on the bias and the doping concentration in the semiconductor. Tunneling through this triangular barrier is referred to as *Fowler-Nordheim tunneling*. Except at high temperatures where the electron concentration is appreciable at higher energies, most of the tunnel current occurs for $E \approx E_{fm}$ or $E_{\max} - E = E_B(0)$. The Fowler-Nordheim tunnel current can then be approximated as [4]

$$J_{FN} \approx \frac{q^2 |\mathcal{E}^2|}{16\pi^2 \hbar E_B(0)} e^{-4\sqrt{2m^*(E_B(0))^{3/2}}/3\hbar q|\mathcal{E}|} \quad (\text{S2.40})$$

where \mathcal{E} is the field in the semiconductor and m^* is the effective mass of the tunneling electrons.

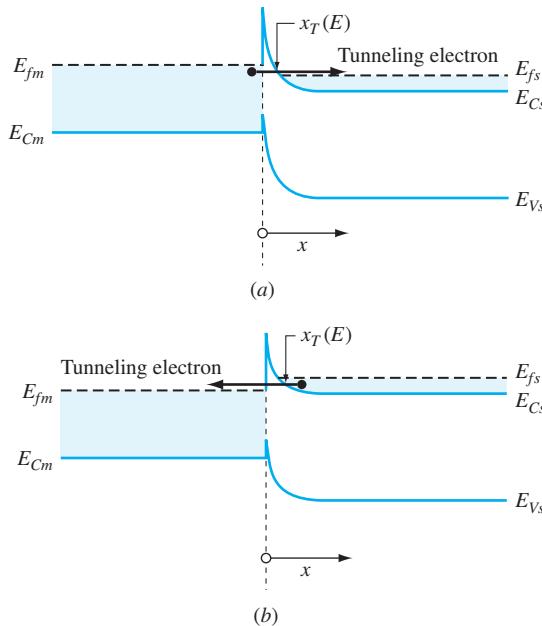


Figure S2.9 For a metal-semiconductor junction in which the semiconductor is degenerate, the tunnel distance is short, resulting in a large current for small reverse (a) or forward (b) bias. This results in a small contact resistance.

Tunneling in Low-Resistance Contacts For a metal-semiconductor contact in which the semiconductor is degenerate, the Fermi level in the semiconductor is within the conduction band, and the depletion region at this energy is narrow enough to result in a small x_T and thus a large tunneling probability at any energy. In this case, the current is large for both forward and reverse bias. Figure S2.9 indicates the tunneling for a small reverse bias (a) and for small forward bias (b). In (a), a large current flows for a small reverse bias (in the millivolt range), while in (b) a large current flows for a small forward bias. The heavier the doping, the smaller the tunneling distance and the smaller the junction resistance.

S2.4.2 BARRIER LOWERING IN SCHOTTKY DIODES DUE TO THE IMAGE EFFECT

In addition to the tunneling described above, the current in a Schottky barrier is also modified from the first-order prediction by the image effect, which lowers the barrier. This arises because an electron in the conduction band of an n-type Schottky diode a distance x from the metal interface (Figure S2.10a). There will be an attractive

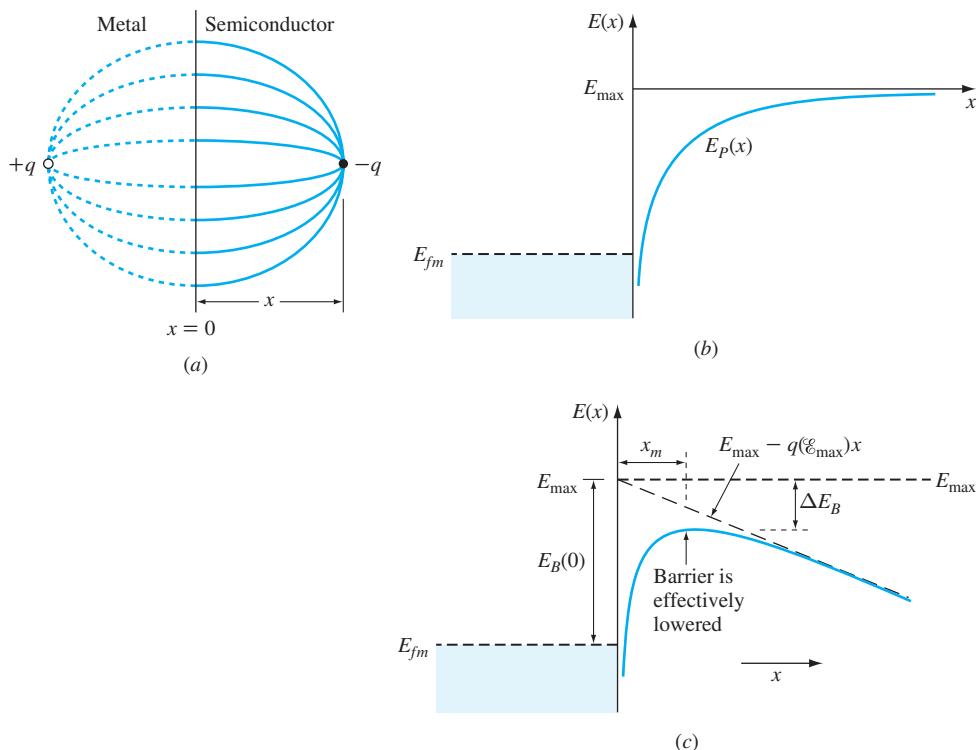


Figure S2.10 An electron in the semiconductor conduction band a distance x from the metal surface creates a positive surface charge in the metal. (a) The field in the semiconductor is equal to that caused by an imaginary $+q$ charge a distance x inside the metal. (b) The resultant potential energy in the semiconductor. (c) This combined with the field in the semiconductor depletion region lowers the metal-semiconductor barrier by an amount ΔE_B .

force between the electron and the metal surface. At the interface the resultant field must be perpendicular to the metal, since the metal is assumed to be a perfect conductor. The field lines in the semiconductor then are equivalent to those produced by considering the metal to be a semiconductor with a charge $+q$ a distance x inside the metal to balance the charge $-q$ of the electron. This $+q$ charge is referred to as an *image*. The equivalent field is indicated in the figure.

The force on the electron by its image charge is given by Coulomb's equation

$$F = \frac{-q^2}{4\pi\epsilon_s(2x)^2} = -\frac{dE_P}{dx} \quad (\text{S2.41})$$

and the electron potential energy due to the image effect is found by integrating E_P from finite x to infinity, which gives

$$E_P(\text{image}) = E_{\max} - \frac{q^2}{16\pi\epsilon_s x} \quad (\text{S2.42})$$

as indicated in Figure S2.10b. The variation of potential energy with distance produces an electric field. There is also an electric field associated with the depletion region, which varies with distance into the semiconductor. The image effect, however, is concentrated near the interface, so we can assume that the electric field due to the semiconductor depletion region is equal to its maximum value $|\mathcal{E}_{\max}|$, as indicated by the dashed line in Figure S2.10c. The potential energy of an electron at position x becomes

$$E_P(x) = -q|\mathcal{E}_{\max}|x - \frac{q^2}{16\pi\epsilon_s x} \quad (\text{S2.43})$$

which has a maximum value at $x = x_m$ where

$$x_m = \left| \frac{q}{16\pi\epsilon_s |\mathcal{E}_{\max}|} \right|^{1/2} \quad (\text{S2.44})$$

and the change in the barrier height due to the image effect is

$$\Delta E_B = q \left| \frac{q|\mathcal{E}_{\max}|}{4\pi\epsilon_s} \right|^{1/2} \quad (\text{S2.45})$$

For a value of $|\mathcal{E}_{\max}| = 10^5$ V/cm = 10^7 V/m and $\epsilon_s = 11.8\epsilon_0$ (e.g., Si), $\Delta E_B \approx 0.036$ eV, and occurs at about 7 nm from the interface. Note that this distance is an order of magnitude greater than the influence of the interface dipoles.

Since this image-induced barrier lowering increases the current density for a Schottky diode, the diode equation becomes

$$J_0 = \frac{qm^*(kT)^2}{2\pi^2\hbar^3} e^{-(E_B(0)-\Delta E_B)/kT} \quad (\text{S2.46})$$

The value of $|\mathcal{E}_{\max}|$ (and thus J_0) depends on doping level and applied voltage:

$$|\mathcal{E}_{\max}| = \left[\frac{2qN_D(V_{bi} - V_a)}{\epsilon_s} \right]^{1/2} \quad (\text{S2.47})$$

S2.5 SUMMARY

In this supplement to Part 2, “Diodes,” we investigated some additional topics in relation to diodes. The dielectric relaxation time is the time required for charge neutrality to be reestablished after a sudden change in carrier concentration. An example is when excess carriers are injected across a junction. For injected minority carriers, within a time constant τ_D , majority carriers move into the region to compensate the charge and level out the band edges. For injected majority carriers, τ_D is the time constant associated with ejecting these carriers from the material via the ohmic contacts.

We also reviewed how capacitance measurements can be used to investigate some junction parameters. For a prototype (step) junction, the $C-V_a$ characteristics can be used to experimentally measure the built-in voltage. They can also

determine the relationship between the doping on either side of the junction. If the junction is one-sided, the doping on the lightly doped side can be found. In nonuniformly doped, one-sided junctions, the doping profile can be obtained.

The stored-charge capacitance of step short-base diodes (the width of the more lightly doped side much less than a minority carrier diffusion length) was found to be orders of magnitude smaller than that of the prototype long-base diode because of the much reduced minority carrier stored charge.

Tunnel current in Schottky diodes was briefly discussed along with the current in low-resistance metal-semiconductor contacts. We saw that tunneling increases the current, and that the image effect lowers the metal-semiconductor barrier, which also increases the current.

S2.6 REVIEW QUESTIONS

1. Explain why the dielectric relaxation time in a semiconductor is greater for injection of minority carriers than for majority carriers.
2. Sketch a circuit for measuring the $C-V_a$ characteristics of a pn junction.
3. Why is the stored-charge capacitance negligible in a reversed pn junction?
4. To make a low-resistance metal contact to an n-type semiconductor, the semiconductor surface is doped degenerately n type, and in a p-type semiconductor the surface is made degenerate p type. Explain why.
5. The image effect in a Schottky diode in an n-type semiconductor lowers the barrier for electrons. What would be the effect in a Schottky diode using a p-type semiconductor?
6. Would an analogous barrier effect be present in an N^+n heterojunction with semiconductors of different dielectric constants? Explain your answer.

S2.7 REFERENCES

1. M. H. Norwood and E. Shatz, "Voltage variable capacitance tuning—A review," *Proc. IEEE*, 56, pp. 788–798, 1968.
2. S. M. Sze, *Physics of Semiconductor Devices*, 2nd ed., Chapter 2, John Wiley & Sons, New York, 1981.
3. Joseph Lindmayer and Charles Y. Wrigley, *Fundamentals of Semiconductor Devices*, Chapter 2, D. Van Nostrand, Princeton, NJ, 1965.
4. M. Lenzlinger and E. H. Snow, "Fowler-Nordheim tunneling into thermally grown SiO_2 ," *J. Appl. Phys.*, 40, pp. 278–283, 1969.

S2.8 PROBLEMS

- S2.1** Explain why the dielectric relaxation time in a semiconductor is greater for injection of minority carriers than for majority carriers.

- S2.2** A prototype (step) junction has $N'_D = 2 \times 10^{16} \text{ cm}^{-2}$, $N'_A = 5 \times 10^{15} \text{ cm}^{-3}$, and $V_{bi} = 0.71 \text{ V}$. Plot the C - V characteristic and $1/C^2$ versus V_a for this junction if the area of the junction is 10^{-4} cm^2 . Plot for $V_a = -5 \text{ V}$ to -0.5 V .
- S2.3** The C - V characteristic is measured for a silicon diode that is heavily doped on the n side and lightly doped on the p side. The experimental results are shown in Figure PS2.1. Find the value of the built-in voltage and N'_A . The junction size is $10^{-4} \times 10^{-4} \text{ cm}^2$.

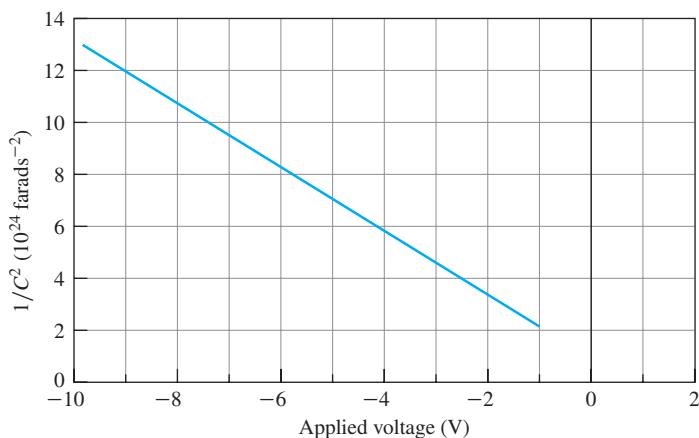
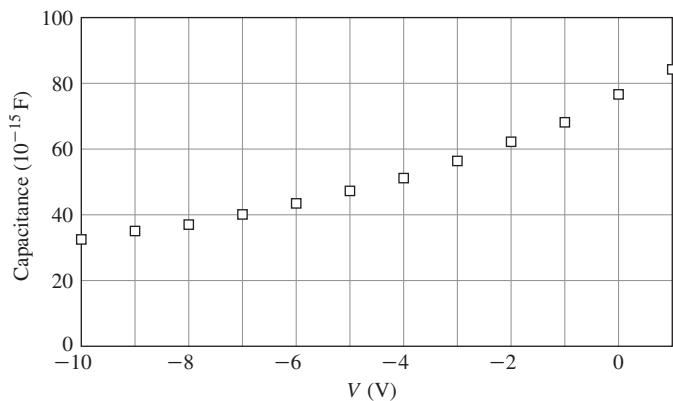


Figure PS2.1

- S2.4** The C - V measurements for a particular junction are given in Figure PS2.2. Plot the junction width as a function of applied voltage, and plot the doping concentration N'_D as a function of junction width w . Let the junction area be $5 \times 10^{-7} \text{ cm}^{-2}$.



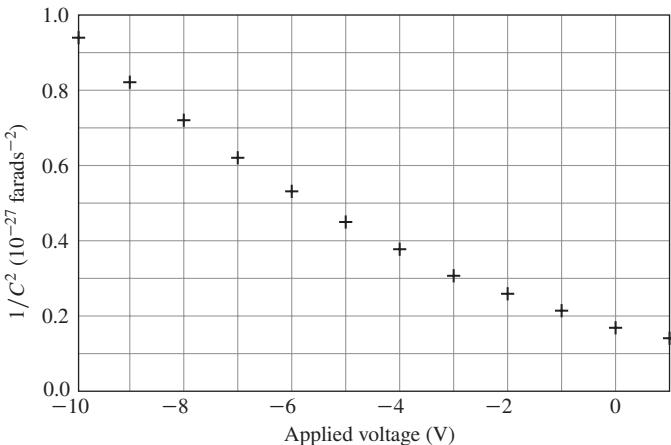


Figure PS2.2

- S2.5** In Section S2.3.3, we discussed varactors and mentioned a specific case of interest in which the resonant frequency of a circuit using an inductor with the tunable capacitance of a junction was proportional to the applied voltage. In some circuits using junction capacitance as a tuning element, however, it is convenient to have the capacitance vary linearly with applied voltage. Find the required doping profile (N'_D versus x) of the deposited epitaxial layer.
- S2.6** A prototype Si pn diode has $N'_A = 10^{17} \text{ cm}^{-3}$ and $N'_D = 10^{16} \text{ cm}^{-3}$. The junction area is 10^{-5} cm^2 . Find the junction capacitance at:
- $V_a = 0$
 - $V_a = +0.6 \text{ V}$
 - $V_a = -5 \text{ V}$.
- S2.7** Find the stored-charge capacitance for a short-base n⁺p diode in which the p-region width is $0.2 \mu\text{m}$, $N'_A = 10^{16} \text{ cm}^{-3}$, and $I = 5 \text{ mA}$. The reclaimable charge in a short-base diode is 2/3.
- S2.8** Find the energy barrier lowering at $V_a = -5 \text{ V}$ of a Si Schottky diode with $N'_A = 10^{17} \text{ cm}^{-3}$ and a built-in voltage of 0.6 V.
- S2.9** Find the dielectric relaxation time for p-type GaAs with $N_A = 2 \times 10^{17} \text{ cm}^{-3}$.
- S2.10** Why is the stored charge capacitance negligible in a reverse biased pn junction?
- S2.11** To make a low resistance metal contact to an n type semiconductor, the semiconductor surface is doped degenerately n type, and in a p-type semiconductor the surface is made degenerate p type. Explain why.
- S2.12** The image effect in a Schottky diode in an n-type semiconductor lowers the barrier for electrons. What would be the effect in a Schottky diode using a p-type semiconductor?

Field-Effect Transistors

Up to now, we have discussed semiconductor materials and two-terminal devices. By far the most important semiconductor devices, however, are transistors, which are three- or four-terminal devices. Transistors have two very useful modes of operation: they can be amplifiers and they can be switches.

When a transistor is used as an amplifier, as in analog circuits, the current or voltage at one terminal controls the current or voltage between the other terminals. A small change in the control signal (the electrical equivalent of turning a knob) can produce large changes in the output signal; thus, the small signal is amplified.

In the digital mode, the signal at the control terminal of the transistor controls the state of the switch. A change in the input is the electrical equivalent of throwing a lever—the input controls whether or not current can pass through the transistor.

Any transistor can operate as either an amplifier or a switch; it depends only on the surrounding circuitry. We leave the discussion of circuit design to another course, but here we focus on the physics of operation of the transistors themselves.

There are two major classes of transistors, based on the physics of their operation. These are the field-effect transistors (FETs) and the bipolar junction transistors (BJTs). The origin of the names will become clear as we develop an understanding of how these devices work.

Interestingly, the field-effect transistor was invented first, but the bipolar junction transistor was the first to be developed into a practical device. For many years, bipolar transistors predominated. More recently, however, FETs have surpassed BJTs in ease of fabrication and low cost, and currently most electronic circuits use FETs as the fundamental circuit elements.

There are many types of field-effect transistors, but they all have in common the element that an electric field across a *gate* structure controls the flow of current between the other two terminals, the *source* and the *drain*. The differences between the various types of FETs are primarily in the structure of the gate and mechanism used to apply the field.

THE GENERIC FET

Before we begin to analyze specific types of field effect transistors, we examine a generic device to understand the basic principles of operation. A simplified perspective view of a field effect transistor is shown in Figure III.1. There are four terminals; the first three are called the source (S), drain (D), and gate (G). There is also a fourth terminal, the body (B) or substrate, which is often connected to the source and thus not always shown. There is an electrically conducting *channel* extending from the source to the drain. The gate is above but electrically isolated from the channel. As we will see, the voltage on the gate terminal with respect to the source, V_{GS} , is used to control the passage of carriers through the channel by varying the field in the insulating region. The mechanism of this control depends on the particular type of FET.

In Figure III.1a the source, channel, and drain are n type and are fabricated in a p-type semiconductor (substrate or body). This device is referred to as an n-channel field-effect transistor, or NFET. In the p-channel device (PFET) in Figure III.1b the source, channel and drain are p type, and the substrate is n type. The figure shows an insulated gate as an example, since this structure is the most common. The insulated gate FET is covered in Chapter 7.

Note that in both cases a pn junction exists between the FET and the substrate. In practice, this junction must never be forward biased. This is because the current should flow between the drain and the source, not into the substrate. In addition, if multiple FETs are put onto the same substrate, as long as the

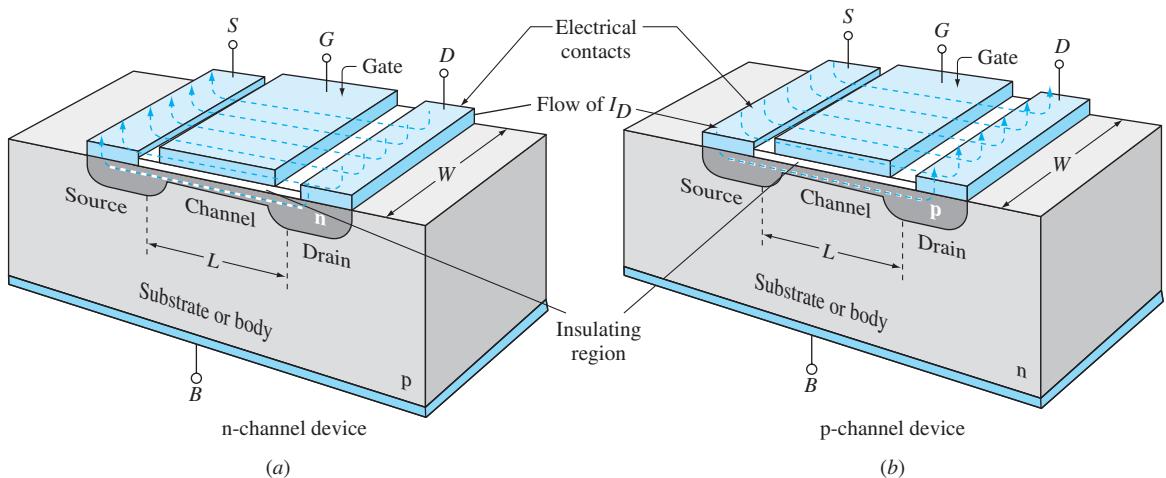


Figure III.1 A generic field-effect transistor contains a source S , a drain D , and a channel controlled by a gate electrode G . Although this is a generic device, the gate is shown for a particular class: the insulated gate field transistors. Other specific gate structures will be discussed later. (a) An n-channel device; (b) p-channel FET. Direction of conventional current flow through the channel is in the direction of the dashed arrows. Here W is the channel width and L is the channel length.

junctions are not forward biased, then in effect the FETs are all electrically isolated from each other.

We will see later that virtually no dc current flows into or out of the gate. The application of a voltage on the gate produces an electric field that affects the conductance of the channel, but the gate itself does not conduct.

Since the carriers (electrons or holes) cannot flow into the substrate and cannot flow through the gate, they are confined to the channel, and under appropriate bias conditions, these carriers can flow between the source and the drain, producing a current in the channel. We refer to this current as I_D , the current at the drain, but it is also the channel current and the source current.¹ The dashed arrows in Figure III.1 indicate the path of the current flow. In an n-channel FET the channel carriers are electrons and the drain voltage with respect to the source, V_{DS} , is positive. Electrons flow from source to drain, so the direction of current is from drain to source, as shown in the figure.

In a p-channel FET, the carriers are holes and the drain-to-source voltage is negative. Therefore, the holes flow from the source to the drain. The current I_D by convention is defined as positive going from drain to source and is therefore negative.

An example of a FET used in a simple circuit is shown in Figure III.2. The FET, M1, is an n-channel metal-oxide-semiconductor field-effect transistor (MOSFET) discussed in more detail in Chapter 7. The drain supply voltage V_{DD} is in series with the load resistor R_L and the channel connects the drain D and

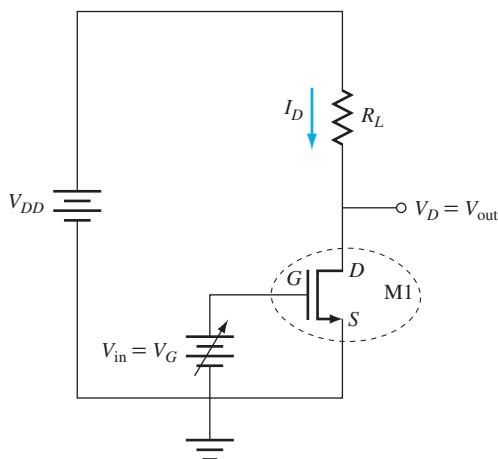


Figure III.2 Simple (inverter) circuit for an NFET. The input voltage V_G controls the channel current I_D and the output voltage V_D .

¹Of course, some leakage current does flow between channel and gate and between channel and substrate, but this leakage current is normally small compared with the channel current between drain and source.

source S . Recall that the conductivity of a semiconductor is controlled by the concentration of carriers. Since in this illustration, the source is at ground (zero) potential, the source voltage $V_S = 0$, the gate-to-source voltage $V_{GS} = V_G$, and the drain-to-source voltage $V_{DS} = V_D$. The electron concentration in the channel is controlled by the gate (input) voltage $V_{in} = V_G$. The output voltage of the circuit is $V_{out} = V_D$. From Kirchhoff's voltage law,

$$V_{out} = V_{DS} = V_D = V_{DD} - I_D R_L$$

For V_G such that $I_D = 0$, $V_{out} = V_{DD}$. For V_G such that I_D is very large, $V_{out} \approx 0$. (We will see later that when the drain current I_D is very large, the voltage between the drain and the source is small.) For intermediate values of V_G , $0 < V_{out} < V_{DD}$.

Figure III.3a shows the typical electrical characteristics of an NFET. Here the current through the channel, I_D , is plotted as a function of the voltage across the channel V_{DS} , for various values of the controlling gate voltage V_{GS} . There are three regions of operation: the sublinear region, the saturation region, and the subthreshold region. The subthreshold region is the horizontal line marked $V_{GS} \leq V_T$. We define the threshold voltage V_T as the value of V_{GS} required to initiate a given current flow through the channel, often taken (in the current saturation region) as $I_{Dsat} = 40W/L$ nA, although other definitions are used. For all gate voltages below this threshold, I_D is small and often considered to be zero, regardless of the value of the drain-source voltage V_{DS} . That means all the I_D - V_{DS} curves for $V_{GS} < V_T$ lie close to the horizontal axis.

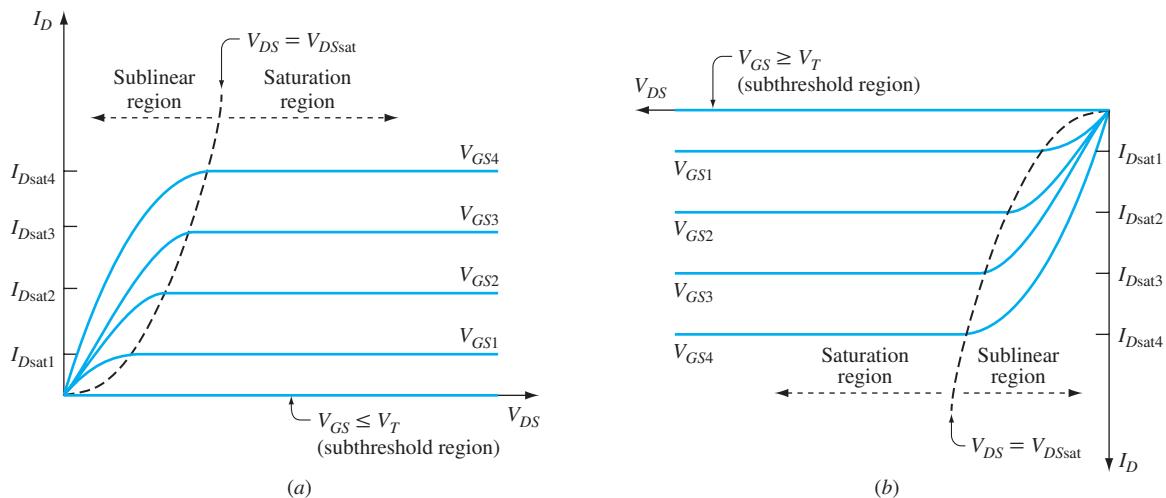


Figure III.3 The I_D - V_{DS} characteristics for a typical NFET (a) and PFET (b) for five values of V_{GS} . The dashed line separates the sublinear and saturation regions at $V_{DS} = V_{DSsat}$. In the NFET, the drain-to-source voltage and drain current are positive, and the gate-to-source voltage must be higher than the threshold voltage V_T for appreciable current to flow. In the PFET, the drain-to-source voltage and drain current are negative, and the gate-to-source voltage must be lower than the threshold for current to flow.

When the gate voltage V_{GS} is above threshold, current can flow. In an NFET, for $V_{GS} > V_T$ and V_{DS} positive, electrons flow from source to drain, or current flow is from drain to source.

Next, look at one of the curves on the plot. Notice that for a given V_{GS} above threshold ($V_{GS} > V_T$), as V_{DS} increases, the I_D - V_{DS} relation is sublinear and eventually I_D saturates. The value of saturated I_D is called $I_{D\text{sat}}$ while the value of V_{DS} at which the current saturates is $V_{DS\text{sat}}$. These are both functions of the gate voltage V_{GS} and the properties of the FET.

The region for $V_{DS} > V_{DS\text{sat}}$ is referred to as the *current-saturation region*, or simply the *saturation region*. The region for $V_{DS} < V_{DS\text{sat}}$ is called the *linear region* (although it is sublinear), or the *sublinear region*. The dashed line in Figure III.3a indicates the division between these two regions. For higher gate voltages, the value of $I_{D\text{sat}}$ and $V_{DS\text{sat}}$ are both larger.

Typical characteristics for a PFET are shown in Figure III.3b. Here holes flow (and thus current flows) from the source to the drain for negative values of V_{DS} . For the PFET, the gate voltage is less than (more negative than) the threshold voltage for a finite I_D .

In Part 3 of this book, we examine the I - V characteristics in some detail. As an introduction here, however, let us consider a geographical analogy to help understand why the curves are shaped as they are. Suppose we have two deep lakes, connected to each other by a shallow canal as illustrated schematically in Figure III.4a. The bottom of the lake system is somewhat analogous to potential energy for electrons as a function of position while the depth of the water represents the electron concentration. The lake on the left (S , representing source) is considered to be at constant depth. Four cases for the lake on the right (D , drain) are shown. In case 1, the surface of D is the same height as the surface of S and no water flows between the lakes. This is indicated by position 1 in Figure III.4b, which shows a plot of water flow versus the difference in surface levels of the two lakes. When the surfaces are equal, the flow is zero. In case 2, the surface of D is slightly below that of S and so some water flows from S to D . Case 3 represents the situation in which the surface of D is at the same height as the bottom of the canal. The slope of the

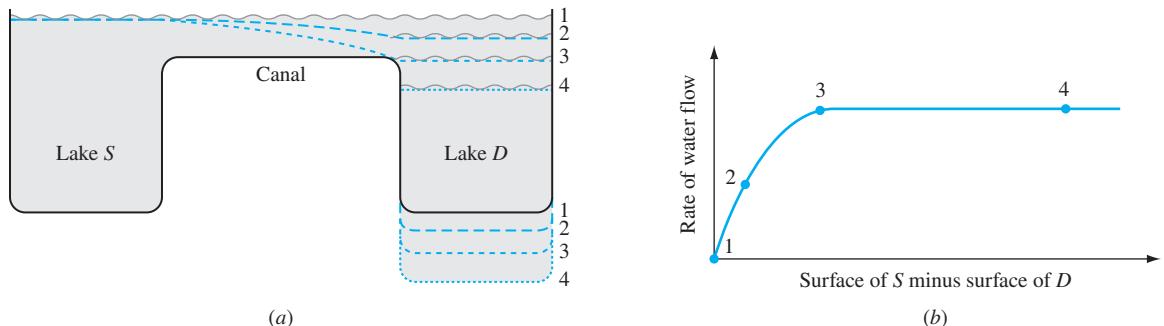


Figure III.4 Lake analogy to FET operation. (a) Two lakes are connected by a canal. The water level of lake S is constant. Four values of water level of lake D are shown. (b) The corresponding rates of water flow.

water surface in the canal is increased and so is the rate of flow. For the surface of D below the bottom of the canal (case 4), the slope of the surface in the canal is not affected. Since the rate of water flow is determined by the slope of the water surface in the canal, the rate of flow saturated at its value at 3. As the level of surface D continues to drop, the rate of water flow through the canal stays the same.

TRANSISTORS IN CIRCUITS

Most of Part 3 of this book will be devoted to understanding and deriving the shapes of the I_D - V_{DS} characteristics of FETs. First, though, we will briefly investigate how these transistors can be used as amplifiers and as switches.

We said earlier that in a digital circuit, changing the gate voltage would be like throwing a switch. Figure III.5a shows an NFET inverter circuit with a resistive load. This is the circuit of Figure III.2 repeated. Suppose that a positive voltage supply V_{DD} is applied to the drain through a load resistor R_L . If the input gate voltage V_{GS} is below threshold (logic low), little current can flow through the channel—the switch is open-circuited. Since $I_D \approx 0$, the voltage drop across the resistor $I_D R_L$ is also zero and the output voltage is $V_{DS} \approx V_{DD}$ (logic high) as indicated at the “switch open” position in Figure III.5b. If the gate voltage is changed to a value above threshold (for the NFET), then current flows across the channel—the switch is closed. The amount of current that can flow is now determined by the external circuit as well as by the transistor. The circuit determines the *load line* (dashed in the figure). The resistance in the external circuit causes some voltage drop in V_{DS} as I_D increases. The output voltage is then $V_{DS} = V_{DD} - I_D R_L$.

For the analog case, the gate voltage is kept above threshold, but is varied by a small amount within some range, Figure III.5c. As the gate voltage varies, the current I_D and the voltage V_{DS} also vary proportionally. Usually the variation in the gate voltage is quite small (the difference in gate voltage from one curve to the next in the figure may be a fraction of a volt), while the variation in V_{DS} is considerably larger—perhaps about a volt. Hence, the transistor acts as an amplifier, magnifying the small change in V_{GS} to a large change in V_{DS} .

To optimize the transistor design, then, the engineer will want to have a thorough understanding of the I_D - V_{DS} characteristics and how to shape them. Here in this introduction, we will outline the basic approach for analyzing the drain current as a function of the gate and drain voltages. Then, in Chapter 7 we execute and refine that approach.

THE BASIS FOR DERIVING THE I_D - V_{DS} CHARACTERISTICS OF A FET

To begin our understanding of the drain current, we recall from elementary physics that current is defined as the amount of charge passing through a given area per unit time. We consider the case of an n-channel FET, Figure III.6. In an NFET, electrons carry the current by moving in the positive y direction. We take the x coordinate to be downward, across the channel.

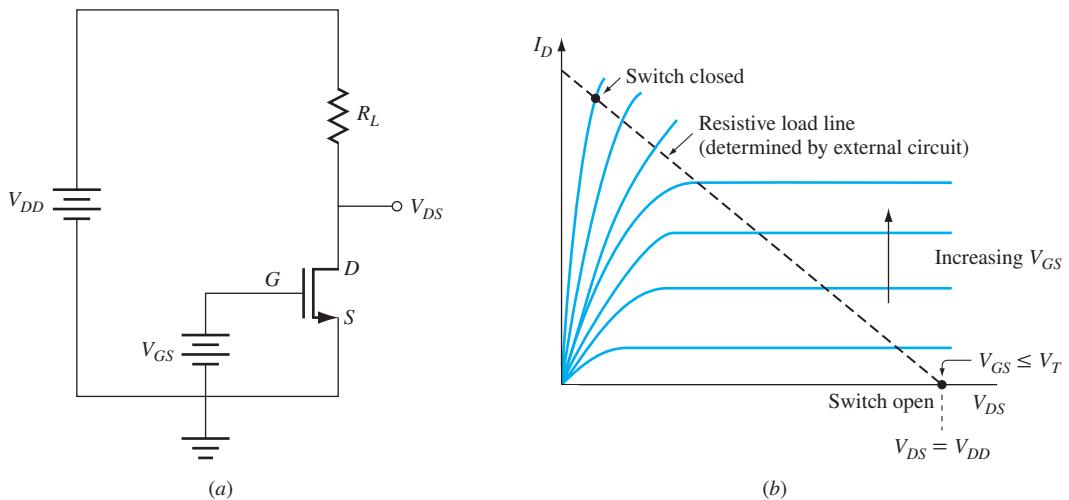


Figure III.5 The I_D - V_{DS} characteristics determine the transistor operation in a circuit. For example, (a) shows a circuit, an inverter. In digital operation, (b) the gate voltage is switched between two values, one above threshold and one below. Current is either near zero, or some value determined by both the transistor (solid lines) and the circuit (dashed line). In an analog circuit (c), the gate voltage remains above threshold, but varies. As V_{GS} changes, the current through the transistor also changes, as well as the voltage V_{DS} .

Let us consider the current flowing across the shaded region. Suppose there are n charges per unit volume in the channel where $n = n(x, y)$. The charge per unit area in the channel at position y is

$$Q_{ch}(y) = -q \int_0^{x_c} n(x, y) dx \quad (\text{III.1})$$

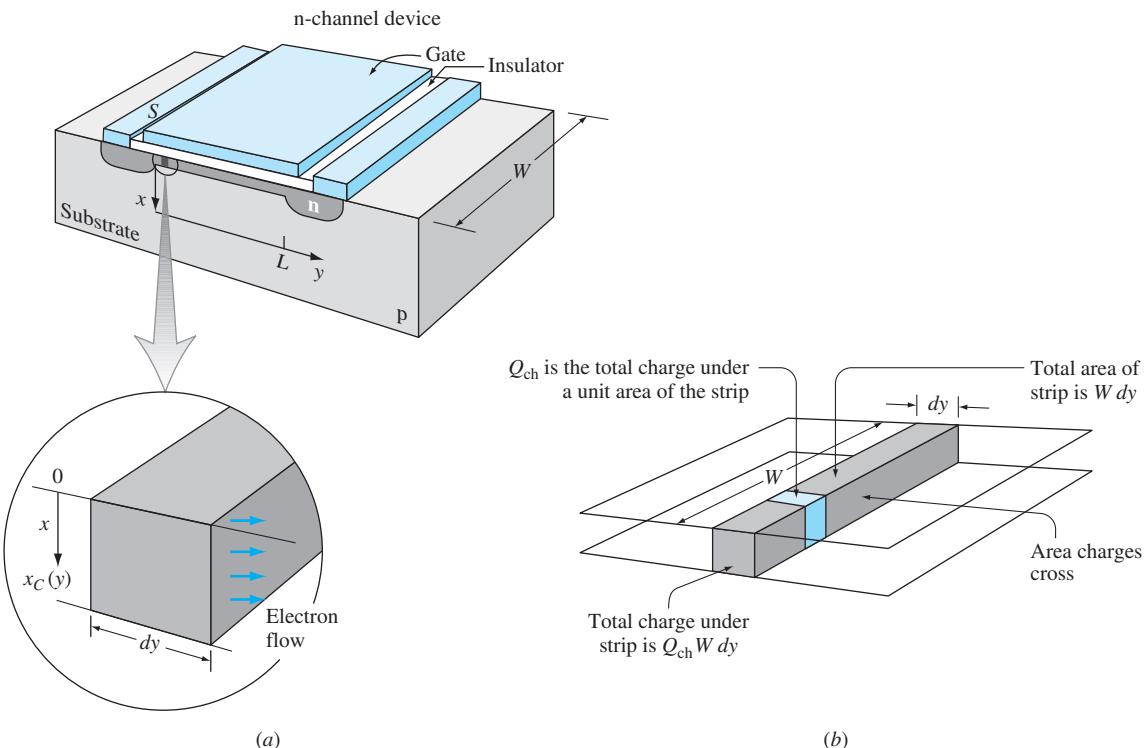


Figure III.6 The geometry of the NFET used for deriving the current. (a) The big picture; (b) the channel charge is the charge per unit area under the strip.

and the integral is taken across the channel depth, x_C as indicated in Figure III.6a. In an incremental length dy at position y in the channel, the total channel charge is

$$(\text{Channel charge in } dy) = Q_{ch}(y)W dy \quad (\text{III.2})$$

Since the current is equal to the charge passing a given area per unit time, the channel current I_D at a given y is

$$I_D = -WQ_{ch}(y)v(y) \quad (\text{III.3})$$

where $v(y)$ is the average channel electron velocity at position y . Note that in Equation (III.3), the velocity v is positive, the charge Q_{ch} is negative, and the negative sign ensures that I_D (from drain to source) is a positive quantity. Note also that I_D is independent of y , so $Q_{ch}(y)$ and $v(y)$ are inversely proportional.

Equation (III.3) is the fundamental equation for current flow in all FETs and is the starting point for deriving the electrical characteristics for any given structure. What is needed, then, are analytical expressions for $Q_{ch}(y)$ and $v(y)$ as functions of applied voltages for a particular FET structure. Once these parameters are modeled, the mathematical analysis is similar for all classes of FETs.

Note that the charge per unit area in the channel, Q_{ch} , Figure III.6b, is analogous to the quantity of water per unit area in the canal of Figure III.4a; i.e., the amount of water per unit volume times the depth of the water. The term v is the average velocity of the water flow at position y . From the water analogy, it is clear that the depth of the water varies along the canal, and similarly the amount of charge available for conduction varies with position in a FET.

In general, current in the channel flows by a combination of drift and diffusion. In practice, however, we are primarily concerned with the operation of FETs in which drift current predominates (*drift transport model* or *drift model*).

The drift current is driven by the electric field along the channel. We call this the longitudinal field \mathcal{E}_L shown in Figure III.7. In an n-channel device, the drain is more positive than the source. The longitudinal electric field then is in the negative y direction. The electrons are accelerated toward the positive terminal, so their velocity is in the opposite direction to the field. Thus for electrons we have

$$v(y) = -\mu(y)\mathcal{E}_L(y) \quad (\text{III.4})$$

where $\mu(y)$ is the average channel mobility (a positive quantity) at position y . Substituting into Equation (III.3), we find the channel current is

$$I_D = WQ_{ch}(y)\mu(y)\mathcal{E}_L(y) \quad (\text{III.5})$$

Electric field is, by definition, $\mathcal{E} = -dV/dy$. The incremental voltage across the channel increment dy is dV_{ch} , as shown in Figure III.7b. Then

$$\mathcal{E}_L = -\frac{dV_{ch}}{dy} \quad (\text{III.6})$$

Equation (III.5) can then be expressed in the form

$$I_D = -WQ_{ch}(y)\mu(y)\frac{dV_{ch}(y)}{dy} \quad \text{or} \quad I_D dy = -WQ_{ch}(y)\mu(y)dV_{ch}(y) \quad (\text{III.7})$$

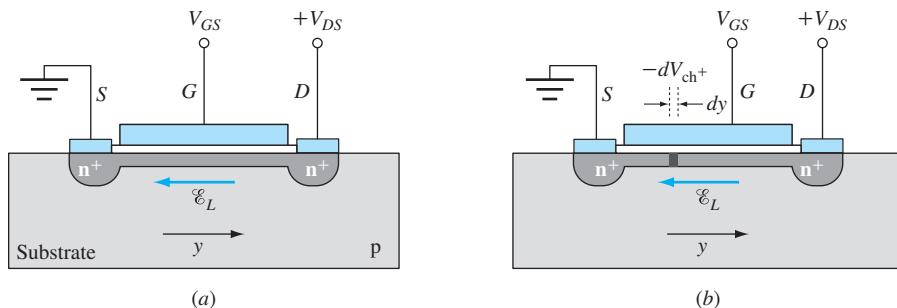


Figure III.7 (a) The longitudinal field \mathcal{E}_L for an NFET. (b) The incremental channel voltage dV_{ch} .

We will use this equation as the starting point for deriving the electrical characteristics of all FETs.

The organization of Part 3, “Field-Effect Transistors,” is as follows:

Chapter 7, “The MOSFET,” is limited to the physical principles of operation of MOSFETs, using n-channel MOSFETs (NFETs) as an example, although the results are readily adapted to p-channel MOSFETs (PFETs).

Chapter 8, “Other Field-Effect Transistors,” begins with a discussion of complementary MOS transistors (CMOSFETs or simply CMOS), which utilizes both NFETs and PFETs manufactured in the same chip. A CMOS inverter circuit is described for digital switching and analog applications. Other FETs are discussed. These include non volatile MOSFETs, or floating gate transistors, silicon on insulator (SOI) MOSFETs, FinFETs (a 3D variation of the planar MOSFET discussed in Chapter 7), heterojunction field-effect transistors (HFETs), metal semiconductor field-effect transistors (MESFETs), junction field-effect transistors (JFETs), and tunnel field-effect transistors (TFETs).

The Supplement to Part 3, “Field-Effect Transistors,” discusses the measurement of electrical parameters used in the description of FET characteristics. Since modern FETs have very short channel lengths (in the tens of nanometers), this makes the devices operate faster, but that affects the physics of operation, too. The short channel effects and their effects on the operating characteristics are discussed. A procedure is described for reducing the size (scaling) of MOSFETs while keeping functional electrical characteristics without exceeding permissible power dissipation and thus device chip temperature. Degradation mechanisms for MOSFETs and chip interconnects are also discussed.

Additional FET topics are covered in online modules. These include low-temperature operation (OM8). Applications of SPICE to MOSFETs are also briefly discussed (OM9). Some examples of MOSFET digital circuits are described (OM7). These include NAND and NOR gates and static random access memories (SRAMs) and dynamic random access memories (DRAMs).

The MOSFET

7.1 INTRODUCTION

In this chapter, we discuss the basic operation of the most important class of FETs: the Si-based insulated-gate field-effect transistor (IGFET). The gate material in these devices was originally a metal (aluminum) and the insulator was silicon dioxide (SiO_2). That is the origin of the term metal-oxide-semiconductor field-effect transistor, or MOSFET. For ease of fabrication and reproducibility, however, a current practice is to use degenerately doped polycrystalline Si (poly-Si), which is highly conductive, instead of metal for the gate (although in modern MOSFETs with very short channels, metals are often used). In n-channel devices, n^+ poly-Si is used for the gates, and p^+ poly-Si is used for p-channel devices. Silicon dioxide was historically used for the insulator, but now nitrogen is often incorporated into the SiO_2 to better passivate the bulk Si surface, increase the dielectric constant, and thus improve device performance. This silicon oxynitride layer is SiO_xN_y or SiON. Other oxides with higher dielectric constants are sometimes used. These insulators are commonly referred to as just “oxide.” While the term IGFET is a more accurate description, in this book we adopt the practice common in the industry, which is to use the more common term MOSFET to describe this class of devices.

7.2 MOSFETS (QUALITATIVE)

In this section, we explore, qualitatively, the basic principles of operation of the MOSFET. To illustrate how the MOSFET channel is formed, a brief qualitative description of MOS capacitors is presented. A more detailed description of MOS capacitors appears in Supplement to Part 3: Additional Considerations for MOSFETs.

7.2.1 INTRODUCTION TO MOS CAPACITORS

In this section we consider an ideal MOS capacitor (MOSC) with the structure of Figure 7.1a. This capacitor consists of a degenerate n^+ Si gate, a thin layer of (insulating) oxide separating a gate from a p-type substrate.

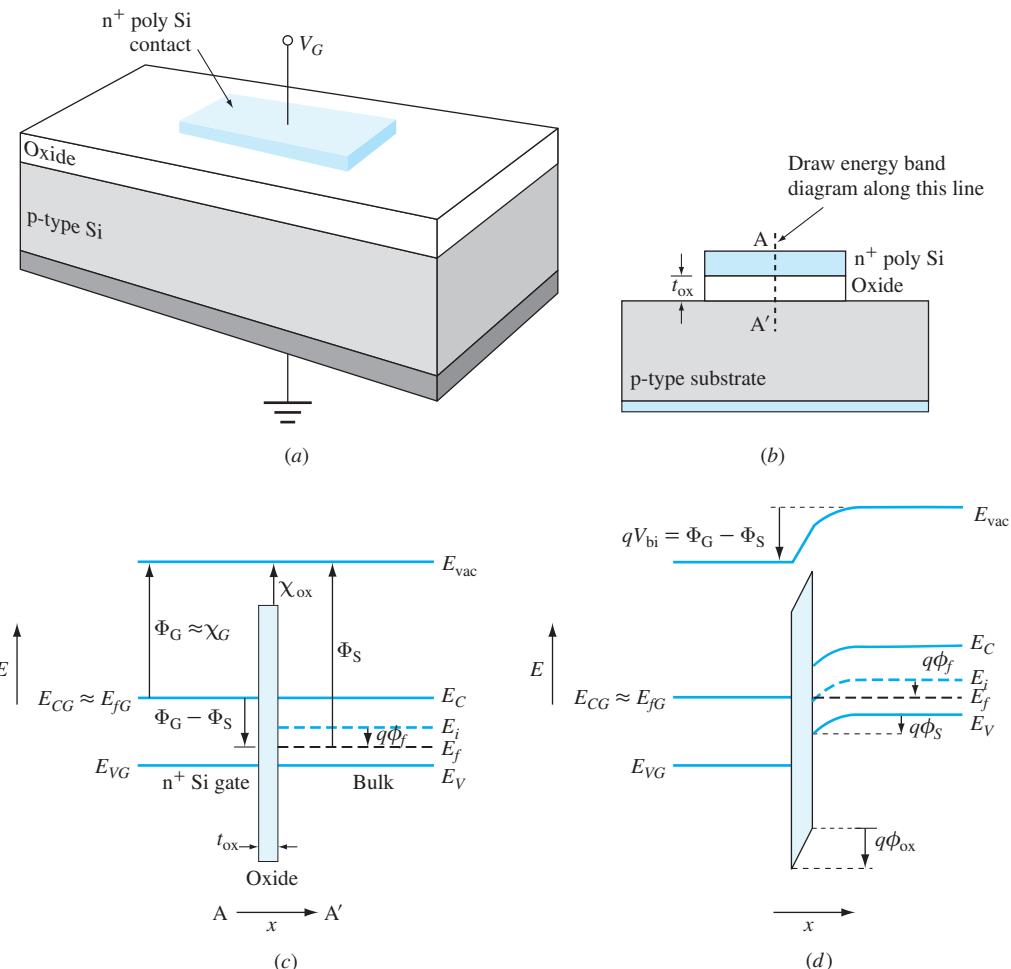


Figure 7.1 The MOS capacitor. (a) Physical structure of an n^+ -Si/oxide/p-Si MOS capacitor; (b) cross section; (c) the energy band diagram under charge neutrality; (d) the energy band diagram at equilibrium (note that the surface of the p-type substrate near the oxide interface has become weakly inverted).

Figure 7.1b shows the cross section of the MOSC. The energy band diagram along the cut A-A' slicing through the gate in the x direction is shown in (c) and (d). We determine the energy band diagrams by following the procedure used in Chapter 6 for heterojunctions. In this case, however, we have two heterojunctions, one between the polysilicon gate and the insulator, and another between the insulator and the semiconductor substrate.

We assume electrical neutrality in every macroscopic region to start, in which case E_{vac} is constant. The result is shown in Figure 7.1c. Here Φ_G and Φ_S

are the work functions¹ (in eV) of the gate and the semiconductor respectively.² Note that Φ_G is approximated as being equal to χ_G because the gate material is degenerately doped and its Fermi level in the gate is thus near the conduction band edge. The quantity ϕ_f is the potential difference (in volts) between the Fermi level and the intrinsic level. The uppercase Φ 's and the χ 's are energies, expressed in eV, and the lowercase ϕ 's are potentials, expressed in volts. On the energy band diagrams, the ϕ 's will always be expressed as $q\phi$, in energy units.

To achieve equilibrium, electrons flow (through an external circuit) from the gate to the semiconductor substrate, causing the Fermi levels to line up. The resulting (equilibrium) energy band diagram is shown in Figure 7.1d. We see that the bands bend, and a built-in voltage results. Some of the voltage is dropped across the oxide and some across the semiconductor. Furthermore, the bulk semiconductor now has a depletion region near its surface—it is depleted of majority carriers, which are holes in this case, since the material is p type.

The total built-in voltage is

$$V_{bi} = \frac{\Phi_G - \Phi_S}{q} = \phi_{ox} + \phi_s \quad (7.1)$$

The voltage drop across the oxide is ϕ_{ox} . The voltage across the Si depletion region, often referred to as the surface potential (i.e., the voltage at the Si surface at its oxide interface relative to that in the neutral bulk), is designated as ϕ_s . Notice that the electric field ($\mathcal{E} = (1/q)(dE_{vac}/dx)$) is discontinuous at both interfaces of the insulator. This is caused by the difference in the permittivities of the materials (Gauss's law, $\mathcal{E}_1\epsilon_1 = \mathcal{E}_2\epsilon_2$).³ The fraction of the built-in voltage appearing across the oxide increases with increasing oxide thickness and with increasing doping level in the Si.

Notice that for this example, at equilibrium (Figure 7.1d), the Fermi level actually crosses the intrinsic level. This means that, while the substrate is *doped* p type, near its surface it is *effectively* n type. The Fermi level near the surface is closer to the conduction band edge than to the valence band edge. At the interface, there is a higher concentration of electrons than holes because of the band bending. The Si surface region is not only depleted, it is *inverted*. We note here that the region consisting of the conductive gate, the insulating oxide, the depletion region, and the substrate can be considered to be a capacitor with

¹Since the original gate material in a MOSFET was a metal, the term Φ_M is also often used for the work function of the gate.

²Note that the gate material is degenerately doped to a degree such that band-gap narrowing occurs. This causes the conduction band edge of the gate material, E_{CG} , to be at a slightly lower energy than that of the substrate, E_{Csub} . The gate material band gap is slightly smaller than the substrate band gap. This band-gap shrinkage is, however, normally less than 0.1 eV and we ignore it. For simplicity, we also ignore any charge trapped in the oxide or at the oxide/Si interface.

³For SiO_2 dielectric and Si substrate, $\epsilon_{\text{Si}} \approx 3\epsilon_{\text{SiO}_2}$, thus $\mathcal{E}_{\text{SiO}_2} \approx 3\mathcal{E}_{\text{Si}}$. In Figure 7.1d and hereafter, for visual clarity the fields at the oxide-semiconductor interface are not drawn to scale.

two different dielectric layers between the electrodes (those of the oxide and of the depletion region), with two different dielectric constants. Note that the built-in voltage is divided between oxide and substrate. If a voltage is applied between gate and substrate, it too will be divided between oxide and substrate and the resultant energy band diagram and charges will be altered.

Figure 7.2 shows the energy band diagrams and charge distributions for various values of gate-substrate voltage V_G for a n^+ Si/oxide/p-type semiconductor. In (a), the case for equilibrium is indicated. Because electrons from the gate transfer to the substrate, with the resultant creation of a depletion region in the Si at the Si/oxide interface, the gate charge is positive and the Si depletion region charge is negative as indicated.

If a negative voltage is applied to the gate, this voltage is divided between oxide and semiconductor. The bands in the semiconductor bend up, causing an accumulation of holes in the semiconductor at the oxide interface (b).

If a positive step voltage (e.g., 2 V) is applied to the gate, a depletion region will be established in the Si within about 3 dielectric relaxation times ($\sim 10^{-12}$ s) resulting in the energy band and charge distribution indicated in (c). However, with time, electrons will be thermally excited into the conduction band where they will be trapped in the potential energy well at the interface. This negative trapped charge (Q_i) raises the energy band in this region until the steady-state condition is reached as indicated in (d), a process that requires on the order of 10 to 100 ms in Si. In steady state, electron generation and recombination rates are equal. Note that in the steady state, the depletion region width is independent of the dc voltage, or Q_B is constant. The charge dependent on voltage resides in the interface charge Q_i . Note also that in (a), (b), and (d), the Fermi level extends to the oxide surface. Since these are steady-state conditions, here E_f represents the quasi-Fermi level for electrons. Because the quasi-Fermi levels for electrons and holes are equal, however (there are no excess carriers), we simply refer to E_f as the Fermi level.

The capacitance-voltage characteristic of a MOS capacitor (MOSC) is an important diagnostic tool for monitoring the fabrication of MOSFETs. Here we present a brief description of the MOSC C - V characteristics.

A simple circuit for measuring the C - V characteristics of a MOSC is shown in Figure 7.3a. A DC voltage V_{DC} and a small-signal ac voltage v_{ac} are applied between gate and substrate, and the ac current i_{ac} is measured. The ratio $i_{ac}/v_{ac} = 2\pi fC$, and C can be calculated.

A typical C - V plot is indicated in (b). Recall that the (differential) capacitance is

$$C = \frac{dQ}{dV} = \frac{\epsilon A}{W}$$

where dQ is the variation of charge on either side of the oxide with a change dV and W is the spacing between the regions of changing charge. For a reverse dc bias (Figure 7.3b) a small change in dV causes a small change in dQ on either side of the oxide and $C = C_{ox} = \epsilon A/t_{ox}$. Thus for this condition, with the area of the gate known, the oxide thickness can be determined.

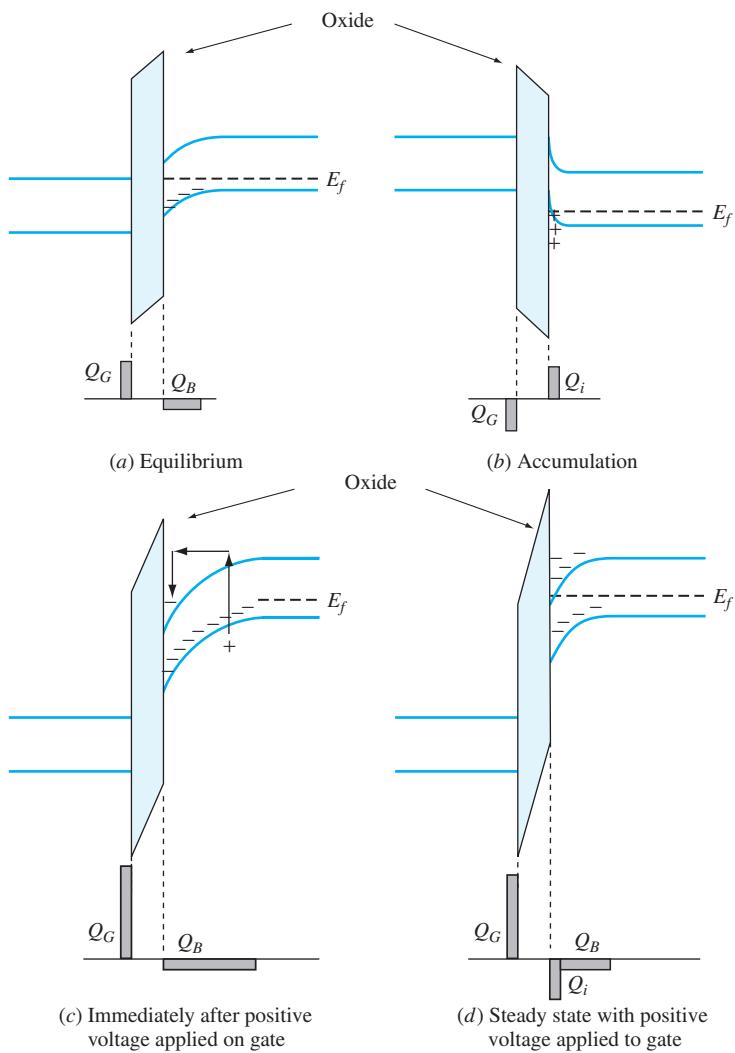


Figure 7.2 Energy band diagrams for the n^+ -Si/oxide/ p -Si capacitor of Figure 7.1 along with the charge distributions for three bias conditions. In (a) the case for equilibrium is indicated. Electrons from the n^+ gate transfer to the p -Si substrate, resulting in a positive gate and a negative depletion region in the substrate. The accumulation condition is indicated in (b). Here a negative voltage is applied to the gate with respect to the substrate such that holes accumulate at the silicon-to-oxide interface. The situation for a positive 2 V step voltage is shown in (c) immediately after the application of the voltage. With time, electrons generated in the transition region are trapped in the potential well at the interface until steady state is reached as indicated in (d).

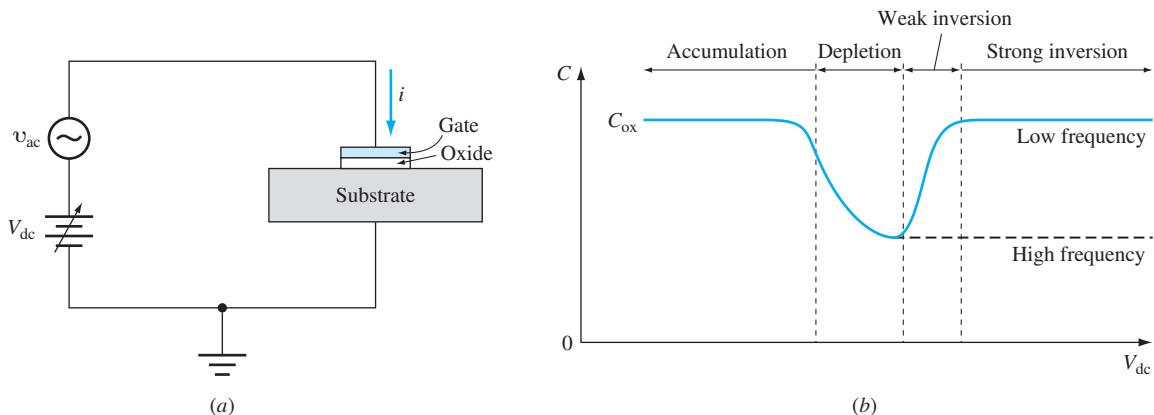


Figure 7.3 (a) Circuit for measuring the capacitance of a MOS capacitor. (b) Capacitance-voltage characteristic for a MOS capacitor at low and high frequencies.

At intermediate voltages, a voltage-dependent depletion region exists in the semiconductor and the MOSC capacitance consists of two capacitors in series.

$$C = \frac{C_{\text{ox}} C_s}{C_{\text{ox}} + C_s}, \quad \frac{1}{C} = \frac{1}{C_{\text{ox}}} + \frac{1}{C_s}$$

where C_s is the semiconductor depletion region capacitance and $C < C_{\text{ox}}$.

At sufficient positive voltage, an inversion layer exists in the semiconductor and for steady-state dc, the depletion layer width is independent of V_{dc} . This is because the Si band bending is enough to locate the Fermi level at the Si-oxide interface slightly within the Si conduction band (see Fig. 7.2d). Because of the high density of states in the conduction band, the Fermi level in the Si at the Si-oxide interface does not shift appreciably with voltage, and to good approximation E_f can be considered constant for strong inversion. Thus in this situation the depletion region width is independent of applied voltage.

An ac voltage creates charge at the edge of the depletion region. If the frequency is low enough (on the order of 1 to 10 Hz), those created electrons have time to enter the inversion region, and the ac influence on the depletion region width is negligible.⁴ Thus on the semiconductor side dQ is at the interface $dQ = dQ_i$ and again $C = \epsilon A / t_{\text{ox}}$

At a high-frequency ac voltage (~100 MHz), the generated electrons have insufficient time to enter the inversion region before the polarity changes, dQ is at the edge of the depletion region, the capacitance is constant, and $C = C_{\text{ox}} C_s / (C_{\text{ox}} + C_s)$. Here $C_s = \epsilon A / w$, where w is the depletion region width and is a measure of the doping level in the substrate.

The MOS capacitance is treated in more detail in Supplement 3.

⁴The low frequency voltage can be thought of as a slowly varying dc voltage which has no effect on the depletion region width.

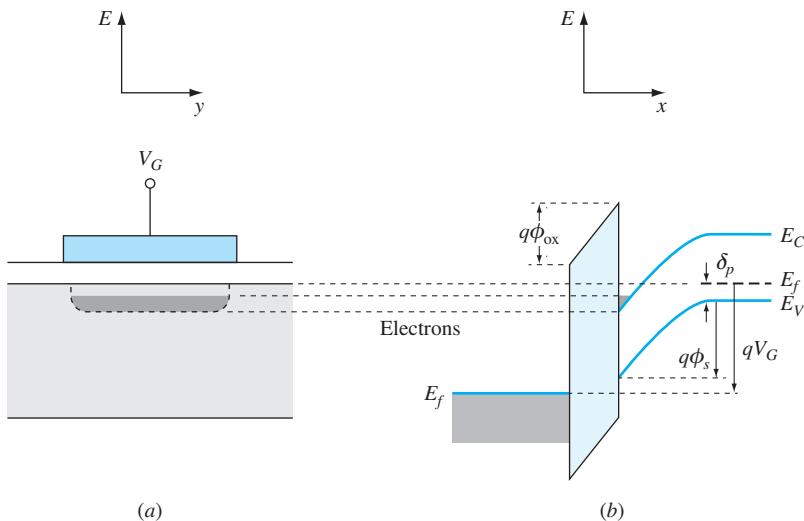


Figure 7.4 (a) Hybrid diagram and (b) corresponding energy band diagram for a MOS capacitor. The shaded areas represent the electron charges in the potential well.

7.2.2 MOS CAPACITOR HYBRID DIAGRAMS

In devices incorporating MOS capacitors, it is often convenient to use “hybrid” diagrams. A hybrid diagram is a combination of the physical diagram and the energy band diagram as indicated in Figure 7.4 for a well partially filled with electrons. While the energy band diagram plots electron energy versus position x , the direction normal to the gate, and the physical diagram plots the lateral structure y versus x , the hybrid diagram, Figure 7.4a, plots lateral position y versus electron energy. The regular energy band diagram is shown in Figure 7.4b.

To construct a hybrid diagram:

- The Fermi level in the bulk semiconductor of the energy band diagram (Figure 7.4b) is aligned with the oxide/Si interface of the physical diagram (a) as indicated.
- The energy at the bottom of the well in (b) is indicated in (a) (the dashed line is the shape of the well).
- The concentration of electrons in the well is indicated in the shaded region of the hybrid diagram.

We use a hybrid diagram to illustrate how the MOS capacitor charges with time. In Figure 7.5a, the hybrid diagram and energy band diagram are shown for a p-type MOS capacitor with no gate voltage applied. At some time $t = 0$, a positive step function voltage V_G is applied to the gate of the MOS capacitor (b). The total voltage across the device then is

$$V_j = V_{bi} + V_G = \phi_{ox} + \phi_s$$

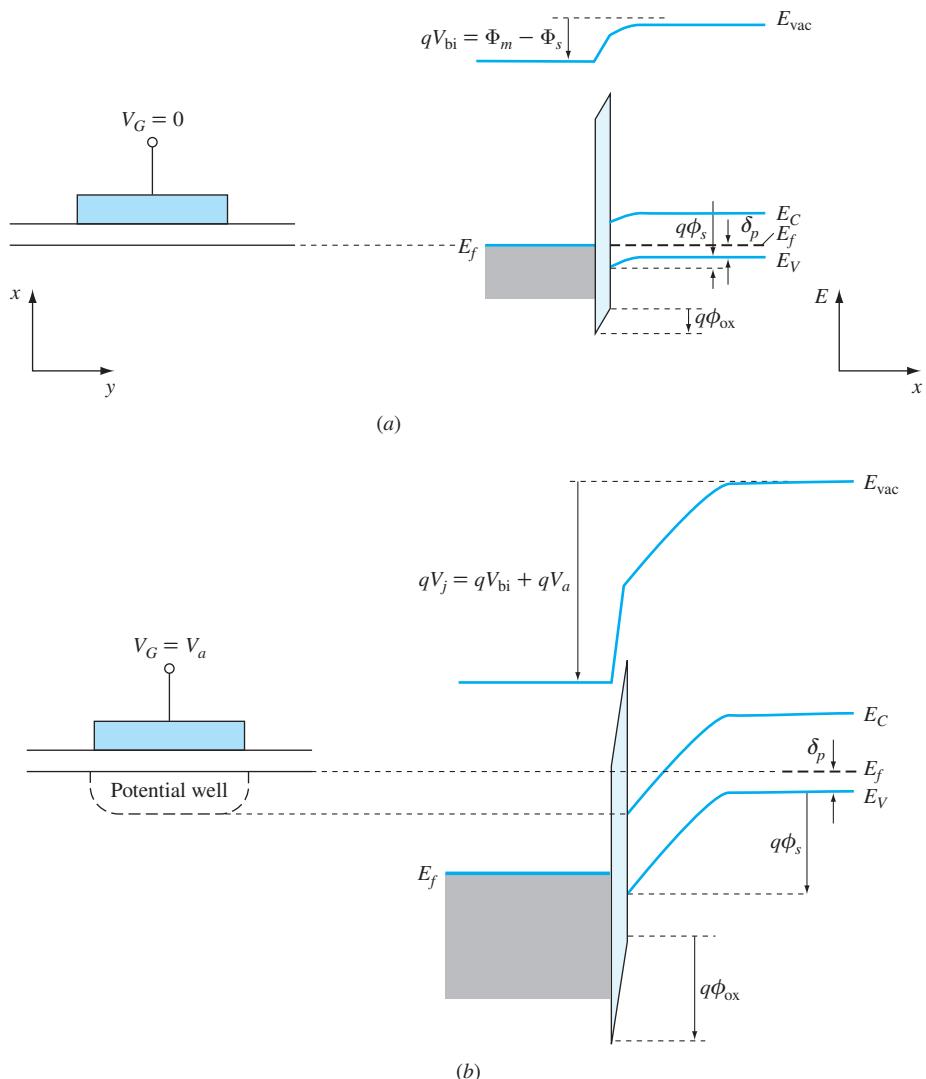


Figure 7.5a-b Hybrid diagram and corresponding energy band diagram for a MOS capacitor with a step voltage applied to the gate. (a) Equilibrium. (b) A positive voltage V_G is applied to the gate. Some of the voltage is dropped across the oxide and some across the semiconductor. At $t = 0$, the capacitor has not yet charged and the potential well is empty.

Immediately after the application of V_G , the potential well is empty, as the capacitor has not yet charged. However, electrons thermally generated in and near the Si depletion region become trapped in the well. The accumulating negative charge causes the energy band diagram near the interface (where the negative charges are) to move upward (higher electron energies). This movement both decreases the surface potential ϕ_s and increases the oxide potential ϕ_{ox} as indicated in Figure 7.5c. When the capacitor is fully charged, the well is filled as

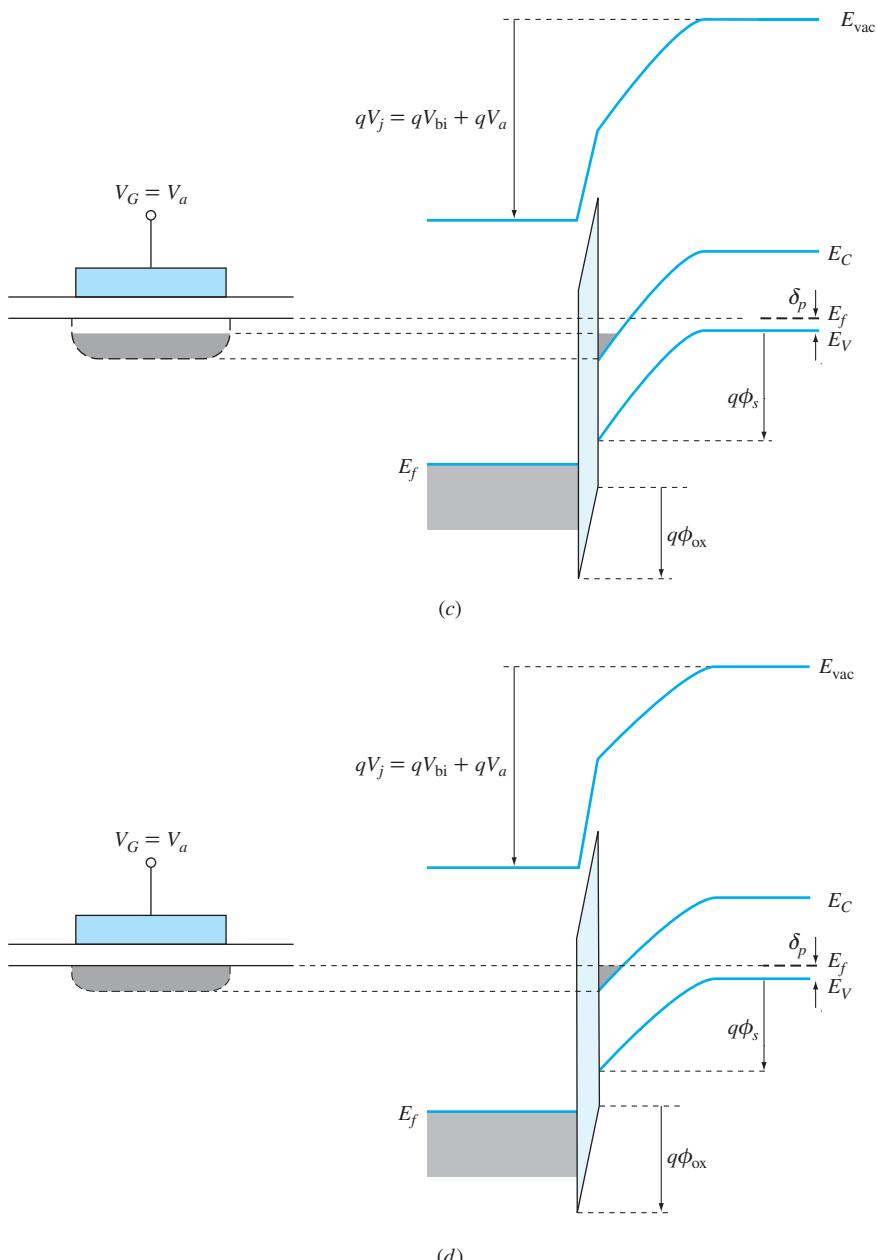


Figure 7.5c-d (c) After a short time the well begins to fill, raising the energy band diagram locally at the semiconductor-oxide interface. Thus ϕ_{ox} increases and ϕ_s decreases. (d) The well is completely full and the capacitor is charged.

pictured in (d). Here the barrier has been reduced such that the rate of electrons entering the well (by thermal generation) is equal to that of those escaping (by recombination).

7.2.3 MOSFETS AT EQUILIBRIUM (QUALITATIVE)

A schematic of an n-channel MOSFET is shown in Figure 7.6. It resembles the MOS capacitor just discussed except that there are source and drain regions of n^+ Si at opposite sides of the gate region. Since electrical connections are made to gate, source, drain, and substrate, a MOSFET is a four-terminal device. Often, however, the substrate is connected to the source, and the MOSFET is then considered to be a three-terminal device. The symbols W and L represent the width and length of the channel.

Figure 7.7a shows the cross section of the MOSFET of Figure 7.6. The equilibrium energy band diagram along the cut A-A' normal to the gate is shown in (b). From the figure, it appears as though there is no n-type channel from source to drain in this device. From the energy band diagram for this structure, we can determine whether a channel in fact exists. For the specific example shown, the band bending in the substrate ($q\phi_s$) is about a half of an electron volt. The equilibrium energy band diagram of (b) is repeated in Figure 7.8, which also indicates the charge in the Si. It can be seen from the figure that the Fermi level is still close to the intrinsic level near the semiconductor-oxide interface. In other words, while a channel does exist, it has a low conductance, since it contains few electrons. Because the electron concentration in the channel is so low, for an applied drain-to-source voltage, only a minuscule current can flow between source and drain. This device is said to be in the *subthreshold* region.

The source and drain are doped n^+ , but the induced channel is only weakly n type. The resulting difference in the electron concentration creates a modest (n^+n) potential energy barrier at either end of the channel. Figure 7.9 shows the equilibrium energy band diagram *along* the channel instead of across it. Since the

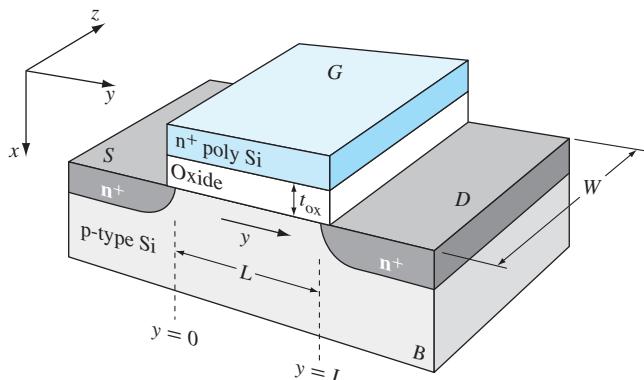


Figure 7.6 Schematic diagram of the structure of an n-channel silicon-based MOSFET. The channel width W , length L , and oxide thickness t_{ox} are shown. The symbols S , G , D , and B represent the source, gate, drain, and substrate (body) respectively.

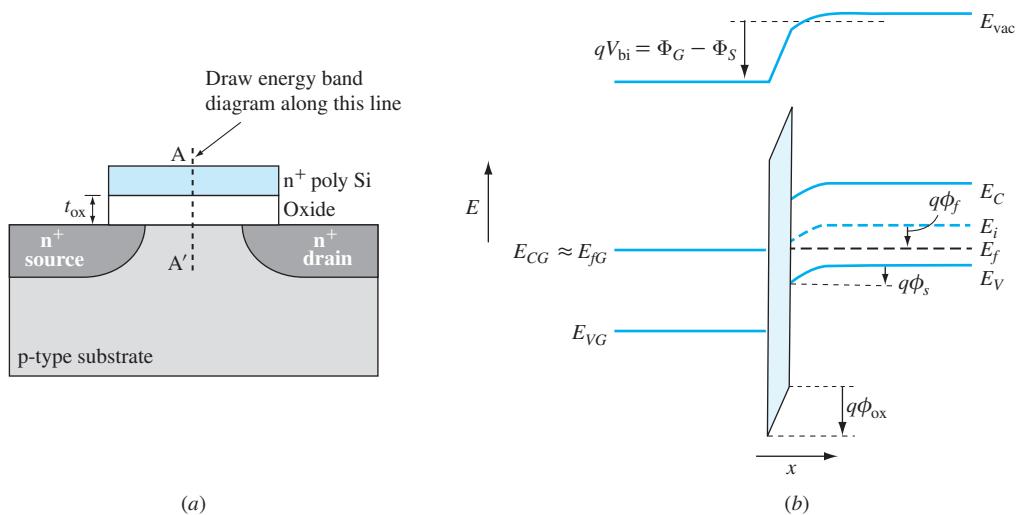


Figure 7.7 (a) Cross section of an n-channel FET; (b) the energy band diagram at equilibrium.

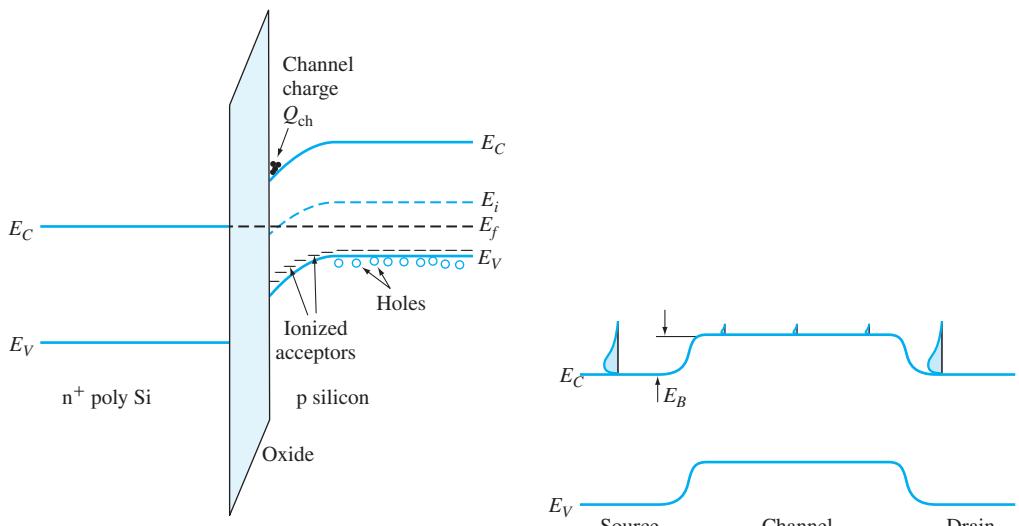


Figure 7.8 The channel charge accumulates in the bulk near the oxide interface. In this case the channel charge consists of electrons.

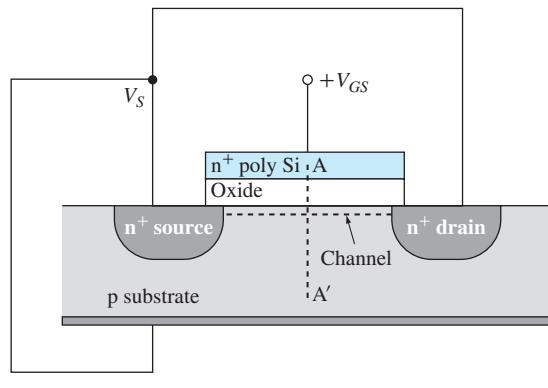
Figure 7.9 The energy band diagram along the channel of the device of Figure 7.8

electron density function $n(E)$ decreases exponentially with increasing energy, the barrier E_B is still large enough that only a small electron concentration exists in the channel. Thus, the induced channel is n type but the channel conductivity is negligible. In the next section, we will see how changing the gate voltage affects the barrier height and thus the channel conductance.

7.2.4 MOSFETS NOT AT EQUILIBRIUM (QUALITATIVE)

So far, we have considered the device to be at equilibrium. Now let us examine the physics of MOSFET operation. We will take the substrate to be connected to the source (a common arrangement). There are still two voltages that can be varied, the gate-source voltage V_{GS} and the drain-source voltage V_{DS} .

The Case for $V_{DS} = 0$ We begin by connecting the source to the drain electrically, such that those two terminals are at the same potential, as shown in Figure 7.10a. Let us then apply a voltage to the gate with respect to the source. This applied voltage is also divided between oxide and substrate, just as the built-in potential was in Figure 7.7. The effect of applying a positive gate voltage is to



(a)

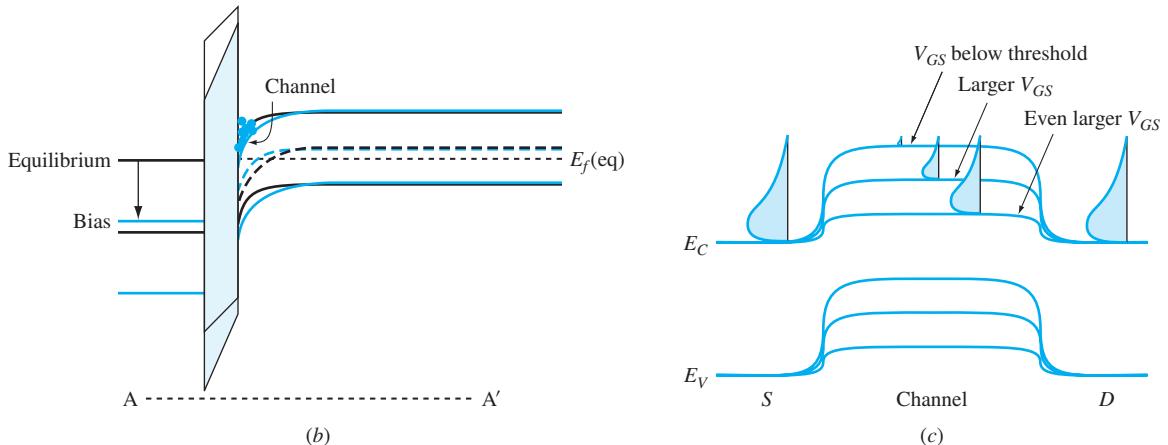


Figure 7.10 A particular MOSFET example, (a) with source, substrate, and drain connected; (b) the energy band diagram along cross section A-A' at equilibrium (black) and under bias (colored); (c) the energy band diagram along the channel for three values of V_{GS} .

lower the channel energy (the conduction band edge), as shown in Figure 7.10b. Here we have drawn the energy band diagram along the line A-A', perpendicular to the gate. The equilibrium diagram is in black, and the energy band diagram under bias is in color. Under bias, the conduction band edge bends down toward the Fermi level. The surface is now more strongly n type than before, and thus more strongly inverted. There are now more electrons in the channel, increasing its conductance. We say that the channel has been “enhanced.”

Another way to look at it is via Figure 7.10c, which shows the energy band diagram along the channel for three different gate voltages. For V_{GS} near threshold, the energy barrier between source and channel at the Si surface, E_B , is fairly high. Few electrons appear in the channel and its conductance is low. As V_{GS} increases above threshold, the barrier decreases. More electrons are able to enter the channel, and thus its conductivity increases.

The channel charge Q_{ch} in the MOSFET channel is analogous to the interface charge Q_i in a capacitor. However, for the capacitor, Q_i is a result of thermal generation and recombination, a relatively slow process. For a MOSFET, the charge is determined by the barriers between source and channel and between drain and channel. The electrons enter and exit the channel by a combination of conduction and diffusion, typically on the order of 10^{-12} to 10^{-11} s, a process normally considered to be instantaneous.

Definition of Threshold We have indicated that, when the gate voltage is below some threshold—i.e., in the subthreshold region—the channel conductance is small. We look at the carrier distributions, and remember that the electron concentration in the semiconductor varies as

$$n = N_C e^{-(E_C - E_f)/kT} \quad (7.2)$$

The electron concentration and thus the conductance of the channel varies exponentially with $E_C - E_f$. This quantity is dependent on the gate-source voltage V_{GS} . An exponential is a smoothly (albeit rapidly) varying function, so it is not clear what value of gate-source voltage should be called *threshold*. A commonly used criterion is that the threshold voltage is the gate-source voltage required to induce an electron concentration at the Si surface that is equal to the hole concentration in the neutral substrate (N'_A).⁵ This is equivalent to saying that a channel exists if the Fermi level in the n channel is as far above the intrinsic level as the Fermi level in the bulk is below the intrinsic level. This condition is shown Figure 7.11. Then

$$\phi_s = 2\phi_f \quad (7.3)$$

where ϕ_s is called the *surface potential*, i.e., the voltage at the Si surface relative to that in the bulk.

⁵As indicated earlier, another often-used criterion is that V_T is the value of V_{GS} required to produce a saturation current $I_{Dsat} = 40 \text{ W/L nA}$.

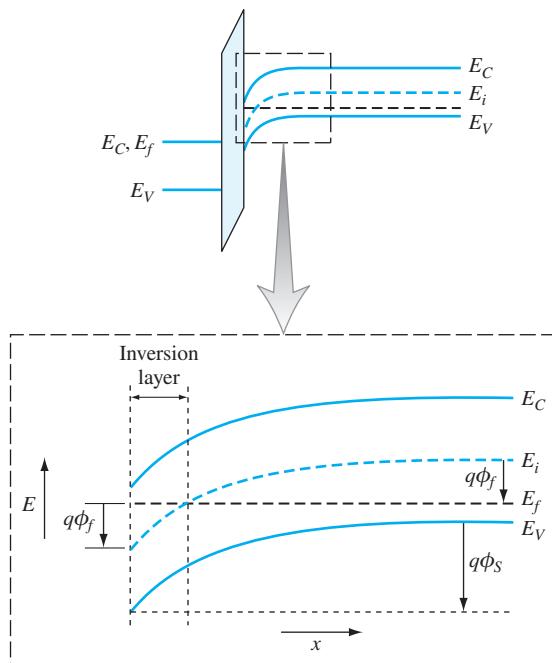


Figure 7.11 The band bending at the surface of the semiconductor. At threshold, the Fermi level is as far above the intrinsic level at the surface (left-hand edge) as it is below the intrinsic level in the bulk.

Enhancement- and Depletion-Type MOSFETs In our earlier example (Figure 7.7), the band bending caused the surface to be n type even with no bias applied. Depending on the doping in the substrate, the surface of the semiconductor may or may not be inverted at equilibrium. If a channel does exist, it may or may not be strong enough to be considered conductive. In the earlier example, the surface was inverted at equilibrium but not enough to be considered a proper channel. With positive gate-source voltage, the channel conductance increased. As mentioned before, the conductance of the channel is *enhanced* by the application of a positive gate voltage. An *enhancement-type* FET is normally **off**. It does not conduct appreciably until a channel is created by the application of a gate voltage.

Figure 7.12a and b shows an enhancement-type NFET and an enhancement-type PFET (i.e., n-channel FET and p-channel FET respectively). Since most FETs currently in use are MOSFETs, we use the terms NFET and PFET respectively for n-type and p-type MOSFETs where appropriate. In the NFET, a positive gate-source voltage must be applied for the transistor to conduct appreciably. For the PFET, the conduction band edge must be bent *upward* to more fully invert the surface such that more holes can enter the channel. This requires

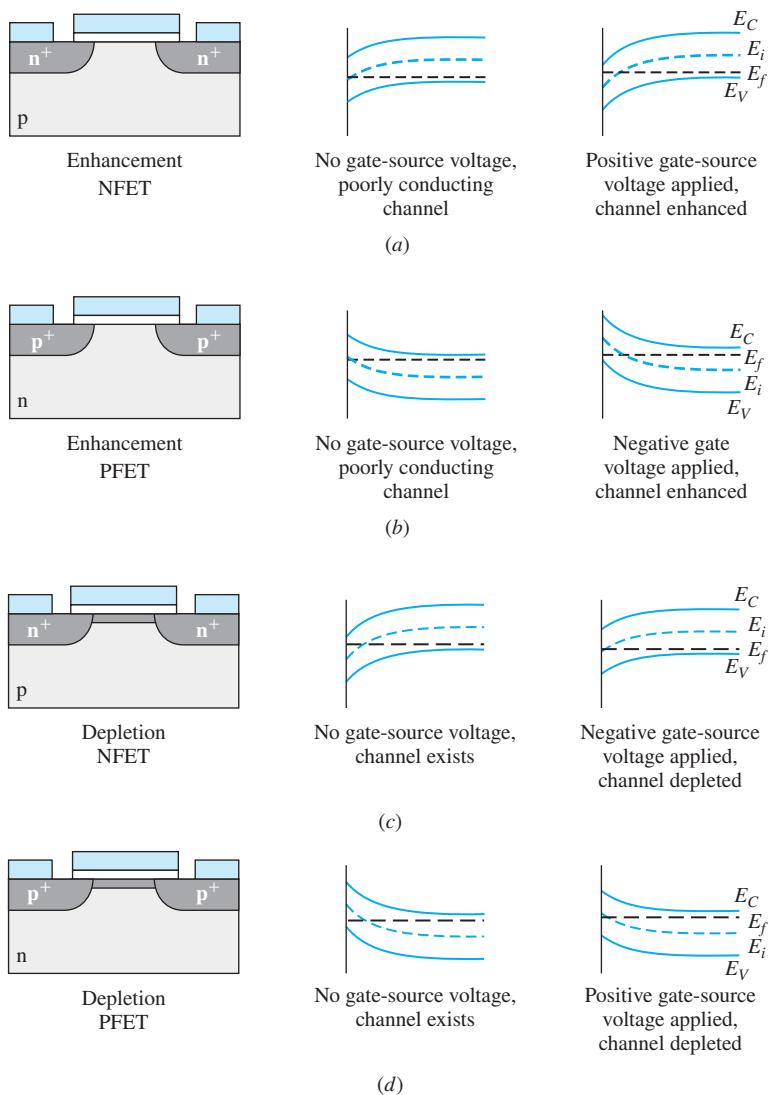


Figure 7.12 The energy bands in the semiconductor for (a) the enhancement NFET; (b) the enhancement PFET; (c) the depletion NFET; (d) the depletion PFET.

a negative voltage on the gate with respect to the source. Thus, for enhancement devices, the threshold voltage of an NFET is positive and the threshold voltage of a PFET is negative. Increasingly negative gate voltage in the PFET causes increasing channel conductance.

It is possible, however, for the device to be fabricated such that a good channel does exist even with no gate-source voltage applied, as shown in Figure 7.12c

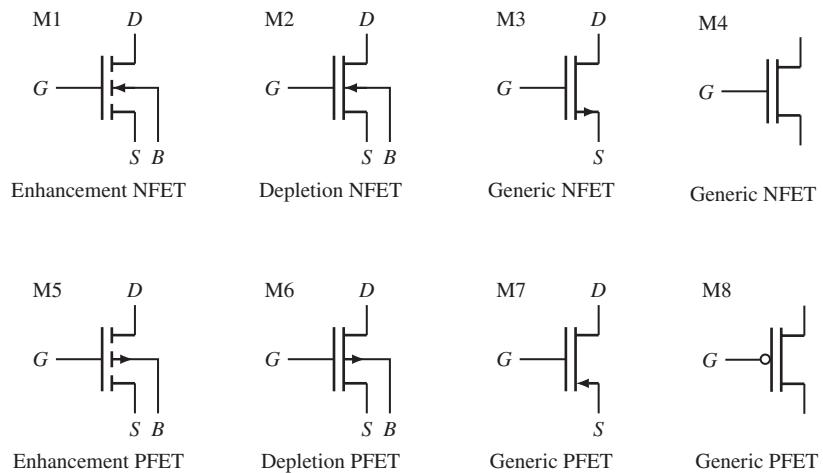


Figure 7.13 Schematic representations (circuit symbols) for MOSFETs. M1 is an enhancement-mode NFET, M2 is a depletion-mode NFET, and M3 and M4 can be used for either type. P-channel MOSFETs are represented by M5 (enhancement), M6 (depletion), and M7 and M8 (either mode).

and d. In other words, a conducting channel exists even at equilibrium. These transistors are normally **on**. In these devices, one has to apply a gate-source voltage to decrease the band bending, deplete the channel, remove the carriers, and thus turn off the conduction. These are called *depletion devices*. While depletion FETs are used in analog circuits, enhancement-type MOSFETs are used in most digital circuits.

Various symbols are used to represent MOSFETs in circuit schematics. Figure 7.13 illustrates some common symbols; M1, M2, M3, and M4 represent n-channel MOSFETs,⁶ while M5, M6, M7, and M8 represent p-channel MOSFETs.

The devices M1 and M5 represent NMOS and PMOS enhancement devices respectively. The broken line representing the channel between source and drain indicate that no conducting channel exists for $V_{GS} = 0$. The arrow between substrate (also called the body, B) and channel points in the direction from p to n as for a diode. Depletion NMOS and PMOS devices are indicated by M2 and M6 respectively. The solid (nonbroken) channel indicates that a conducting channel exists for $V_{GS} = 0$.

Devices M3 and M4 represent NMOS devices while M7 and M8 represent PMOS devices. Symbols M4 and M8 represent n-channel and p-channel enhancement devices respectively, in which current can flow in either direction in the channel depending on circuit conditions. In this case each output terminal acts as either source or drain (e.g., static random access memories, SRAM).

⁶It is common practice to designate MOSFETs by the letter M.

The “bubble” in the gate terminal of M8 indicates a p-channel device. The substrate or body connection is not normally shown in these symbols. It is assumed to be connected to a voltage that does not forward bias any pn junctions (source to body and drain to body) to minimize current flow. The body voltage also affects the threshold voltage, and this is a factor in its choice.

More About Threshold Let us look at the threshold conditions more closely. The concentration of electrons at the surface of the semiconductor is

$$n_s = N_C e^{-(E_{Cch} - E_f)/kT} \quad (7.4)$$

where E_{Cch} is the energy of the conduction band edge in the channel at the Si surface, Figure 7.14. We see also that the barrier height is equal to $E_B = E_{Cch} - E_f$ since in the heavily doped source $E_f \approx E_C$ (see Figure 7.14b). The concentration of electrons in the channel at the surface can be expressed as

$$n_s = N_C e^{-E_B/kT} \quad (7.5)$$

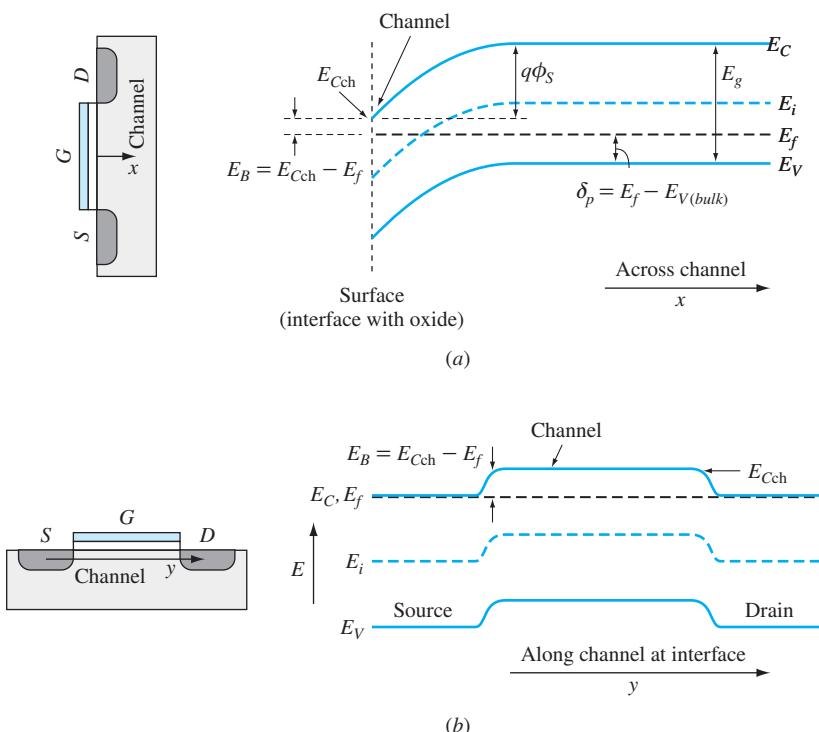


Figure 7.14 The energy band diagram of the NFET across the channel (a) and along the channel (b), with zero voltage between the drain and the source.

From this, we can solve for the barrier height:

$$E_B = kT \ln \frac{N_C}{n_s} \quad (7.6)$$

From Figure 7.14a the surface potential ϕ_s can be written

$$\phi_s = \frac{1}{q} [E_g - E_B - \delta_p] \quad (7.7)$$

where δ_p is the energy difference between the Fermi level and the valence band edge in the bulk (neutral) Si.

According to Equation (7.6), E_B varies slowly (logarithmically) with n_s ; therefore, from Equation (7.7), ϕ_s also varies slowly. Since ϕ_s is varying slowly, the approximation is often used that above threshold, the band bending ϕ_s remains equal to its threshold value. The gate voltage is dropped partly across the oxide and partly through the band-bending region in the semiconductor (the depletion region). Because the amount of band bending ϕ_s is (approximately) constant, any change in gate voltage above threshold is assumed to be dropped across the oxide. We write this as

$$V_{GS} - V_T = \phi_{ox} - \phi_{ox}^{\text{th}} \quad V_{GS} > V_T \quad (7.8)$$

where ϕ_{ox}^{th} is the oxide voltage at threshold.

EXAMPLE 7.1

Show that the approximation in Equation (7.8) is valid. That is, how realistic is it that any additional gate voltage above threshold is dropped across the oxide and not the semiconductor?

Solution

Consider the n-channel MOSFET of Figure 7.6. Let the net substrate doping be $N'_A = 10^{16} \text{ cm}^{-3}$. We know that threshold occurs when the electron concentration at the surface of the channel, n_s , is equal to N'_A , because that is when the surface has the same number of electrons as the bulk has holes.

From Equation (7.6), we have

$$\begin{aligned} E_{B(\text{threshold})} &= kT \ln \frac{N_C}{n_{s(\text{threshold})}} = (0.026 \text{ eV}) \ln \left(\frac{2.86 \times 10^{19} \text{ cm}^{-3}}{10^{16} \text{ cm}^{-3}} \right) \\ &= 7.96kT = 0.207 \text{ eV} \end{aligned}$$

The maximum value that n_s can realistically attain in silicon MOSFETs is about 10^{19} cm^{-3} at which point

$$\begin{aligned} E_{B(\text{way above threshold})} &= kT \ln \frac{N_C}{n_{s(\text{way above threshold})}} = (0.026 \text{ eV}) \ln \left(\frac{2.86 \times 10^{19} \text{ cm}^{-3}}{10^{19} \text{ cm}^{-3}} \right) \\ &= 1.05kT = 0.027 \text{ eV} \end{aligned}$$

In other words, between threshold and way above threshold, the barrier height E_B and thus $q\phi_s$ vary by only about $6.9 kT$, or 180 meV, at room temperature. From Equation (7.7), over this same range the surface potential ϕ_s varies by the value $\Delta E_B/q = 0.180$ V.

From Equation (7.3), at threshold, $\phi_{s(\text{threshold})} = 2\phi_f$. Since $\phi_f = (E_i - E_f)/q$ and $E_i - E_f = kT \ln(N'_A/n_i)$, we have

$$\phi_f = \frac{kT}{q} \ln \frac{N'_A}{n_i} = 0.026 \ln \frac{10^{16}}{1.08 \times 10^{10}} = 0.357 \text{ V}$$

Therefore the band bending at threshold is

$$\phi_{s(\text{threshold})} = 2\phi_f = 2(0.357) = 0.714 \text{ V}$$

Now we find the band bending at a surface concentration of 10^{19} cm^{-3} , which is

$$\phi_{s(\text{way above threshold})} = 0.714 + 0.180 = 0.894 \text{ V}$$

or about 25 percent above its value at threshold. The voltage drop across the semiconductor, then, is not exactly constant above threshold but it is changing slowly. Therefore, the approximation that the surface potential ϕ_s is constant above threshold normally is adequate.

The Case for $V_{DS} > 0$ In the previous section, we looked at the effect of the gate-source voltage on the energy band diagram, barrier heights, and carrier concentrations in the channel. The carrier concentration in the channel relates directly to the conductance in the channel. We assumed there that both ends of the channel were at the same voltage.

In this section, we will allow the drain voltage to be different from the source voltage, thus producing a longitudinal electric field in the channel. This field will induce current to flow along the channel. Just as the current that flows through a resistor depends on the voltage across it, the current from the drain to the source depends on the drain-to-source voltage. In a MOSFET, however, the conductance (and thus resistance) of the channel depends on the gate-source voltage.

Figure 7.15a shows an n-channel MOSFET. The drain is at a positive voltage V_{DS} with respect to the source. The gate has some applied voltage above threshold, so the channel is conducting. If we could take an imaginary voltmeter and somehow measure the voltage across the oxide at the source end of the channel, it would be $V_G - V_S = V_{GS}$. At the drain end, the voltage across the oxide is $V_G - V_D$. Since the channel voltage varies along the channel and the substrate voltage is constant, this varying channel voltage means also that the depletion region width varies from one end to the other. While the depletion region is wider at the drain than at the source, the actual channel is narrower.

Since V_{DS} is positive, electrons entering the potential well from the source end will drift down the channel to the lower electron energy at the drain end. Figure 7.15b shows the energy band diagrams normal to the gate at source and

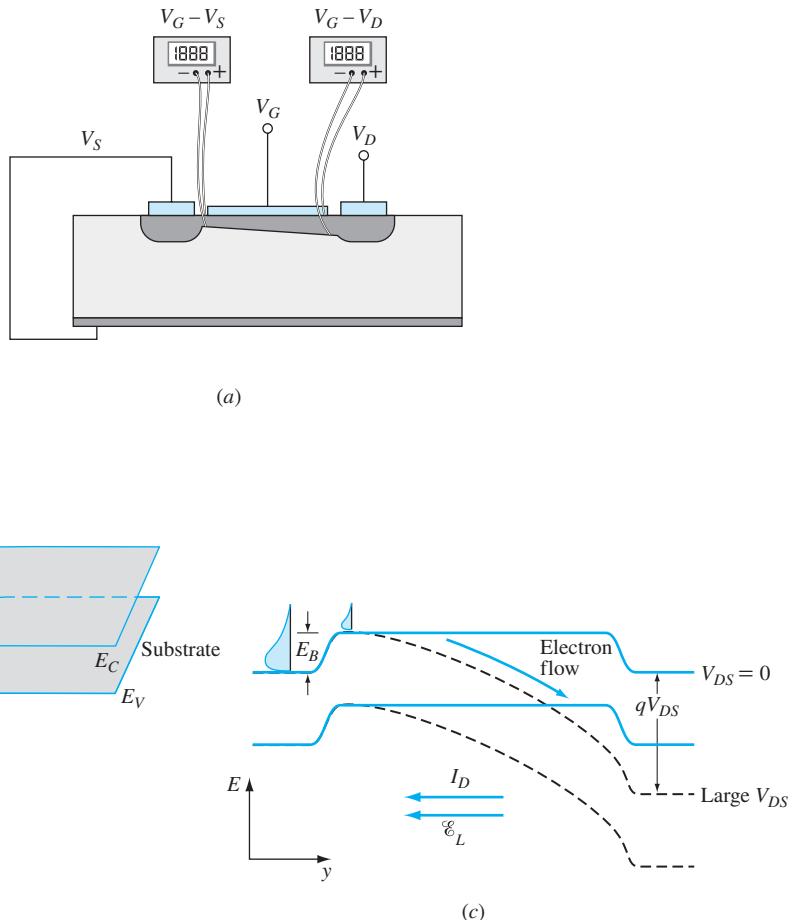


Figure 7.15 (a) With a voltage on the drain with respect to the source, the depletion region width varies along the channel. So does the voltage across the channel at any given point. (b) The energy band diagrams normal to the gate at source and at drain. Here the drain voltage is higher than the source voltage, so the “depth” of the channel varies along its length. (c) The energy band diagram along the channel with no voltage on the drain with respect to the source, and with positive bias applied drain to source.

at drain for a positive drain voltage. The channel potential energy (E_C) decreases along the channel from source to drain.

Figure 7.15c shows the energy band diagram along the channel. When no drain-source voltage is applied (solid line), electrons in the channel do not drift, since the longitudinal field $\mathcal{E}_L = 0$. When $V_{DS} > 0$ is applied, the electron energy at the drain end is lowered and electrons are moved toward the drain. Since electrons are negatively charged, the actual current I_D flows from drain to source.

Let us examine the current-voltage characteristics of a typical NFET in Figure 7.16. We expect that when the gate voltage is below threshold ($V_{GS} \leq V_T$),

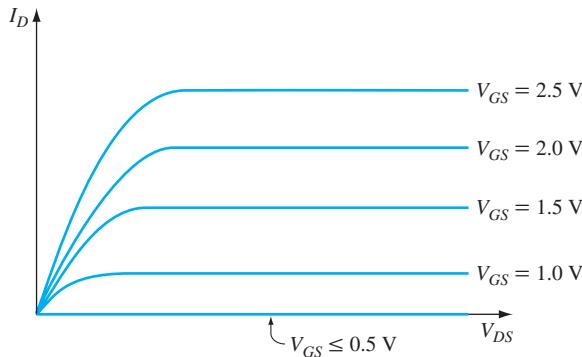


Figure 7.16 The I_D - V_{DS} characteristics of a typical MOSFET. The threshold voltage for this MOSFET is 0.5 V.

the channel will be weakly conductive and the current $I_{ch} = I_D$ will be negligible regardless of the value of V_{DS} . The threshold voltage for the transistor in this example is taken to be $V_T = 0.5 \text{ V}$, so for any gate-source voltage below 0.5 V the transistor does not conduct appreciably.

As V_{GS} increases above the threshold voltage, the barrier height E_B for electrons entering the channel decreases (recall Figure 7.10c), which results in more electrons entering the channel. The channel conductance increases and thus the current also increases and becomes appreciable for $V_{GS} > V_T$.

There is an interesting feature of Figure 7.16 that bears investigation. We might expect that for a given gate voltage, and thus a given channel conductivity (fixed resistor), we would see the current vary linearly with the voltage V_{DS} across the “resistor” (channel). We see this linear behavior in Figure 7.16 for very small values of V_{DS} , but then the current levels off and saturates. We can explain the saturation effect as follows.

Figure 7.17 shows a plot of the energy band diagram along the channel from source to drain for a given value of V_{GS} above threshold. There are three values shown—for $V_{DS} = 0$, for small V_{DS} , and for a larger value of V_{DS} . When the drain voltage is the same as the source voltage ($V_{DS} = 0$), the longitudinal field in the channel is also zero, so there is no slope to the conduction band edge along the channel ($\mathcal{E}_L = (1/q)(dE_C/dy)$). Thus $I_D = 0$ [Equation (III.5) in the Introduction to Part 3]. For small V_{DS} , the energy band diagram tilts slightly. The electrons that enter the channel from the source are accelerated toward the drain by the channel field. At first, as the drain voltage increases, the longitudinal field increases and thus I_D increases (look again at the low V_{DS} end of the I_D - V_{DS} characteristics, Figure 7.16). We will show in the next section that, with increasing V_{DS} , the longitudinal field \mathcal{E}_L increases faster near the drain end than the source end (this is shown in the figure—remember that field is proportional to the slope of the conduction band edge), and most of the incremental drain voltage is dropped near the drain. In other words, with increasing V_{DS} there are increasingly smaller

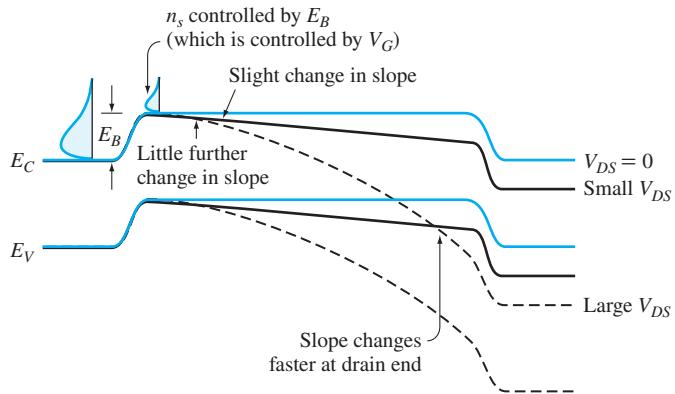


Figure 7.17 The energy band along the channel for three different values of V_{DS} . The current saturates because, as the drain voltage increases, the slope (and thus the electric field) increases faster at the drain end, but at the source end, there is little change. Thus the current is limited by the field at the source end.

changes in the longitudinal field at the source end, as shown in the figure. Eventually, at some value of V_{DS} , the field at the source reaches a limiting value. Since the current at one end of the channel must be the same as the current at the other end, the current I_D is limited to what can be supported by this field at the source end. The drain current $I_D \propto n_s \mathcal{E}_{L(\text{source})}$, and n_s (the carrier concentration in the channel at the source) is controlled by the barrier height E_B and thus by V_{GS} . When the field \mathcal{E}_L ($y = 0$) saturates, so does I_D . (Compare with the analogy with the two lakes and the canal discussed in the Introduction to Part 3.)

Note also that E_B , the barrier from source to channel, depends on the gate-source voltage V_{GS} . Therefore, the number of electrons in the channel n_s , which depends on the barrier height, also depends on V_{GS} but not on V_{DS} .⁷

To summarize our qualitative discussion of MOSFET behavior, a MOSFET is a voltage-controlled resistor. The resistor is between the source and the drain. We can control the conductance of the channel between source and drain by controlling the number of channel carriers available for conduction. In the MOSFET, that control results from adjusting the gate voltage, which in turn controls the band bending in the semiconductor. The gate voltage forces the conduction band edge to bend closer to or farther away from the Fermi level.

Current does not flow into the gate terminal, because there is an insulating oxide layer between the gate and the source, channel and drain. The gate voltage induces an electric field in the oxide, which in turn influences the energy bands

⁷As we discuss in Section 7.6 for very short (submicrometer) channel lengths, V_{DS} does, in fact, affect the value of E_B .

in the semiconductor. The electric field from which the FET gets its name is the field induced by the gate voltage.

In the next section, we will apply our physical understanding of these processes to derive expressions for the I_D - V_{DS} characteristics of the transistors.

7.3 DRIFT MODEL FOR MOSFETS (QUANTITATIVE)

Now that we have a physical understanding of how MOSFETs work, we can be more quantitative. First we derive expressions for the I_D - V_{DS} characteristics of an NFET using the drift model. While the current in the channel is a function of drift and diffusion, for $V_{GS} > V_T$ drift dominates. Next the electrical characteristics are derived for the ballistic model in which the electrons make no collisions in the channel between source and drain. The derivations of the electrical characteristics of MOSFETs in the drift model are presented in three steps. First, we consider a formulation, in which the carrier mobility is assumed constant along the channel. [1] This is the simplest model, called the *long-channel* model, and it predicts the general form of the I_D - V_{DS} characteristics. It is useful for obtaining insight into the general behavior of MOSFETs but does not closely reproduce the results for modern devices. Therefore, we then modify the simple model to account for variation in mobility. The mobility is affected by two things: the transverse and the longitudinal electric fields in the channel. Accounting for these are the second and third steps in our development.

Recall that electron mobility is given by the equation $\mu_n = \frac{q\bar{t}}{m_{ce}^*}$ where \bar{t} is the mean free time between collisions (in this context, within the channel) and m_{ce}^* is the electron conductivity effective mass. In the case of very short channel lengths (e.g., 10–20 nm in Si), the electrons make no (or few) collisions between source and drain. This model is referred to as the *ballistic transport model* or simply *ballistic model*. It is discussed in Section (7.5).

We consider the enhancement-type NFET device of Figure 7.6, which is repeated as Figure 7.18. The figure indicates channel width W , the channel length, L , and the oxide thickness t_{ox} . The direction of the longitudinal field \mathcal{E}_L is shown, along with the field component perpendicular to the channel, called the transverse field, \mathcal{E}_T . Recall from Figure III.7 that the channel voltage V_{ch} is the voltage at a given point along the channel with respect to the source, and is a function of position along the channel. At the drain end of the channel the channel voltage V_{ch} is equal to the drain-source voltage V_{DS} .

Our starting point will be Equations (III.3) to (III.7), which we repeat here for convenience:

$$I_D = -WQ_{ch}(y)v(y) \quad (\text{III.3})$$

$$v(y) = -\mu(y)\mathcal{E}_L(y) \quad (\text{III.4})$$

$$I_D = WQ_{ch}(y)\mu(y)\mathcal{E}_L(y) \quad (\text{III.5})$$

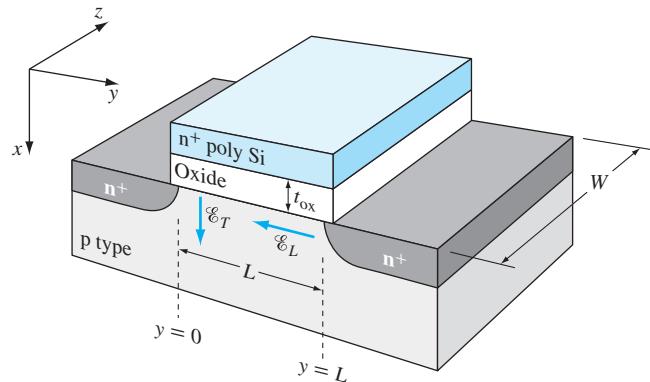


Figure 7.18 The NFET. Longitudinal and transverse electric field directions are indicated.

$$\mathcal{E}_L = -\frac{dV_{ch}}{dy} \quad (\text{III.6})$$

$$I_D dy = -W Q_{ch}(y) \mu(y) dV_{ch}(y) \quad (\text{III.7})$$

Since our goal is to obtain an expression for I_D , we see from Equation (III.7) that we need the following: an analytical expression for μ , the carrier mobility in the channel; an expression for Q_{ch} , the charge per unit area in the channel (this is the mobile charge, in this case electrons in the conduction band); and an expression for V_{ch} as a function of position y . We will derive these first for the simple long-channel model, to illustrate the physics. Later we will add more realism (and complication) to the equations.

7.3.1 LONG-CHANNEL DRIFT MOSFET MODEL WITH CONSTANT CHANNEL MOBILITY

There are several models used to describe μ and Q_{ch} . We will start with an oversimplified model for these quantities. While this model gives realistic results only for the I_D - V_{DS} characteristics for MOSFETs with very *long channels* ($L > 5$ to $10 \mu\text{m}$), it is mathematically simple and does illustrate the general principles of operation. This formulation is employed in the SPICE Level 1 model.

Channel Charge Density Since the channel current depends on the channel charge density, our first task will be to analyze the charge in the channel. The charge will depend on the bias conditions. When the gate voltage is below threshold, the conductance is small because the number of electrons available for conduction is small. For simplicity we approximate:

$$Q_{ch} \approx 0 \quad \text{and} \quad I_D \approx 0 \quad (V_{GS} \leq V_T) \quad (7.9)$$

For $V_{GS} > V_T$, of course, Q_{ch} is nonzero. We can find how much charge is present by recognizing that the region under the gate acts as a capacitor. Two

conductive plates (the heavily doped polysilicon gate electrode and the conductive channel) are separated by an insulator (the oxide). The capacitance of a parallel plate capacitor is given by

$$C = \epsilon \frac{A}{t} \quad (7.10)$$

where ϵ is the permittivity of the insulator, A is the area of the plate, and t is the thickness of the dielectric layer. In our device, we use the oxide thickness t_{ox} . The area of the gate electrode is $W \times L$.

On a given integrated circuit, different transistors may have different widths and lengths. The oxide thickness, on the other hand, is usually a constant for a given process and therefore common to all devices on the chip. It is therefore useful to define an *oxide capacitance per unit area* C'_{ox} :

$$C'_{\text{ox}} = \frac{\epsilon_{\text{ox}}}{t_{\text{ox}}} \quad (7.11)$$

where ϵ_{ox} is the permittivity of the oxide $= \epsilon_{\text{ox}} = \epsilon_r \epsilon_0$, and ϵ_r is the relative permittivity (dielectric constant) of the dielectric. The dielectric constant is $\epsilon_r = 3.9$ for SiO_2 and about 5 for SiON . (The symbol k is often used for the dielectric constant.)

We will need to know the voltage across the capacitor, which is the voltage across the oxide. We don't know what the voltage across the oxide is, exactly, but we do recall from Example 7.1 that to reasonable approximation, for $V_{GS} > V_T$ any *change* in gate voltage appears across the oxide. Since capacitance is $C = |dQ/dV|$, then

$$C'_{\text{ox}} = -\frac{dQ_{\text{ch}}}{dV_{GS}} = -\frac{\Delta Q_{\text{ch}}}{\Delta V_{GS}} = -\frac{[Q_{\text{ch}}(V_{GS}) - Q_{\text{ch}}(V_T)]}{V_{GS} - V_T} = -\frac{[Q_{\text{ch}}(V_{GS}) - 0]}{V_{GS} - V_T} \quad (7.12)$$

where we have used the information that Q_{ch} is negative and for $V_{GS} = V_T$, $Q_{\text{ch}} = 0$ [Equation (7.9)]. Letting $Q_{\text{ch}} = Q_{\text{ch}}(V_{GS})$, we have

$$Q_{\text{ch}} = -C'_{\text{ox}}(V_{GS} - V_T) \quad V_{DS} = 0 \quad (7.13)$$

When V_{DS} is no longer zero but is positive, the voltage on the lower plate of the capacitor between the channel and ground, V_{ch} , is a function of position y along the channel. The voltage dropped across the oxide will thus vary along y and affect Q_{ch} :

$$Q_{\text{ch}}(y) = -C'_{\text{ox}}(V_{GS} - V_T - V_{\text{ch}}(y)) \quad V_{GS} - V_T > V_{\text{ch}}(y) \quad (7.14)$$

At the source end of the channel, where $V_{\text{ch}} = 0$ since V_{ch} is the channel voltage with respect to the source, this reduces to Equation (7.13).

We now have expressions for the channel charge—but there is a problem. Equation (7.14) is valid only for $(V_{GS} - V_T) > V_{\text{ch}}(y)$. This will always be true at the source end of the channel, provided the gate voltage is above threshold, since $V_{\text{ch}} = 0$. At the drain end, however, $V_{\text{ch}}(y = L) = V_{DS}$. Thus, if $V_{DS} > (V_{GS} - V_T)$, then at some position y in the channel, the channel voltage must be equal to

$V_{ch} = (V_{GS} - V_T)$. At that point along the channel, Equation (7.14) implies that $Q_{ch} = 0$. However, since V_{GS} is above threshold and V_{DS} is not zero, we know that a current I_D is flowing through the channel. Then from Equation (III.3), since $I_D > 0$, this implies that the electron velocity is infinite. Furthermore, at y greater than this value, $V_{ch} > (V_{GS} - V_T)$, implying that the channel charge is positive. That would mean that the channel current is carried by holes. From physical arguments, however, we reject both of these scenarios. We know that the maximum possible velocity is the saturation velocity v_{sat} , and from Figure 7.14a holes clearly cannot enter the channel. Thus, since I_D must be a constant at every position in the channel, from Equation (III.3) when the velocity has a maximum the charge has a minimum. The minimum value of Q_{ch} is

$$Q_{ch\ min} = -\frac{I_{D\ sat}}{Wv_{sat}} \quad (7.15)$$

where $I_{D\ sat}$ is the saturation current, which is indicated in Figure 7.19 for several values of V_{GS} . According to this model, the current cannot exceed this amount for a given value of gate voltage V_{GS} . The saturation current is reached when there is a position in the channel for which $V_{ch} = V_{GS} - V_T$. This happens first at the drain end, when $V_{DS} = V_{GS} - V_T$. We call this the drain saturation voltage $V_{DS\ sat}$.

Above this saturation point, Equation (7.14) no longer applies. We will take this point up again later. In the region where it does apply, though, we now have a model for the channel charge.

Channel Mobility Next, we examine the channel mobility μ . In this long-channel simple model, we will take the mobility to be constant. In reality, the mobility depends on the longitudinal electric field \mathcal{E}_L (and thus on V_{ch}). For example, we saw in Chapter 3 that, under high fields, the velocity saturates. Further, the transverse field \mathcal{E}_T will have an effect. We will handle these dependencies of mobility on the fields explicitly later, however. Here we will consider the simplest model, the low-field case, in which the mobility is constant.

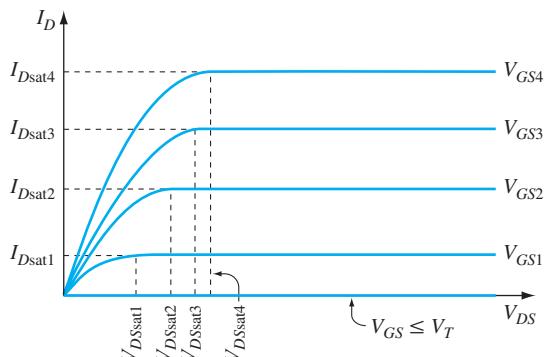


Figure 7.19 The saturation current and saturation voltage are defined.

Long-Channel (Square Law) Model, Constant Mobility In this model, for mathematical simplicity we assume μ to be constant, and of value somewhere in the neighborhood of one-third to one-sixth of its bulk value. We use the term *bulk* to distinguish the mobility in a large crystal from the mobility experienced by a carrier in a thin layer such as the channel of a MOSFET. The reason for assuming one-third to one-sixth of the bulk value is that the small thickness of the channel will tend to slow the carriers down from the bulk value, as we will see later.

To obtain the I_D - V_{DS} characteristics for this model, we integrate both sides of Equation (III.7):

$$\int_0^L I_D dy = \int_0^{V_{DS}} -WQ_{ch}(y)\mu(y)dV_{ch} \quad (7.16)$$

Note the limits of integration. Over the length of the channel, the channel voltage varies from the source voltage $V_{ch} = 0$ to the drain-source voltage V_{DS} .

As long as the drain voltage is less than $(V_{GS} - V_T)$, the channel voltage V_{ch} will also satisfy $V_{ch} < (V_{GS} - V_T)$, so we can use Equation (7.14) for the channel charge. Since the mobility μ is constant in this model, it comes out of the integral. The current I_D cannot vary with position along the channel, so it is also a constant and it also comes out of its integral. Thus Equation (7.16) becomes

$$I_D \int_0^L dy = -W\mu \int_0^{V_{DS}} Q_{ch} dV_{ch} = -W\mu \int_0^{V_{DS}} [-C'_{ox}(V_{GS} - V_T - V_{ch})] dV_{ch} \quad (7.17)$$

Integrating, the result is, for $V_{DS} \leq (V_{GS} - V_T)$,

$$I_D = \frac{WC'_{ox}\mu}{L} \int_0^{V_{DS}} (V_{GS} - V_T - V_{ch}) dV_{ch} \quad V_{DS} \leq (V_{GS} - V_T) \quad (7.18)$$

or

$$I_D = \frac{WC'_{ox}\mu}{L} \left[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2} \right] \quad V_{DS} \leq (V_{GS} - V_T) \quad (7.19)$$

There are two independent variables in this equation, the drain voltage and the gate voltage. Figure 7.20 shows the drain current calculated from Equation (7.19) versus the drain voltage for two different values of gate-source voltage V_{GS} . We chose the parameters of the NFET to be $t_{ox} = 4$ nm, $k = \epsilon_r = 3.9$ (SiO_2), $W/L = 5$, and $\mu = 500 \text{ cm}^2/\text{V} \cdot \text{s}$.

We see from the plot that the current reaches a maximum ($dI_D/dV_{DS} = 0$) for $V_{DS} = (V_{GS} - V_T)$. This is also the limit of validity of Equation (7.19). For $V_{DS} > (V_{GS} - V_T)$, from Equation (7.19), I_D would be expected to decrease as indicated by the colored dashed line of Figure 7.20. We will show later that the simple model predicts that once the curve in Figure 7.20 reaches its peak, the current

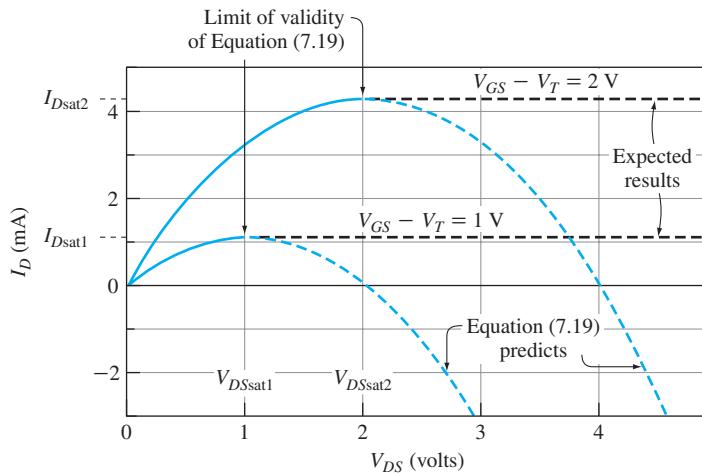


Figure 7.20 The current predicted using Equation (7.19) (solid lines) is only valid up to the point where $V_{DS} = V_{GS} - V_T$. After that the current saturates (black dashed lines), whereas Equation (7.19) would predict a decrease and eventually a sign reversal in the current (colored dashed lines).

remains (essentially) constant for larger V_{DS} . Thus, the current saturates at some value of I_{Dsat} as indicated.

The value of V_{DSsat} at which I_D saturates is found from taking $\partial I_D / \partial V_{DS}$ in Equation (7.19) and setting it to zero:

$$\frac{\partial I_D}{\partial V_{DS}} = 0 = \frac{WC'_ox\mu}{L}(V_{GS} - V_T - V_{DSsat}) \quad (7.20)$$

or

$$V_{DSsat} = (V_G - V_T) \quad (7.21)$$

Above this value, I_D remains constant.

We can use this result in Equation (7.17). By setting the limit of integration to V_{DSsat} , we obtain an expression for the saturation current:

$$I_D = I_{Dsat} = \frac{WC'_ox\mu}{L} \left[\left(V_{GS} - V_T - \frac{V_{DSsat}}{2} \right) V_{DSsat} \right] \quad (7.22)$$

But, since in this model $V_{DSsat} = V_{GS} - V_T$, we can write:

$$I_{Dsat} = \frac{WC'_ox\mu}{2L} (V_{GS} - V_T)^2 = \frac{WC'_ox\mu}{2L} V_{DSsat}^2 \quad (7.23)$$

Since I_{Dsat} is proportional to V_{DSsat}^2 , this model is sometimes referred to as the *square law model*. It results from the simple long-channel model, assuming

constant mobility, and uses Equation (7.14) to represent the channel charge in the region below threshold.

Thus, we can describe the I_D - V_{DS} characteristics for this model with the following three equations:

$$I_D = \frac{WC'_\text{ox}\mu}{L} \left[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2} \right] \quad V_{DS} \leq V_{DS\text{sat}}, V_{GS} \geq V_T \quad (7.24)$$

$$\begin{aligned} I_{D\text{sat}} &= \frac{WC'_\text{ox}\mu}{L} \left[\left(V_{GS} - V_T - \frac{V_{DS\text{sat}}}{2} \right) V_{DS\text{sat}} \right] \\ &= \frac{WC'_\text{ox}\mu}{2L} (V_{GS} - V_T)^2 \quad V_{DS} \geq V_{DS\text{sat}}, V_{GS} \geq V_T \end{aligned} \quad (7.25)$$

$$V_{DS\text{sat}} = (V_{GS} - V_T) \quad (7.26)$$

EXAMPLE 7.2

Using the simple long-channel model, assuming constant mobility, plot the I_D - V_{DS} characteristics for an NFET with $W/L = 5$ and $t_\text{ox} = 4 \text{ nm}$. Take the constant mobility for electrons in the channel to be $500 \text{ cm}^2/\text{V}\cdot\text{s}$. Plot for $V_{GS} - V_T = 1, 2, 3, \text{ and } 4 \text{ V}$, and V_{DS} from 0 to 5 V. Assume $\epsilon_r = 3.9$ (SiO_2).

Solution

From Equation (7.11), we have

$$C'_\text{ox} = \frac{\epsilon_\text{ox}}{t_\text{ox}} = \frac{\epsilon_r(\text{ox})\epsilon_0}{t_\text{ox}} = \frac{3.9(8.85 \times 10^{-14} \text{ F/cm})}{4 \times 10^{-7} \text{ cm}} = 8.6 \times 10^{-7} \text{ F/cm}^2$$

For each value of $V_{GS} - V_T$, we must find the saturation point to know whether to use Equation (7.24) or (7.25). For example, from Equation (7.26), we have for $V_{GS} - V_T = 1 \text{ V}$, $V_{DS\text{sat}} = V_{GS} - V_T = 1 \text{ V}$. For V_{DS} between 0 and 1 V, then, we use Equation (7.24)

$$\begin{aligned} I_D &= \frac{W}{L} C'_\text{ox} \mu \left[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2} \right] \\ &= (5)(8.6 \times 10^{-7} \text{ F/cm}^2)(500 \text{ cm}^2/\text{V}\cdot\text{s}) \left[(1) \cdot V_{DS} - \frac{V_{DS}^2}{2} \right] \\ &= 2.15 \times 10^{-3} \left[V_{DS} - \left(\frac{V_{DS}}{2} \right)^2 \right] \end{aligned}$$

At $V_{DS} = V_{DS\text{sat}} = 1 \text{ V}$, I_D reaches its saturation value of [Equation (7.25)]

$$I_{D\text{sat}} = \frac{WC'_\text{ox}\mu}{2L} (V_{GS} - V_T)^2 = (5) \frac{(8.6 \times 10^{-7})(500)}{2} (1)^2 = 1.07 \text{ mA}$$

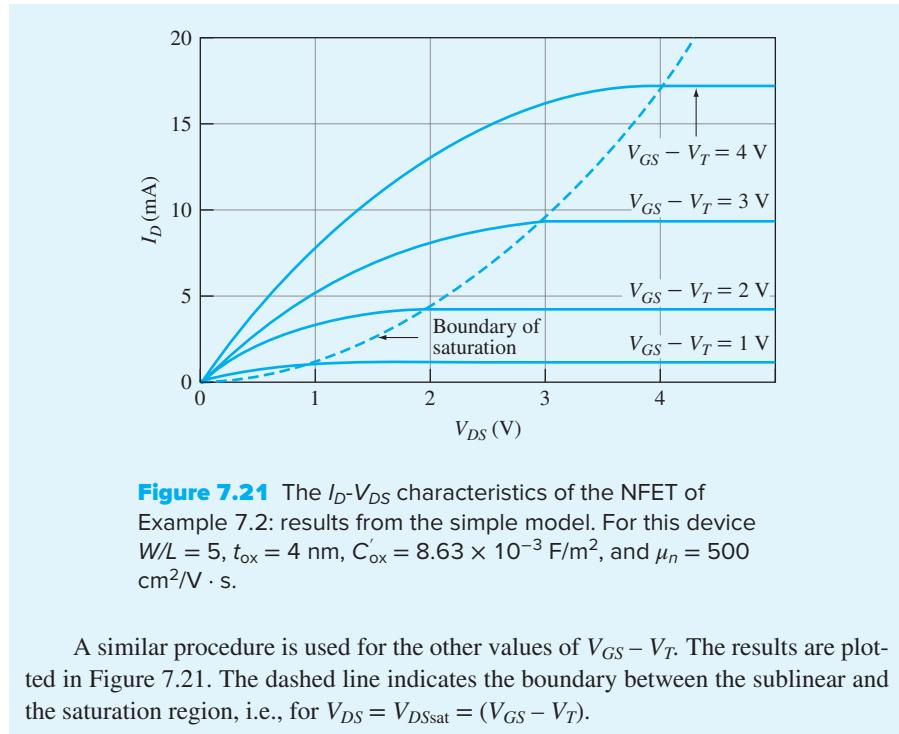


Figure 7.21 The I_D - V_{DS} characteristics of the NFET of Example 7.2: results from the simple model. For this device $W/L = 5$, $t_{ox} = 4 \text{ nm}$, $C'_\text{ox} = 8.63 \times 10^{-3} \text{ F/m}^2$, and $\mu_n = 500 \text{ cm}^2/\text{V} \cdot \text{s}$.

A similar procedure is used for the other values of $V_{GS} - V_T$. The results are plotted in Figure 7.21. The dashed line indicates the boundary between the sublinear and the saturation region, i.e., for $V_{DS} = V_{DS\text{sat}} = (V_{GS} - V_T)$.

Current Saturation Revisited Earlier we claimed that the simple long-channel model (constant μ) predicts that once the drain voltage reaches $V_{DS\text{sat}} = V_{GS} - V_T$, the current saturates and remains constant for all higher drain voltages. This saturation was discussed qualitatively by analogy with water flow between two lakes via a canal as discussed in Part 3, Field-Effect Transistors. Here we discuss current saturation more quantitatively.

When we discussed Figure 7.16, we indicated that for low values of V_{DS} , the longitudinal field \mathcal{E}_L is constant along the channel. At larger drain voltages, the field increases appreciably at the drain end, but not much at the source end, as was shown in Figure 7.17. Let us consider the effect of this point analytically.

For a given V_{GS} and V_{DS} with constant μ , Equation (III.5) becomes

$$I_D = W\mu Q_{ch}(y)\mathcal{E}_L(y)$$

The current is constant and proportional to the $Q_{ch}\mathcal{E}_L$ product at any value of y . It is convenient to determine this product and thus the I_D - V_{DS} relation near the source end of the channel ($y = 0$). We can find \mathcal{E}_L from Equation (III.6) ($\mathcal{E}_L = -dV_{ch}/dy$) if we have an expression for V_{ch} as a function of y . We can find $V_{ch}(y)$ by integrating Equation (III.7) from 0 to y . From Equations (III.7) and (7.14), then

$$I_D \int_0^y dy = WC'_\text{ox}\mu \int_0^{V_{ch}(y)} (V_{GS} - V_T - V_{ch}) dV_{ch} \quad (7.27)$$

or

$$I_D = \frac{WC_{\text{ox}}' \mu}{y} \left(V_{GS} - V_T - \frac{V_{\text{ch}}(y)}{2} \right) V_{\text{ch}}(y) \quad (7.28)$$

Solving for $V_{\text{ch}}(y)$

$$V_{\text{ch}}(y) = (V_{GS} - V_T) - \sqrt{(V_{GS} - V_T)^2 - \frac{2I_D y}{WC_{\text{ox}}' \mu}} \quad (7.29)$$

where the negative sign associated with the square root is used, since for $y = 0$, $V_{\text{ch}}(0) = 0$.

This gives us an expression for the channel voltage as a function of distance along the channel. The potential energy (E_C) has the same shape as the potential, but inverted ($dE_C/dy = -qdV_{\text{ch}}/dy$), which means the conduction band edge has the shape

$$E_C(y) = E_C(0) - qV_{\text{ch}}(y) \quad (7.30)$$

Therefore, we can substitute Equation (7.29) into Equation (7.30) to obtain

$$E_C(y) = E_C(0) - q \left[(V_{GS} - V_T) - \sqrt{(V_{GS} - V_T)^2 - \frac{2I_D y}{WC_{\text{ox}}' \mu}} \right] \quad (7.31)$$

The conduction band edge along the channel is plotted in Figure 7.22a for several values of V_{DS} , with $W/L = 10$, $\mu = 500 \text{ cm}^2/\text{V} \cdot \text{s}$, $(V_{GS} - V_T) = 2 \text{ V}$ and $C_{\text{ox}}' = 6.9 \times 10^{-7} \text{ F/cm}^2$ ($t_{\text{ox}} = 5 \text{ nm}$ and is taken to be SiO_2), and with I_D obtained from Equation (7.19). For $V_{DS} = 0$, $I_D = 0$, and there is no voltage drop along the channel and E_C is flat. As V_{DS} increases, the band bends increasingly as seen in the figure. We note that the magnitude of the slope of the E_C -y plot at the source ($y = 0$) increases with increasing V_{DS} and tends toward saturation as V_{DS} approaches $V_{GS} - V_T$ (2 V).

The electric field is proportional to the slope of E_C . The longitudinal field at some point y is

$$\mathcal{E}_L(y) = -\frac{dV_{\text{ch}}}{dy} = \frac{1}{q} \frac{dE_C}{dy} = -\frac{\frac{I_D}{WC_{\text{ox}}' \mu}}{\sqrt{(V_{GS} - V_T)^2 - \frac{2I_D y}{WC_{\text{ox}}' \mu}}} \quad (7.32)$$

The field $\mathcal{E}_L(y)$ is plotted in Figure 7.22b for the same device, for various values of drain voltage. Notice that near the source end, the magnitude of the field $\mathcal{E}_L(0)$ varies rapidly with V_{DS} for small values of V_{DS} , but approaches a constant (saturates) as V_{DS} approaches $V_{DS,\text{sat}} = V_{GS} - V_T$.

Combining Equations (III.5) and (7.13) at $y = 0$, we have as our final result

$$I_D = -WC_{\text{ox}}'(V_{GS} - V_T)\mu \mathcal{E}_L(0) \quad (7.33)$$

Again, the current is proportional to the electric field at the source end. Since the field is saturating, the current also saturates.

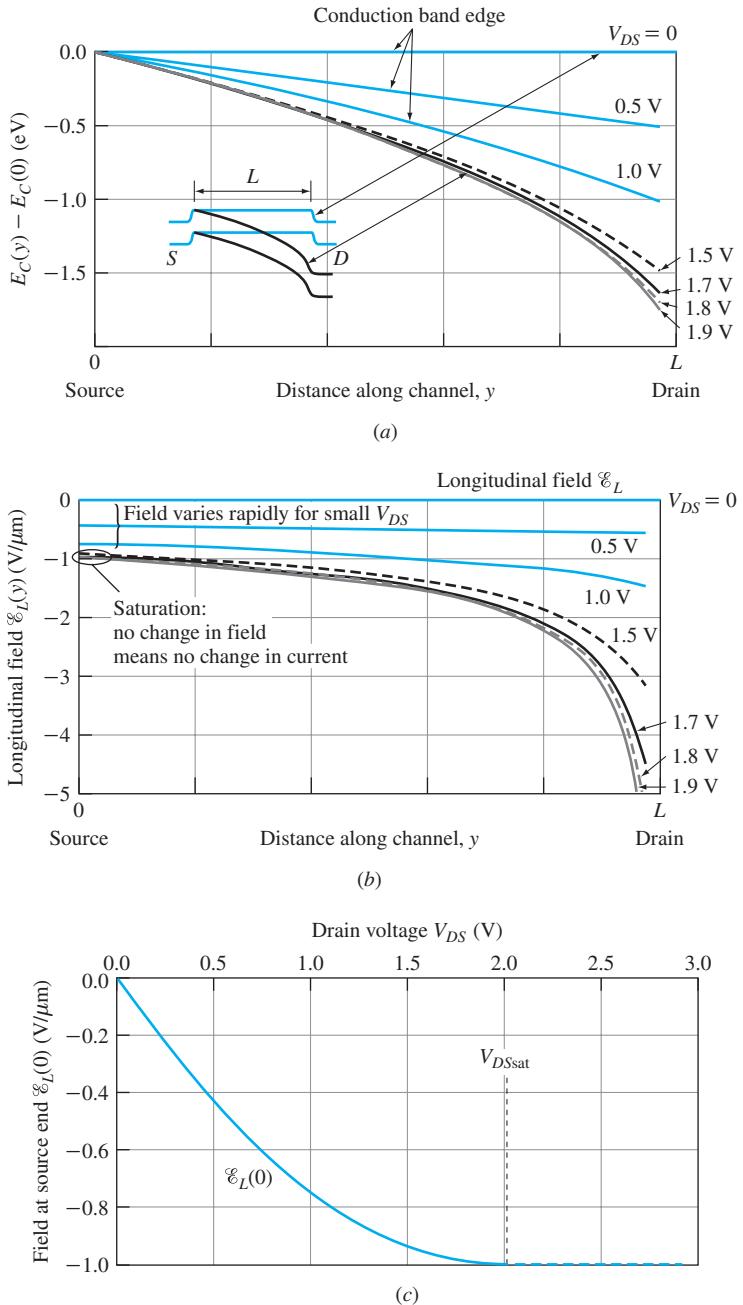


Figure 7.22 Illustration for current saturation. (a) The conduction band edge along the channel bends more at the drain end than at the source end for large drain voltage V_{DS} . (b) Since the longitudinal field is proportional to the slope of E_C the field changes rapidly at the drain end for increasing values of V_{DS} but not at the source end. (c) The field at the source end is constant as V_{DS} increases beyond a certain point ($V_{DS} = V_{DSsat}$); thus the current is constant as well.

Finally, let us examine the rate at which the last term, the field at the source end $\mathcal{E}_L(0)$, varies with V_{DS} . We use Equation (7.32) with $y = 0$:

$$\mathcal{E}_L(0) = -\frac{\frac{I_D}{WC_{ox}'\mu}}{(V_{GS} - V_T)} = -\frac{I_D}{WC_{ox}'\mu(V_{GS} - V_T)} \quad (7.34)$$

Substituting for I_D from Equation (7.19) yields

$$\mathcal{E}_L(0) = -\frac{\left(V_{GS} - V_T - \frac{V_{DS}}{2}\right)V_{DS}}{L(V_{GS} - V_T)} = -\frac{1}{L} \left[1 - \frac{V_{DS}}{2(V_{GS} - V_T)}\right]V_{DS} \quad (7.35)$$

For $V_{GS} - V_T = 2$ V, we have

$$\mathcal{E}_L(0) = -\frac{\left(1 - \frac{V_{DS}}{4}\right)V_{DS}}{L} \quad (7.36)$$

For a channel length of $L = 1$ μm , that produces a field of

$$\mathcal{E}_L(0) = -\left(1 - \frac{V_{DS}}{4}\right)V_{DS} \frac{\text{V}}{\mu\text{m}} \quad (7.37)$$

This is plotted in Figure 7.22c. We observe that the magnitude of the field at the source increases with V_{DS} and then levels off. For saturation, we know that the slope is zero, or

$$\frac{d\mathcal{E}_L(0)}{dV_{DS}} = 0 = -\left(1 - \frac{V_{DS}}{2}\right) \quad (7.38)$$

which occurs for $V_{DS} = (V_{GS} - V_T) = 2$ V, as seen in the figure.

EXAMPLE 7.3

Estimate the time it takes a charge (current) spike to travel from the source to the drain for a small spike in gate voltage. Consider a MOSFET with $L = 2$ μm and $\mu = 400 \text{ cm}^2/\text{V}\cdot\text{s}$

Solution

The electron velocity is given by $v = \frac{dy}{dt} = -\mu\mathcal{E}$, so we can write $dt = -\frac{dy}{\mu\mathcal{E}}$. Integrating both sides gives

$$\int_0^T dt = \int_0^L \frac{dy}{-\mu\mathcal{E}} \quad \text{or} \quad T \approx \frac{L}{-\mu\langle\mathcal{E}\rangle}$$

For simplicity, let us choose $V_{DS} = 1$ V. From Figure 7.22(b), we find the electric field is $\langle\mathcal{E}\rangle \approx -1 \text{ V}/\mu\text{m} = -10^4 \text{ V}/\text{cm}$, and thus $T \approx \frac{2 \times 10^{-4} \text{ cm}}{(400 \text{ cm}^2/\text{V}\cdot\text{s})(10^4 \text{ V}/\text{cm})} = 5 \times 10^{-11} \text{ s}$ or 50 ps. For greater V_{DS} the transit time is even smaller.

EXAMPLE 7.4

Show that for constant μ , evaluating Equation (III.5) at $y = 0$ gives Equation (7.19) for I_D .

Solution

Equation (III.5) is repeated here:

$$I_D = WQ_{\text{ch}}(y)\mu(y)\mathcal{E}_L(y)$$

At $y = 0$, we have $V_{\text{ch}} = 0$. From Equation (7.14), then, $Q_{\text{ch}}(0) = C'_{\text{ox}}(V_{GS} - V_T)$. Combining these with the expression for $\mathcal{E}_L(0)$ from Equation (7.35) into Equation (III.5), we obtain

$$I_D = W\mu Q_{\text{ch}}(0)\mathcal{E}_L(0) = W\mu \left[-C'_{\text{ox}}(V_{GS} - V_T) \right] \left[-\frac{\left(V_{GS} - V_T - \frac{V_{DS}}{2}\right)V_{DS}}{L(V_{GS} - V_T)} \right]$$

Cancelling $(V_{GS} - V_T)$ gives

$$I_D = \frac{W\mu C'_{\text{ox}}}{L} \left(V_{GS} - V_T - \frac{V_{DS}}{2} \right) V_{DS}$$

which is Equation (7.19).

Channel Length Modulation In the above long-channel MOSFET model, in the sublinear region the drain current is given by Equation (7.19):

$$I_D = \frac{WC'_{\text{ox}}\mu}{L} \left[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2} \right] \quad V_{DS} \leq V_{DS\text{sat}} \quad (7.19)$$

and in saturation combining Equations (7.25) and (7.26) gives

$$I_D = I_{D\text{sat}} = \frac{WC'_{\text{ox}}\mu}{2L} (V_{GS} - V_T)^2 = \frac{WC'_{\text{ox}}\mu}{2L} V_{DS\text{sat}}^2 \quad V_{DS} \geq V_{DS\text{sat}} \quad (7.39)$$

which is constant. In real devices, however, as the drain voltage increases above the saturation point $V_{DS} > V_{GS} - V_T = V_{DS\text{sat}}$, I_D continues to increase slowly with increasing V_{DS} , as indicated in Figure 7.23. This figure represents the experimental results for an n-channel MOSFET with $t_{\text{ox}} = 4.7$ nm, $L = 0.27$ μm , and $V_T = 0.3$ V. There are two physical reasons for this increase in drain current that were not taken into account in the simple model: (1) increasing drain voltage V_{DS} reduces the *effective* channel length L , which we will discuss next, and (2) increasing V_{DS} reduces the value of the threshold voltage. The second of these effects is important for very short channels and is discussed later where short-channel effects are handled.

Let us examine qualitatively why the channel length is effectively shortened as V_{DS} increases above saturation. Figure 7.24 (a to e) shows the energy band

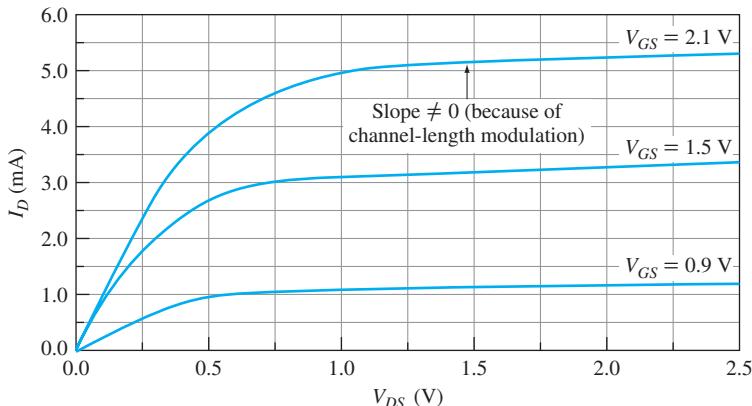


Figure 7.23 Experimental I_D - V_{DS} characteristics for an n-channel MOSFET for three values of gate voltage. The current actually increases with increasing V_D in the “current saturation” region because of channel-length modulation. For this device, $t_{ox} = 4.7 \text{ nm}$, $L = 0.27 \mu\text{m}$, $W = 8.6 \mu\text{m}$, and $V_T = 0.3 \text{ V}$.

diagrams (top) and corresponding hybrid diagrams (bottom) for different bias conditions. In the energy band diagrams, the edges of the depletion regions are indicated by solid lines. The dashed lines represent the maximum energy between source and channel and drain and channel. Note that the maximum energy at the Si-oxide interface extends into the channel region by a small distance from the source and drain. This corresponds to a small barrier to electrons as indicated in Figures 7.14, 7.15 and 7.17. In the first case of Figure 7.24, the gate voltage is below threshold ($V_{GS} < V_T$) so there is no charge in the channel. In parts (b)-(e), the transistor is on, at various values of $V_{GS} > V_T$. Figure 7.24f shows the corresponding currents. In (a) since $V_{GS} < V_T$, negligible charge exists in the channel. In (b), where $V_{GS} > V_T$ and $V_{DS} = 0$, the channel charge Q_{ch} is constant in y and no current flows. In (c), for small V_{DS} , Q_{ch} decreases with increasing y and current flows as indicated. For $V_{DS} \geq (V_{GS} - V_T) = V_{DSsat}$, (d) the current saturates as discussed earlier. For $V_{DS} > V_{DSsat}$, the channel voltage V_{ch} reaches V_{DSsat} somewhere before the end of the channel [part (e) of the figure]. The effective channel length then is shorter than the physical channel by some amount ΔL :

$$L_{eff} = L - \Delta L \quad (7.40)$$

Here we will treat ΔL as an empirical quantity (established from measurements).

Now, substituting Equation (7.40) into Equation (7.39), we get for the drain current

$$I_D = \frac{WC'_{ox}\mu}{2(L - \Delta L)} V_{DSsat}^2 = \frac{WC'_{ox}\mu}{2L(1 - \frac{\Delta L}{L})} V_{DSsat}^2 = \frac{I_{Dsat}}{\left(1 - \frac{\Delta L}{L}\right)} \quad (7.41)$$

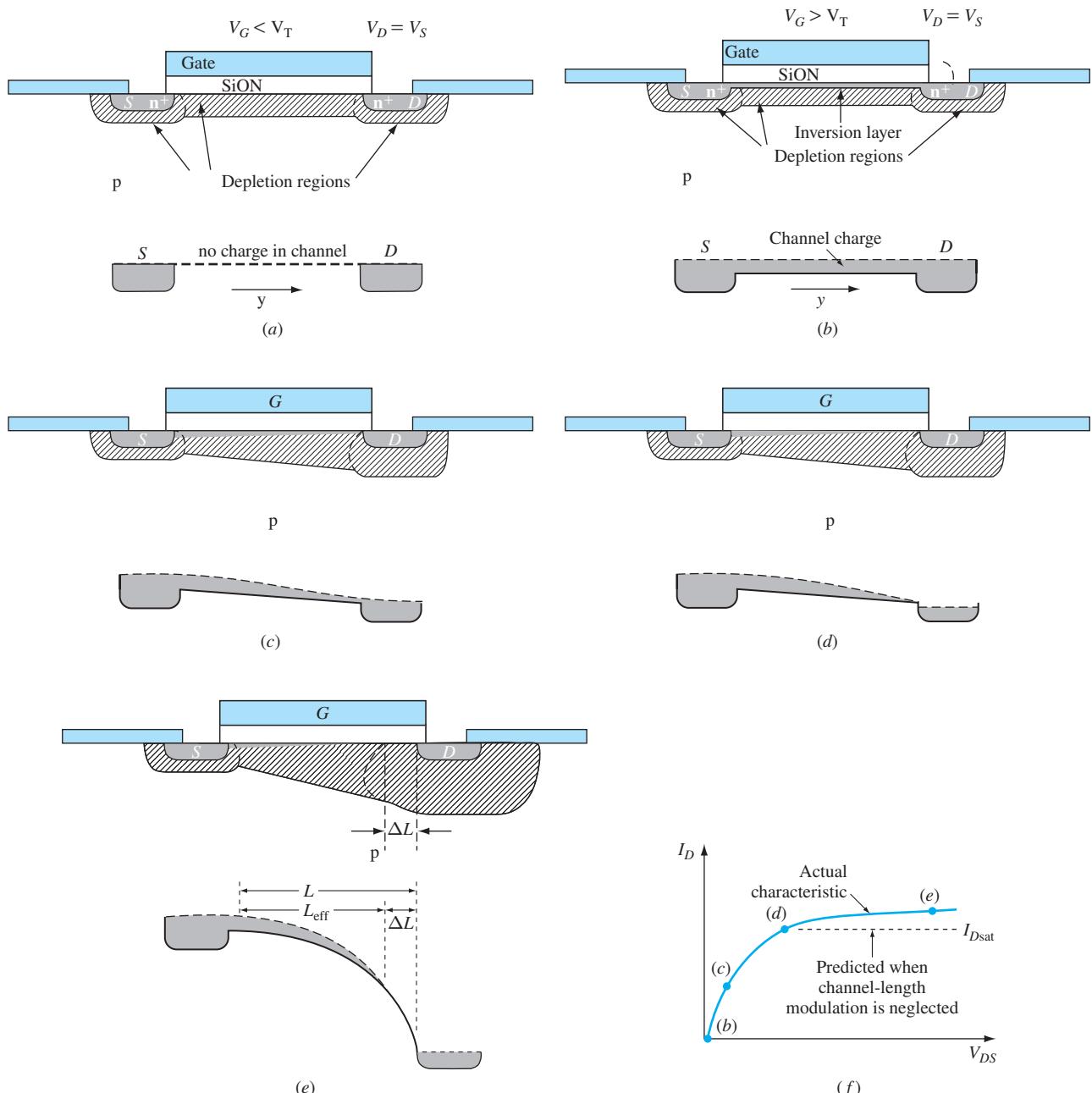


Figure 7.24 Qualitative explanation for channel-length modulation. Parts (a) to (e) show the energy bands (top) and hybrid diagrams (bottom). Parts (a) to (d) repeat the explanation of the simple long-channel model. In (e), as the drain voltage continues to increase, the point at which the channel charge approaches 0 (shaded region), or the point at which $V_{ch} = V_{GS} - V_T$, moves along the channel toward the source. The channel becomes effectively shorter. (e) The corresponding points on the I_D - V_{DS} characteristics are shown in (f). From point (d) on, the simple model predicts constant current (dashed line).

where ΔL is a function of V_{DS} . For small $\Delta L/L$,

$$\frac{1}{\left(1 - \frac{\Delta L}{L}\right)} \approx \left(1 + \frac{\Delta L}{L}\right)$$

and to first approximation, for $V_{DS} > V_{DS\text{sat}}$ the fractional change in channel length is proportional to $V_{DS} - V_{DS\text{sat}}$:

$$\frac{\Delta L}{L} = \lambda(V_{DS} - V_{DS\text{sat}}) \quad (7.42)$$

The quantity λ is known as the *channel length modulation parameter* (a SPICE parameter).⁸ For $V_{DS} > V_{DS\text{sat}}$ then

$$I_D \approx I_{D\text{sat}}[1 + \lambda(V_{DS} - V_{DS\text{sat}})] \quad (7.43)$$

The slope $\partial I_D / \partial V_{DS}$ in the I_D - V_{DS} characteristic in saturation is the differential output conductance. The output conductance is thus proportional to λ . From Equation (7.42) it is seen that λ increases with decreasing L .

Extrapolation of the I_D - V_{DS} plots to $I_D = 0$ occurs at a voltage V_A , often referred to as the “Early voltage” analogous to a similar effect in the electrical characteristics in bipolar transistors.

EXAMPLE 7.5

Find the SPICE parameter λ for the device of Figure 7.23.

Solution

From Equation (7.43), for $V_{DS} > V_{DS\text{sat}}$,

$$I_D = I_{D\text{sat}}[1 + \lambda(V_{DS} - V_{DS\text{sat}})]$$

Then

$$\lambda = \frac{1}{I_{D\text{sat}}} \frac{\partial I_D}{\partial V_{DS}} = \frac{1}{I_{D\text{sat}}} \frac{\Delta I_D}{\Delta V_{DS}}$$

For $V_{GS} = 2.1$ V, from Figure 7.21 the saturation voltage is $V_{DS\text{sat}} \approx 1.2$ V and $I_{D\text{sat}} \approx 5.1$ mA. To find the slope, we extrapolate the straight-line portion of the I_D - V_{DS} curve from 2.5 V to 0 V, as shown in Figure 7.25. We obtain

$$\frac{\Delta I_D}{\Delta V_{DS}} = \frac{5.36 - 4.9}{2.5 - 0} = \frac{0.46 \text{ mA}}{2.5 \text{ V}}$$

Then

$$\lambda = \left(\frac{1}{5.1 \text{ mA}}\right) \left(\frac{0.46 \text{ mA}}{2.5 \text{ V}}\right) = 0.036 \text{ V}^{-1}$$

corresponding to an Early voltage (V_A) of 28 V.

⁸In SPICE Level 1, the expression used is $I_D = I_{D\text{sat}} (1 + \lambda V_{DS})$.

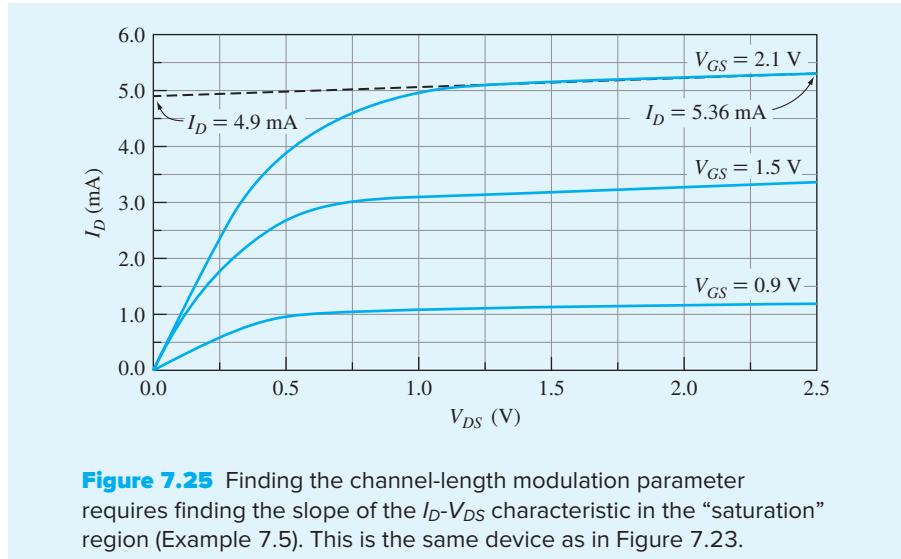


Figure 7.25 Finding the channel-length modulation parameter requires finding the slope of the I_D - V_{DS} characteristic in the “saturation” region (Example 7.5). This is the same device as in Figure 7.23.

7.3.2 MORE REALISTIC LONG-CHANNEL MODELS: EFFECT OF FIELDS ON THE MOBILITY

In the above long-channel model for finding the I - V_{DS} characteristics of a MOSFET, we assumed that the mobility was a constant. In reality, the carrier mobility is dependent on the transverse and longitudinal fields \mathcal{E}_T and \mathcal{E}_L . Thus, the electrical characteristics of FETs depend on the strengths of these channel fields. For example, for large longitudinal field \mathcal{E}_L , the carrier velocities saturate, which limits the current actually obtainable to something lower than predicted by the simple model. At low longitudinal fields, the mobility is independent of \mathcal{E}_L but it does depend on \mathcal{E}_T . We will consider the effects of these two fields separately.

When the longitudinal field (along the channel) is small enough that the velocity is proportional to \mathcal{E}_L , the carriers have what is called their *low-field mobility*, μ_{lf} . The value of this low-field mobility, however, is influenced by the transverse field (across the channel) \mathcal{E}_T .

We first consider the effects of \mathcal{E}_T on the low-field mobility, and then examine how that affects the electrical characteristics of a FET. In the next section, we repeat this to account for the effects of \mathcal{E}_L .

Effect of the Transverse Field \mathcal{E}_T on the Low-Field Mobility In general, the rate of change of the transverse electric field, $\partial\mathcal{E}_x/\partial x$, and longitudinal field $\partial\mathcal{E}_y/\partial y$, are both functions of gate-source voltage and drain-source voltage. To obtain an expression for the depletion layer width, Poisson’s equation must be solved in two dimensions. To simplify the mathematics, the assumption is made that $\frac{\partial\mathcal{E}_x}{\partial x} \gg \frac{\partial\mathcal{E}_y}{\partial y}$ so that a one-dimensional equation can be solved. This implies that the change in depletion width is a function of the gate-source voltage only.

This assumption is known as the gradual-channel approximation or GCA. In this section, we assume that the longitudinal field is low enough that the carrier velocity is small compared with the saturation velocity for electrons moving along the channel. Thus, the electrons have their low-field mobility μ_{lf} . The transverse field \mathcal{E}_T , however, influences the value of the low-field mobility, [2] which earlier we took to be constant ($\mu = \mu_{lf}$). Let us examine the origin of this effect.

In addition to the scattering mechanisms in bulk semiconductors, e.g., lattice and impurity scattering, the electron in the channel of a FET is additionally scattered by collisions with the walls of the channel, as shown in Figure 7.26a. This reduces the mean free time between collisions \bar{t} , and thus μ , from the bulk values. As electrons travel from source to drain, they are restricted to the channel region by the potential barrier at the Si/oxide interface and the barrier in E_C in the Si, Figure 7.26b. Note that most of the electrons in the channel are near the bottom of the potential well formed by these barriers, where the channel is extremely narrow. In practical MOSFETs, this additional scattering mechanism reduces the low-field mobility μ_{lf} by a factor of about 3 to 6 from the bulk value.

Let us examine this effect more analytically. We showed in Chapter 3 that in bulk Si the mean free time between collisions for electrons was on the order of 2×10^{-13} s. The mean free path then is approximately

$$\bar{l} \approx \bar{v} \bar{t}$$

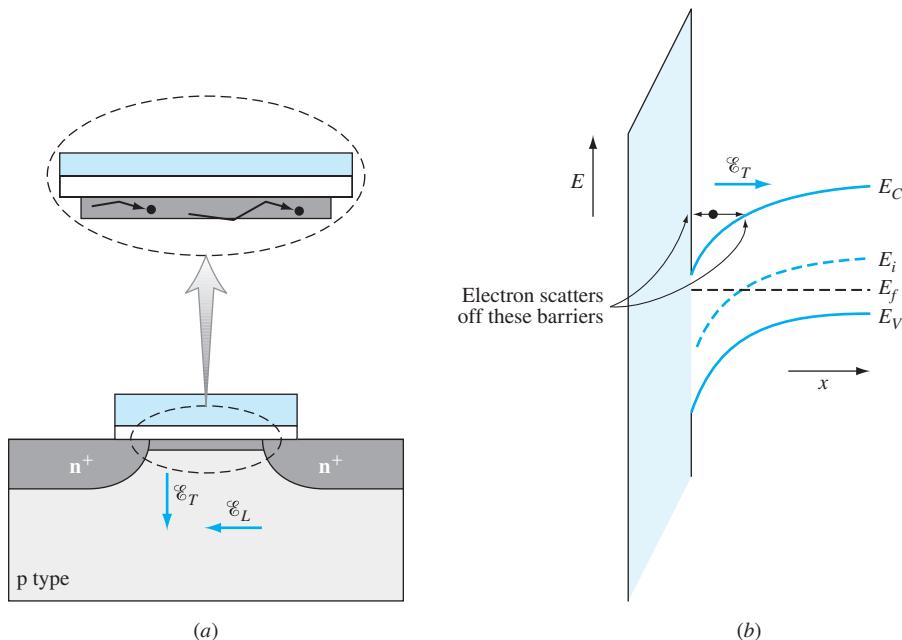


Figure 7.26 The effect of the transverse electric field on the mobility. (a) The electrons in the channel collide with the “walls” of the channel. (b) The energy band diagram shows that the walls are potential barriers at the oxide interface and the barrier of the depletion region in the semiconductor.

where \bar{v} is the average thermal speed. At room temperature, this speed is on the order of $10^7 \text{ cm/s} = 10^5 \text{ m/s}$. Thus for bulk Si, the mean free path is about $\bar{l} \approx 2 \times 10^{-13} \times 10^5 = 2 \times 10^{-8} \text{ m} = 20 \text{ nm}$, and is independent of direction.

In a MOSFET, the mean free path \bar{l} and the mean free time between collisions \bar{t} for electrons traveling along the channel (in the y and z directions with no x -directed or transverse velocity) should be about the same as for bulk Si. For electrons with a v_x component, however, there is the additional scattering from collisions with the channel walls just discussed. This additional scattering reduces \bar{t} and the mobility μ , as seen in the following example.

EXAMPLE 7.6

Estimate the time \bar{t}_x between collisions for a channel electron traveling in the x direction, perpendicular to the gate. Compare this with the mean free time \bar{t} of an electron in bulk silicon.

Solution

Consider an electron with energy $\frac{3}{2}KT$ above the channel floor, and suppose the transverse field is $10^5 \text{ V/cm} = 10^7 \text{ V/m}$ and assumed constant with x . Consider the electron to have just made a collision at the Si/oxide interface at $t = 0$, where its kinetic energy is $\frac{3}{2}KT = m^* v_{\max}^2/2$. This is its maximum velocity because, as the electron goes across the channel in Figure 7.27, its total energy is constant but the potential energy is increasing, so the kinetic energy (and thus the electron velocity) is decreasing. The force on the electron is $F = -q\mathcal{E}_T = m^*dv/dt$ and tends to decelerate it.

We can write

$$-q\mathcal{E}_T \int_0^{\bar{t}_x} dt = m^* \int_{v_{\max}}^0 dv$$

and

$$\bar{t}_x = \frac{m^* v_{\max}}{q\mathcal{E}_T}$$

But since

$$\frac{m^* v_{\max}^2}{2} = \frac{3}{2}kT$$

we have for v_{\max} :

$$v_{\max} = \sqrt{\frac{3kT}{m^*}}$$

Thus,

$$\bar{t}_x = \frac{m^*}{q\mathcal{E}_T} \sqrt{\frac{3kT}{m^*}} = \frac{\sqrt{3m^* kT}}{q\mathcal{E}_T}$$

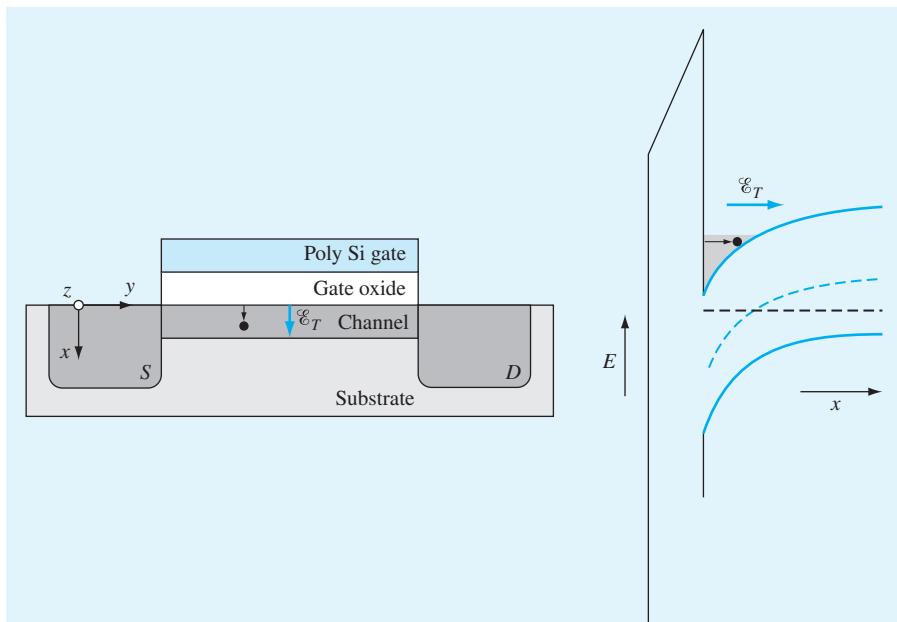


Figure 7.27 Geometry for Example 7.6. We consider only the transverse component of the electron's motion.

Expressing kT in eV gives

$$\bar{t}_x = \frac{1}{\mathcal{E}_T} \sqrt{\frac{3m^*kT(\text{eV})}{q}}$$

Letting m^* be the conductivity effective mass, $m^* = 0.26m_0$, we have

$$\bar{t}_x = \frac{1}{10^7} \sqrt{\frac{3 \times 0.26 \times 9.11 \times 10^{-31} \times 0.026}{1.6 \times 10^{-19}}} = 0.34 \times 10^{-13} \text{ s}$$

which is approximately a factor of 6 smaller than the value of \bar{t} in bulk Si. As a result, sidewall scattering is the predominant scattering mechanism for the x -directed electrons.

EXAMPLE 7.7

For Example 7.6, find the distance \bar{l}_x between electron collisions.

Solution

For constant \mathcal{E}_T , $E_K(\text{eV}) = \mathcal{E}_T \bar{l}_x$

$$\bar{l}_x = \frac{E_K(\text{eV})}{\mathcal{E}_T} = \frac{\frac{3}{2} \times 0.026}{10^7} = 3.9 \text{ nm}$$

This is appreciably less than the 20 nm for the mean free path \bar{l} in bulk silicon.

Note that electrons near the bottom of the channel have smaller kinetic energies and that the transverse field is higher there, reducing their time between collisions even further. Since electrons in the channel have a range of v_x , however, some average \bar{t}_x must be used, and it must also be averaged with the mean free times in the y and z directions to obtain an overall \bar{t} . The point is, \bar{t} and μ are appreciably reduced from their bulk values.

The electron mean free path and thus \bar{t} are therefore dependent on \mathcal{E}_T , which can be seen another way in Figure 7.28. There we have plotted part of the energy band diagram in the region of the channel. The oxide interface is on the left, and two possible conduction band edges are shown on the right. The steeper the slope, the higher the transverse electric field and thus the smaller the mean free time and the mobility.

The value of \mathcal{E}_T in the channel depends on the slope of the conduction band edge. This is a function of two things. First, there is the charge per unit area, Q_B , in the Si depletion region adjacent to the channel. In an n-channel device, these are the fixed, negatively charged ionized acceptors in the p-type substrate. Second, there is the mobile charge in the channel, Q_{ch} . In an NFET the channel charges are electrons, which tend to raise the potential energy of the channel, affecting the slope. Therefore, the transverse field strength varies with doping, bias conditions, and depth in the channel.

The effect of the transverse field on the low-field mobility is discussed further in the Supplement to Part 3. However, experimentally the low-field mobility can be expressed as

$$\mu_{lf} = \frac{\mu_0}{1 + \theta(V_{GS} - V_T - V_{ch})} \quad (7.44)$$

where μ_0 is the channel mobility at the source ($V_{ch}=0$) at threshold ($V_{GS}=V_T$). Here the quantity θ is a measured empirical parameter, on the order of 0.03 to 0.2 V⁻¹,

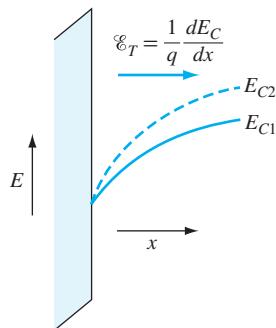


Figure 7.28 With increased band bending, the transverse field \mathcal{E}_T increases. This in turn reduces \bar{t} , \bar{t} , and μ_{lf} .

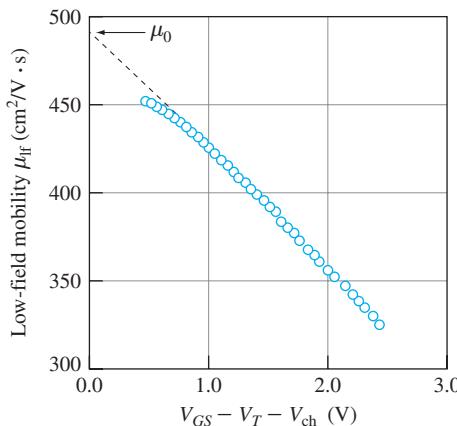


Figure 7.29 Variation of the low-field mobility as a function of $V_{GS} - V_T - V_{ch}$ for an n-channel silicon MOSFET. The low-field mobility can be expressed as $\mu_{if} = \mu_0[1 - \theta(V_{GS} - V_T - V_{ch})]$.

and depends on the processing parameters, including the substrate doping, substrate bias, and oxide thickness.

An experimental plot [3] of μ_{if} versus $(V_{GS} - V_T - V_{ch})$ for a silicon MOSFET is shown in Figure 7.29. We see there that μ_{if} varies about 30 percent over about a 2-V variation in gate voltage.

We saw that the low-field mobility in the channel, μ_{if} , increases from source to drain. It can be seen in the figure that μ_{if} decreases approximately linearly with $(V_{GS} - V_T - V_{ch})$:

$$\mu_{if} = \mu_0[1 - \theta(V_{GS} - V_T - V_{ch})]$$

where μ_0 is the zero-voltage extrapolation and θ the negative of the slope. Note that the maximum mobility is less than $500 \text{ cm}^2/\text{V} \cdot \text{s}$, considerably less than the bulk value of about 1330.

Since $\theta(V_G - V_T - V_{ch})$ is small compared with unity, and since for $x \ll 1$, $1 - x \approx 1/(1 + x)$, μ_{if} can be expressed in the more customary form of Equation (7.44).

EXAMPLE 7.8

Find the value for θ for the device of Figure 7.29.

Solution

Since $\mu_{if} = \mu_0[1 - \theta(V_{GS} - V_T - V_{ch})]$, we can find a formula for θ by taking the derivative of this expression. Using points on the graph gives

$$\theta = -\frac{d\frac{\mu_{if}}{\mu_0}}{d(V_{GS} - V_T - V_{ch})} = -\frac{1}{\mu_0} \frac{d\mu_{if}}{d(V_{GS} - V_T - V_{ch})} = -\frac{1}{480} \frac{(355 - 480)}{(2 - 0)} = 0.13 \text{ V}^{-1}$$

Effect of \mathcal{E}_T on the I_D - V_{DS} Characteristics Now that we have considered the effect of \mathcal{E}_T on μ_{lf} , the next step is to see how that affects the I_D - V_{DS} characteristics compared with the constant-mobility model. We correct Equation (7.17) to include the variation of the mobility with V_{ch} [Equation (7.44)]. The expression becomes

$$I_D \int_0^L dy = WC'_\text{ox} \mu_0 \int_0^{V_{DS}} \frac{V_{GS} - V_T - V_{ch}}{1 + \theta(V_{GS} - V_T - V_{ch})} dV_{ch} \quad (7.45)$$

which yields

$$I_D = \frac{WC'_\text{ox} \mu_0}{\theta^2 L} \left\{ \theta V_{DS} + \ln \left[\frac{1 + \theta(V_{GS} - V_T - V_{DS})}{1 + \theta(V_{GS} - V_T)} \right] \right\} \quad (7.46)$$

Equation (7.46) is, however, somewhat unwieldy and is not amenable to physical interpretation. Consequently, Equation (7.44) is normally approximated as

$$\mu_{lf} = \frac{\mu_0}{1 + \theta(V_{GS} - V_T)} \quad (7.47)$$

In this case the dependence on V_{ch} has been conveniently removed, so that $\mu = \mu_{lf}$ can be moved back outside the integral of Equation (7.16) with the result

$$\begin{aligned} I_D &= \frac{WC'_\text{ox} \mu_0}{L[1 + \theta(V_{GS} - V_T)]} \left(V_{GS} - V_T - \frac{V_{DS}}{2} \right) V_{DS} \\ &= \frac{WC'_\text{ox} \mu_{lf}}{L} \left(V_{GS} - V_T - \frac{V_{DS}}{2} \right) V_{DS} \quad V_{DS} \leq V_{DS\text{sat}} \end{aligned} \quad (7.48)$$

$$I_{D\text{sat}} = \frac{WC'_\text{ox} \mu_{lf}}{L} \left(V_{GS} - V_T - \frac{V_{DS\text{sat}}}{2} \right) V_{DS\text{sat}} \quad V_{DS} \geq V_{DS\text{sat}} \quad (7.49)$$

These equations are similar to Equations (7.24) and (7.25) with constant μ replaced by μ_{lf} . The difference is that here the mobility is a function of V_{GS} .

EXAMPLE 7.9

Compare the I_D - V_{DS} characteristics for a MOSFET using the constant mobility model, and then taking the transverse field into account. Let $W/L = 5$, $\theta = 0.13 \text{ V}^{-1}$, $t_{ox} = 5 \text{ nm}$, $V_T = 1 \text{ V}$ and $\mu_0 = 480 \text{ cm}^2/\text{V} \cdot \text{s}$. The oxide is SiO_2 .

Solution

We use Equations (7.48) and (7.49) where $\mu_{lf} = \mu_0$ in the constant mobility model and $\mu_{lf} = \mu_0/[1 + \theta(V_{GS} - V_T)]$ to include the effect of \mathcal{E}_T on μ_{lf} .

The oxide capacitance per unit area is

$$C'_\text{ox} = \frac{\epsilon_\text{ox}}{t_{ox}} = \frac{3.9(8.85 \times 10^{-12})}{5 \times 10^{-9}} = 6.9 \times 10^{-3} \text{ F/cm}^2 = 6.9 \times 10^{-7} \text{ F/cm}^2$$

The low-field mobility depends on $V_{GS} - V_T$. For $V_{GS} - V_T = 0$ V,

$$\mu_{lf} = \frac{\mu_0}{1 + 0} = \mu_0 = 480 \text{ cm}^2/\text{V} \cdot \text{s}$$

For $V_{GS} - V_T = 1$ V,

$$\mu_{lf} = \frac{\mu_0}{1 + \theta(V_{GS} - V_T)} = \frac{480}{1 + (0.13)(1)} = 425 \text{ cm}^2/\text{V} \cdot \text{s}$$

For $V_{GS} - V_T = 2$ V,

$$\mu_{lf} = \frac{\mu_0}{1 + \theta(V_{GS} - V_T)} = \frac{480}{1 + (0.13)(2)} = 380 \text{ cm}^2/\text{V} \cdot \text{s}$$

For $V_{GS} - V_T = 3$ V,

$$\mu_{lf} = \frac{\mu_0}{1 + \theta(V_{GS} - V_T)} = \frac{480}{1 + (0.13)(3)} = 345 \text{ cm}^2/\text{V} \cdot \text{s}$$

For $V_{GS} - V_T = 4$ V,

$$\mu_{lf} = \frac{\mu_0}{1 + (0.13)(4)} = 316 \text{ cm}^2/\text{V} \cdot \text{s}$$

Figure 7.30 compares the constant mobility curves with the curves obtained by considering the effect of the transverse field. Notice that the drain currents are noticeably smaller when the transverse field is accounted for. Note, however, that neglecting the effect of V_{ch} , i.e., using Equation (7.47) instead of (7.44), overestimates the effect of V_{GS} on the low-field mobility.

Effect of the Longitudinal Field \mathcal{E}_L on Channel Mobility We have seen that the transverse field has an effect on the mobility and thus affects the values of the saturation current for a given gate voltage. In this section we examine the effect of the longitudinal field \mathcal{E}_L on the mobility and thus on the I_D - V_{DS} curves.

In semiconductors, carrier velocities increase with increasing electric field and eventually saturate. This velocity saturation effect can be significant for carriers in the channel of a FET. In modern devices, the gate lengths are very small (a fraction of a micrometer), resulting in very high fields over a significant fraction of the channel length.

For many semiconductors, including Si, the carrier velocity in the channel of a FET can be empirically expressed as

$$|v| = \frac{\mu_{lf} |\mathcal{E}_L|}{1 + \frac{\mu_{lf} |\mathcal{E}_L|}{v_{sat}}} \quad (7.50)$$

where μ_{lf} is the low-field mobility (channel carrier mobility at low \mathcal{E}_L) and v_{sat} is the carrier saturation velocity in the channel. Equation (7.50) is used for both electrons in Si n-channel FETs and for holes in p-channel devices.

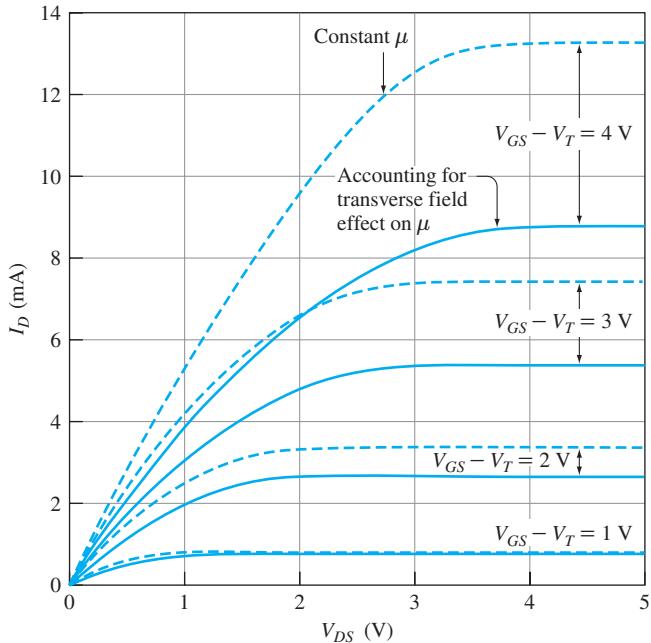


Figure 7.30 Comparison of I_D - V_{DS} characteristics computed by using the constant mobility model ($\theta = 0$, dashed lines) and taking into account the effect of the transverse field (solid line) for $\theta = 0.13 \text{ V}^{-1}$. The transverse field tends to reduce the currents.

Since we know that

$$|v| = \mu |\mathcal{E}_L| \quad (7.51)$$

the mobility can be expressed as

$$\mu = \frac{\mu_{lf}}{1 + \frac{\mu_{lf} |\mathcal{E}_L|}{v_{sat}}} \quad (7.52)$$

From Equation (7.52) we see that with increasing field $|\mathcal{E}_L|$ along the channel, μ decreases. This is a result of the reduction in the mean free time between collisions primarily due to optical phonon scattering. Figure 7.31 shows experimental data of electron mobility as a function of \mathcal{E}_L in an n-channel MOSFET. The data are matched to Equations (7.51) and (7.52) (solid lines). From the figure, we see that velocity saturates at $\mathcal{E}_L \approx 4 \times 10^4 \text{ V/cm}$ and the saturation velocity is $v_{sat} \approx 4 \times 10^6 \text{ cm/s}$ for this device. Although v_{sat} depends somewhat on μ_{lf} —which depends on temperature, transverse field, and substrate doping concentration—for carriers in the channel of a Si MOSFET, we will use $v_{sat} \approx 10^7 \text{ cm/s}$ for both electrons and holes, as is common in the literature. [3] These values are those found for bulk Si.

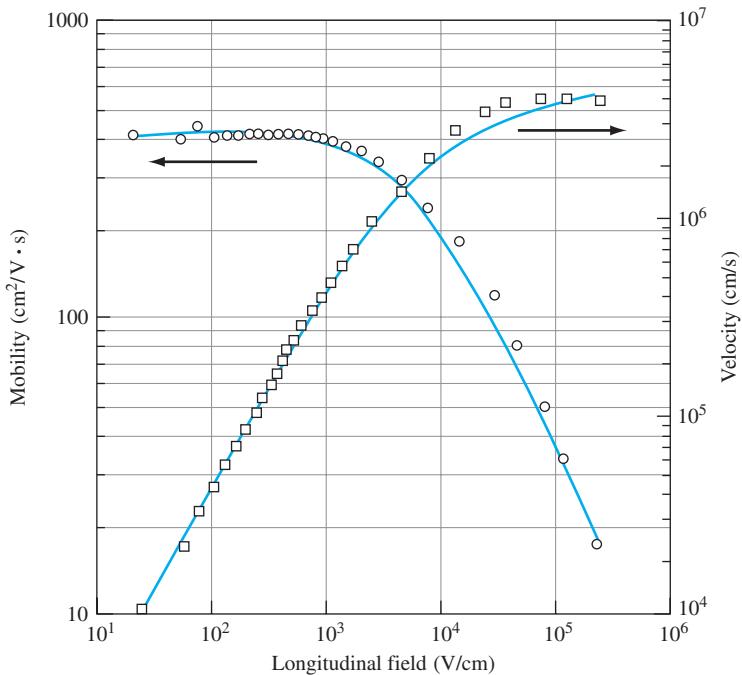


Figure 7.31 Channel electron mobility and velocity ($v = \mu \mathcal{E}$) as a function of longitudinal field for $V_{GS} = 1.42$ V.

Effect of \mathcal{E}_L on the I_D - V_{DS} Characteristics of MOSFETs We have seen that μ varies with longitudinal electric field, so now we will revise the simple long-channel model to account for this influence of the carrier mobility on the I_D versus V_{DS} characteristics. If we substitute Equation (7.52) into Equation (III.7), we can write

$$I_D dy = \frac{-WQ_{ch}\mu_{lf} dV_{ch}}{1 + \frac{\mu_{lf}|\mathcal{E}_L|}{v_{sat}}} \quad (7.53)$$

The electric field can be expressed as $|\mathcal{E}_L| = dV_{ch}/dy$, so, substituting into the denominator, we have

$$I_D dy = -\frac{WQ_{ch}\mu_{lf} dV_{ch}}{1 + \frac{\mu_{lf}}{v_{sat}} \frac{dV_{ch}}{dy}} \quad (7.54)$$

Multiplying both sides of Equation (7.54) by the denominator and rearranging gives

$$I_D dy + \frac{I_D \mu_{lf}}{v_{sat}} dV_{ch} = -WQ_{ch}\mu_{lf} dV_{ch} \quad (7.55)$$

Now we can integrate both sides. For the sublinear region, we write

$$I_D \int_0^L dy + I_D \int_0^{V_{DS}} \frac{\mu_{lf}}{v_{sat}} dV_{ch} = -W \int_0^{V_{DS}} Q_{ch} \mu_{lf} dV_{ch} \quad V_{DS} \leq V_{DSsat} \quad (7.56)$$

and for the saturation region we have

$$I_{Dsat} \int_0^L dy + I_D \int_0^{V_{DSsat}} \frac{\mu_{lf}}{v_{sat}} dV_{ch} = -W \int_0^{V_{DSsat}} Q_{ch} \mu_{lf} dV_{ch} \quad V_{DS} \geq V_{DSsat} \quad (7.57)$$

Integrating, the results are

$$I_D = -\frac{W \mu_{lf} \int_0^{V_{DS}} Q_{ch} dV_{ch}}{L + \frac{\mu_{lf} V_{DS}}{v_{sat}}} \quad 0 \leq V_{DS} \leq V_{DSsat} \quad (7.58)$$

$$I_{Dsat} = -\frac{W \mu_{lf} \int_0^{V_{DSsat}} Q_{ch} dV_{ch}}{L + \frac{\mu_{lf} V_{DSsat}}{v_{sat}}} \quad V_{DS} \geq V_{DSsat} \quad (7.59)$$

Using the same model for the channel charge Q_{ch} as in the long-channel model [Equation (7.14)], we obtain [4]

$$I_D = \frac{WC'_{ox} \mu_{lf}}{L + \frac{\mu_{lf} V_{DS}}{v_{sat}}} \left[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2} \right] \quad V_{DS} \leq V_{DSsat} \quad (7.60)$$

$$I_{Dsat} = \frac{WC'_{ox} \mu_{lf}}{L + \frac{\mu_{lf} V_{DSsat}}{v_{sat}}} \left[(V_{GS} - V_T)V_{DSsat} - \frac{V_{DSsat}^2}{2} \right] \quad V_{DS} \geq V_{DSsat} \quad (7.61)$$

These are similar to the expressions from the previous model of Equations (7.48) and (7.49), except that L is replaced by $L + \mu_{lf} V_{DS}/v_{sat}$ for $V_{DS} \leq V_{DSsat}$ and by $L + \mu_{lf} V_{DSsat}/v_{sat}$ for $V_{DS} \geq V_{DSsat}$. In effect, the inclusion of the longitudinal field \mathcal{E}_L on the mobility causes the channel length to appear longer by the amount $\mu_{lf} V_{DS}/v_{sat}$ (or $\mu_{lf} V_{DSsat}/v_{sat}$) than for the simpler model. To reflect this effect, Equations (7.58) and (7.59) are normally written⁹

$$I_D = \frac{-W \mu_{lf} \int_0^{V_{DS}} Q_{ch} dV_{ch}}{L \left(1 + \frac{\mu_{lf} V_{DS}}{L v_{sat}} \right)} \quad V_{DS} \leq V_{DSsat} \quad (7.62)$$

⁹Because of the dependence of mobility on the channel field as expressed in Equation (7.52), both μ_{lf} and v_{sat} appear in Equations (7.62) and (7.63) for current, for all values of V_{DS} .

$$I_{D\text{sat}} = \frac{-W\mu_{lf}\int_0^{V_{DS\text{sat}}} Q_{ch} dV_{ch}}{L\left(1 + \frac{\mu_{lf}V_{DS\text{sat}}}{Lv_{sat}}\right)} \quad V_{DS} \geq V_{DS\text{sat}} \quad (7.63)$$

As we shall see shortly, $V_{DS\text{sat}} \neq (V_{GS} - V_T)$, unlike the case for the simple long-channel model.

The quantity in the parentheses in the denominator of Equation (7.63), then, represents the influence of the velocity saturation effect on the I_D - V_{DS} characteristics, [4] and

$$I_D = \frac{I_D(\text{no velocity saturation model})}{1 + \frac{\mu_{lf}V_{DS}}{Lv_{sat}}} \quad V_{DS} \leq V_{DS\text{sat}} \quad (7.64)$$

$$I_{D\text{sat}} = \frac{I_{D\text{sat}}(\text{no velocity saturation model})}{1 + \frac{\mu_{lf}V_{DS\text{sat}}}{Lv_{sat}}} \quad V_{DS} \geq V_{DS\text{sat}} \quad (7.65)$$

where $I_{D\text{sat}}$, without accounting for velocity saturation, is given by Equation (7.49), but $V_{DS\text{sat}} \neq (V_{GS} - V_T)$.

From the preceding equation, it is seen that the reduction of current from that given by the long-channel model is greater as the channel lengths get shorter. This results from \mathcal{E}_L being large, causing the saturation velocity effects to be important over a larger fraction of the channel.

Since we will present some numerical illustrations of n-channel and p-channel MOSFETs, we present in Table 7.1 values for some parameters for room temperature operation of typical Si MOSFETs. We will also use these values in the following example.

EXAMPLE 7.10

Find the value of L for an n-channel Si MOSFET for which the velocity saturation effect reduces the subsaturation current by a factor of 2 for a drain-source voltage $V_{DS} = 2$ V.

Solution

We recognize from Equation (7.64) that we want to set

$$\left(1 + \frac{\mu_{lf}V_{DS}}{Lv_{sat}}\right) = 2$$

From Table 7.1, for $\mu_{lf} = 250 \text{ cm}^2/\text{V} \cdot \text{s}$ and $v_{sat} = 10^7 \text{ cm/s}$, solving for L , we find a channel length of

$$L = \frac{\mu_{lf}V_{DS}}{v_{sat}} = \frac{(250 \text{ cm}/\text{V} \cdot \text{s}) \times (2 \text{ V})}{10^7 \text{ cm/s}} = 5 \times 10^{-5} \text{ cm} = 0.5 \mu\text{m}$$

For a more recent device with $L = 50 \text{ nm}$ we find that the current obtained using the long-channel model is off by a factor of

$$\frac{I_D(\text{long-channel model})}{I_D} = 1 + \frac{\mu_{lf} V_{DS}}{L v_{sat}} = 1 + \frac{(250 \text{ cm}^2/\text{V} \cdot \text{s}) \times (2\text{V})}{(5 \times 10^{-16} \text{ cm}) \times (10^7 \text{ cm/s})} = 10$$

Clearly, the effect of velocity saturation must be taken into account in realistic FETs.

Continuing the derivation of the I_D - V_{DS} characteristics of the FET, we rewrite Equations (7.64) and (7.65) in the standard form:

$$I_D = \frac{W C'_{ox} \mu_{lf} \left(V_{GS} - V_T - \frac{V_{DS}}{2} \right) V_{DS}}{L \left(1 + \frac{\mu_{lf} V_{DS}}{L v_{sat}} \right)} \quad V_{DS} \leq V_{DSsat} \quad (7.66)$$

$$I_{Dsat} = \frac{W C'_{ox} \mu_{lf} \left(V_{GS} - V_T - \frac{V_{DSsat}}{2} \right) V_{DSsat}}{L \left(1 + \frac{\mu_{lf} V_{DSsat}}{L v_{sat}} \right)} \quad V_{DS} \geq V_{DSsat} \quad (7.67)$$

Table 7.1 Some parameters for typical Si MOSFETs

Parameter	n-channel MOSFET	p-channel MOSFET
μ_{lf} (low-field mobility)	$250 \text{ cm}^2/\text{V} \cdot \text{s}$	$100 \text{ cm}^2/\text{V} \cdot \text{s}$
v_{sat} (carrier saturation velocity)	10^7 cm/s	10^7 cm/s
t_{ox} (gate oxide thickness)	2 nm	2 nm

We need to find an expression for V_{DSsat} . We can use the same approach we took in Section 7.3.1. By setting $\partial I_D / \partial V_{DS} = 0$ in Equation (7.66):

$$V_{DSsat} = \frac{v_{sat}}{\mu_{lf}} L \left[\left(1 + \frac{2\mu_{lf}(V_{GS} - V_T)}{v_{sat} L} \right)^{1/2} - 1 \right] \quad (7.68)$$

Multiplying and dividing Equation (7.68) by

$$\left[\left(1 + \frac{2\mu_{lf}(V_{GS} - V_T)}{v_{sat} L} \right)^{1/2} + 1 \right] \text{ and recalling that } (x + 1) \cdot (x - 1) = (x^2 - 1),$$

Equation (7.68) takes an alternate form:

$$V_{DSsat} = \frac{2(V_{GS} - V_T)}{\left[\left(1 + \frac{2\mu_{lf}(V_{GS} - V_T)}{v_{sat} L} \right)^{1/2} + 1 \right]} \quad (7.69)$$

Note from Equation (7.69), that for large L , V_{DSsat} approaches $(V_{GS} - V_T)$ and Equation (7.67) approaches Equation (7.25) for the long-channel model.

EXAMPLE 7.11

Estimate the minimum channel length such that the long-channel formulation gives a reasonable description for an n-channel MOSFET.

■ Solution

From Equation (7.69), $V_{Dsat} \approx (V_{GS} - V_T)$ for $\frac{2\mu_{lf}(V_{GS} - V_T)}{v_{sat}L} \ll 1$

Choose $\frac{2\mu_{lf}(V_{GS} - V_T)}{v_{sat}L} < 0.1$. Then $L > \frac{2\mu_{lf}(V_{GS} - V_T)}{0.1 v_{sat}}$

Letting $\mu_{lf} = 250 \text{ cm}^2/\text{V} \cdot \text{s}$, $v_{sat} = 10^7 \text{ cm/s}$, and $(V_{GS} - V_T) = 1 \text{ V}$

$$\text{we find } L > \frac{(2 \times 250 \text{ cm}^2/\text{V} \cdot \text{s})(1 \text{ V})}{0.1 \times 10^7 \text{ cm/s}} = 5 \times 10^{-4} \text{ cm} = 5 \mu\text{m}$$

So, for L greater than about $5 \mu\text{m}$ the long-channel model is a reasonable approximation.

For small L such that $\frac{\mu_{lf}V_{Dsat}}{v_{sat}L} \gg 1$, Equation (7.67) becomes

$$I_{Dsat} = WC'_{ox} v_{sat} \left(V_{GS} - V_T - \frac{V_{Dsat}}{2} \right) \quad (7.70)$$

and is only slightly dependent on L (through V_{Dsat}). In the limit as $L \rightarrow 0$ in Equation (7.68), $V_{Dsat} \rightarrow 0$ and

$$I_{Dsat} = WC'_{ox} v_{sat} (V_{GS} - V_T) \quad (7.71)$$

Note that in this formulation, for $L = 0$, I_{Dsat} is finite. Here, however, v_{sat} is not a physically meaningful quantity since it is dependent on carrier collisions in the channel. Since for L less than one mean free path, there are negligible collisions in the channel, a different transport mechanism must be employed. This is called *ballistic transport* and is treated in Section 7.5.

It is of interest to determine the value of L satisfying the above assumption that $\frac{\mu_{lf}V_{Dsat}}{L v_{sat}} \gg 1$, and permitting the use of Equation (7.70). Assuming $\frac{\mu_{lf}V_{Dsat}}{L v_{sat}} \gg 10$, we obtain $L \ll \frac{\mu_{lf}V_{Dsat}}{10 v_{sat}}$. To find V_{Dsat} as a function of L we use Equation (7.68). Since except for very small $(V_{GS} - V_T)$ the unity terms can be ignored, and

$$V_{Dsat} \approx \left(\frac{2 v_{sat} L (V_{GS} - V_T)}{\mu_{lf}} \right)^{1/2} > \frac{10 v_{sat} L}{\mu_{lf}}.$$

Solving for L ,

$$L < \frac{2 \mu_{lf} (V_{GS} - V_T)}{100 v_{sat}} = 5 \times 10^7 \text{ cm} = 5 \text{ nm}$$

For $L < 5 \text{ nm}$, the drift model presented here is not valid, since an electron's mean free path is on the order of 10 to 20 nm, and thus few electrons make collisions in the channel. The model in which electrons make no (or few) collisions in the channel (ballistic transport model) is more appropriate for channel lengths less than about 10 nm. For $10 \text{ nm} < L < 20 \text{ nm}$, a combination of the drift and ballistic models is appropriate.

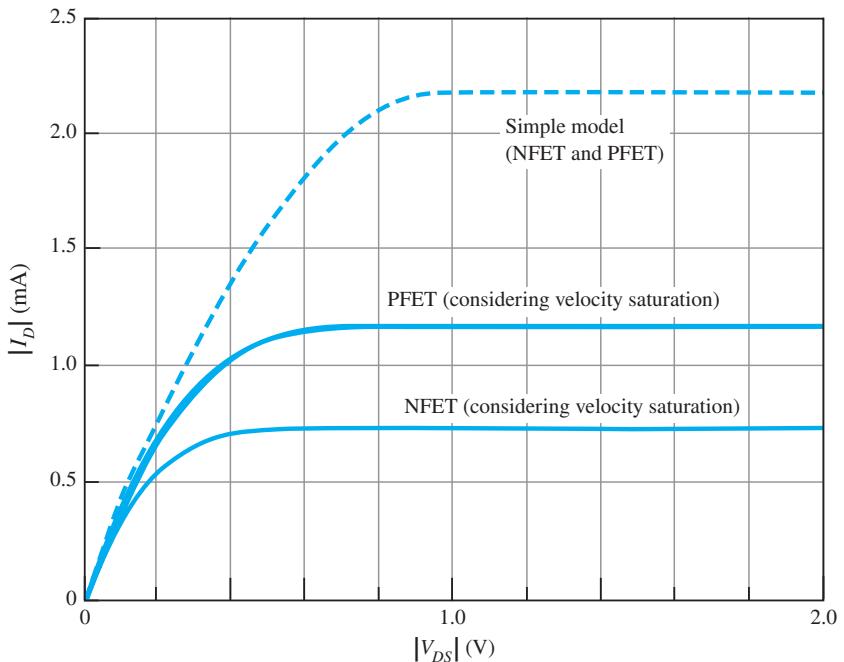


Figure 7.32 The calculated $I-V_{DS}$ characteristics for the simple model and for NMOS and PMOS with carrier velocity saturation accounted for. The case of $V_{GS} - V_T = 1$ V and $L = 100$ nm considered. The W/L ratios of the NFET and PFET have been scaled to produce the same $I-V_{DS}$ curve as predicted from the simple model. The differing mobilities cause the scaled devices to be different.

Let us compare the $I_D = V_{DS}$ characteristics as obtained from the simple drift model and the model that includes velocity saturation. For this example we choose the parameters of Table 7.1 with $V_{GS} - V_T = 1$ V and a channel length of $L = 100$ nm. First, we plot the results from the long-channel model using constant mobility in Figure 7.32 (dashed line).

Before we continue, we notice that both the PFET and the NFET have the same result as for the earlier model. Since the low-field mobilities for electrons and holes are different ($\mu_{lfn} \approx 2.5\mu_{lfp}$), the currents in the two FETs would be significantly different if the devices were otherwise identical. In many circuits using both NFETs and PFETs (e.g., CMOS, as discussed in Chapter 8), it is desirable to have equal saturation currents of both devices. To achieve this for long-channel devices, the W/L ratios of PFETs are made 2.5 times that of NFETs. This is what has been done here. We choose $W = 2 \mu\text{m}$ and $5 \mu\text{m}$ respectively for the NFET and the PFET. This will equate the characteristics for the two devices as predicted from the long-channel model, i.e., Equation (7.48).

The characteristics from the model that considers velocity saturation are also shown in the figure. From Equations (7.66) and (7.67), it is evident that even with the width-to-length ratio corrected as above, the currents in the velocity

saturation model are not the same for the NFET and PFET. We can see from Figure 7.32 that at small V_{DS} , both models predict the same slope. In this region, the longitudinal field is not yet large enough for velocity saturation effects to be important. However, the inclusion of velocity saturation effects causes the saturation current to decrease and the voltage at which current saturates ($V_{DS\text{sat}}$) to decrease also. Since μ_f is larger for the NFET than for the PFET, $I_{D\text{sat}}$ and $V_{DS\text{sat}}$ are smaller for the NFET. As we will see in Chapter 8, this means that the velocity saturation effect reduces the performance of the field-effect transistors.

We indicated earlier that, when velocity saturation is accounted for, the saturation voltage is no longer equal to $V_{GS} - V_T$. As the channel length gets shorter, from Equation (7.68) the saturation voltage $V_{DS\text{sat}}$ also decreases. For example, Figure 7.33 shows that for a channel length of $L = 2 \mu\text{m}$ and $|V_{GS} - V_T| = 2.6 \text{ V}$, the saturation voltages are $V_{DS\text{sat}} = 2.28$ and 2.45 V respectively for the NFET and the PFET. For a shorter channel device, e.g., $L = 65 \text{ nm}$, these values reduce to 0.93 and 1.3 V respectively. These values are appreciably smaller than the $V_{DS\text{sat}} = (V_{GS} - V_T) = 2.6 \text{ V}$ we would get from the constant-mobility long-channel model. Again, this is because the simple long-channel model does not consider velocity saturation. The figure also shows $V_{Ds\text{at}}$ for $V_g - V_t = 1 \text{ V}$.

While the saturation voltage $V_{DS\text{sat}}$ depends on the channel length L (but not the channel width W), the saturation current $I_{D\text{sat}}$ depends on both W and L . Figure 7.34 shows $I_{D\text{sat}}$ as a function of L for a constant W/L ratio of 10. The figure shows results for both n- and p-channel MOSFETs. As expected, for a given W/L ratio, $I_{D\text{sat}}$ decreases with decreasing channel length, again because of velocity saturation.

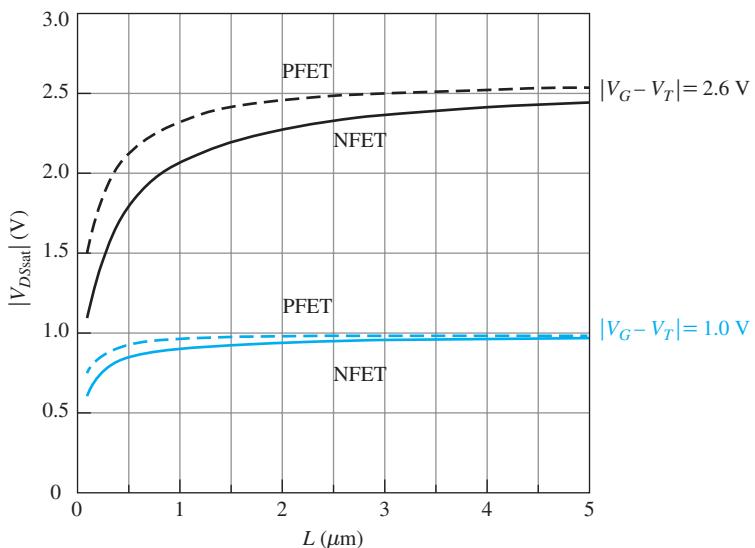


Figure 7.33 The saturation voltage as a function of channel length for $|V_{GS} - V_T| = 2.6 \text{ V}$ and $|V_{GS} - V_T| = 1.0 \text{ V}$.

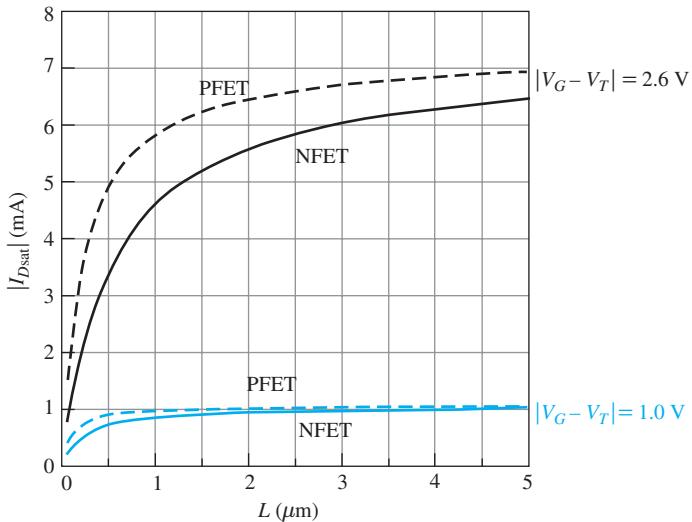


Figure 7.34 Saturation current $|I_{D\text{sat}}|$ as a function of L for n-and p-channel silicon MOSFETs. Here $|(V_{GS} - V_T)| = 1 \text{ V}$ and $|(V_{GS} - V_T)| = 1.0 \text{ V}$, the width-to-length ratio is $W/L = 10$, and $W_p = 2.5 W_n$.

*7.3.3 SERIES RESISTANCE

Next, we look at the resistance of a FET, which is the resistance between the source and the drain. This includes the resistance along the channel itself, plus the resistances between the source and drain contacts and the channel, as shown in Figure 7.35 for an n-channel MOSFET. The total resistance R_{tot} in a FET is

$$R_{\text{tot}} = \frac{V_{DS}}{I_D} = R_S + R_{\text{ch}} + R_D \quad (7.72)$$

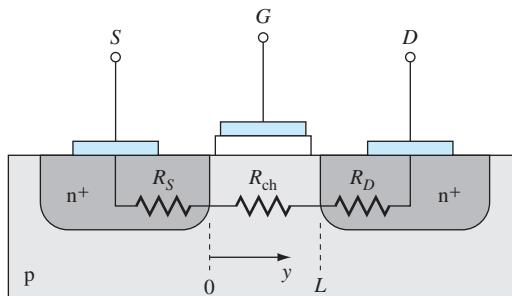


Figure 7.35 Schematic of an NMOS indicating the channel resistance R_{ch} , the source resistance R_S , and the drain resistance R_D .

where R_S is the resistance from the source contact to the source end of the channel, R_{ch} is the resistance of the channel, and R_D is the resistance from the drain end of the channel to the drain contact.

The region $0 \leq y \leq L$ can be treated as an embedded transistor with external resistances R_S and R_D connected in series to the source and drain contacts respectively. The voltage between the drain and source of the “intrinsic” transistor is then given by the drain-source voltage less the voltage drop in the “external” resistors, $V_{DS} - [I_D(R_S + R_D)]$. At the same time, the voltage between the gate and source of the intrinsic transistor is given by $V_{GS} - I_D R_S$. Then Equation (7.66) becomes, below saturation

$$I_D = \frac{WC'_ox\mu_{lf}\left[V_{GS} - I_D(R_S) - V_T - \frac{V_{DS} - I_D(R_S + R_D)}{2}\right]\left[V_{DS} - I_D(R_S + R_D)\right]}{L\left\{1 + \frac{\mu_{lf}[V_{DS} - I_D(R_S + R_D)]}{Lv_{sat}}\right\}} \quad V_{DS} \leq V_{DSsat} \quad (7.73)$$

Because of the symmetry of the MOSFET, the series resistances are normally equal, so that $R_S \approx R_D$ and Equation (7.73) becomes

$$I_D = \frac{WC'_ox\mu_{lf}\left(V_{GS} - V_T - \frac{V_{DS}}{2}\right)(V_{DS} - 2I_D R_S)}{L\left[1 + \frac{\mu_{lf}(V_{DS} - 2I_D R_S)}{Lv_{sat}}\right]} \quad V_{DS} \leq V_{DSsat} \quad (7.74)$$

In saturation, Equation (7.67) is adapted in a similar manner to obtain

$$I_{Dsat} = \frac{WC'_ox\mu_{lf}\left(V_{GS} - V_T - \frac{V_{DSsat}}{2}\right)(V_{DSsat} - 2I_{Dsat} R_S)}{L\left[1 + \frac{\mu_{lf}(V_{DSsat} - 2I_{Dsat} R_S)}{Lv_{sat}}\right]} \quad V_{DS} \geq V_{DSsat} \quad (7.75)$$

Equations (7.74) and (7.75) can be solved for the current, algebraically or iteratively.

We can neglect the series resistances for devices in which the channel resistance far exceeds the series resistances, or $R_{ch} \gg 2R_S$. The series resistance depends on the processing details as well as on the channel width W . Typical values for R_S and R_D are about 10 to 100 Ω .

7.4 COMPARISON OF MODELS WITH EXPERIMENT

We have examined several long-channel models for the I_D - V_{DS} characteristics of MOSFETs. These are the constant-mobility model, the model in which transverse field is included, and the model in which velocity saturation is included. We also modified these to account for series resistance. We now wish to explore the question: How good are these models?

Let us consider an actual n-channel MOSFET designed to operate with a supply voltage of 1.8 V. The measured parameters are:

$$L = 0.25 \mu\text{m}$$

$$W = 9.9 \mu\text{m}$$

$$t_{\text{ox}} = 4.7 \text{ nm}$$

$$\mu_{\text{lf}} = 400 \text{ cm}^2/\text{V} \cdot \text{s}$$

$$R_s = R_D = 19.9 \Omega$$

$$V_T = 0.30 \text{ V}$$

The gate-source voltage is taken equal to the supply voltage, 1.8 V.

The top line in Figure 7.36 gives the results calculated from the simple long-channel model [Equations (7.48) and (7.49)], that is, the model that assumes constant mobility. Since we are considering only a single gate voltage, the appropriate value of low-field mobility has been selected, and we do not need to consider the effect of the transverse field variation with gate voltage. Also plotted are the results obtained by using the same device parameters but considering the effects of the longitudinal field by accounting for velocity saturation [Equations (7.66) and (7.67)]. Also indicated in Figure 7.36 are the results of the model taking into account the source and drain series resistances (R_s and R_D) in addition to velocity saturation. [5] The saturation velocity was not measured for this device but was assumed to be $v_{\text{sat}} = 7 \times 10^6 \text{ cm/s}$ as measured on a similar device. Finally, the actual measured data are plotted.

It can be seen from Figure 7.36 that the current predicted by the velocity saturation model is appreciably smaller than that predicted by the constant mobility

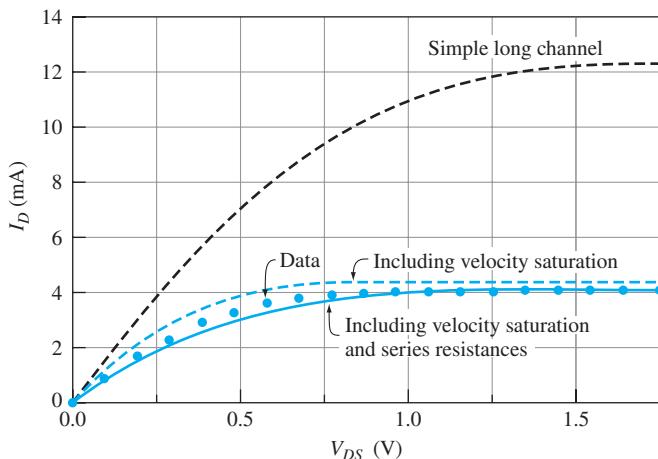


Figure 7.36 Comparison of the simple long-channel model, the model including velocity saturation, the model including both velocity saturation and the series resistances R_S and R_D , and the actual measured data. For this NFET device, $L = 0.25 \mu\text{m}$, $W = 9.9 \mu\text{m}$, and $V_T = 0.3 \text{ V}$. The gate-source voltage is 1.8 V.

long-channel model, and is in reasonable agreement with experiment. Taking drain and source resistance into account results in quite good agreement between theory and experiment.

7.5 BALLISTIC MODEL FOR MOSFETS

In the formulation for n-channel MOSFET current in which the channel is long enough that electrons make many collisions between source and drain, the current is predominantly from drift (drift model) and is proportional to the longitudinal field at the source as expressed in Equation (7.33). In modern devices, however, the channels can be short enough that electrons make few collisions in the channel, so the theory must be modified. Here we discuss a model in which the electrons make *no* collisions between source and drain. This is referred to as a *ballistic transport model* or simply, *ballistic model* [6].

Figure 7.37a shows the energy band diagram for a MOSFET for $V_{GS} > V_T$ and $V_{DS} = 0$. For a MOSFET, the doping in source and drain are equal. In the channel there is no longitudinal field and the electron concentration is uniform;

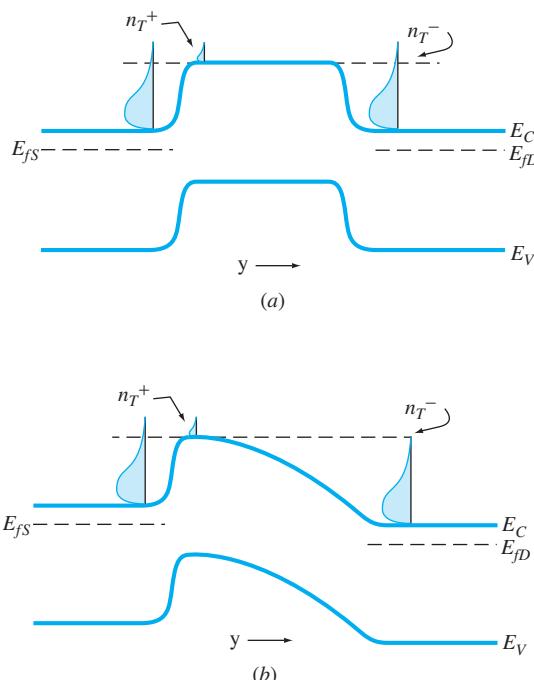


Figure 7.37. (a) MOSFET with $V_{GS} > V_T$ and $V_{DS} = 0$. (b) MOSFET with $V_{GS} > V_T$ and $V_{DS} > 0$. The number of electrons per unit area with energy higher than the barrier at the source end, n_T^+ and at the drain end, n_T^- .

thus the drift and diffusion currents are both zero. The current resulting from electrons flowing from source to drain (I_D^+) is equal to that flowing from drain to source (I_D^-) or

$$I_D = (I_D^+ - I_D^-) = 0$$

Figure 7.37b shows the energy band diagram for $V_{GS} > V_T$ and $V_{DS} > 0$. At the source end, there is some number of electrons n_T^+ per unit area with energy higher than the barrier, traveling to the right. They cross the channel at constant energy (no collisions) and thermalize down to E_C at the drain. (It is assumed that the source and drain are both at equilibrium so that the Fermi level has meaning.) The number of electrons crossing the barrier per unit area supplied by the drain is n_T^- . The drain current then is

$$I_D = I_D^+ - I_D^- \quad (7.76)$$

These currents can be expressed as $I_D^+ = Wq n_T^+ v_T$ and $I_D^- = Wq n_T^- v_T$, where v_T is the magnitude of the average positive (or negative) electron velocity in the y -direction at the top of the barrier adjacent to the source. It is given by

$$v_T = \left(\frac{2kT}{\pi m^*} \right)^{\frac{1}{2}} \quad (7.77)$$

EXAMPLE 7.12

Find the value of v_T for silicon at $T = 300$ K using the conductivity effective mass.

Solution

From Equation (7.77)

$$v_T = \left(\frac{2kT}{\pi m^*} \right)^{\frac{1}{2}} = \left(\frac{2 \times (1.38 \times 10^{-23}) \times (300)}{3.14 \times (0.26 \times 9.11 \times 10^{-31})} \right)^{\frac{1}{2}} = 1.05 \times 10^5 \text{ m/s} = 1.05 \times 10^7 \text{ cm/s}$$

which is close to the saturation velocity for electrons in silicon ($v_{\text{sat}} \approx 10^7$ cm/s) employed in the drift model. This is a coincidence in Si, however; for other materials they are not the same.

From Equation (7.76), the current is,

$$I_D = Wq v_T (n_T^+ - n_T^-) = Wv_T q n_T^+ \left(1 - \frac{n_T^-}{n_T^+} \right) \quad (7.78)$$

Assuming Boltzmann statistics,

$$\begin{aligned} n_T^+ &= N_C \langle t \rangle e^{-\left(\frac{E_T - E_{fS}}{kT}\right)} \\ n_T^- &= N_C \langle t \rangle e^{-\left(\frac{E_T - E_{fD}}{kT}\right)} \end{aligned}$$

where $\langle t \rangle$ is the average thickness of the occupied channel, E_{fS} is the Fermi energy in the source, and E_{fD} is the Fermi level in the drain. Then

$$\frac{n_T^+}{n_T^-} = e^{\left(\frac{E_{fD} - E_{fS}}{kT}\right)} = e^{-\left(\frac{qV_{DS}}{kT}\right)}$$

$$\text{and } I_D = Wv_T q n_T^+ \left(1 - e^{-\left(\frac{qV_{DS}}{kT}\right)} \right) \quad (7.79)$$

The charge per unit area, Q_T , at the top of the barrier is

$$Q_T = -q(n_T^+ + n_T^-) = -qn_T^+ \left(1 + \frac{n_T^-}{n_T^+} \right) = -qn_T^+ \left(1 + e^{-\left(\frac{qV_{DS}}{kT}\right)} \right) \quad (7.80)$$

Thus,

$$qn_T^+ = \frac{-Q_T}{1 + e^{-\left(\frac{qV_{DS}}{kT}\right)}} \quad (7.81)$$

and

$$I_D = -WQ_T v_T \frac{1 - e^{-\left(\frac{qV_{DS}}{kT}\right)}}{1 + e^{-\left(\frac{qV_{DS}}{kT}\right)}} \quad (7.82)$$

When the transistor is **on**, however, $V_{GS} > V_T$, and $Q_T = -C'_\text{ox}(V_{GS} - V_T)$, so Equation (7.82) becomes

$$I_D = WC'_\text{ox} v_t (V_{GS} - V_T) \frac{\left(1 - e^{-\left(\frac{qV_{DS}}{kT}\right)} \right)}{\left(1 + e^{-\left(\frac{qV_{DS}}{kT}\right)} \right)} \quad (7.83)$$

Equation (7.83) resembles the current equation for the conventional (drift) model) repeated here:

$$I_D = \frac{WC'_\text{ox} v_{\text{sat}} \left(V_{GS} - V_T - \frac{V_{DS}}{2} \right) V_{DS}}{L \left(1 + \frac{\mu_{lf} V_{DS}}{L v_{\text{sat}}} \right)} \quad (7.66)$$

with

$$v_t (V_{GS} - V_T) \frac{\left(1 - e^{-\left(\frac{qV_{DS}}{kT}\right)} \right)}{\left(1 + e^{-\left(\frac{qV_{DS}}{kT}\right)} \right)} \text{ replacing } v_{\text{sat}} \left(\frac{\left(V_{GS} - V_T - \frac{V_{DS}}{2} \right) V_{DS}}{L \left(1 + \frac{\mu_{lf} V_{DS}}{L v_{\text{sat}}} \right)} \right).$$

Note that in the ballistic model, the current is independent of L . This is because in this model all electrons from the source that are injected over the barrier arrive at the drain independent of channel length.

For $V_{DS} > \frac{kT}{q}$, the current saturates at a value

$$I_{D\text{sat}} = WC'_\text{ox} v_T (V_{GS} - V_T) \quad (7.84)$$

We compare this ballistic transport model with that for the drift formulation, Equations (7.66), (7.67), and (7.68). Figure 7.38 shows the calculated characteristics of I_D/W for the two models for silicon MOSFETs with $L = 100$ nm (10^{-5} cm), $\epsilon_r = 5$ (assuming SiON as the gate dielectric), $\mu_{lf} = 250$ cm²/V·s, $v_{\text{sat}} = 10^7$ cm/s, $v_T = 1.05 \times 10^7$ cm/s, $V_T = 0.2$ V, and $T = 300$ K. For the ballistic

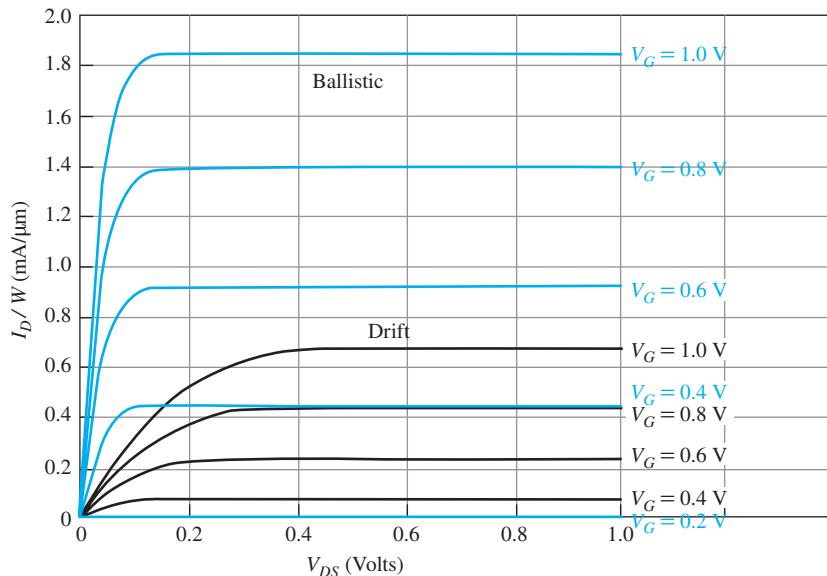


Figure 7.38. The I - V characteristics for a MOSFET using the ballistic versus the normal (drift) model. It is assumed that $V_G = V_{GS}$.

model, the initial slope is much greater than for the drift model, and for a given value of V_G , the current saturates at a value appreciably greater and at a smaller value of drain voltage.

7.6 SOME SHORT-CHANNEL EFFECTS

While the discussed electrical characteristics of MOSFETs depend on channel length, in this section we discuss some additional parasitic effects that become important for short channels, particularly in the submicrometer range.

7.6.1 DEPENDENCE OF EFFECTIVE CHANNEL LENGTH ON V_{DS}

Earlier, we discussed the effect of channel-length modulation in saturation. Figure 7.24 showed that once saturation was reached, any additional drain voltage effectively moved the point along the channel where saturation occurred, effectively shortening the channel. You will recall that this caused a nonzero slope in the I_D - V_{DS} characteristics in the saturation region (the channel-length modulation effect). For very short (submicrometer) channels, however, the drain voltage can modulate the channel length even for $V_{DS} < V_{DS\text{sat}}$.

For long-channel MOSFETs, we took the channel length L to be the distance L_m between the metallurgical junctions of source-channel and drain-channel, as shown in Figure 7.39a for an NMOS transistor.¹⁰ The source and drain are much

¹⁰We define the metallurgical junction as the channel edge of the degenerate source or drain.

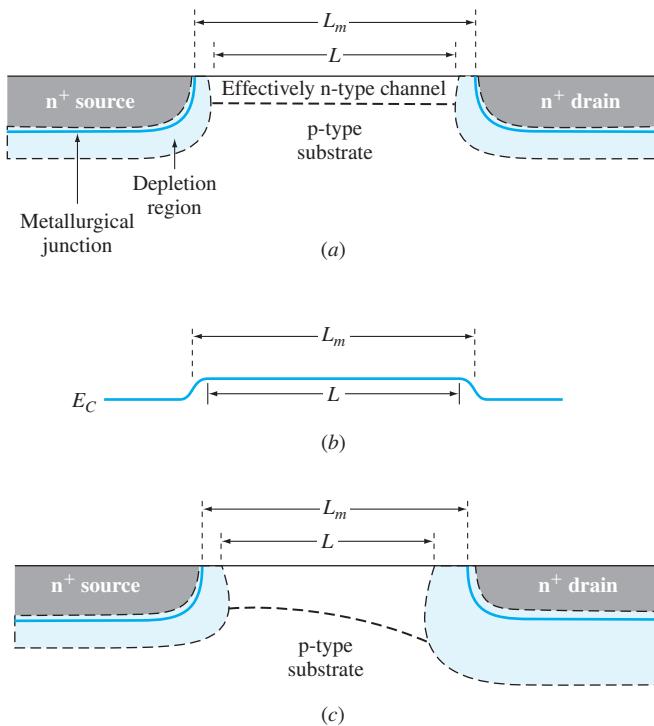


Figure 7.39 Illustration of the depletion regions between the source and the substrate, and between the drain and the substrate (a) for $V_{DS} = 0$. The effective channel length L is reduced from that of the metallurgical channel length L_m . (b) The $E_C - y$ diagram along the channel for $V_{DS} = 0$. (c) With increasing V_{DS} , the effective channel length L decreases.

more heavily doped than the substrate, however, such that the depletion regions between n^+ source and drain and p-type substrate extend into the substrate. This is shown in the figure for $V_{DS} = 0$. The depletion region thickness decreases near the channel, because the source-substrate and drain-substrate voltages are reduced there, as a result of the positive surface potential. Remember, the channel can be effectively n type, even under no bias, because of the band bending. In this region, then, the junctions are n^+n , and the transition regions are primarily in the n region. The energy band diagram along the channel for this case is shown in Figure 7.39b. In effect, the channel length L is less than the metallurgical channel length L_m .

If now a positive drain voltage is applied, the drain-channel depletion region width will increase as shown in Figure 7.39c. Thus, the effective channel length L decreases with increasing V_{DS} . Recall, however, that the channel current increases as the effective channel length gets smaller [as we expect from

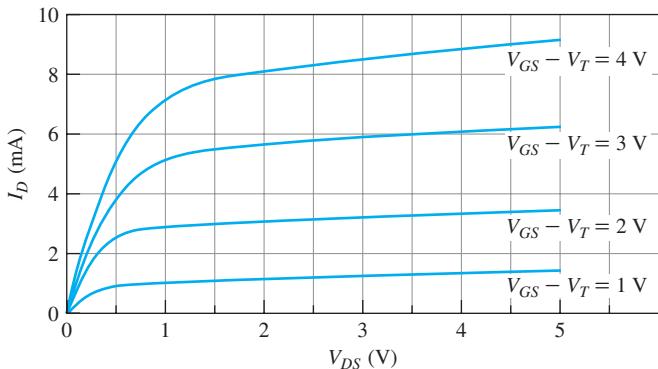


Figure 7.40 For short-channel devices, the reduction of the effective channel length with increasing V_{DS} results in an increase in I_D .

Equations (7.66) and (7.67), where the *effective* channel length is the entire quantity in the denominator]. Thus, we expect that as V_{DS} increases, the channel length decreases, with a resultant increase in the current, both above and below saturation. This is shown in the I_D - V_{DS} characteristics of Figure 7.40. The long-channel (simple) model, and the velocity saturation model (Figure 7.32), as well as the ballistic model (Figure 7.37) showed the I_D - V_{DS} characteristic as flat in the saturation region.

Why do we call this a short-channel effect? It does occur in long-channel devices too. The reason is that for long-channel devices, the depletion region thickness is small compared with the overall channel length, and $L \approx L_m$. The fractional change in L with V_{DS} is small by comparison. In short-channel devices, the shortening becomes significant, and the slope of the I_D - V_{DS} characteristics is affected in both the sublinear and saturation regions. The channel-length modulation discussed earlier happens in both long- and short-channel devices, but only in saturation.

7.6.2 DEPENDENCE OF THRESHOLD VOLTAGE ON THE DRAIN VOLTAGE

The second short-channel effect we discuss is that the value of the threshold voltage can be affected by the drain voltage. [7] Recall that earlier we drew the energy band diagram along the channel. This is shown again in Figure 7.41a.

There is a barrier at the source-channel interface and another barrier at the drain-channel interface. These two barriers were treated as independent in the long-channel devices. For short-channel devices, however, the drain voltage can influence the barrier height at the source end of the channel, as illustrated in Figure 7.41b. If the drain voltage is large enough and the channel short enough,

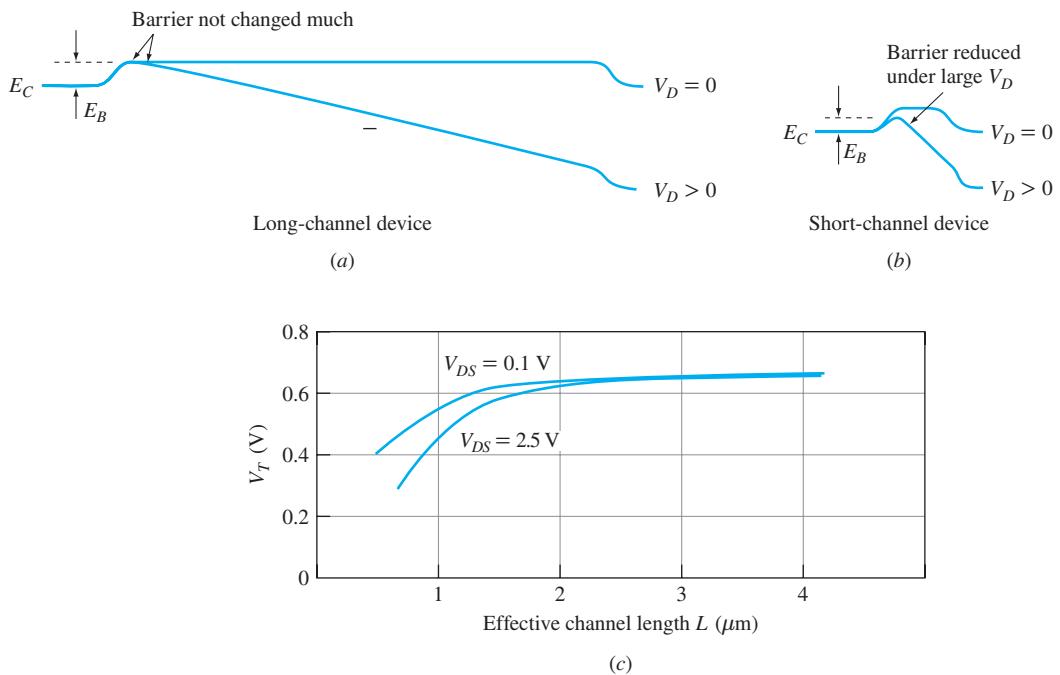


Figure 7.41 (a) For long channels the drain voltage has negligible effect on the barrier at the source-channel interface. (b) For short-channel devices the drain voltage tends to reduce the barrier at the source end. (c) The result is that the threshold voltage is decreased. The effect is more pronounced as the channels get shorter.

the barrier E_B from the source to the channel is reduced. This effect then reduces the threshold voltage. The threshold voltage is plotted as a function of effective channel length in Figure 7.41c for two values of drain voltage. Note that as the channel gets shorter, the effect gets more pronounced. This effect is referred to as the *drain-induced barrier lowering* or *DIBL* effect. The DIBL effect is a major cause of the finite slope in the I_D - V_{DS} characteristics (g_{dsat}) of submicrometer devices (e.g., the device of Figure 7.23 with $L = 0.27 \mu\text{m}$).

7.7 SUBTHRESHOLD LEAKAGE CURRENT

The experimental I_D - V_{GS} characteristics of an n-channel MOSFET with $V_{DS} = 0.1 \text{ V}$ are shown in Figure 7.42 on a semilogarithmic scale. The $[\log I_D]$ - V_{GS} characteristic approximates a straight line below threshold. This means that I_D decreases approximately exponentially with decreasing V_{GS} below threshold.

We can't neglect this subthreshold leakage current for the following reason. In switching circuits, a device is **on** when it is operating well above threshold ($V_{GS} \approx V_{DD}$, where V_{DD} is the supply voltage). The device is in the **off** state when

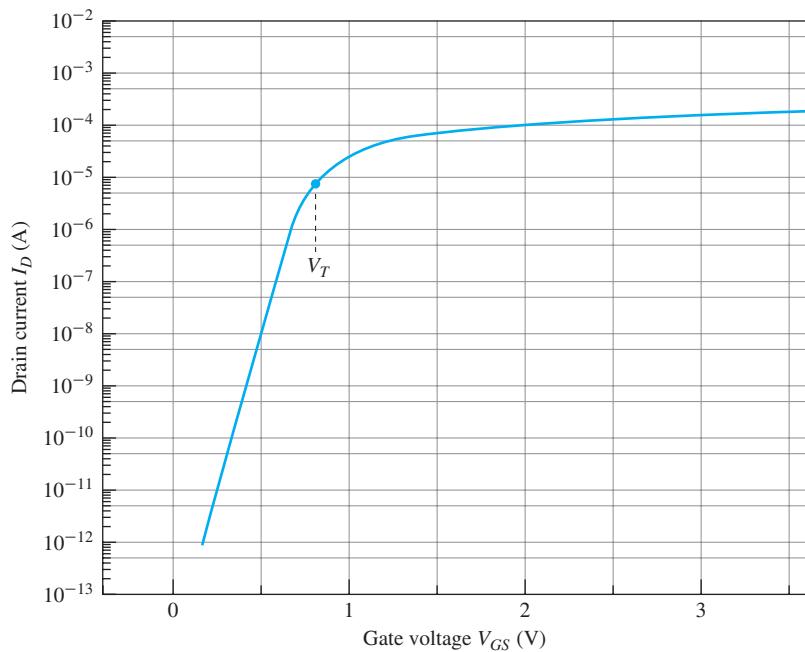


Figure 7.42 The experimental I_D - V_{GS} characteristics with $I_{DS} = 0.1$ V on a semilog plot. The current varies approximately exponentially with V_{GS} below threshold.

it is operating well below threshold ($V_{GS} \approx 0$).¹¹ Even in the **off** state, a small current does flow. In an integrated circuit with millions or billions of transistors, this off-state current can result in considerable power dissipation and a resulting temperature rise. Thus, it is worthwhile to evaluate the value of the subthreshold leakage current.

We know that the electron concentration in the channel at the source varies exponentially with source-channel barrier height E_B . Therefore the current, which is determined by the number of carriers available, has the same dependence, or

$$I_D = I_0 e^{q(V_{GS} - V_T)/nkT} \quad (V_{GS} < V_T) \quad (7.85)$$

where I_0 is the current at threshold ($V_{GS} = V_T$) and $1/n$ is the fraction of ($V_{GS} - V_T$) that affects the source-channel barrier. As the gate voltage is varied, some of the change in voltage is dropped across the oxide and some is dropped across the semiconductor. This can be written as

$$\frac{1}{n} = \frac{\Delta V_{ch}(y=0)}{\Delta V_{GS}} \quad (7.86)$$

¹¹We are discussing enhancement devices, which are common in digital circuits.

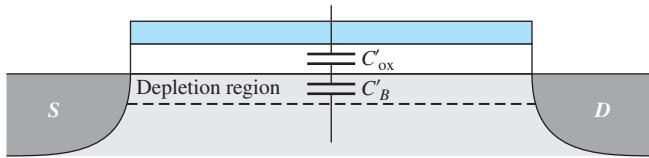


Figure 7.43 The capacitances of the MOSFET gate oxide and substrate. The substrate capacitance C'_B is the depletion layer capacitance.

where ΔV_{ch} ($y = 0$) is the channel voltage with respect to the source at the source end of the channel.

We can find n as follows. Figure 7.43 shows that the gate structure can be viewed as two capacitors in series. These are C'_{ox} and C'_B , where C'_{ox} is the oxide capacitance per unit area and C'_B is the substrate (bulk) capacitance per unit area, caused by the depletion layer at the semiconductor surface. Below threshold, neglecting the small charge in the channel, C'_{ox} and C'_B act as a voltage divider between gate and substrate. That means that if the gate voltage is changed by ΔV_{GS} , that change is also divided between C'_{ox} and C'_B . Then

$$\frac{\Delta V_{ch}}{\Delta V_{GS}} = \frac{1}{n} = \frac{\frac{1}{C'_B}}{\frac{1}{C'_B} + \frac{1}{C'_{ox}}} = \frac{C'_{ox}}{C'_{ox} + C'_B} \quad (7.87)$$

or the constant n in Equation (7.86) is

$$n \approx \frac{C'_{ox} + C'_B}{C'_{ox}} = 1 + \frac{C'_B}{C'_{ox}} \quad (7.88)$$

Figure 7.42 showed the variation of I_D with V_{GS} for a constant $V_{DS} = 0.1$ V with the value of V_T indicated. For gate voltages well below threshold, because of the exponential dependence, the current is in fact small enough not to be a problem. However, as the gate-source voltage gets close to V_T , the current becomes appreciable. To minimize off-state leakage current, when the transistors are **off**, they should be strongly **off**. On the other hand, when the transistor is **on** the gate voltage must be changed enough to produce a strong change in I_D . We therefore ask, “What variation in gate voltage would produce a factor of 10 change in the drain current?” To answer this, we define the gate voltage variation per decade of current in the straight-line region as the “swing,” or “subthreshold swing” S . It is found from

$$S = \frac{\partial V_{GS}}{\partial \log I_D} \quad (7.89)$$

Taking this derivative using the expression for V_{GS} as a function of I_D in Equation (7.85),

$$S = \frac{2.3kTn}{q} \quad (7.90)$$

The factor 2.3 comes from converting from the natural log to log base 10.

For the device shown, $S = 84 \text{ mV/decade}$ of current. We note from Equation (7.88) that the parameter n depends on the ratio of C'_B/C'_{ox} . Ideally, we would like to have the voltage below threshold vary as little as possible and still get large current change. For small S , we want to have a small C'_B (created by a large transition region width or, equivalently, low substrate doping) and a large C'_{ox} (thin oxide, oxide with high dielectric constant). The value of n approaches its minimum value of unity as C'_{ox} approaches infinity or t_{ox} approaches zero. To minimize S , the value of n should be minimized. The minimum value for S is then, for unity n at room temperature,

$$S_{\min} = 2.3 \frac{kT}{q} \approx 60 \text{ mV/decade} \quad (7.91)$$

For logic circuits, the threshold voltage is often chosen to be approximately 20 percent of the supply voltage, or $V_T \approx 0.2V_{DD}$. Minimizing S can minimize the required threshold voltage V_T and thus the supply voltage for the chip, V_{DD} . Decreasing V_{DD} lowers the power dissipation during switching.

EXERCISE 7.12

Consider a chip with 2 million transistors, half of which are off at a given time. We wish to keep the total off-state subthreshold current for the chip below $10 \mu\text{A}$. For a single device, the current at threshold is $1 \mu\text{A}$ and the subthreshold slope S is 80 mV per decade of current. Estimate the minimum power supply voltage V_{DD} . The input gate voltage varies between 0 and V_{DD} .

■ Solution

The total current at V_{GS} (below threshold, the off state) for the chip is $10 \mu\text{A}$, so the maximum leakage current allowable per transistor is $I_D = (10^{-5} \text{ A})/10^6$ transistor = $10^{-11} \text{ A/transistor}$. At threshold the current is $I_D = 10^{-6} \text{ A/transistor}$, so between $V_{GS} = 0$ and $V_{GS} = V_T$, there are five decades of current. At 80 mV/decade , that results in a threshold of $V_T = 80 \times 5 \text{ mV} = 0.4 \text{ V}$. Since often it is chosen that $V_{DD} \approx 5V_T$, then $V_{DD} = 2 \text{ V}$.

The power dissipation associated with the leakage current in the previous example is $2 \text{ V} \times 10^{-5} \text{ A} = 2 \mu\text{W}/\text{chip}$. This is not the total power dissipation, however. As will be discussed in Chapter 8, the power dissipation associated with switching is normally much greater than that associated with subthreshold leakage.

7.8 SUMMARY

In this chapter, we discussed the physical principles of operation of the Si-based MOSFET. These transistors have four terminals, source (S), gate (G), drain (D), and substrate or body (B). However, often the substrate is connected to the

source, making it, in effect, a three-terminal device. The voltage on the gate electrode is used to control the resistance along the channel, between the drain and the source. These are called *field-effect* transistors because the gate controls the channel conductance via an electric field. This field appears across the insulating layer between gate and substrate. This layer is often silicon oxinitride (SiON), often referred to as simply “oxide.” Note that only displacement current flows into the gate, because the oxide is an insulator. The voltage applied to the gate creates the electric field that in turn bends the bands in the substrate, inverting the channel (enhancement MOSFETs) or uninverting it (depletion MOSFETs).

We began by considering a simple (long-channel) model for the current-voltage characteristics in a MOSFET, assuming constant channel mobility. We saw that when the gate voltage is greater than some threshold, current can flow in the channel. The greater the voltage across the channel (between the drain and the source), the more current should flow. We found that this is true, but only up to a point. At some V_{DS} , the longitudinal field \mathcal{E}_L is large enough that it sweeps the carriers along the channel from source to drain as fast as they can be supplied by the source, and the current saturates. The saturation current $I_{D\text{sat}}$ and the saturation voltage $V_{DS\text{sat}}$ (the value of V_{DS} at which the current saturates) both depend on the gate voltage.

The equations for the simple long-channel model are:

SIMPLE LONG-CHANNEL MODEL

$$I_D \approx 0 \quad V_{GS} < V_T \quad (7.9)$$

$$I_D = \frac{WC'_\text{ox}\mu}{L} \left[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2} \right] \quad V_{DS} \leq V_{DS\text{sat}}, \quad V_{GS} \geq V_T \quad (7.24)$$

$$I_{D\text{sat}} = \frac{WC'_\text{ox}\mu}{L} \left[\left(V_{GS} - V_T - \frac{V_{DS\text{sat}}}{2} \right) V_{DS\text{sat}} \right] \quad V_{DS} \geq V_{DS\text{sat}}, \quad V_{GS} \geq V_T \quad (7.25)$$

$$V_{DS\text{sat}} = V_{GS} - V_T \quad (7.26)$$

The mobility is taken to be constant for all bias conditions and is equal to the low-field mobility in the channel. This mobility is considerably less than the bulk value.

Next, we considered the effects of channel length modulation, and we saw that under large drain voltages the voltage in the channel reaches saturation $V_{DS\text{sat}}$ somewhere before the drain, producing a nonzero slope in the saturation characteristics. That slope was modeled empirically by

CHANNEL MODULATION EFFECT (SATURATION ONLY)

$$I_D \approx I_{D\text{sat}}(1 + \lambda(V_{DS} - V_{DS\text{sat}})) \quad (7.43)$$

with

$$\frac{\Delta L}{L} = \lambda(V_{DS} - V_{DS\text{sat}}) \quad (7.42)$$

where the channel is effectively shortened by ΔL .

Then we considered the effects of the transverse field. The channel is quite thin, and carriers will reflect off the potential barriers at the interface with the oxide and the band bending in the substrate. As a result, carrier velocities are reduced even more than accounted for earlier, with the result that the currents are smaller as well. The carrier low-field mobility can be modeled in terms of the gate voltage:

EFFECT OF TRANSVERSE FIELD (EMPIRICAL)

$$\mu_{lf} = \frac{\mu_0}{1 + \theta(V_{GS} - V_T)} \quad (7.47)$$

where the parameters μ_0 (the low-field mobility at the source end at threshold) and θ are determined experimentally. The influence on the I_D - V_{DS} characteristics is to replace constant mobility μ in the simple long-channel model by the low-field mobility μ_{lf} , which is dependent on gate voltage.

These simple long-channel models ignore velocity saturation (i.e., it is assumed that the carriers could attain an arbitrary high velocity such that the $Q_{ch}v$ product is constant). However, velocity saturation does occur in real devices. The mobility is reduced by the high longitudinal field that results from the short channels. When we consider the reduced mobility, we saw that in a FET, the current could be expressed as

LONG-CHANNEL MODEL WITH VELOCITY SATURATION

$$I_D = \frac{I_D(\text{neglecting velocity saturation})}{\left(1 + \frac{\mu_{lf}V_{DS}}{Lv_{sat}}\right)} \quad V_{DS} \leq V_{DSsat} \quad (7.64)$$

$$I_{Dsat} = \frac{I_{Dsat}(\text{neglecting velocity saturation})}{\left(1 + \frac{\mu_{lf}V_{DSsat}}{Lv_{sat}}\right)} \quad V_{DS} \geq V_{DSsat} \quad (7.65)$$

The drain-source voltage at which the current saturates is

$$V_{DSsat} = \frac{v_{sat}L}{\mu_{lf}} \left[\left(1 + \frac{2\mu_{lf}(V_{GS} - V_T)}{v_{sat}L}\right)^{1/2} - 1 \right] \quad (7.68)$$

Or in the alternate form,

$$V_{DSsat} = \frac{2(V_{GS} - V_T)}{\left[\left(1 + \frac{2\mu_{lf}(V_{GS} - V_T)}{v_{sat}L}\right)^{\frac{1}{2}} + 1\right]} \quad (7.69)$$

It is found that the model involving velocity saturation effects agrees reasonably well with experiment.

Then we considered the effect of the source and drain series resistances on the I_D - V_{DS} characteristics. The results are, including velocity saturation and assuming $R_S = R_D$,

VELOCITY SATURATION AND SERIES RESISTANCE

$$I_D = \frac{WC'_\text{ox}\mu_{\text{lf}}\left(V_{GS} - V_T - \frac{V_{DS}}{2}\right)(V_{DS} - 2I_DR_S)}{L\left[1 + \frac{\mu_{\text{lf}}(V_{DS} - 2I_DR_S)}{Lv_{\text{sat}}}\right]} \quad V_{DS} \leq V_{DS\text{sat}} \quad (7.74)$$

$$I_{D\text{sat}} = \frac{WC'_\text{ox}\mu_{\text{lf}}\left(V_{GS} - V_T - \frac{V_{DS\text{sat}}}{2}\right)(V_{DS\text{sat}} - 2I_{D\text{sat}}R_S)}{L\left[1 + \frac{\mu_{\text{lf}}(V_{DS\text{sat}} - 2I_{D\text{sat}}R_S)}{Lv_{\text{sat}}}\right]} \quad V_{DS} \geq V_{DS\text{sat}} \quad (7.75)$$

For very short channel lengths ($L < 20$ nm) electrons make few collisions in the channel traveling from source to drain. A model was presented for the electrical characteristics of such a *ballistic* MOSFET. The resultant equations are

$$I_D = WC'_\text{ox}v_t(V_{GS} - V_T) \frac{\left(1 - e^{-\left(\frac{qV_{DS}}{kT}\right)}\right)}{\left(1 + e^{-\left(\frac{qV_{DS}}{kT}\right)}\right)} \quad (7.83)$$

where v_t is the magnitude of the average velocity of electrons traveling from source to drain at the source end of the channel.

7.9 REFERENCES

1. C. Y. Sah, “Evolution of the MOS transistor—from concept to VLSI,” *Proc. IEEE*, 76, pp. 1280–1326, 1988.
2. G. Baccarani and M. R. Wordman, “Transconductance degradation in thin-oxide MOSFETs,” *IEEE Trans. Electron Devices*, ED-30, pp. 1295–1304, 1983.
3. Dennis Hoyniak, Edward Nowak, and Richard L. Anderson, “Channel electron mobility dependence on lateral electric field in field-effect transistors,” *J. Appl. Phys.*, 87, pp. 876–881, 2000.
4. B. T. Murphy, “Unified field-effect transistor theory including velocity saturation,” *IEEE J. Solid-State Circuits*, SC-15, pp. 325–327, 1980.
5. Dac C. Pham, “Selective device-temperature scaling for optimum power-delay product in MOSFET circuit design,” dissertation, University of Vermont, 1998. Unpublished.
6. Mark Lundstrom, “ECE 612: Nanoscale Transistors (Fall 2008),” <https://nanohub.org/resources/5328> Lecture 7.
7. R. R. Troutman, “VLSI limitations from drain-induced barrier lowering,” *IEEE Journal of Solid State Circuits*, SC-14, pp. 389–391, 1979.

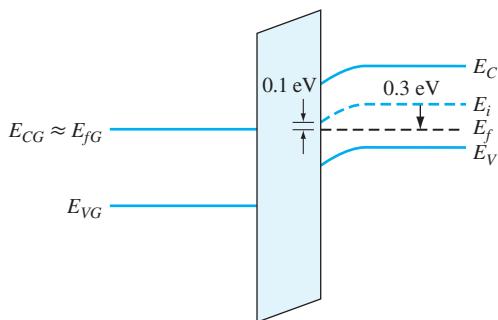
7.10 REVIEW QUESTIONS

1. Explain why virtually no current flows into the gate of a MOSFET. If no current flows into this electrode, how can a signal on the gate have an effect on the operation of the rest of the transistor?

2. Explain why the current (rate of flow) between source and drain saturates as the potential difference between the drain and the source increases.
3. Summarize in words the steps used to derive the I_D - V_{DS} characteristics of a FET using the long-channel model as an example.
4. Explain in words how applying a voltage to the gate can, in effect, change the material at the surface of the channel from p type to n type.
5. Is the device in Figure 7.7 an enhancement or a depletion FET?
6. Why is the carrier mobility in the channel much less than in the bulk?
7. Explain why the threshold voltage depends on the drain voltage for short-channel but not for long-channel MOSFETs.
8. Why is the saturation current calculated from the ballistic model of a MOSFET greater than that calculated using the drift model?

7.11 PROBLEMS

- 7.1 In any circuit, the transistor operation can be understood by examining the superposition of the transistor characteristics and the conditions imposed by the circuit (the load line). What differentiates, then, a digital transistor circuit from an analog one?
- 7.2 In modern FETs, the gate is often degenerately doped silicon, whose Fermi level is essentially at the bottom of the conduction band (for an NFET) or at the top of the valence band (for a PFET).
 - a. Draw an energy band diagram similar to Figure 7.7b, except making the transistor a PFET.
 - b. Suppose the device is an NFET, but the gate is made of metal. Draw the energy band diagram. Take $\Phi_M < \Phi_S$.
- 7.3 For the transistor of Figure P7.1, by how much should the gate voltage be changed to produce inversion? Threshold? Assume 10 percent of the applied voltage appears across the oxide and 90 percent across the semiconductor. If the device in the figure is at equilibrium, is this an enhancement or a depletion FET?
- 7.4 For each of the transistors of Figure 7.12.
 - a. What is the polarity of the threshold voltage V_T ?
 - b. What is the polarity of V_{DS} that should be used?
 - c. When $V_{GS} = 0$ (equilibrium), is the transistor on or off?
 - d. Is the current I_D carried by electrons or holes?
 - e. For $V_{GS} > V_T$, is the current I_D carried primarily by drift or diffusion?
- 7.5 An NFET is fabricated with a degenerately doped n-type gate. What doping concentration in the p-type substrate is needed for there to be a channel even with no voltage applied? Assume half the built-in voltage is dropped across the oxide and half across the silicon. Let the definition of “a channel exists” be that the surface of the silicon is inverted such that the electron concentration at the Si/oxide interface is equal to the hole

**Figure P7.1**

concentration in the p-type bulk. Is the doping you found a minimum or a maximum required to create a depletion-type device?

- 7.6 Consider two silicon MOSFETs, one n channel and the other p channel, with substrate dopings of 10^{16} cm^{-3} . The NMOS has an n⁺ gate and the PMOS has a p⁺ gate, both doped to 10^{19} cm^{-3} . Find the built-in voltage V_{bi} for each, and draw the energy band diagram. Neglect the band-gap narrowing effects discussed in Chapter 2.
- 7.7 In the MOS process, structures like the gate of a transistor are used to make capacitors as well. If the oxide thickness is 2 nm, what area is needed to achieve a capacitance of 1 pF? The permittivity of silicon dioxide is $3.9\epsilon_0$.
- 7.8 In MOS processing, the W/L ratio is often intentionally made different from transistor to transistor. Plot the I_D - V_{DS} characteristics for the device of Figure 7.21 with the W/L ratio changed to 10 instead of 5. Compare this with the results for $W/L = 5$.
- 7.9 Explain why the electron affinity model can be used to good approximation to determine the band lineup normal to the gate in a MOSFET. That is, why is the tunneling-induced dipole effect negligible?
- 7.10 Plot the I_D - V_{DS} characteristics for an NFET, using the long-channel model, for which $W = 10 \mu\text{m}$, $L = 1 \mu\text{m}$, $t_{ox} = 2 \text{ nm}$, $V_T = 0.25 \text{ V}$, and the channel length modulation parameter is $\lambda = 0.04 \text{ V}^{-1}$. Use $V_{GS} = 1 \text{ V}$, 2 V , 3 V , and 4 V . Find the output conductance in saturation for $V_{GS} = 3 \text{ V}$.
- 7.11 An enhancement NFET with the characteristics in Table 7.1 has a threshold voltage of $V_T = 1 \text{ V}$, a channel length of $1 \mu\text{m}$, and a width of $5 \mu\text{m}$. Considering velocity saturation, with $v_{sat} = 10^7 \text{ cm/s}$, find the current I_D for
 - a. $V_{GS} = 0 \text{ V}$, $V_{DS} = 1 \text{ V}$
 - b. $V_{GS} = 2 \text{ V}$, $V_{DS} = 1 \text{ V}$
 - c. $V_{GS} = 3 \text{ V}$, $V_{DS} = 1 \text{ V}$
- 7.12 An NFET is made with $t_{ox} = 2 \text{ nm}$, $L = 1 \mu\text{m}$, $W = 10 \mu\text{m}$, $V_T = 1 \text{ V}$, and $\mu_{lf} = 250 \text{ cm}^2/\text{V} \cdot \text{s}$. If the simple model is used, what should the width of the PFET be to get the same saturation current (apart from polarity)? Let the low-field mobility for holes be $100 \text{ cm}^2/\text{V} \cdot \text{s}$.

- 7.13** An NFET and a PFET are made on the same chip, using the same process. The NFET has $C'_{\text{ox}} = 1.73 \times 10^{-6} \text{ F/cm}^2$, $t_{\text{ox}} = 2 \text{ nm}$, $L = 0.2 \mu\text{m}$, $W = 15 \mu\text{m}$, $V_T = 1.5 \text{ V}$, and $\mu_{\text{lf}} = 250 \text{ cm}^2/\text{V} \cdot \text{s}$. If the PFET is identical except for its mobility ($100 \text{ cm}^2/\text{V} \cdot \text{s}$) and its width W ,
- What should W be for the PFET to make the characteristics the same as for the NFET, as predicted by the simple model?
 - Find $V_{D\text{Ssat}}$ and $I_{D\text{sat}}$ for $V_{GS} - V_T = 1 \text{ V}$.
 - If velocity saturation is considered, how different are $V_{D\text{Ssat}}$ and $I_{D\text{sat}}$ for the NFET and the PFET compared with the simple model results? Express your result as a ratio (e.g., $I_{D\text{satNFET}}/I_{D\text{satPFET}}$ and $V_{D\text{SsatNFET}}/V_{D\text{SsatPFET}}$). Assume $v_{\text{sat}} = 10^7 \text{ cm/s}$.
- 7.14** A good way to check the validity of a derivation is to verify that it reduces to the expected result for a particular known case. For example, in the simple model, we neglected velocity saturation, and in the later model we considered it. Since the high field that causes velocity saturation occurs in short-channel devices, we would expect that the later model would reduce to the simple model for long-channel devices.
- Show that in the limit of large L , $V_{D\text{Ssat}}$ as given by Equation (7.68) reduces to $V_{D\text{Ssat}} = (V_{GS} - V_T)$ as given by Equation (7.26) for the simple (long-channel) model.
 - Show that in the limit of large L , $I_{D\text{sat}}$ as given by Equation (7.61) reduces to Equation (7.25) for the simple model.

You may find the following information useful: For $x < 1$

$$(1 \pm x)^n = 1 \pm nx + \frac{n(n-1)x^2}{2!} \pm \dots$$

$$(1 \pm x)^{-n} = 1 \mp nx + \frac{n(n+1)x^2}{2!} \mp \dots$$

- 7.15** Find an approximate closed-form expression for the drain current, accounting for series resistance and velocity saturation. That is, solve for I_D in Equation (7.74). Neglect the term in I_D^2 .
- 7.16** Consider an n-channel MOSFET with $L = 65 \text{ nm}$ and $\mu_{\text{lf}} = 250 \text{ cm}^2/\text{V}\cdot\text{s}$. Find the ratio of the saturation current calculated from the ballistic model ($I_{D\text{sat(B)}}$) to that from the drift model ($I_{D\text{sat(D)}}$).
- 7.17** Depletion regions exist in the substrate adjacent to the source and the drain. If these two depletion regions overlap, the source-to-channel barrier decreases, causing excess current to flow between drain and source. This condition is referred to as punch-through.

A silicon n-channel MOSFET has source and drain doping concentration $N'_D = 10^{19} \text{ cm}^{-3}$ and substrate doping concentration $N'_A = 10^{17} \text{ cm}^{-3}$ respectively. For, $L = 0.4 \mu\text{m}$, find the punch-through voltage V_{pt} . Assume source is connected to substrate.

Other Field-Effect Transistors

8.1 INTRODUCTION

In Chapter 7, the basic static characteristics of MOSFETs were discussed. In this chapter, we look at some additional considerations for MOSFETs and discuss other types of FETs.

First, we will develop our understanding of the parameters that control MOSFET behavior. Specifically, up to now the threshold voltage and the low-field mobility were treated as known (measured) quantities. In this chapter, we show how these parameters can be measured for particular MOSFETs.

After that, we will look at MOSFETs in action. We begin with a simple CMOS device. CMOS means complementary MOS circuitry in which both n-channel and p-channel devices are used. We choose the inverter logic circuit as a typical circuit and investigate its operation, power dissipation, and delays in switching (propagation delays).

Although the inverter is a classic example of a digital circuit, and while MOSFETs are used extensively in digital circuits, they are also used frequently in analog circuitry and in fact, CMOS can be used as a high-gain, highly linear amplifier.

We then discuss three variations of the Si-based MOSFET, silicon on insulator (SOI), nonvolatile (floating gate) MOSFETs, and FinFETs.

Other types of field-effect transistors (MESFETs, JFETs, HFETs, and TFETs) are briefly discussed. These devices, while important, currently are used less frequently than MOSFETs.

8.2 MEASUREMENT OF THRESHOLD VOLTAGE AND LOW-FIELD MOBILITY

In Chapter 7, we took the value of the threshold voltage V_T to be known. It is difficult to predict accurately the threshold voltage of a MOSFET during the design phase. That is because it depends on some process characteristics (for example, charges trapped in the oxide and defects at the silicon-oxide interface). Thus in practice, the threshold voltage is measured for a sample device produced in a particular fabrication process.

The threshold voltage can be measured by comparing the experimental current-voltage characteristics with those predicted by theory. To understand these measurements, we begin by recalling from Chapter 7 that above threshold but below saturation, the drain current has the form

$$I_D = \frac{WC'_{\text{ox}}\mu_{\text{lf}}\left(V_{GS} - V_T - \frac{V_{DS}}{2}\right)V_{DS}}{L\left(1 + \frac{\mu_{\text{lf}}V_{DS}}{Lv_{\text{sat}}}\right)} \quad (8.1)$$

where, from Equation (7.47),

$$\mu_{\text{lf}} = \frac{\mu_0}{1 + \theta(V_{GS} - V_T)} \quad (8.2)$$

Equation (8.1) is a model including the effect of the transverse field (in μ_{lf}) and longitudinal field (in velocity saturation). For V_{DS} small enough that the saturation velocity (i.e., the second term in the denominator) can be neglected, or $(\mu_{\text{lf}}V_{DS}/Lv_{\text{sat}} \ll 1)$ Equation (8.1) simplifies to

$$I_D = \frac{WC'_{\text{ox}}\mu_{\text{lf}}\left(V_{GS} - V_T - \frac{V_{DS}}{2}\right)V_{DS}}{L} \quad (8.3)$$

Then substituting Equation (8.2) into Equation (8.3) gives

$$I_D = \frac{WC'_{\text{ox}}\mu_0\left(V_{GS} - V_T - \frac{V_{DS}}{2}\right)V_{DS}}{L[1 + \theta(V_{GS} - V_T)]} \quad (8.4)$$

For small $(V_{GS} - V_T)$ this can be approximated

$$I_D = \frac{WC'_{\text{ox}}\mu_0\left(V_{GS} - V_T - \frac{V_{DS}}{2}\right)V_{DS}}{L} \quad (8.5)$$

This is the equation of a straight line of I_D against V_{GS} . The intercept ($I_D = 0$) is at

$$V_{GS}(0) = V_T + \frac{V_{DS}}{2} \quad (8.6)$$

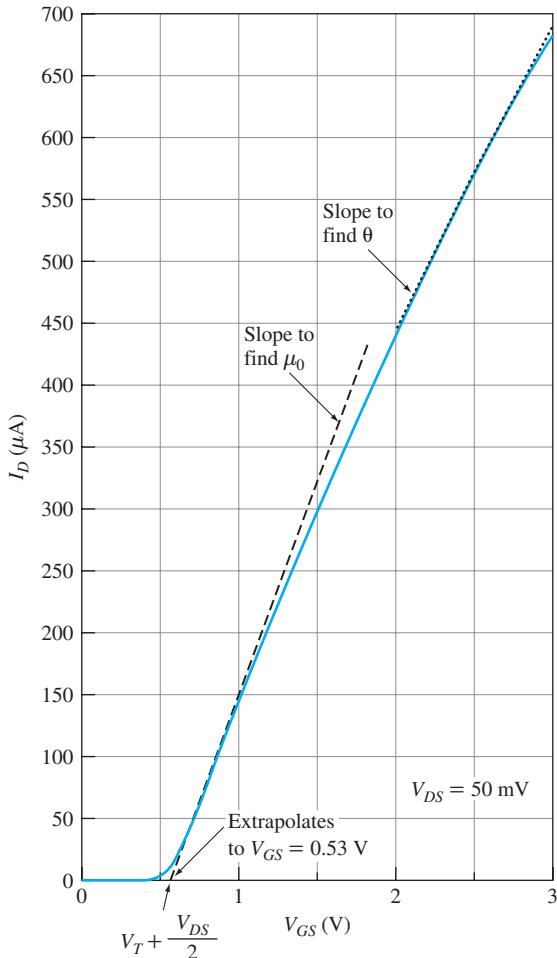


Figure 8.1 Plot of I_D - V_{GS} characteristics for $V_{DS} = 50$ mV (solid line) for a MOSFET with $W = 10 \mu\text{m}$, $L = 0.5 \mu\text{m}$, and $t_{\text{ox}} = 5 \text{ nm}$. The lower dashed line is the tangent to the maximum slope region of the curve. The upper dotted line is tangent at a chosen value of V_{GS} . From the plot, the threshold voltage V_T , mobility at the source at threshold μ_0 , and the gate voltage induced-mobility modulation factor θ can be determined.

and the slope is

$$\frac{dI_D}{dV_{GS}} = \frac{WC'_{\text{ox}}\mu_0 V_{DS}}{L} \quad (8.7)$$

Figure 8.1 shows how the threshold voltage can be determined from an experimental plot of I_D versus V_{GS} . The dashed line has its intercept given by Equation (8.6). From the extrapolated intercept at $I_D = 0$,

$$V_T = V_{GS}(0) - \frac{V_{DS}}{2} \quad (8.8)$$

Further, the slope of the dashed line can be used to find the value of μ_0 . Since the values of W , C'_{ox} , ($C'_{\text{ox}} = \epsilon_{\text{ox}}/t_{\text{ox}}$) are known from the fabrication process, then the low-field mobility at threshold, μ_0 , can be determined. [1] From Equation 8.7, we can solve for μ_0 , the mobility at threshold

$$\mu_0 = \frac{L}{WC'_{\text{ox}}} \cdot \frac{dI_D}{dV_{GS}} \quad (8.9)$$

where $\frac{dI_D}{dV_{GS}}$ is the slope at small V_{GS} , the slope of the lower dashed line in Figure 8.1.

We see, however, that the experimental I_{DS} - V_{GS} curve is not a straight line as predicted from Equation (8.3). This is because the low-field mobility, μ_{lf} , is also dependent on V_{GS} . We can obtain μ_{lf} from Equation 8.2, recognizing from Figure 8.1 that θ is related to the slope of the I_D - V_{GS} curve at higher V_{GS} . This is shown by the upper dotted line in the figure. But, since θ is small (on the order of 0.1 V^{-1}), for small $(V_{GS} - V_T)$

$$\mu_{lf} \approx \mu_0 [1 - \theta(V_{GS} - V_T)] \quad (8.10)$$

and

$$I_D \approx \frac{WC'_{\text{ox}} V_{DS} \mu_0 [1 - \theta(V_{GS} - V_T)] \left(V_{GS} - V_T - \frac{V_{DS}}{2} \right)}{L} \quad (8.11)$$

Figure 8.1 shows the I_D - V_{GS} characteristics for a MOSFET device with $V_{DS} = 50 \text{ mV}$, $W = 10 \mu\text{m}$, $L = 0.5 \mu\text{m}$, and $t_{\text{ox}} = 5 \text{ nm}$. We can find V_T and μ_0 from the line tangent to the curve at maximum slope. The value of θ can be deduced from the deviation of the actual characteristics from the straight-line approximation.

For $(V_{GS} - V_T)$ large enough that $(V_{GS} - V_T) \gg V_{DS}/2$, the slope of the I_D - V_{GS} characteristic from Equation (8.11) is

$$\frac{dI_D}{dV_{GS}} = \frac{WC'_{\text{ox}} \mu_0 V_{DS}}{L} [1 - 2\theta(V_{GS} - V_T)] \quad (8.12)$$

The utility of this is that the parameter θ can be determined by comparing the slope of the I_D - V_{GS} characteristic at a given $(V_{GS} - V_T)$ with the maximum slope.

EXAMPLE 8.1

Find the values of V_T , μ_0 , μ_{lf} and θ from the I_D - V_{GS} characteristics of Figure 8.1. The drain-source voltage is $V_D = 0.05 \text{ V}$.

Solution

From Figure 8.1 the straight-line $V_{GS}(0)$ intercept is at $V_{GS} = 0.53 \text{ V}$. Since $V_{DS} = 0.05 \text{ V}$, from Equation (8.8),

$$V_T = V_{GS}(0) - \frac{V_{DS}}{2} = 0.53 - \frac{0.05}{2} \approx 0.5 \text{ V}$$

The value of θ can be found by comparing the slope of the I_D - V_{GS} characteristic at threshold with that at some arbitrary point. Suppose we choose $V_{GS} = 2.5$ V. From Equation (8.7),

$$\left. \frac{dI_D}{dV_{GS}} \right|_{\text{threshold}} = \frac{WC'_{\text{ox}}\mu_0 V_{DS}}{L}$$

and from Equation (8.12) we have

$$\left. \frac{dI_D}{dV_{GS}} \right|_{V_{GS} = 2.5 \text{ V}} = \frac{WC'_{\text{ox}}\mu_0 V_{DS}}{L} [1 - 2\theta(V_{GS} - V_T)]$$

Thus

$$\frac{\text{Slope of data at } V_{GS} = 2.5 \text{ V}}{\text{Slope of straight line at threshold}} = 1 - 2\theta(V_{GS} - V_T)$$

The slope of the data at $V_{GS} = 2.5$ V is $236 \mu\text{A/V}$ and that of the straight-line section is $333 \mu\text{A/V}$. Thus

$$\theta = \frac{1 - \frac{236}{333}}{2(2.5 - 0.5)} = \frac{0.29}{4} = 0.073 \text{ V}^{-1}$$

From Equation (8.7),

$$\mu_0 = \left. \frac{dI_D}{dV_{GS}} \right|_{\text{threshold}} \cdot \frac{L}{WC'_{\text{ox}} V_{DS}} = 483 \text{ cm}^2/\text{V} \cdot \text{s}$$

And from Equation (8.10),

$$\mu_{\text{lf}} = 483 \text{ cm}^2/\text{V} \cdot \text{s} (1 - 0.73 V^{-1}(2.5 - 0.5)V) = 412 \text{ cm}^2/\text{V} \cdot \text{s}$$

We remind the reader that this development is valid only for small V_{DS} .

EXAMPLE 8.2

Find the maximum value of V_{DS} such that the preceding method yields valid results.

■ Solution

We used the condition of small V_{DS} to approximate Equation (8.1) by

$$I_D = \frac{W\mu_{\text{lf}}C'_{\text{ox}}}{L} \left(V_{GS} - V_T - \frac{V_{DS}}{2} \right) V_{DS}$$

saying in effect that

$$\frac{\mu_{\text{lf}} V_{DS}}{L v_{\text{sat}}} \ll 1$$

Choosing a condition of “maximum acceptable V_{DS} ” to be such that the error is 10 percent, we write

$$\frac{\mu_{lf} V_{DS}}{L v_{sat}} < 0.1$$

Rearranging gives us our result:

$$V_{DS\max} \leq \frac{0.1 v_{sat}}{\mu_{lf}} L$$

For $v_{sat} = 10^7$ cm/s and $\mu_{lf} = 250$ cm²/V · s, the maximum drain voltage for which this graphical method of finding the threshold voltage is valid is:

$$V_{DS\max} \leq 4 \times 10^3 L \text{ cm} = 0.4 L \mu\text{m}$$

For a channel length of 0.25 μm , that gives

$$V_{DS\max} \leq 0.1 \text{ V}$$

This voltage decreases with decreasing L .

8.3 COMPLEMENTARY MOSFETS (CMOS)

So far, we have discussed n-channel MOSFETs and p-channel MOSFETs independently. Currently, most integrated circuits use both n-channel and p-channel devices, hence the term complementary MOSFETs, or CMOS. Figure 8.2a shows the schematic cutaway view of a CMOS inverter using the so-called n-well technology. The n well is ion-implanted into a p-type substrate. The p-channel device is fabricated in the n well while the n-channel FET is fabricated directly into the p substrate. The inverter is a fundamental building block in digital integrated circuits.

A cross-sectional schematic view of the inverter is shown in Figure 8.2b. Note that the n⁺ source is connected to the substrate (body). There is a pn junction between the source of the n-channel device and the substrate. It has zero bias across it, however, preventing current from flowing between the source and the substrate. The p-channel source is connected to the n well. This ensures that current cannot flow out of the p source into the n well. Note that the n-well–p-substrate junction is reverse biased so that the well-substrate current is small. Note also that both gates (inputs) are connected to each other, and the two drains (outputs) are connected to each other.

8.3.1 OPERATION OF THE CMOS INVERTER

A circuit diagram of a digital CMOS inverter is shown in Figure 8.3a. Both channel regions are indicated by broken lines (the lines parallel to the gate electrode)

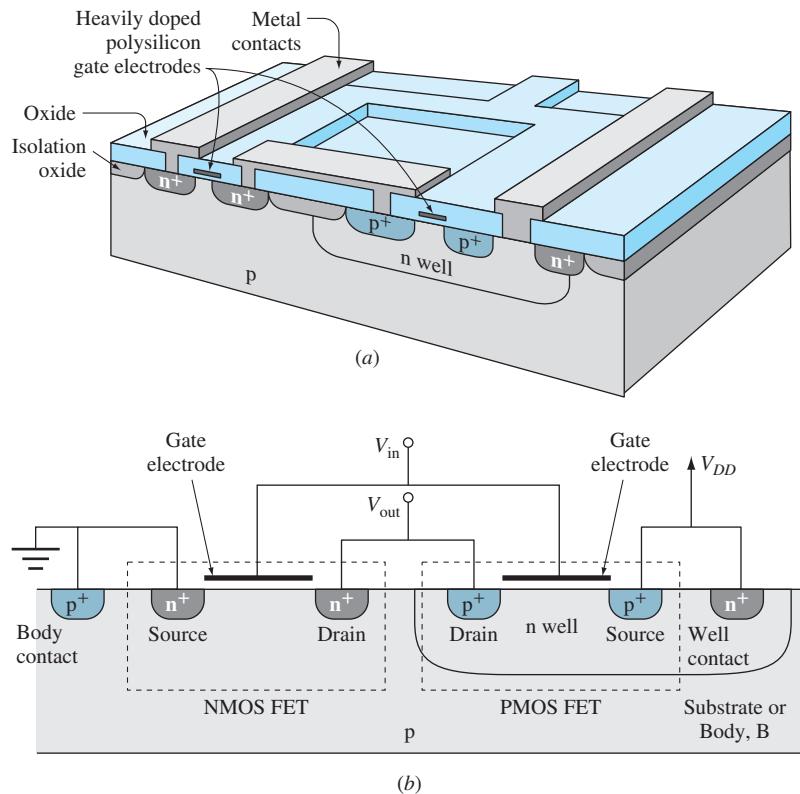


Figure 8.2 The CMOS inverter: (a) physical structure, (b) cross-sectional diagram.

to indicate that these are enhancement devices; i.e., for zero volts between gate and source, they are **off**. Let us assume a power supply voltage of V_{DD} and an input square wave of 0 to V_{DD} as indicated in Figure 8.3b. When the input voltage is 0 V, the gate-to-source voltage of the NFET is zero and the device is **off**. Its channel does not conduct. That transistor acts as an open circuit. In the PFET, however, the gate is negative with respect to the source, and so this device is **on**. The channel acts as a conductor between the power supply V_{DD} and the output port. As a result, the output voltage is also V_{DD} , or at a logic **1** state.

When, however, the gate voltage $V_{in} = V_{DD}$, the NFET is **on** and the PFET is **off**. Thus, the output port is effectively tied to ground, or is at logic **0**. The circuit inverts the input signal. Note that when the NMOS gate voltage is $V_{GSn} = 0$, the PMOS gate-to-source voltage is $V_{GSp} = -V_{DD}$, and when $V_{GSn} = V_{DD}$ then $V_{GSp} = 0$. In either state, one device is **off**, and so in the steady state no current can flow from V_{DD} to ground.¹

¹Actually, of course, the subthreshold leakage current of the off device does flow.

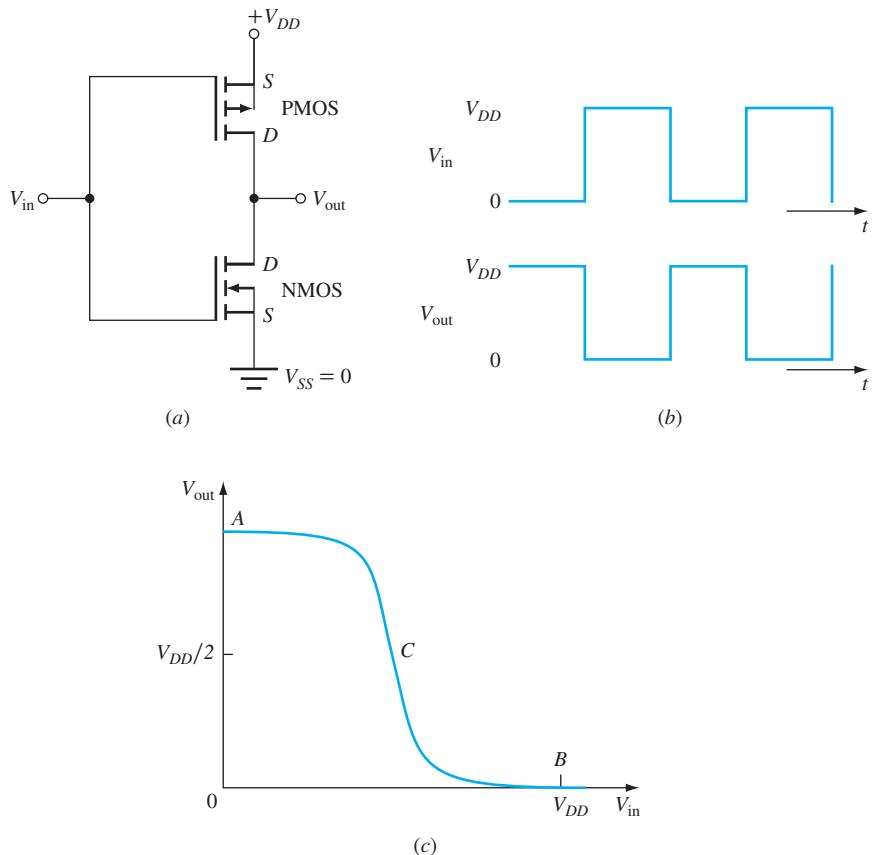


Figure 8.3 (a) CMOS logic inverter circuit diagram; (b) input and output waveforms; (c) the transfer characteristics of a CMOS inverter.

The voltage transfer characteristic (V_{out} versus V_{in}) is shown in Figure 8.3c. For $V_{\text{in}} = 0$, the device is operating in region A. Here the NMOS is **off** while the PMOS is **on**. In region B, for $V_{\text{in}} = V_{DD}$, the NMOS is **on** and the PMOS is **off**. In region C, where $V_{\text{in}} \approx V_{DD}/2$, both NMOS and PMOS are operating in their current saturation regions. The steepness in region C is a measure of the switching speed or, in analog circuits, the voltage gain.

Figure 8.4(a) shows a simplified schematic for the CMOS inverter. Generic diagrams are used for the PFET and NFET, and the bubble in the gate indicates a PFET. The standard circuit symbol for an inverter is shown in (b) where the bubble in the output indicates that the output is the inverse of the input. When a CMOS inverter is used as an analog amplifier, the gate-to-source voltage is biased at $V_{GS} = \frac{V_{DD}}{2}$, as indicated in Figure 8.5. Here the CMOS inverter voltage transfer function of Figure 8.3c is repeated, but now a small signal input voltage applied to the gate is amplified (and inverted) as indicated. In practical inverters the transition region is much sharper than shown here for clarity, and voltage gains well in the hundreds are obtained.

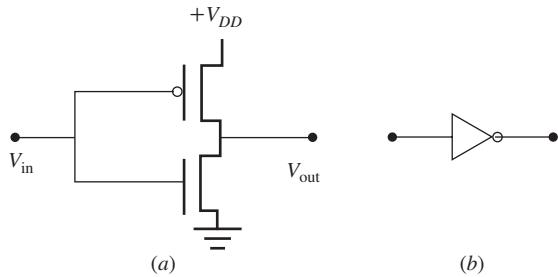


Figure 8.4 Simplified versions of the CMOS Inverter of Figure 8.3a.

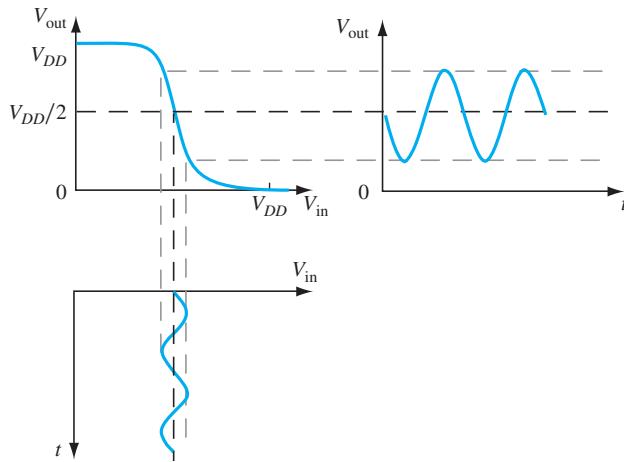


Figure 8.5 Analog CMOS Inverter indicating voltage gain

8.3.2 MATCHING OF CMOS DEVICES

In CMOS logic inverters, for maximum switching speed, it is usually desirable to match the I_D - V_{DS} characteristics of the NMOS and PMOS devices as closely as possible. Since the I_D - V_{DS} characteristics depend in part on the mobilities [see, for example, Equations (7.66) and (7.67)], and the mobilities of electrons and holes are different, the characteristics would be different for otherwise identical devices. The NFETs and PFETs must therefore be “matched.” This is typically done by matching their saturation currents. From Equation (7.67),

$$I_{D\text{sat}} = \frac{WC'_\text{ox}\mu_\text{lf}}{L} \left[\frac{\left(V_{GS} - V_T - \frac{V_{DS\text{sat}}}{2} \right) V_{DS\text{sat}}}{\left(1 + \frac{\mu_\text{lf} V_{DS\text{sat}}}{L v_\text{sat}} \right)} \right] \quad (8.13)$$

If we assume that the channel lengths are the same for both the NFETs and PFETs, and that $v_{\text{satn}} = v_{\text{satp}}$ and $C'_{\text{oxn}} = C'_{\text{oxp}}$, then to make $I_{D\text{satn}} = I_{D\text{satp}}$, we obtain

$$\frac{W_p}{W_n} = \frac{\mu_{\text{lfn}} \left[\left(V_{GS} - V_T - \frac{V_{DS\text{satn}}}{2} \right) V_{DS\text{satn}} \right] \left[1 + \frac{\mu_{\text{lfp}} V_{DS\text{satp}}}{v_{\text{sat}} L} \right]}{\mu_{\text{lfp}} \left[\left(V_{GS} - V_T - \frac{V_{DS\text{satp}}}{2} \right) V_{DS\text{satp}} \right] \left[1 + \frac{\mu_{\text{lfn}} V_{DS\text{satn}}}{v_{\text{sat}} L} \right]} \quad (8.14)$$

From Equation (7.68), we have

$$V_{DS\text{sat}} = \frac{v_{\text{sat}}}{\mu_{\text{lf}}} L \left[\left(1 + \frac{2\mu_{\text{lf}}|V_{GS} - V_T|}{v_{\text{sat}} L} \right)^{1/2} - 1 \right] \quad (8.15)$$

All quantities in Equation (8.15) are the same for the n- and p-channel devices except for the low-field mobilities μ_{lf} . Recall from Table 7.1 that typical values are $\mu_{\text{lfn}} = 250 \text{ cm}^2/\text{V} \cdot \text{s}$ and $\mu_{\text{lfp}} = 100 \text{ cm}^2/\text{V} \cdot \text{s}$. Thus, for a given $V_{GS} - V_T$, the ratio W_p/W_n depends on the difference of $\mu_{\text{lfn}} V_{DS\text{satn}}$ and $\mu_{\text{lfp}} V_{DS\text{satp}}$. Figure 8.6 plots the ratio (W_p/W_n) required to match the transistor currents as a function of channel length L for $(V_{GS} - V_T) = 1 \text{ V}$. The widths of the NFETs and PFETs vary from $W_p/W_n = 1.31$ for a channel length of $L = 30 \text{ nm}$ to $W_p/W_n = 2.34$ for $L = 2 \mu\text{m}$. This latter compares with $W_p/W_n = 2.5$ predicted by the long-channel model, neglecting velocity saturation effects.

Note, however, that although the saturation currents have been matched, the voltage at which saturation begins, $V_{DS\text{sat}}$, is independent of W . That means that below saturation, the I_D - V_{DS} characteristics for the NMOS and PMOS still do

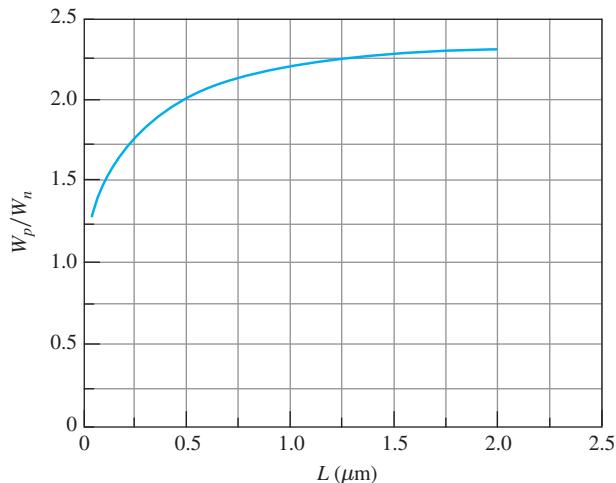


Figure 8.6 The ratio of W_p/W_n needed to match saturation currents ($I_{D\text{satn}} = I_{D\text{satp}}$), as a function of channel length, $L_{n\text{FET}} = L_{p\text{FET}}$.

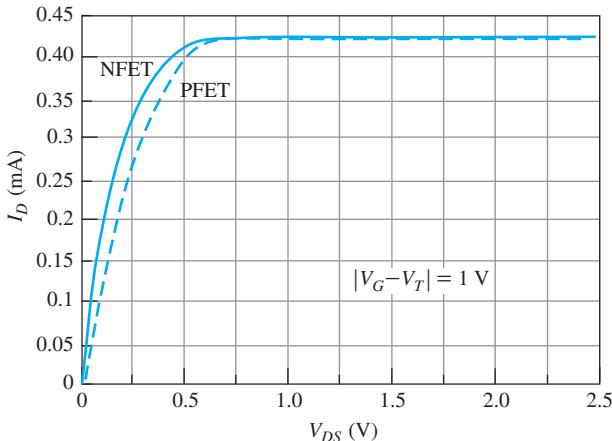


Figure 8.7 Comparison of I_D - V_{DS} characteristics for NMOS and PMOS with matched saturation current. $L_{nFET} = L_{pFET} = 0.2 \mu\text{m}$; $W/L = 10$ for the NMOS and $W/L = 17.7$ for the PMOS.

not match. This is illustrated in Figure 8.7 for a channel length of $0.2 \mu\text{m}$, where from Figure 8.6, $W_p/W_n = 1.77$. The width-to-length ratios for the NMOS and PMOS are $W/L = 10$ and 17.7 respectively. The device characteristics can be approximated more closely below threshold by making the channel lengths $L_p < L_n$ and readjusting W_p/W_n .

8.4 SWITCHING IN CMOS INVERTER CIRCUITS

Up to now, we have discussed the steady-state aspects of a CMOS logic inverter switch. In this section, we consider the transient switching effects.

8.4.1 EFFECT OF LOAD CAPACITANCE

Consider the circuit of Figure 8.8a. This is the inverter circuit shown earlier in Figure 8.3, but we have added a load capacitance. The capacitor C_L includes stray wiring capacitance, the output capacitance of the circuit, and the capacitance of the input to the next stage. Unfortunately, it takes some current to charge or discharge the capacitance when the circuit switches from one logic state to the other. This charging and discharging introduces a time delay. In addition, a current flows to ground during the transition, creating additional power dissipation.

When the capacitance is being charged (output goes low to high), current flows from the power supply through the PMOS to the capacitance. During discharging, the current flows from the capacitance through the NMOS to ground. The net result is the transfer of some amount of charge to ground from the power supply, which increases the overall power consumption. The output voltage takes

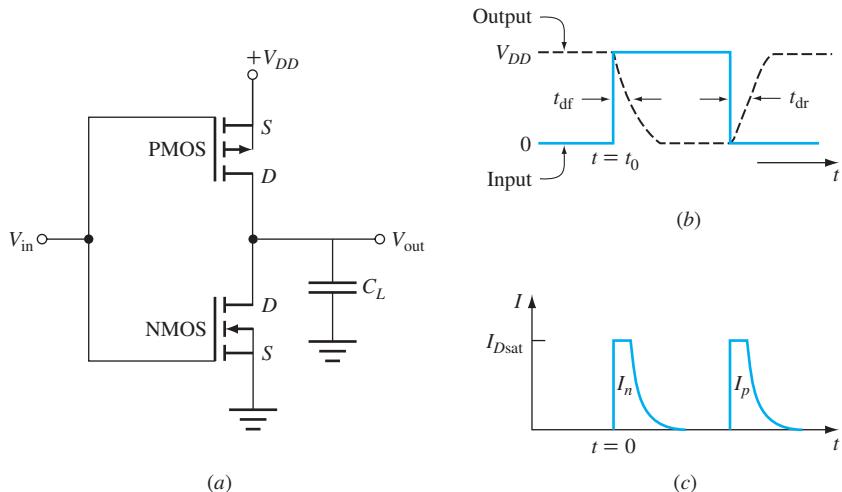


Figure 8.8 (a) The inverter circuit; (b) the input and output voltage signals. The solid line is the input waveform; the dashed line is the output. The output cannot change instantaneously because it requires some time for the load capacitance to charge and discharge. (c) The charging and discharging current that flows during a cycle. One pulse flows through the n-channel device, the other pulse flows through the p-channel device.

a finite time to change state as indicated in Figure 8.8b. The current, indicated in (c), flows only during switching. (The current waveform and the fall and rise time delays, t_{df} and t_{dr} , are discussed in the next section.)

During each cycle, then, a charge $Q = C_L V_{DD}$ flows from the power supply to ground. The energy transferred per cycle is:

$$\text{Energy/cycle} = Q V_{DD} = C_L V_{DD}^2 \quad (8.16)$$

The power dissipated is the energy per cycle times the number of cycles per second, or frequency. The dynamic power dissipation or the power associated with charging and discharging the load capacitance during switching is then²

$$P_{\text{dynamic}} = C_L V_{DD}^2 f \quad (8.17)$$

For high-speed circuits the frequency f is large. To minimize the power dissipation, and thus the temperature of the chip, it is required to minimize C_L and V_{DD} . The minimum supply voltage, however, is dictated by the small leakage current in the off state. This leakage current contributes to static power dissipation.

²For a more elegant derivation, see Reference 2.

EXAMPLE 8.3

(a) Find the dynamic power dissipation for a CMOS inverter operating with $V_{DD} = 2.5$ V at 100-MHz frequency. Assume that the load capacitance is 0.1 pF. (b) Repeat for a CMOS circuit with 20 million CMOS pairs operating at 15 GHz with $V_{DD} = 1.1$ V.

■ Solution

a. From Equation (8.17),

$$P_{\text{dynamic}} = C_L V_{DD}^2 f = 10^{-13} \times (2.5)^2 \times 10^8 = 6.25 \times 10^{-5} \text{ W}$$

$$\text{b. } P_{\text{dynamic}} = C_L V_{DD}^2 f \times \text{number of CMOS pairs} = 10^{-13} \times (1.1)^2 \times (15 \times 10^9) \times (20 \times 10^6) = 36.3 \text{ kW}$$

The dissipation of this much power illustrates a problem.

8.4.2 PROPAGATION (GATE) DELAY IN CMOS SWITCHING CIRCUITS

The speed of operation of a switching circuit depends on the time required for a change in the input to be reflected in the output. In other words, once a change in signal is applied at the input, it takes some time for that new information to propagate “through” the circuit. One measure of the time delay between input and output is the propagation delay time t_d , sometimes referred to as gate delay. It is defined as the time it takes from the input being at 50 percent of its voltage swing to the output being at 50 percent of its voltage swing. As an example, we consider again the CMOS inverter of Figure 8.8.

For our purposes, an estimate of the delay time is adequate. To get this estimate, we assume a voltage step function series of pulses between zero and V_{DD} is applied to the input at $t = t_0 = 0$ as in Figure 8.8b.

The input and output voltages range between zero and V_{DD} , the power supply voltage. The propagation delay times (there are two, one for rising and one for falling) are the times needed to switch between $V_{\text{in}} = V_{DD}/2$ and $V_{\text{out}} = V_{DD}/2$; these are indicated as t_{dr} for rising and t_{df} for falling in Figure 8.8b. Since these times may differ, the propagation delay is defined as the average of the two:

$$t_d = \frac{t_{\text{df}} + t_{\text{dr}}}{2} \quad (8.18)$$

On the load capacitor, the charge is $Q = C_L V_{\text{out}}$. When the output changes from V_{DD} to $V_{DD}/2$, the charge on the capacitor changes by ΔQ , where

$$\Delta Q = -C_L \frac{V_{DD}}{2} = - \int_0^{t_{\text{df}}} I_{Dn} dt \quad (8.19)$$

The current flows through the NFET during this transition from high toward low at the output, hence I_{Dn} .

Consider an integrated CMOS circuit for which $V_{GS} - V_T = 2.6$ V, corresponding to $V_{DD} = 3.3$ V. Knowing I_{Dn} as a function of time, t_{df} can be calculated from Equation (8.19). The value of t_{df} can be estimated with the aid of Figure 8.9. Here we have plotted the current I_{Dn} in the NMOS as the input voltage is switched from logic **0** to logic **1**. Initially the NMOS is **off** and the PMOS is **on**, so the output voltage (also the voltage $V_{D(NFET)}$) is high. At $t = 0$, the NMOS is turned **on** and the PMOS is turned **off** by a step voltage input as indicated in Figure 8.8a. This corresponds to I_{Dn} going from zero (point 1 in Figure 8.9) to I_{Dsat} (point 2). As the capacitor discharges through the NMOS at a rate proportional to I_{Dn} , V_{DSn} decreases with time from V_{DD} (point 2) to V_{DSsat} (point 3) and to zero (point 4). For $V_{DSsat} \leq V_{DD}/2$ as indicated, the current is constant at I_{Dsat} during the time t_{df} required for the output to decrease from V_{DD} to $V_{DD}/2$. For this case, from Equation (8.19),

$$t_{df} \approx \frac{C_L V_{DD}}{2 I_{Dsatn}} \quad (8.20)$$

In this analysis, we are considering the propagation delay for an inverter discharging its load capacitance through the NFET. Therefore, the values for the parameters (L , W , μ_{lf} , V_{DSsat} , v_{sat} , C'_{ox}) used to calculate I_{Dsat} in Equation (8.20) are those of the NFET. The load capacitance is *charged*, however, through the PFET when the output goes low to high. Thus t_{dr} is

$$t_{dr} \approx \frac{C_L V_{DD}}{2 I_{Dsatp}} \quad (8.21)$$

For this, the parameters used to find I_{Dsat} pertain to the PFET. From Equations (8.18), (8.20), and (8.21) we can write

$$t_d = \frac{1}{2} \left[\frac{C_L V_{DD}}{2 I_{Dsatn}} + \frac{C_L V_{DD}}{2 I_{Dsatp}} \right] \quad (8.22)$$

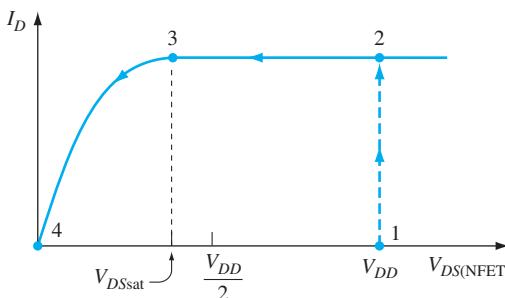


Figure 8.9 The current in the NFET as the input is switched from a logic **low** to a logic **high**. Initially the NFET output (drain) is **high** (point 1). When the input is switched to **high**, the NFET turns on and current can flow (point 2). The flowing current I_{Dn} discharges the load capacitance (point 3), sending the output voltage to zero (point 4).

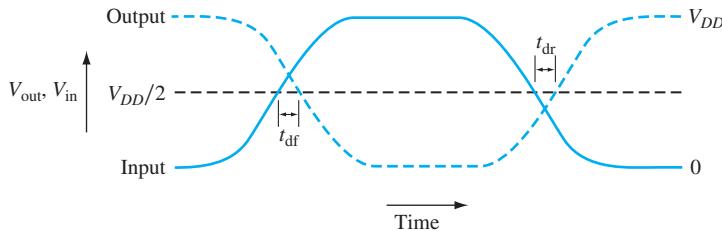


Figure 8.10 The definition of rise and fall delay times when the input is not a perfect step function.

We now investigate the validity of Equations (8.20) and (8.21), which rely on the condition that $|V_{DSsat}| \leq V_{DD}/2$. In the previous chapter, Figure 7.33 indicated the values of $|V_{DSsat}|$ as functions of channel length for NFETs and PFETs with the value of $V_{GS} - V_T = 2.6$ V. The value for the required saturation voltage is $|V_{DSsat}| \leq V_{DD}/2 = 3.3/2 = 1.65$ V. This value of V_{DSsat} can be obtained by making the channel length $L < 1.8 \mu\text{m}$ for the NFET, and $L < 0.8 \mu\text{m}$ for the PFET. For L in the submicrometer range then, a reasonable approximation is to let $I_D = I_{Dsat}$.

Ideally, to minimize gate delay we would want the rise and fall propagation delay times to be equal. To equate them, we make I_{Dsatn} and I_{Dsatp} approximately equal in Equation (8.22). We do this by appropriately adjusting the (W_p/W_n) ratios of the transistors as discussed earlier.

In the preceding discussion of propagation delay, it was assumed that a step voltage function was applied to the input. In reality, of course, the input voltage takes some time to change state, since it is coming from the output of a previous circuit. For this case, t_{df} and t_{dr} are indicated in Figure 8.10 as the time between the input crossing $V_{DD}/2$ and the output crossing $V_{DD}/2$ in the falling and rising cases, respectively.

In short-channel MOSFETs the expressions for the currents are more complex than discussed here, and the load capacitance depends on V_{DS} . In such cases, for accurate results, t_d is normally calculated numerically by using a circuit simulator.

EXAMPLE 8.4

Determine the propagation delay time for the CMOS inverter of Figure 8.8. Assume $I_{Dsatn} = I_{Dsatp} = 1 \text{ mA}$, $C_L = 0.1 \text{ pF}$, and $V_{DD} = 2.5 \text{ V}$.

■ Solution

From Equation (8.22), with $I_{Dsatn} = I_{Dsatp} = I_{Dsat}$,

$$t_d = \frac{C_L V_{DD}}{2 I_{Dsat}} = \frac{10^{-13} \text{ F} \times 2.5 \text{ V}}{2 \times 10^{-3} \text{ A}} = 1.25 \times 10^{-10} \text{ s}$$

or $t_d = 125 \text{ ps}$.

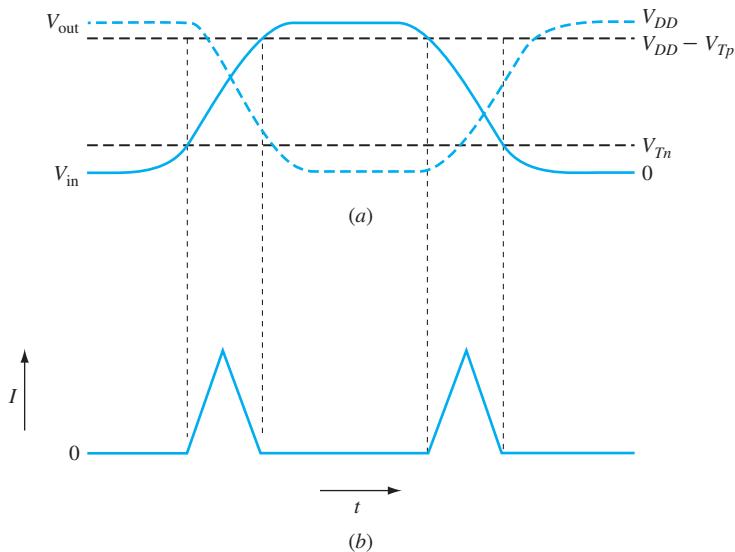


Figure 8.11 During the time that both devices are **on**, current flows from V_{DD} to ground. (a) The voltage waveforms; (b) the current waveform.

8.4.3 PASS-THROUGH CURRENT IN CMOS SWITCHING

The switching waveforms from Figure 8.10 are repeated in Figure 8.11a with some details added. The threshold voltages for the NMOS and PMOS are V_{Tn} and V_{Tp} respectively. We can see that when the input is switching, and passing through the range $V_{Tn} < V_{in} < (V_{DD} - |V_{Tp}|)$, both transistors are briefly on at the same time. During this interval, current flows from the power supply through the transistors to ground as indicated in part (b) of the figure. This results in dynamic power dissipation. The average value of this current decreases as the transition time of the input voltage is reduced. The power dissipation associated with this *pass-through* current can approach 50 percent of that given by Equation (8.17).

8.5 OTHER MOSFETs

Up to now, we have considered standard n-channel and p-channel MOSFETs. In this section we briefly discuss three variations of MOSFET structures. The first is silicon on insulator (SOI) MOSFETs, in which the device is separated from the substrate by an insulating buried oxide layer (BOX). Next we discuss FinFETs, a type of three-dimensional MOSFET in which the entire transistor structure is above the Si substrate. Then we consider floating gate (FG) MOSFETs, which retain their state (**0** or **1**) if the power is removed.

8.5.1 SILICON ON INSULATOR (SOI) MOSFETs

A variation of the MOSFET structure previously discussed is the so-called silicon on insulator (SOI) device. One method to fabricate these devices is the SIMOX process (separation by implanted oxygen). For an n-channel MOSFET,

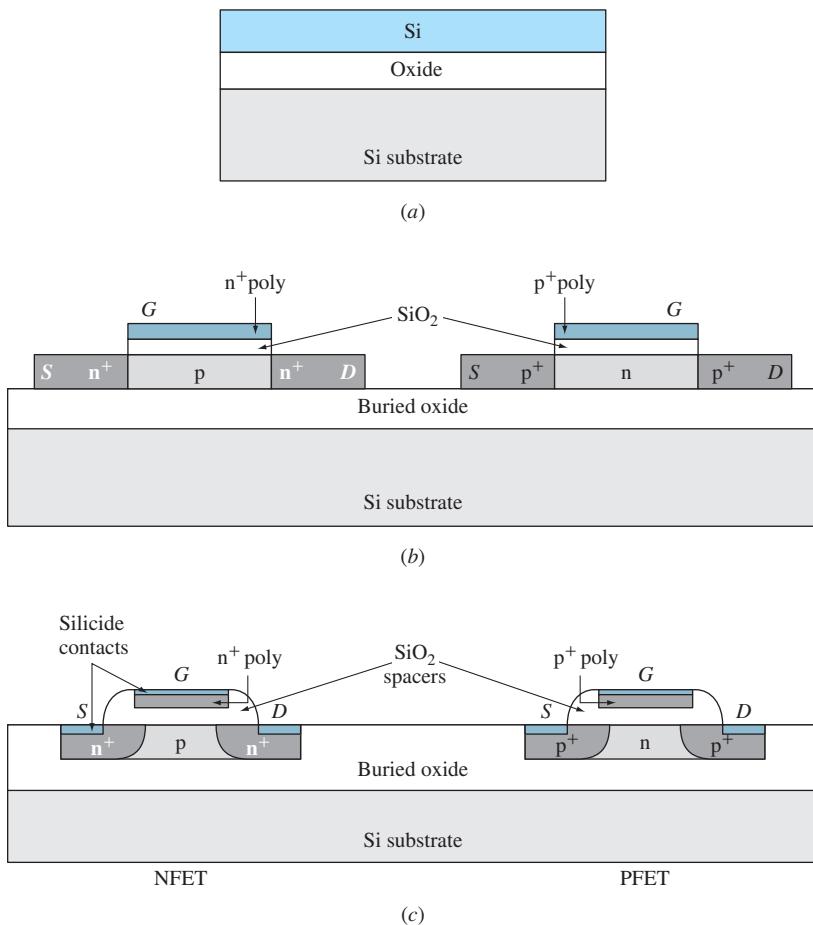


Figure 8.12 Cross section of (a) silicon on insulator; (b) an SOI NMOS and PMOS transistor; (c) a CMOS SOI structure. In (c) the gate, drain, and source contacts are metal silicides. The parasitic junction capacitances and substrate leakage currents are reduced from those of conventional MOSFETs.

oxygen is ion implanted beneath the surface of a p-type substrate. The wafer is then annealed at a high temperature, so a layer of SiO_2 is formed that effectively isolates the surface crystalline Si layer from that of the substrate (Figure 8.12a). The devices are then formed in the surface Si by standard means. Finally, the surface Si is selectively etched to isolate the devices as shown in Figure 8.12b, which shows NMOS and PMOS devices for a CMOS technology. With further processing, including SiO_2 deposition, the resultant CMOS structure is indicated in (c). The “metal” contacts are shown as refractory metal silicides (e.g., TiSi_2). The oxide spacers serve to confine the silicide contacts to source and drain regions. Such SOI devices can be manufactured with higher switching speeds than standard MOSFETs because the isolation of the devices results in reduced parasitic capacitance. [3]

There are two versions of SOI, partially depleted (PD SOI) and fully depleted (FD SOI). In the partially depleted SOI, the Si thickness under the gate is greater than the Si depletion region adjacent to the channel, Figure 8.13a. This results in an electrically “floating” neutral Si region between the oxide and the channel. This floating region can affect the device electrical characteristics. In the high-field depleted region near the drain, impact ionization can occur. In an n-channel device the depletion region field is in a direction such that the electrons are collected by the drain, but the holes are stored in the floating body, charging it positively. This in turn reduces the source-channel barrier and increases the channel current. This effect is often referred to as the *floating body* or *kink* effect. In high-speed devices, the time to charge the floating body is longer than the input transition time, resulting in an increased transient current.

In a fully depleted SOI device, Figure 8.13b, the Si layer is thin enough and the p region is lightly doped (or not intentionally doped) such that the depletion region extends from the channel to the SiO_2 region. This has the effect of reducing the subthreshold swing. From Equation (8.23),

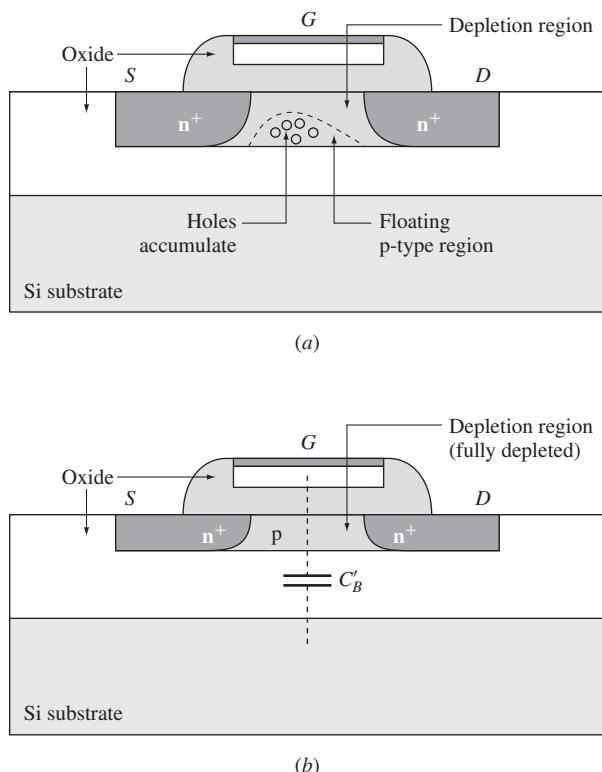


Figure 8.13 Two styles of SOI (an NMOS is shown).

(a) Partially depleted, with a floating semiconductor region; (b) fully depleted.

$$S = 2.3 \frac{kT}{q} \left(1 + \frac{C'_B}{C'_{ox}} \right) \quad (8.23)$$

where C'_B is the substrate capacitance per unit area. In the FD SOI the substrate capacitance C'_B extends from the channel across the p-type body and the SiO_2 layer to the Si substrate, resulting in a small value for C'_B . Thus the subthreshold voltage swing S is very near its ideal value of 60 mV/decade at room temperature. This reduced S permits a lower threshold voltage V_T , and thus a lower supply voltage for a given value of off current. In very short channel devices, the oxide effectively increases the field penetration in the region below the gate and increases the effect of the drain voltage on the threshold voltage (the DIBL effect).

Double-Gate SOI A variation of the SOI MOSFET structure is indicated in Figure 8.14 in simplified form. It is similar to the single-gate SOI MOSFET but

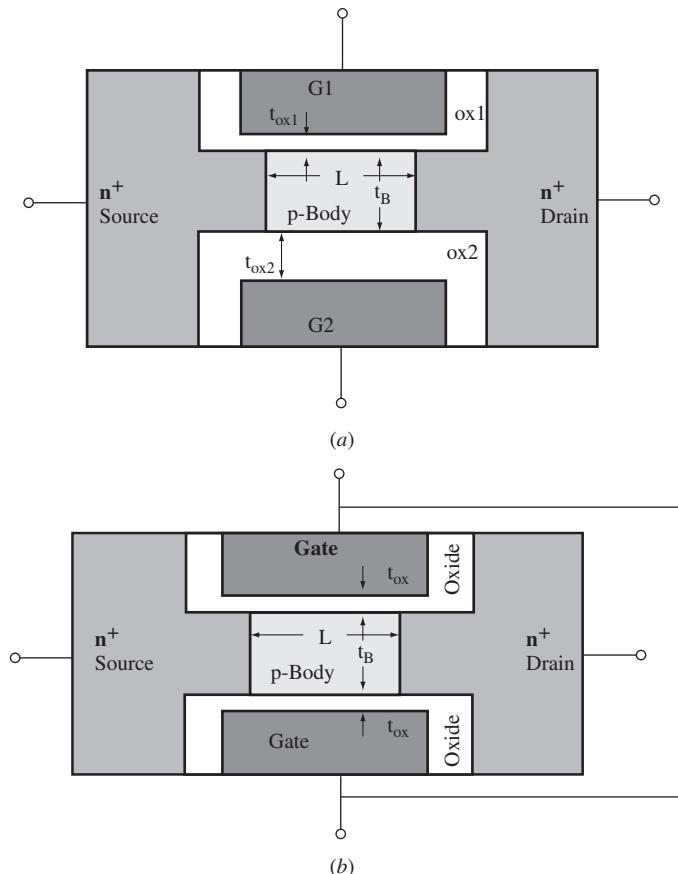


Figure 8.14 Simplified drawing of two double-gate MOSFET structures. An asymmetric SOI MOSFET (a) and a symmetric SOI MOSFET (b).

has two gates: Gate 1 (G1) above and Gate 2 (G2) below a thin silicon body (B) region. Figure 8.14(a) indicates the structure of an asymmetrical double-gate SOI MOSFET, in which the Gate 1 voltage is used to control the device current while Gate 2 is used to set the threshold voltage associated with Gate 1. The oxide for Gate 1 (ox1) is thin and of high quality (normally thermally grown), while the Gate 2 oxide (ox2) is thicker and need not be of high quality. Figure 8.14(b) illustrates the structure of a symmetrical SOI MOSFET. Here both gates are connected together, both contributing to current. Both oxides must be thin and of high quality. This adds considerably to the difficulty and cost of fabrication.

Asymmetrical Dual-Gate SOI MOSFETs Figure 8.15 shows the energy band diagram normal to the surface of a partially depleted (PD) dual-gate SOI device (a) and a fully depleted (FD) dual-gate SOI device (b). In (a), there is a depletion region associated with each gate, separated by an electrically neutral nondepleted region. In (b) the depletion regions are merged and there is no neutral region, or the Si body is fully depleted. In Figure 8.15b, the reference energy is the conduction band energy of the body in the flat-band condition (neutral, or undepleted condition). Assuming that the gate materials are the same, the surface potential in the body, ϕ_{s2} , is less than ϕ_{s1} because t_{ox2} , the oxide thickness for Gate 2, is greater than that for Gate 1. In the PD SOI device of Figure 8.15a, the MOSFETs associated with Gate 1 and Gate 2 are independent and this structure is of little interest. Here we consider only FD dual-gate SOI MOSFETs.

In the fully depleted device the charge density in the silicon body is $Q_B = -qN'_A$ and the total charge per unit cross-sectional area of the channel is $Q'_B = -qN'_A t_B$, where t_B is the thickness of the silicon body. The gate voltages for Gate 1 and Gate 2 with respect to the Si conduction band in the body in the

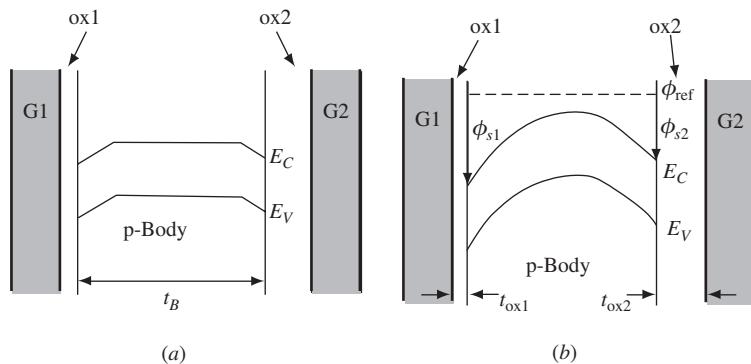


Figure 8.15 Energy band diagram normal to the surface of a partially depleted asymmetrical SOI MOSFET (a) and a fully depleted SOI MOSFET (b). The reference potential ϕ_{ref} is taken to be the conduction band in the body for the electrically neutral condition. The surface potentials are the potentials at the body-oxide surfaces with respect to this reference.

flat-band condition can be determined from Poisson's equation and Gauss's law. The results are [4]

$$V_{G1} = \frac{\Phi_{GB1}}{q} + \phi_{s1} - \frac{Q'_{ch1} + \frac{Q'_B}{2}}{C'_{ox1}} + \frac{C'_B (\phi_{s1} - \phi_{s2})}{C'_{ox1}} \quad (8.24)$$

$$V_{G2} = \frac{\Phi_{GB2}}{q} + \phi_{s2} - \frac{Q'_{ch2} + \frac{Q'_B}{2}}{C'_{ox2}} + \frac{C'_B (\phi_{s2} - \phi_{s1})}{C'_{ox2}} \quad (8.25)$$

where

V_{G1} is the voltage of Gate 1 with respect to the source.

V_{G2} is the voltage of Gate 2 with respect to the source.

Φ_{GB1} is the work function difference between Gate 1 and (neutral) body.

Φ_{GB2} is the work function difference between Gate 2 and (neutral) body.

Q'_{ch1} is the charge per unit area in the channel adjacent to Gate 1.

Q'_{ch2} is the charge per unit area in the channel adjacent to Gate 2.

Q'_B is the charge per unit area in the body ($Q_B = -qN_A t_B$).

t_B is the thickness of the Si body.

C'_{ox1} is the capacitance per unit area between Gate1 and Channel 1.

C'_{ox2} is the capacitance per unit area between Gate 2 and Channel 2.

C'_B is the capacitance per unit area of the Si body ($C'_B = \frac{\epsilon_{si}}{t_B}$).

ϕ_{s1} is the surface potential of the body silicon at Channel 1.

ϕ_{s2} is the surface potential of the body silicon at Channel 2.

In the asymmetrical device, current flow is in the top channel (Channel 1), while the Gate 2 voltage is used to adjust the threshold voltage of Channel 1. Figure 8.16 shows the potential profile across the device normal to the surface for three values of V_{G2} . Increasing V_{G2} lowers the electron energy everywhere

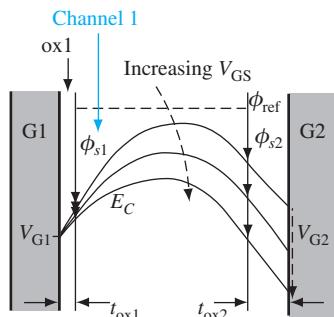


Figure 8.16 Potential profiles between Gate 1 and Gate 2 for different values of V_{G2} . Increasing V_{G2} decreases the threshold voltage of Gate 1.

between Gate 2 and Gate 1 as indicated. It can be seen that increasing V_{G2} increases ϕ_{s1} , thus reducing the body barrier height and the threshold voltage of the Gate 1–body structure.

Figure 8.17 shows a plot of the threshold voltage V_{T1} as a function of V_{G2} . Since it is desired to have no current flowing from source to drain in channel 2, the range of V_{G2} that can be applied to vary the threshold voltage of Gate 1 is limited to that for which the body adjacent to ox2 is depleted. That is because under either accumulation or inversion, channel carriers are present that can carry current.

Symmetrical Double-Gate SOI MOSFET A variation of the double-gate fully depleted (FD) SOI MOSFET is the symmetrical SOI structure indicated earlier in Figure 8.14(b). Here the two oxide layers have the same thickness and the gates are connected together. Thus there are two channels in parallel. In this case,

$$V_{G1} = V_{G2} = V_G, \quad \phi_{s1} = \phi_{s2} = \phi_s, \quad C'_{ox1} = C'_{ox2} = C'_{ox} \text{ and } Q'_{ch1} = Q'_{ch2} = \frac{Q'_{ch}}{2}$$

where Q'_{ch} is the charge in both channels combined. Then from Equation (8.24) [or (8.25)]

$$V_G = \frac{\Phi_{GB}}{q} + \phi_s - \frac{Q'_{ch} + Q'_B}{2C'_{ox}} \quad (8.26)$$

Threshold Voltage of a Double-Gate Fully Depleted SOI MOSFET In a conventional MOSFET the threshold voltage is defined as $V_T = 2\phi_s$ where ϕ_s is determined by N_A . For a symmetrical FD SOI, however, at threshold $V_G = V_T$, $Q'_{ch} = 0$ and Equation (8.26) becomes³

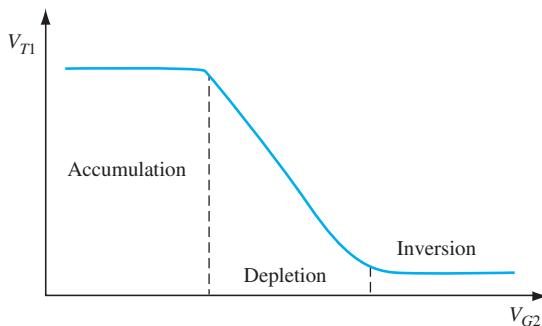


Figure 8.17 Variation of V_{T1} with V_{G2} . To avoid drain-to-source current in channel 2, the interface between the body and oxide 2 must be between accumulation and inversion.

³Other definitions for threshold voltages in FD SOI MOSFETs exist.

$$V_T = \frac{\Phi_{GB}}{q} + \phi_s - \frac{Q'_B}{2C'_{ox}} \quad (8.27)$$

Solving for Q'_{ch} in Equations (8.26) and (8.27)

$$Q'_{ch} = -2C'_{ox}(V_G - V_T) \quad (8.28)$$

With an applied drain-to-source voltage the drain current can be determined as for a conventional MOSFET. We start with Equation (7.58), repeated here

$$I_D = \frac{-W\mu_{lf}\int_0^{V_{DS}} Q_{ch} dV_{ch}(y)}{L\left(1 + \frac{\mu_{lf}V_{DS}}{Lv_{sat}}\right)} \quad (8.29)$$

With $V_{DS} > 0$, from Equation (8.28), Q_{ch} becomes

$$Q'_{ch}(y) = -2C'_{ox}(V_{GS} - V_T - V_{ch}(y)) \quad (8.30)$$

The drain current is then

$$I_D = 2 \frac{WC'_{ox}\mu_{lf}\left[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2}\right]}{L\left(1 + \frac{\mu_{lf}V_{DS}}{Lv_{sat}}\right)} \quad V_{DS} \leq V_{Dsat} \quad (8.31)$$

$$I_{Dsat} = 2 \frac{WC'_{ox}\mu_{lf}\left[(V_{GS} - V_T)V_{DSSat} - \frac{V_{DSSat}^2}{2}\right]}{L\left(1 + \frac{\mu_{lf}V_{DSSat}}{Lv_{sat}}\right)} \quad V_{DS} \geq V_{DSSat} \quad (8.32)$$

The current is twice that of a conventional MOSFET because of the two gates.

Subthreshold Swing in Double-Gate FD SOI MOSFET A major advantage of a FD SOI MOSFET is the reduced subthreshold swing compared to a conventional MOSFET.

The ideality factor of a MOSFET, n , is defined as

$$n = \frac{dV_G}{d\phi_s} \quad (8.33)$$

From Equation (8.26)

$$n = \frac{dV_G}{d\phi_s} = \frac{d\left(\frac{\Phi_{GB}}{q} + \phi_s - \frac{Q'_{ch} + Q'_B}{2}\right)}{d\phi_s} \quad (8.34)$$

In the subthreshold region, $Q'_{ch} \approx 0$ and Φ_{GB} , Q'_B and C'_{ox} are all independent of ϕ_s . Thus the ideality factor is unity and the subthreshold swing has its minimum

value of $2.3 \frac{kT}{q}$ or 60 mV/decade of subthreshold current at room temperature. For a very thin body, ultra-thin body, or UTB SOI device ($t_B < 5$ nm), the potential is nearly constant across the body. Here the structure resembles the quantum electron-in-a-box problem, and rather than having well-defined conduction and valence bands as in a bulk semiconductor, minibands are formed above E_C and below E_V as indicated in Figure 8.18. In effect, the forbidden band has increased, which increases the threshold voltage.

This structure is difficult to fabricate. A structure that has the advantages of a symmetrical double-gate SOI structure but is easier to fabricate is the FinFET, which is discussed in Section 8.5.2.

It is interesting to compare the operation of a fully depleted SOI CMOS and a bulk CMOS inverter. For matched devices such that $I_{D\text{satn}} = I_{D\text{satp}} = I_{D\text{sat}}$, from Equation (8.22) the gate delay for either case is

$$t_d = \frac{V_{DD}}{2I_{D\text{sat}}} C_L \quad (8.35)$$

where C_L is the load capacitance. In practice, a CMOS inverter is normally used to drive one or more following circuits. In this case the load capacitance is

$$C_L = C_{\text{out}} + C_w + \text{FO} \times C_{\text{in}} \quad (8.36)$$

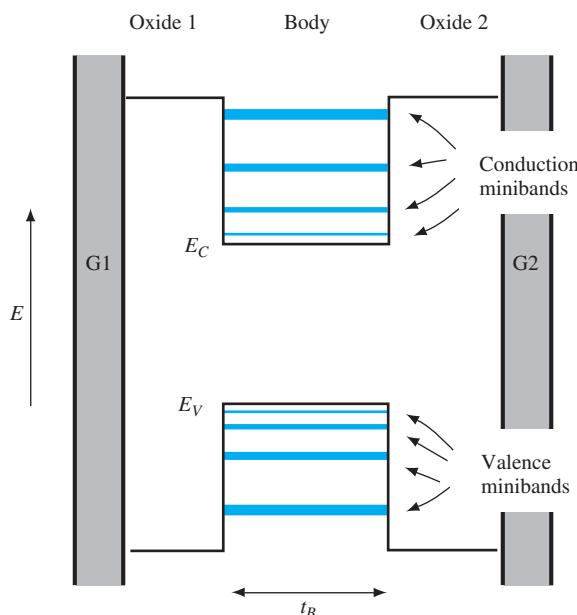


Figure 8.18 In an ultra thin body (UTB) symmetrical two gate SOI device mini bands exist above E_C and below E_V . This increases the threshold voltage.

where C_{out} = output capacitance of the driving circuit

C_w = stray wiring capacitance

FO = fan-out, or the number of following circuits being driven in parallel

C_{in} = input capacitance of one of the driven circuits, assuming identical driven circuits.

The output capacitance C_{out} is the capacitance at the output of the driving stage and consists of the drain junction capacitances and the drain-to-gate capacitances, including the overlap capacitance. The single-stage input capacitance C_{in} includes the gate-to-source, gate-to-drain, and gate-to-substrate (gate-to-channel + channel-to-substrate) capacitances.

In a bulk CMOS inverter, a large fraction of C_{out} is due to the drain junction capacitances, which are negligible in SOI CMOS. The wiring capacitances C_w and input capacitances C_{in} in SOI CMOS are comparable to those in bulk CMOS. Because of its decreased C_{out} , then, the SOI CMOS is inherently faster than bulk CMOS. However, this speed advantage decreases with increased fan-out.

Another advantage of SOI devices is their reduced susceptibility to *soft errors* relative to their bulk counterparts. Soft errors are a result of high-energy particles striking a device, creating a number of electron-hole pairs in the Si body that can be collected and can change a stored logic state. Because of the reduced volume of the bulk Si in SOI devices, the number of electron-hole pairs created is low compared with the case for bulk MOS devices.

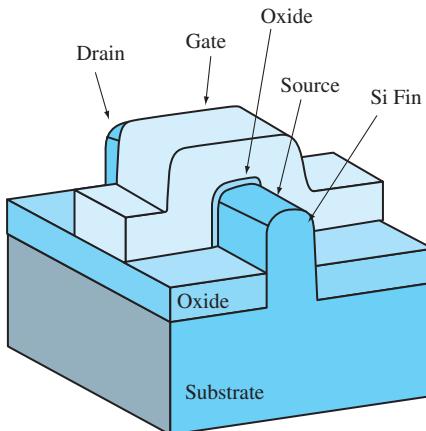
8.5.2 FINFETS

To achieve a higher packing density of MOSFETs, the planar technology is replaced by a three-dimensional technology. Two versions of a tri-gate FinFET are shown in Figure 8.19. The transistors feature a “*Fin*” above the surface of the bulk silicon. The silicon Fin is composed of a doped source and drain, and an undoped (or lightly doped) semiconductor channel region separated by a thin insulating layer under the conducting gate. In (a), the Fin is an extension of the Si substrate. In (b) the Fin is separated from the Si substrate by an oxide layer (SOI FinFET). The term *tri-gate* comes from the gate region on the two sides and the top of the silicon Fin channel. Under gate bias, the channel is fully depleted and the thickness of the channel is that of the silicon Fin; thus the channel capacitance, C'_B , is essentially zero and the subthreshold swing is, from Equation (8.23),

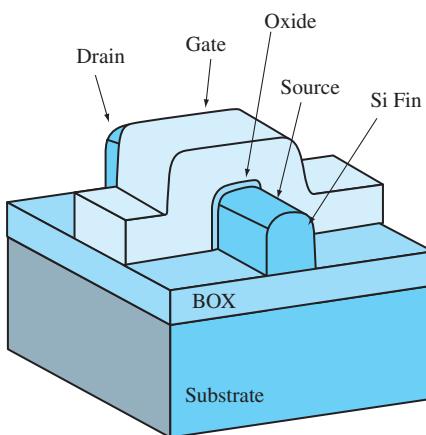
$$S = \frac{2.3kT}{q} \left(1 + \frac{C'_B}{C'_{\text{ox}}} \right) \approx \frac{2.3kT}{q} = 60 \text{ mV/decade} \quad (8.37)$$

which results in a reduced leakage current in the **off** state ($V_{GS} = 0$), as indicated in Figure 8.20.

In a planar MOSFET the **on** current is determined by the gate width/length ratio. The MOSFET width equivalent in a FinFET is (approximately) twice its height plus its width. Since a small Fin width is desired, and because of the difficulty of fabricating large Fin heights, for adequate **on** current, FinFETs often



(a)



(b)

Figure 8.19 Two versions of a tri-gate FinFET.

consist of several Fins in parallel. This is indicated schematically in Figure 8.21 for three parallel Fins having common source and drain. An electron microscope photo of a multiple-channel SOI FinFET is shown in Figure 8.22 for two magnifications. In this device the gate is a metal (TiN) and the gate-channel insulation is HfO_2 . The insulating HfO_2 is chosen because it has a higher dielectric constant (ϵ_r or $k \approx 20$) than SiON (≈ 5) or SiO_2 (3.9). This higher dielectric constant increases the gate-channel capacitance and reduces the subthreshold swing [Equation (8.23)] as well as increasing the $I_{\text{on}}/I_{\text{off}}$ ratio. A variation of the FinFET is a *gate-all-around (GAA) transistor*, Figure 8.23. Here the Fin is replaced by a *nanowire*, a narrow rod of silicon (silicon nanowire) surrounded by the control gate.

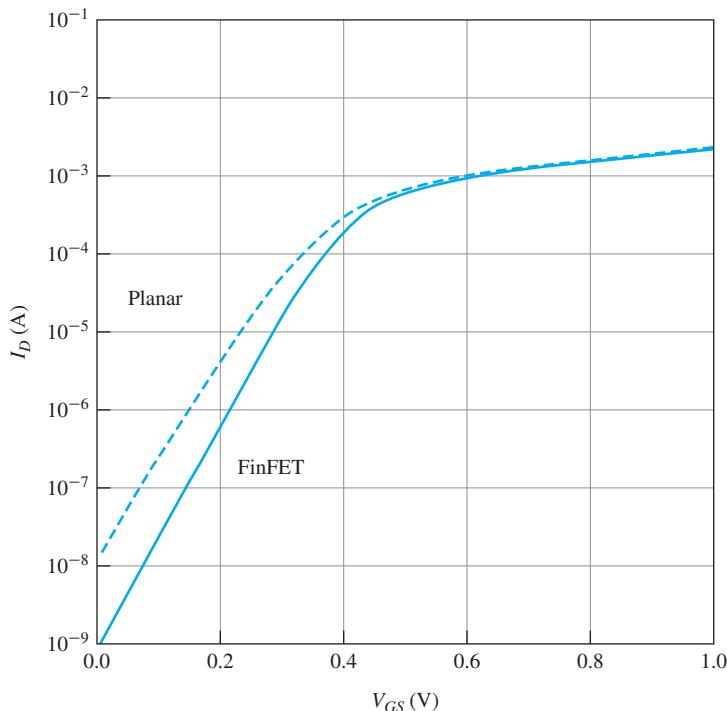


Figure 8.20 Comparison of MOSFET and FinFET. The FinFET has a reduced threshold swing and thus reduced leakage current in the off state for a given on current.

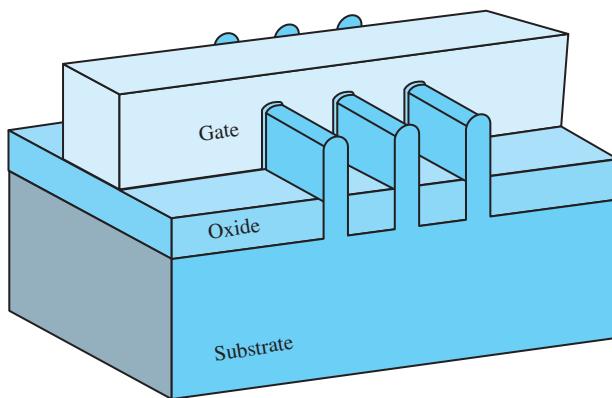


Figure 8.21 Schematic of three FinFETs in parallel.

8.5.3 NONVOLATILE MOSFETS

Nonvolatile silicon semiconductor memories use a MOSFET similar to a standard MOSFET except that it has two gates instead of one, a control gate (CG) and a poly-Si floating gate (FG) as shown in Figure 8.24 for one structure of an

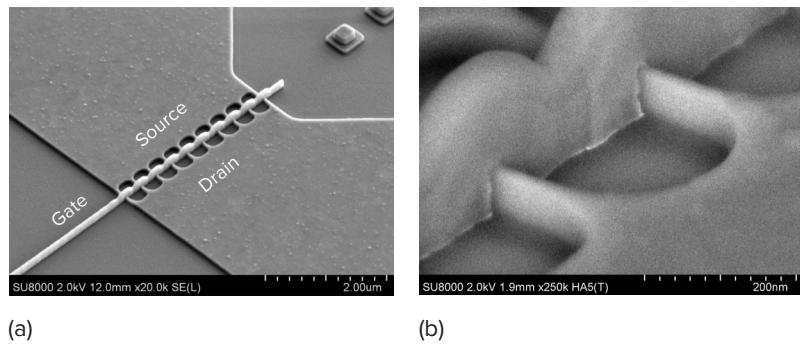


Figure 8.22 Electron micrographs for a SOI FinFET having 10 parallel channels (a) and at increased magnification (b). The Fin width is 20 nm and its height is 65 nm. The channel length (the gate length from source to drain) is 50 nm. (Courtesy IMEC)

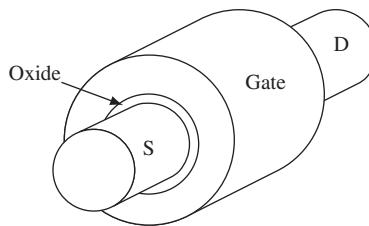


Figure 8.23 Schematic of gate-all-around (GAA) MOSFET.

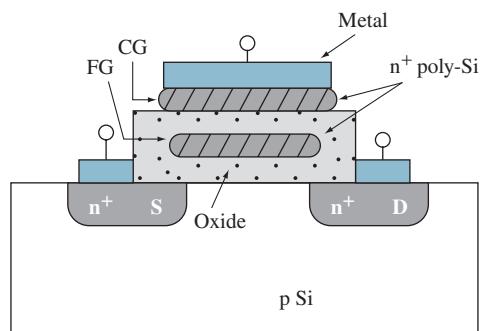


Figure 8.24 Cross-sectional view of an n-channel nonvolatile MOSFET. A Si floating gate (FG) is isolated between the control gate (CG) and Si substrate (not to scale).

n-channel floating gate MOSFET. The FG is electrically isolated by an insulating layer, so any electrons that are trapped there remain there for several years (with no voltages applied). By applying appropriate biases, electrons can be injected into or extracted from the floating gate. The presence of negative charge in the FG reduces the field in the oxide at the oxide-Si body interface, thus increasing the threshold voltage.

For an electrically neutral FG, the transistor is designed to have a threshold voltage V_{T1} . With appropriate electron charge in the FG, the threshold voltage is increased to a value V_{T2} . To determine the state of the transistor (read-out of the memory cell), a drain-source voltage is applied, and a voltage between the values V_{T2} and V_{T1} is applied to the control gate with respect to the source. If the FG is uncharged, the channel conducts, producing a drain current (logical 1). If the FG is charged, no current flows (logical 0).

Programming To charge the FG (i.e., change the device state from **1** to **0**), the CG is biased to a relatively high voltage (V_{PP}), currently about 5 V, and the drain-source voltage ($V_D = V_{DD}$) is sufficient that considerable current flows. A cross-sectional view (for clarity, not to scale) of a floating gate transistor is shown in Figure 8.25a along with the corresponding energy band diagram near

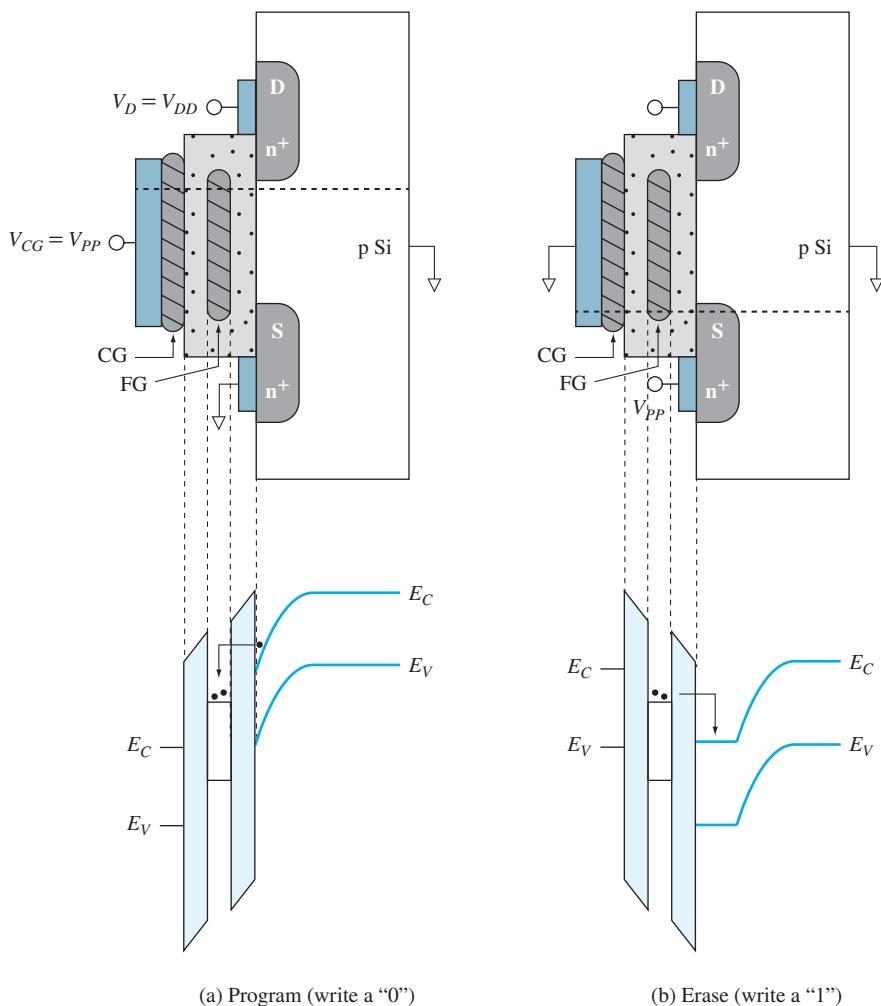


Figure 8.25 View of the MOSFET of Figure 8.24 with voltages applied. (a) To program (write a **0**), V_{PP} is applied to the control gate, increasing the threshold voltage from V_{T1} to V_{T2} . Electrons tunnel through the oxide from the drain end of the channel to the FG, which becomes negatively charged (moves upward on the energy band diagram). (b) To erase (write a **1**), the control gate is grounded and V_{PP} is applied to the source. Electrons tunnel from the FG into the source until the FG is at the potential of the source, decreasing the threshold voltage from V_{T2} to V_{T1} .

the drain (also not to scale). In current devices with a thin FG-channel insulating layer (*e.g.* 2–3 nm), the primary process of charging the FG is by tunnel injection from channel to FG. As the FG is charged negatively, the FG band energies rise (on the diagram) until the FG potential reaches the potential of the channel at the drain, at which point the tunneling stops. The charge on the FG and thus the threshold voltage (V_{T2}) is determined by the CG voltage (V_{PP}) and the drain voltage applied during programming.

To erase the transistor (removing the FG charge, or changing the cell status from **0** to **1**), a positive voltage (V_{PP}) is applied to the source with respect to the substrate and CG. The situation is shown in Figure 8.25b. The FG is discharged by electrons tunneling into the source.

8.6 OTHER FETs

Until now, we have concentrated on Si-based MOSFETs. There are, however, other types of FETs that have important applications. In this section, some specific FETs are briefly discussed.

8.6.1 HETEROJUNCTION FIELD-EFFECT TRANSISTORS (HFETs)

We begin with heterojunction field-effect transistors (HFETs). These have structures similar to that of MOSFETs, except that the gate structure is replaced by a semiconductor with a higher band gap than the rest of the device.

Figure 8.26a shows the cross section of an HFET device⁴ made in the AlGaAs-GaAs system. Figure 8.26b shows the energy band diagram perpendicular to the gate. When the heterojunction is formed, a discontinuity in the conduction band edge E_C results. This is analogous to the barrier between the semiconductor and the oxide in a MOSFET. A channel is thus created at the interface between the two different semiconductors.

To fabricate the device, a semi-insulating GaAs substrate is made by introducing traps near the center of the gap such that the Fermi level lies near midgap. Because the Fermi level is near the center of the gap, there are few free electrons or holes and the substrate is a semi-insulator.

Next, a lightly doped p layer of GaAs is grown epitaxially onto the semi-insulating substrate. This serves as the bulk semiconductor for the device, and the channel at its surface is manipulated by the application of an electric field at the gate (field-effect transistor). Over this bulk layer, a thin layer of near-intrinsic AlGaAs is grown, and then a thicker layer of n AlGaAs. The AlGaAs layers serve as the gate structure.

Finally, a layer of n⁺ GaAs is deposited as a passivation layer for the AlGaAs. A passivator is required because the Al in the AlGaAs reacts chemically with oxygen in the atmosphere. This GaAs “cap” isolates the AlGaAs from

⁴This device is also referred to as a modulation doped field-effect transistor (MODFET) or a high electron mobility transistor (HEMT).

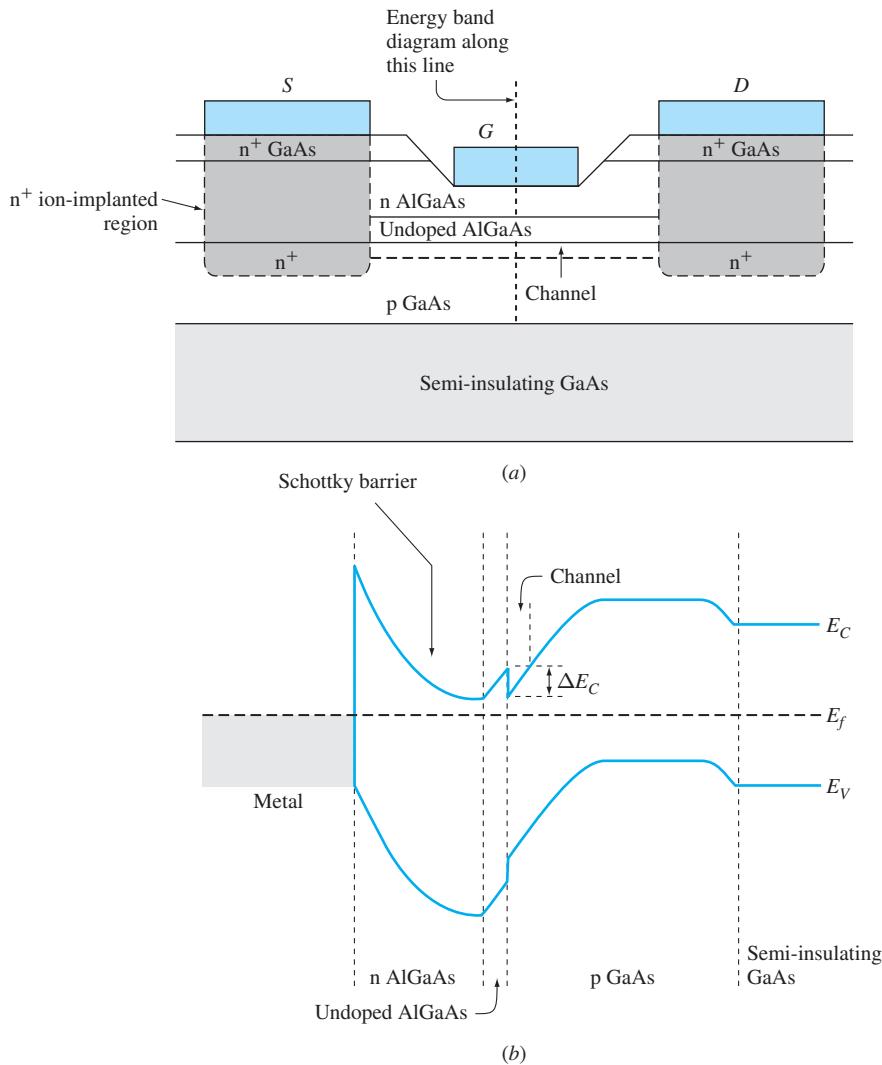


Figure 8.26 (a) The cross-sectional schematic of a GaAs-based HFET; (b) the energy band diagram normal to the gate. The Schottky barriers at the metal-GaAs interfaces (source and drain) are thin enough to be of low resistance because of tunneling.

the atmosphere. This layer is later etched in the region of the gate to make the gate contact directly to the wide-band-gap layer.

To create the source and drain, donor impurities are then implanted through the layers to make n⁺ regions that extend from the surface to inside the p GaAs. A metal is then deposited on the n⁺ GaAs surface to form thin ohmic (tunneling) Schottky barriers. Because the depletion region in the n⁺ GaAs is thin, the tunnel current is large and the contact is low resistance, or ohmic.

Because of the difference in conduction band edges E_C , a channel exists in the p GaAs at its interface with the AlGaAs. Depositing a metal on the n AlGaAs makes the gate contact. A Schottky barrier exists at the metal-AlGaAs contact. This AlGaAs layer is thin enough that at equilibrium it is entirely depleted. Thus there is no conducting path from source to drain through the AlGaAs.

The conduction band diagrams for this case are indicated for three values of gate voltage in Figure 8.27. Note that for the case illustrated, at $V_G = 0$ a conducting channel exists between source and drain and the device is a depletion mode HFET.

The conduction band profile for V_G positive is shown in Figure 8.27b. The positive gate voltage lowers the conduction band edge at the AlGaAs-GaAs hetero junction with respect to the electrically neutral GaAs; thus the electron

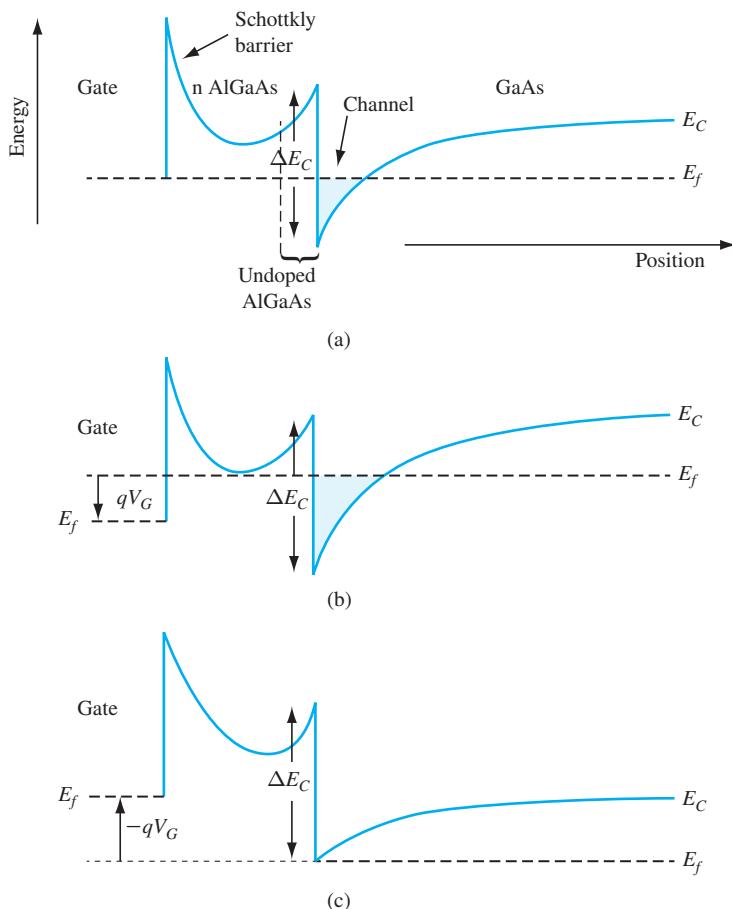


Figure 8.27 Conduction band diagram normal to the gate for the HFET of Figure 8.26a (a) at equilibrium, (b) for positive gate bias and (c) for negative gate bias.

concentration in the channel increases. Note that the positive gate voltage is limited to that which lowers the conduction band minimum in the AlGaAs sufficiently to permit electron occupation, permitting an alternate current path between source and drain. For a negative gate voltage the conduction band edges increase in energy, reducing the electron concentration in the channel. Figure 8.27c shows the case for a depleted channel—i. e. the **off** condition.

EXAMPLE 8.5

If the undoped layer is insulating in the channel, explain why there is a low resistance from source contact to the internal n^+ source region.

■ Solution

This is best illustrated with the aid of the energy band diagram normal to the source contact, Figure 8.28. Because the source region is heavily implanted with donors, the total region between the source contact and the p region is degenerate n type. The barrier between the metal contact and the n^+ GaAs is thin enough to be a low-resistance tunneling junction. The n^+ GaAs region is heavily doped, so it also has low resistance. The next barrier, between the n^+ GaAs and the n^+ AlGaAs, is also thin, again allowing tunneling, while the deeper n^+ AlGaAs layer between them is heavily doped and thus has low resistance. Finally, the undoped AlGaAs layer, which created the channel in the region under the gate, is degenerate because of the implanted ions and here causes a thin barrier that allows tunneling. Similarly, the drain contact–intrinsic drain junction also has low resistance.

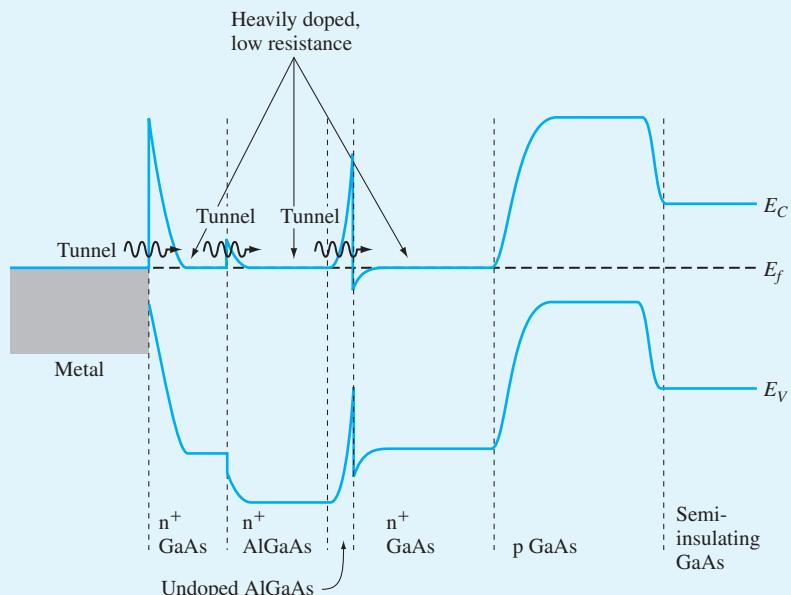


Figure 8.28 The energy band diagram for an HFET perpendicular to the source.

From the energy band diagram of Figure 8.26b, we can see that the built-in voltage between gate and substrate is dropped in part across the undoped AlGaAs and in part across the p GaAs. The energy at the bottom tip of the channel, relative to the Fermi level, is determined by the thickness of the undoped AlGaAs. The AlGaAs must be thick enough that the built-in voltage it supports is greater than $\Delta E_C/q$. The thickness of this layer therefore controls the HFET threshold voltage, an effect similar to the influence of the oxide thickness on V_T in a MOSFET.

The operation of an HFET is similar to that of a MOSFET. For the device in the figure, the channel is above the Fermi level at equilibrium and does not contain many electrons. Thus, the channel does not conduct appreciably. When a positive voltage is applied to the gate, the channel side of the energy band diagram moves down. The bottom of the channel will approach the Fermi level, allowing the channel to fill with electrons and conduct. A more negative gate voltage will deplete the channel. The potential well that forms the channel is very thin, so the conducting layer is often considered to be a two-dimensional sheet of mobile charge, or a two-dimensional electron gas (2DEG).

The advantage of using an HFET structure is high speed. The undoped AlGaAs and the lightly p-doped GaAs provide little scattering of channel electrons by ionized impurities. Since the electron scattering is reduced, their mobility is increased. Thus, high-speed devices can result.

Another type of HFET or HEMT fabricated in the gallium nitride system uses an AlGaN/GaN heterostructure. This HFET is different from the AlGaAs/GaAs HFET just described because there is naturally occurring charge in the channel resulting not from doping but from the crystallographic structure. To see this, we start with Figure 8.29, which shows two possible orientations of GaN's wurtzite structure. Both alternate layers of ABAB, but in one case A and B layers are inverted versions of the other case. Further, because the bonds are somewhat ionic, the center of charge for the negative charge is not the same as for the positive charge, creating an induced dipole. The surface is said to be N-polar when the electric dipole moment points toward the surface and Ga-polar otherwise. The polarization results in a built-in electric field. The number of spontaneous polarization charges per unit area on the surface of GaN is $n_\pi \approx 10^{13} \text{ cm}^{-2}$, so the built-in electric field is

$$\mathcal{E}_\pi = \frac{Q_\pi}{\epsilon} = \frac{qn_\pi}{\epsilon_r \epsilon_0} = \frac{(1.6 \times 10^{-19} \text{ C})(10^{13} \text{ cm}^{-2})}{(9.5)(8.84 \times 10^{-14} \text{ F/cm})} = 1.9 \text{ MV/cm} \quad (8.38)$$

When a heterojunction between GaN and AlGaN is formed, a second effect comes into play. Because of the mismatch in lattice constant, there is some strain at the interface. GaN and its alloys are *piezoelectric*, meaning that compressive or tensile strain can produce a voltage. Increasing the Al content increases both the spontaneous polarization charge and the piezoelectric charge. The sign of the piezoelectric charge, however, can vary depending on whether the material is

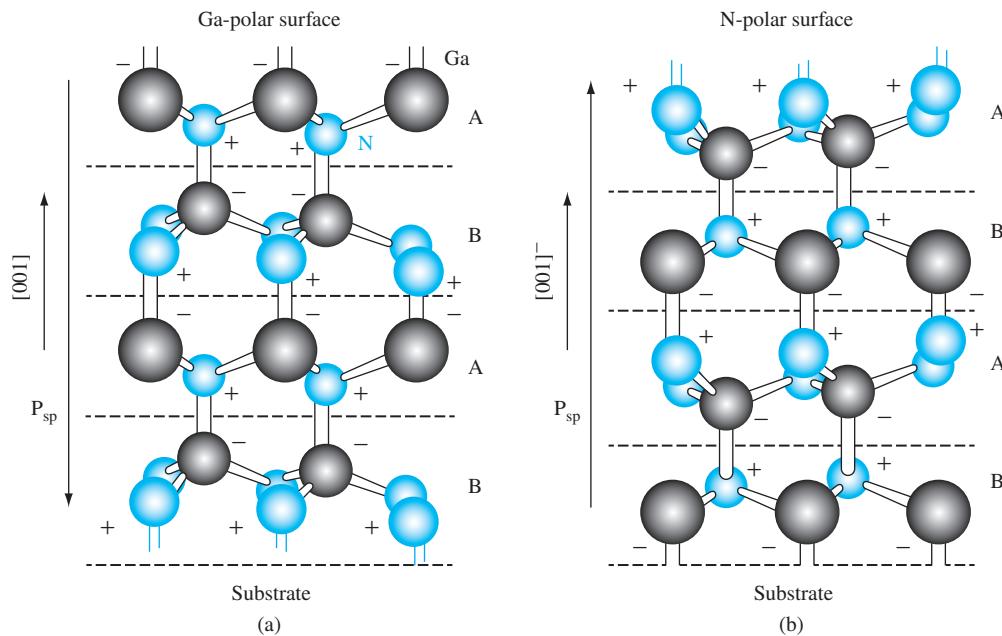


Figure 8.29 GaN in the wurtzite crystal structure. (a) Ga-polar surface, (b) N-polar surface.

Ga-polar or N-polar, and whether the strain is compressive or tensile. Thus the piezoelectrically induced charge can add to or partially screen the spontaneous polarization charge.

In a GaN-based HFET, there is a heterojunction between a layer of AlGaN and GaN, Figure 8.30a. In this device, the AlGaN layer acts as the insulating gate, and a thin channel of charge appears on the GaN side of the AlGaN/GaN heterojunction, Figure 8.30b. The AlGaN polarization charges (the sum of the spontaneous and piezoelectrically-induced charges), given by $\pm Q_{\pi AlGaN}$, shown in part (c) of the figure, create a constant electric field in the AlGaN layer, seen as a constant slope in E_C in that layer, Figure 8.30d. The polarization charge increases with Al content, resulting in a net charge difference between the AlGaN and GaN material at the heterojunction. There are donor-like defects at the interface that contribute electrons, which are trapped by the potential well formed due to the ΔE_C . As with the previous HFET, these electrons are confined to a very narrow region, resulting in a two-dimensional electron concentration at the interface, with a density n_{2DEG} .

At the interface between the GaN and the substrate, many defects are present that can become ionized (Q_{scf}), effectively screening the polarization charge there and resulting in flat bands in the bulk GaN. At the surface of the AlGaN, there are N_{DD} donor-like surface states that partially offset the polarization charges at the surface.

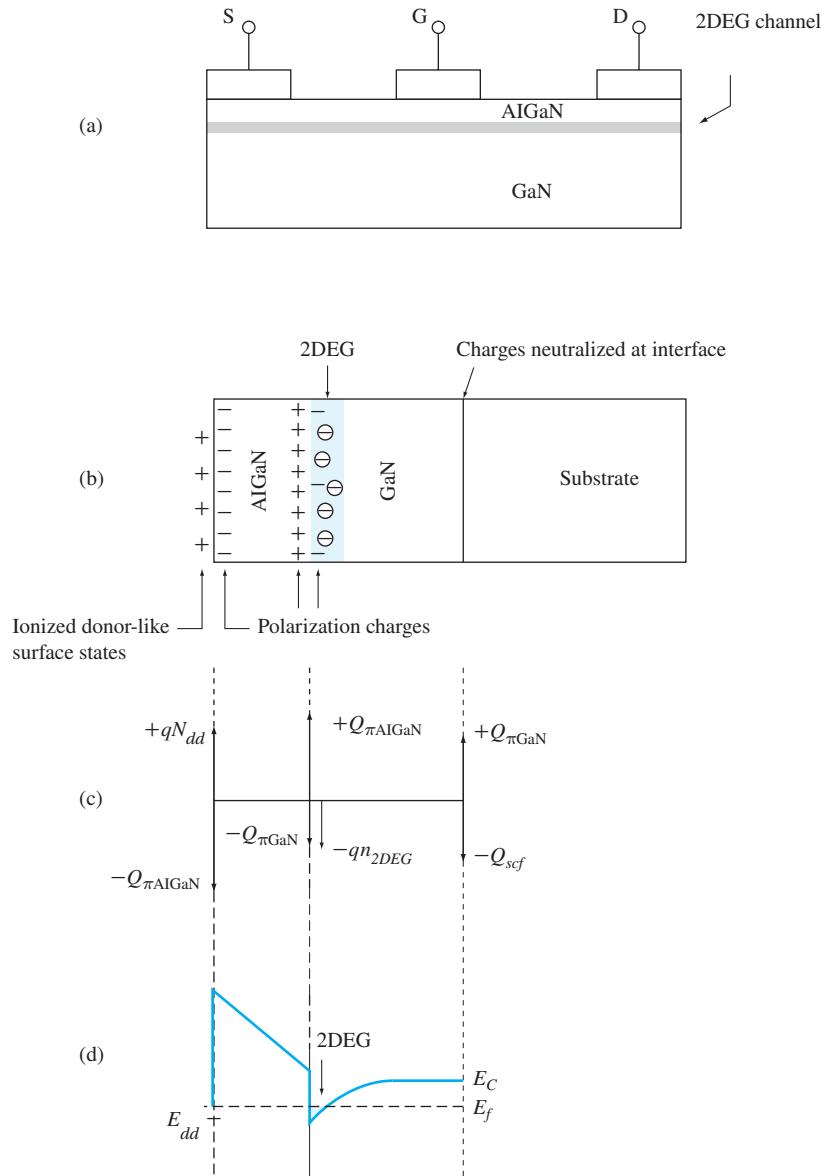


Figure 8.30 (a) Structure of an AlGaN/Gan HFET. (b) Polarization charges appear at the interfaces. Surface states contribute charges creating a two-dimensional electron gas in the GaN near the heterojunction. (c) Sources of charge. (d) equilibrium conduction energy band diagram.

The net result is that at the heterojunction there is a thin sheet of charge $-qn_{2DEG}$ that can carry current in the channel, Figure 8.30d. The difference between the AlGaN/GaN polar HFET and an HFET in the AlGaAs/GaAs system

is that in the polar case, the 2DEG arises from the polarization charges rather than donors, and the sheet charge densities are on the order of 10 to 100 times greater for the GaN-based devices than for AlGaAs/GaAs HFETs. This results in higher current-carrying capacity, with current density greater than 2 A/mm reported [5]. The large band gap of GaN and its alloys also implies high breakdown voltages, which, combined with the high electron mobility, makes GaN an excellent material for high-speed power transistors.

8.6.2 METAL-SEMICONDUCTOR FIELD-EFFECT TRANSISTORS (MESFETS)

Figure 8.31 illustrates the structure of an n-channel GaAs-based metal-semiconductor field-effect transistor (MESFET). A thin film of n-type GaAs is either ion implanted into or epitaxially deposited onto a semi-insulating GaAs substrate. The source and drain are then formed by heavily doping (degenerately) n^+ -type regions by ion implantation. The implant is followed by an annealing step to reduce the structural damage caused by the ion implantation. A metallic gate is then deposited over the n-type region to form a Schottky barrier.

The Schottky barrier causes a depletion region that extends into the GaAs as shown in Figure 8.32a for the case of equilibrium. The energy band diagram perpendicular to the gate (zero drain and gate voltages) is shown in Figure 8.32b. The channel thickness t extends from the edge of the Schottky barrier depletion region to the edge of the n GaAs-substrate depletion region.

This device is different from the FETs discussed earlier, in that in this case the channel does not form directly at the semiconductor surface, but somewhere below it. The depletion region induced by the Schottky barrier is devoid of carriers, and is thus insulating. The depletion region behaves, then, somewhat like the insulating gate in the MOSFET.

We recognize that since the source, channel, and drain are all n type, a channel exists even when no voltage is applied. Thus, the MESFET shown in Figure 8.32 is a depletion-mode MESFET.

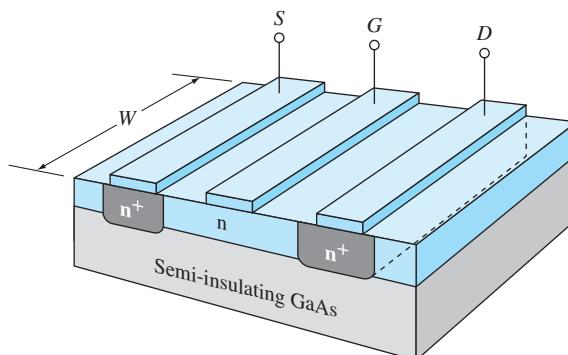
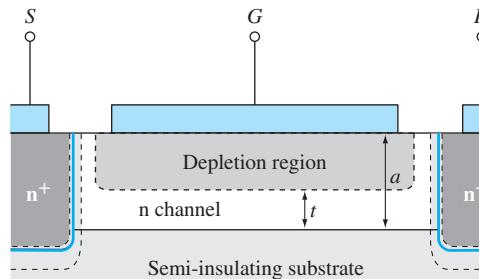


Figure 8.31 Schematic showing the structure of a GaAs MESFET.



(a)

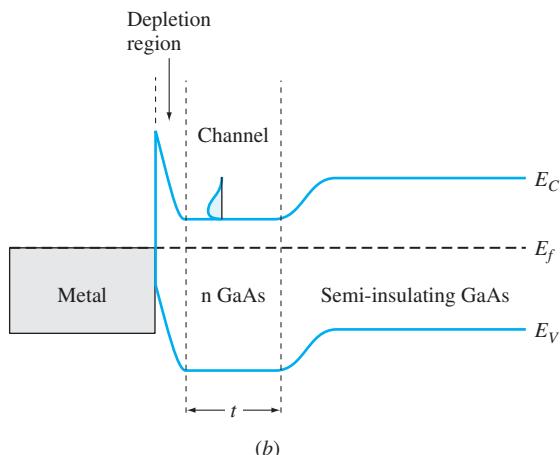


Figure 8.32 (a) Cross section of a MESFET at equilibrium indicating the depletion regions; (b) the energy band diagram perpendicular to the gate. The channel thickness is t .

Next, we recall that the channel charge per unit area in a field-effect transistor depends on the barrier height from source to channel, E_B . This barrier height is equal to the built-in voltage between the n⁺ source and the n channel, Figure 8.33a. The electron concentration n in the channel is equal to the net doping in the channel, N'_D . The value of the channel charge density (charge per unit area) Q_{ch} at the source end is then proportional to the channel thickness t (from $Q_{ch} = -qN'_Dt$, where it is assumed that N'_D is uniform in the channel), and t is controlled by the gate voltage.

We also observe that there are two depletion regions. One is the depletion region caused by the Schottky junction and the other is the depletion region between the channel and the semi-insulating substrate. If a gate voltage V_G is applied such that these two depletion regions overlap, the source-channel barrier increases. This is shown in Figure 8.33b. Increasing this barrier reduces the mobile channel charge Q_{ch} to (almost) zero, which turns off the device. The value

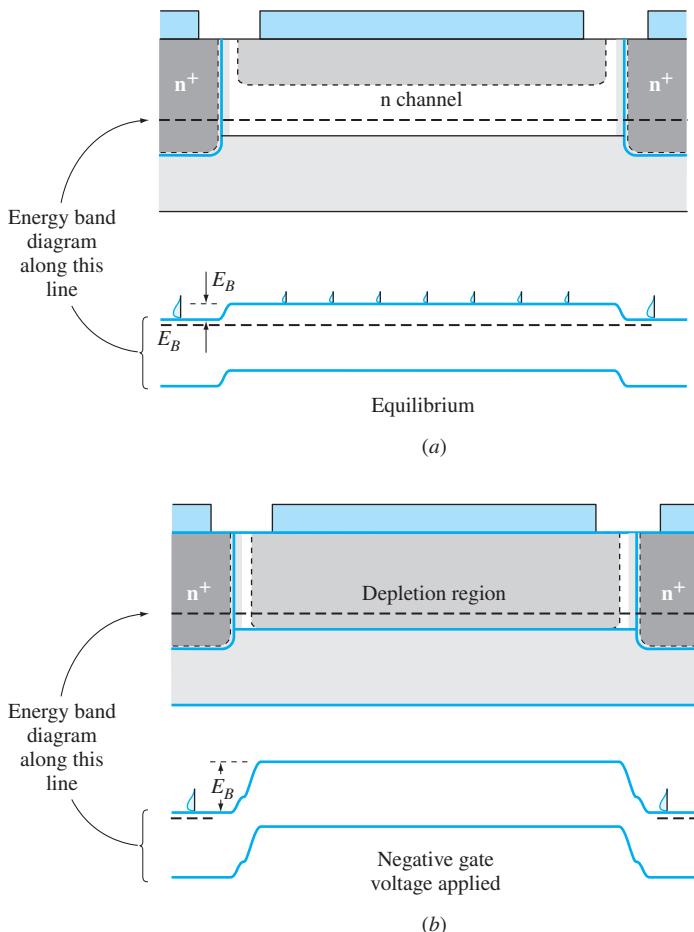


Figure 8.33 MESFET energy band diagram and depletion region (a) at equilibrium and (b) for an applied gate voltage that depletes the channel. In the second case, the channel is still n type, but it is empty of carriers because of the increased barrier height.

of gate voltage required to achieve this condition is called the *threshold voltage*, and in this case is negative.⁵

An enhancement-mode MESFET exists if the depletion regions at the source end in Figure 8.33a overlap with no gate voltage applied. In that case, the source-to-channel barrier is such that a negligible conducting channel exists at equilibrium. Application of a positive gate voltage then reduces the gate-channel depletion region and forms a channel, which enhances the current. In an enhancement-mode FET, the gate-channel Schottky barrier is actually forward

⁵This voltage is sometimes referred to as the pinch-off voltage.

biased to reduce the depletion region width. The gate voltage has to be limited then to about $\frac{3}{4}$ V to avoid excessive gate-channel current. In an enhancement-mode MESFET, the threshold voltage is positive.

The cross-sectional view of a MESFET with a positive drain voltage is shown schematically in Figure 8.34 along with the energy band diagrams perpendicular to the gate near the source and near the drain. At the drain end, the reverse bias across the Schottky barrier is greater than at the source, so the depletion region is wider. As a result, the channel thickness decreases from source to drain.

The I_D-V_{DS} characteristics of the MESFET are qualitatively similar to those of the MOSFET, although the physics is slightly different. Let us consider the case below saturation first.

Recall that the current through the channel is proportional to the charge per unit area, Q_{ch} , which in turn is proportional to the thickness t of the channel. Thus, as the drain voltage increases, the resistance also increases. The current therefore increases sublinearly with V_{DS} below saturation. This can be also be seen with the aid of Figure 8.35, which shows the variation of E_C with x and y for $0 < V_{DS} < V_{DSsat}$ (i.e., operating in the sublinear region) at a given gate voltage V_{GS} . In the channel, the electron concentration is $n_{ch} = N'_D$, or the net doping concentration in the n-type channel region. The gate voltage controls the thickness of the channel near the source, but this also controls the channel charge per unit area, since $Q_{ch} = -qN'_Dt$. Toward the drain end of the channel, the thickness

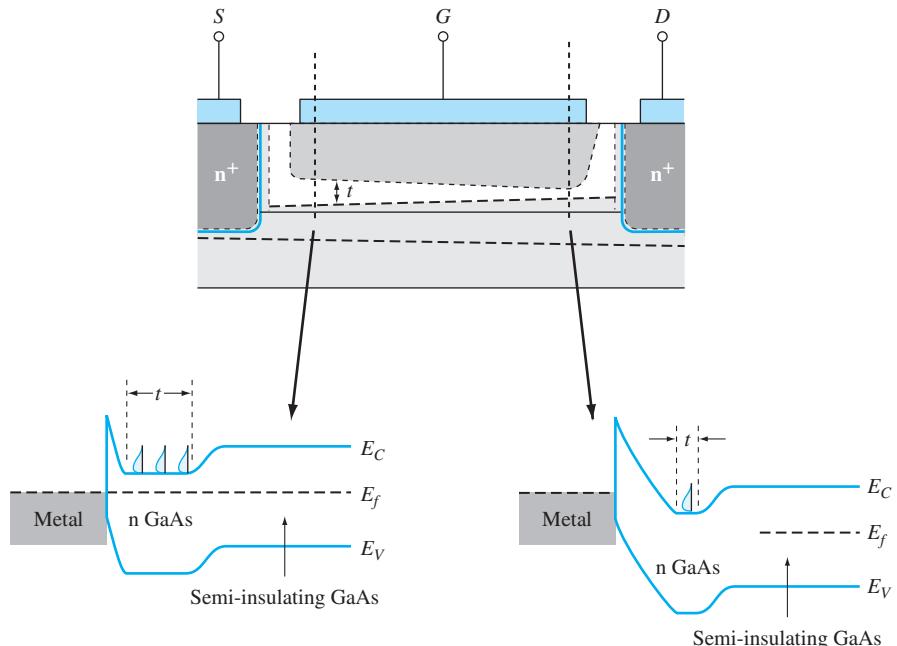


Figure 8.34 Cross section of a MESFET under small V_{DS} bias and the corresponding energy band diagrams at the source end and drain end of the gate.

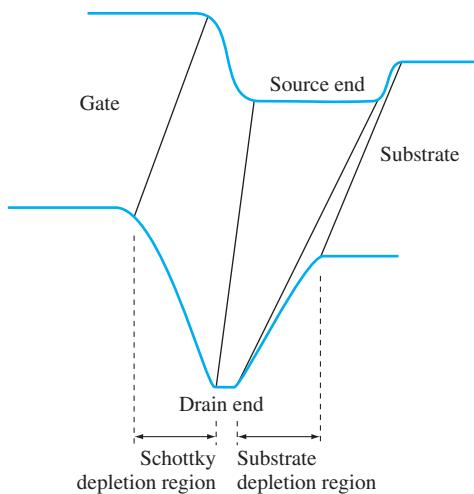


Figure 8.35 The electron potential energy, (E_C) along the channel of a MESFET for $0 < V_{DS} < V_{DSsat}$. The channel thickness decreases with increasing distance along the channel.

t decreases, with a corresponding decrease in Q_{ch} . Since the current along the channel is constant, from $I_D = W Q_{ch}v$ [Equation (III.3)] we can see that if the channel charge decreases, the velocity increases.

We expect the drain current to saturate at some value of V_{DS} . Let us examine how that occurs. With sufficiently high drain voltage, the depletion regions in the channel overlap near the drain as indicated in Figure 8.36, which shows the energy band diagram perpendicular to the junction at the source end and drain end under current saturation. It may seem that current should stop flowing because usually a depletion region has no carriers. In this case, however, although the depletion regions have met, the energy barriers of their walls will form a potential trough, similar to that in Figure 8.35, funneling current from the source into the drain. Thus, current continues to flow.

The current saturates, however, for the same reason as it does in a MOSFET. The number of carriers entering the channel is controlled by the barrier at the source end, as was shown in Figure 7.17, and the current is proportional to channel charge and channel field, $I_D = W Q_{ch}(0)\mu_{lf}\mathcal{E}_L(0)$. The application of a small drain voltage changes the field even at the source end, but as V_{DS} increases, it bends the band increasingly at the drain end but with little additional effect at the source end.

8.6.3 JUNCTION FIELD-EFFECT TRANSISTORS (JFETs)

A JFET is similar to a MESFET except that a p⁺n junction gate replaces the Schottky barrier gate, and the semi-insulating substrate is replaced by a p-type

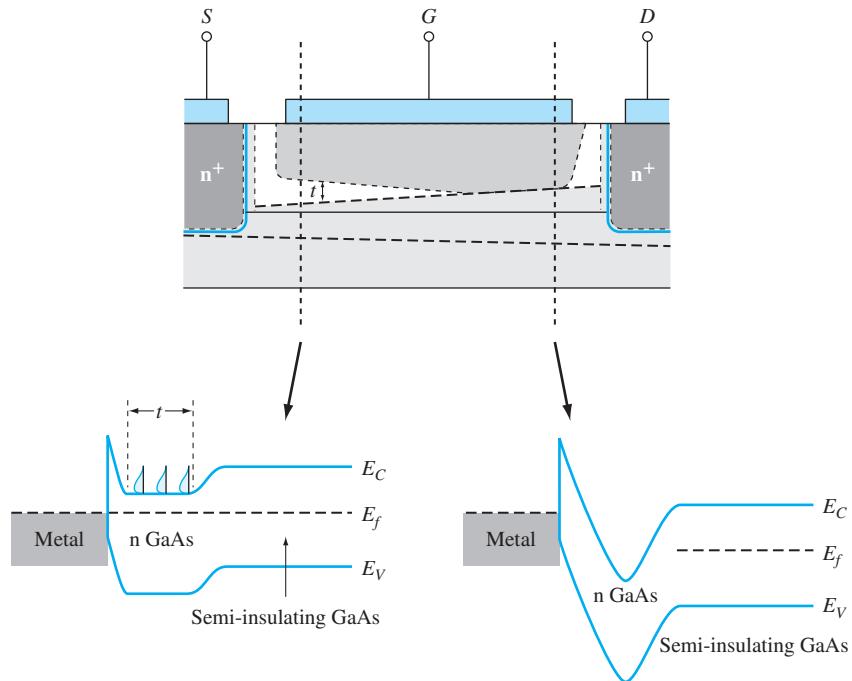


Figure 8.36 The MESFET of Figure 8.34 with $V_{DS} > V_{DSsat}$. At the source, the diagram is the same as for Figure 8.34. At the drain, however, the two depletion regions overlap.

substrate (for an NFET). This is shown schematically in Figure 8.37. As with a MESFET, the control field is applied via a depletion region associated with the reverse-biased gate junction. In the MESFET, that junction is a Schottky barrier; in the JFET, it is a p⁺n junction. The channel of thickness t extends from the edge of the gate channel depletion region to the substrate channel depletion region as indicated Figure 8.37a for zero drain-source and gate-source voltages. The corresponding energy band diagram is shown in Figure 8.37b. (Only E_C is shown here.) Figure 8.37c shows the depletion regions for a nonzero drain voltage that is still below saturation. The metallurgical channel thickness is a , the depletion width is w , and the effective channel thickness is t , where $t = (a - w)$.

The JFET physics of operation and electrical characteristics are similar to those for a MESFET. A reverse-biased gate-channel voltage reduces the current.

8.6.4 TUNNEL FIELD-EFFECT TRANSISTORS (TFETs)

In the FETs discussed, the minimum subthreshold swing S is limited to 60 mV per decade of current at room temperature, since the channel current results from carriers being thermally excited over a barrier between source and channel. This limitation can be overcome by carrier flow from source to channel

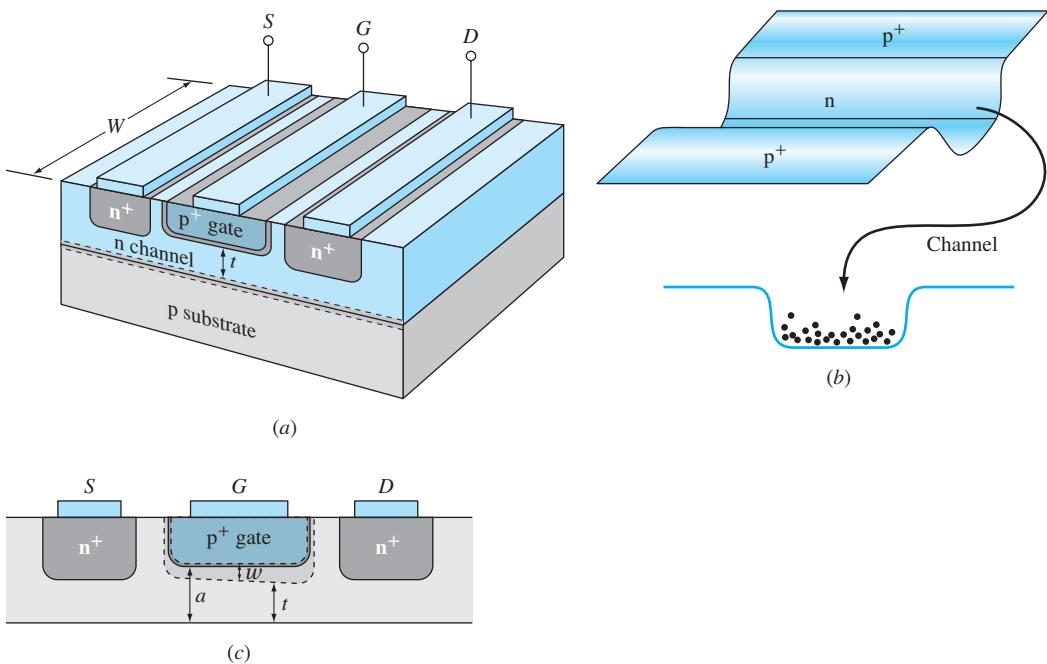


Figure 8.37 The JFET. The shaded areas represent the depletion regions. (a) Cross-sectional diagram; (b) the energy band diagram, (E_C), for $V_{DS} = 0$; (c) when $V_{DS} > 0$ (but not yet in saturation) the depletion region at the drain end increases, narrowing the channel and increasing the channel resistance.

by band-to-band tunneling in the tunnel field effect transistor or *TFET*, similar to a tunnel diode. Figure 8.38 shows an electron micrograph of a Si nanowire gate-all-around (GAA) TFET. The source and gate are p⁺ (boron), the drain is n⁺ (arsenic) and the nanowire channel is undoped. The gate-channel insulator is HfO₂. The thin TiN layer seen in the enlargement at the right is used as a

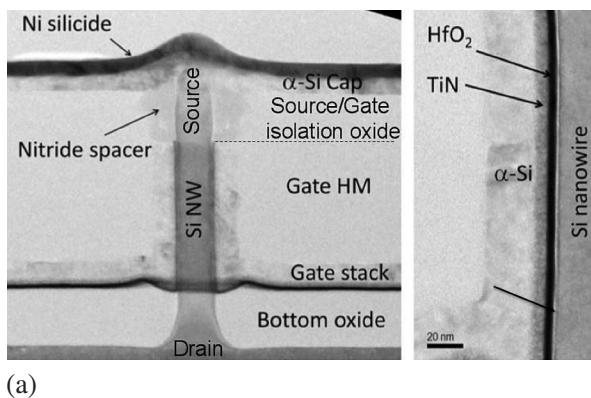


Figure 8.38 (a) Electron micrograph of a gate-all-around Si nanowire TFET. (Courtesy IMEC)

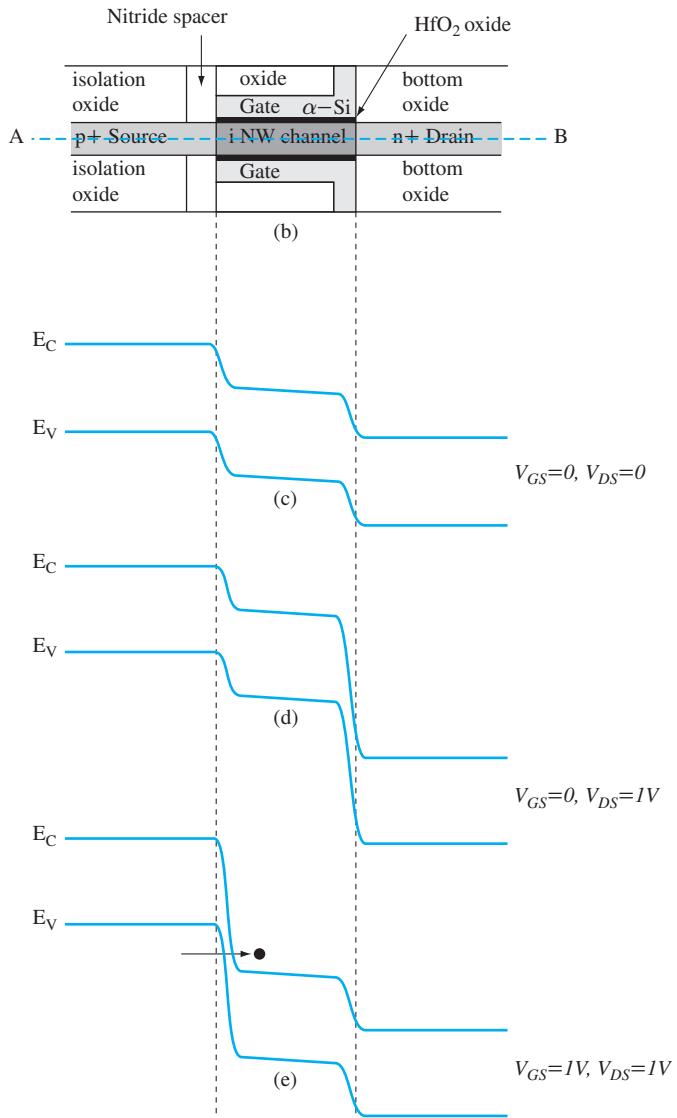


Figure 8.38 (cont.) (b) Schematic diagram; (c) Energy band diagram for $V_{GS} = 0$, $V_{DS} = 0$; (d) $V_{GS} = 0$, $V_{DS} = 1$ V and (e) $V_{GS} = 1$ V, $V_{DS} = 1$ V.

conductive barrier between the gate oxide and its degenerate silicon contact. The TiN film blocks boron diffusion from the Si gate contact into the gate dielectric, but is conductive enough ($\sigma \approx 50 \mu\Omega \cdot \text{cm}$) to provide a good electrical contact

between gate dielectric and Si contact. The α -Si regions are deposited amorphous Si that with subsequent processing become degenerate polycrystalline Si, and are used to connect both gate and source. (Similar geometry gate-all-around MOSFETs have been fabricated in which both source and drain are n^+ .) The “Gate HM” refers to “hard mask,” an oxide layer used as a mask during photolithography. Heterojunction TFETs with the source band gap smaller than that of channel results in a narrower tunneling region and thus an increased **on** current [6].

The energy band diagram from source to drain along the A-B line at equilibrium is shown in Figure 8.38b. The case for $V_{GS} = 0$ and a small V_{DS} (e.g., 1 V) is shown in (c). The device is effectively a reverse-biased pin diode, with the potential of the intrinsic channel being influenced by the gate. A small (negligible) current flows, resulting from electrons generated in the intrinsic channel and accelerated to the drain by the small field in the channel. If a positive gate voltage (e.g., 1 V) is applied to the gate as indicated in (d), electrons can tunnel through the thin transition region between the p^+ valence band to the intrinsic conduction band and thus travel to the n^+ drain. In direct-gap materials such as gallium arsenide or indium phosphide, the tunneling is between states of equal energy and wave vector. In indirect gap semiconductors such as silicon, the tunneling requires emission or absorption of a phonon with an appropriate wave vector to satisfy the laws of conservation of energy and wave vector. As a result, the tunneling probability at a given band overlap is greater for direct-gap materials than for indirect gap materials, resulting in a greater **on/off** current ratio.

The tunneling probability of a normally incident electron to the tunneling region was discussed in the Supplement to Part 1. It is similar to tunneling in tunnel diodes. The tunneling current is $I_{\text{Tunneling}} \propto e^{-\left(\frac{4E_g^{3/2}\sqrt{2m^*}}{3q\delta\hbar}\right)}$ [6], where \mathcal{E} and E_g are the electric field and the band gap respectively at the tunneling energy. It is seen that the probability is larger for materials with smaller effective mass and smaller band gap. Thus direct gap semiconductors with small effective mass and small band gap are desired. The band gap must be large enough, however, that thermal injection of minority carriers in the **off** state contributes to negligible current.

It is seen in Figure 8.38 that the field in the channel region is much smaller than that for a MOSFET (the bands are nearly flat along the channel). This suggests a small **on** current. The primary limitation to the current, however, is the low tunneling probability. As a result, TFETs are not expected to be high-speed devices (at least for indirect gap semiconductors, where tunneling between valence and conduction bands requires the interaction of phonons to satisfy the law of conservation of wave vector). Because of the different mechanisms of current flow compared to a MOSFET, a smaller subthreshold slope (S) than that of a MOSFET is possible [7, 8], making TFETs an attractive technology for low-voltage low-power circuits. Figure 8.38a shows the cross-section for a vertical silicon nanowire TFET; however TFETs have also been fabricated in planar structures and FinFET structures.

8.7 BULK CHANNEL FETs: QUANTITATIVE

We saw that the physics of operation of the MESFET and JFET are similar. Here we will be more quantitative. As with all FETs, the mathematical treatment of MESFETs and JFETs begins with Equation (III.7). For a JFET or a MESFET where $Q_{\text{ch}} = -qn(y)t(y)$,

$$I_D = qW\mu(y)n(y)t(y) \frac{dV_{\text{ch}}(y)}{dy} \quad (8.39)$$

In principle, then, provided we have models for μ , n , and t , Equation (8.39) can be solved for $V_{\text{ch}}(y)$. We can obtain the I_D - V_{DS} characteristics by setting $V_{\text{ch}}(L) = V_{DS}$. We will illustrate the solution for a JFET, but the MESFET solution is similar.

Simple Model for the JFET As we did with the MOSFET, we begin with a simple model for a JFET in which velocity saturation effects are ignored. We start with the I_D - V_{DS} characteristics below saturation ($0 < V_{DS} < V_{DS\text{sat}}$). For this long-channel model, we make the following approximations:

1. The net channel doping is uniform, and $n(y) \approx N'_D$.
2. The electron mobility is the low-field mobility ($\mu = \mu_{\text{lf}}$), and independent of \mathcal{E}_L .
3. The gate-channel depletion width $w(y)$ is entirely on the channel side. (This is reasonable since the p⁺ gate is heavily doped.)
4. The depletion width in the n-type channel region adjacent to the substrate is negligible. This implies that $t(y) = a - w(y)$, where a is the thickness of the n-type region. (See Figure 8.37c.)
5. The channel is long enough ($L \gg a$) such that $\mathcal{E}_L \ll \mathcal{E}_T$ over most of the channel. We also assume that the gate depletion region thickness w can be considered independent of $\mathcal{E}_L(y)$. This is referred to as the *gradual channel approximation*.

We should point out that approximation 4 is not realistic even for this simple model. The actual depletion region in the channel adjacent to the substrate depends on the net doping in the channel compared with that of the substrate, which must be known for an accurate analysis.

With these assumptions, Equation (8.39) becomes

$$I_D dy = qW\mu N'_D a \left[1 - \frac{w(y)}{a} \right] dV_{\text{ch}} \quad (8.40)$$

For a given N'_D , the depletion width depends only on the gate-to-channel voltage $V_G - V_{\text{ch}}(y)$. Adapting Equation (5.39), the expression for the depletion width in an n⁺p junction, we have

$$w(y) = \left[\frac{2\epsilon_s}{qN'_D} (V_{\text{bi}} - V_{GS} + V_{\text{ch}}(y)) \right]^{1/2} \quad (8.41)$$

where we have expressed the junction voltage as $V_j = V_{bi} - V_{GS} + V_{ch}(y)$. Here the applied voltage V_{GS} is negative to ensure that the junction is reverse biased.

It is convenient to express I_D in terms of its threshold voltage (a readily measurable quantity). At threshold, $V_{GS} = V_T$ and at the source end, $w(0) = a$ and $V_{ch} = 0$, so we can rewrite Equation (8.41) at $y = 0$ as

$$w(0) = a = \left[\frac{2\epsilon_s}{qN'_D} (V_{bi} - V_T) \right]^{1/2} \quad (8.42)$$

Solving for the threshold voltage yields:

$$V_T = V_{bi} - \frac{a^2 q N'_D}{2 \epsilon_s} \quad (8.43)$$

The threshold therefore depends on doping and the thickness of the n-type layer. Note that V_T is positive for an enhancement JFET and negative for a depletion JFET.

Substituting the expressions for $w(y)$ and a from Equations (8.41) and (8.29) into the brackets of Equation (8.40) gives

$$I_D dy = qW\mu N'_D a \left[1 - \left(\frac{V_{bi} - V_{GS} + V_{ch}}{V_{bi} - V_T} \right)^{1/2} \right] dV_{ch} \quad (8.44)$$

Now we invoke assumption 2, which is that $\mu = \mu_{lf}$ and is independent of V_{ch} . Integrating Equation (8.44) from source to drain, after some algebra we obtain, below saturation,

$$I_D = \frac{qW\mu_{lf}N'_D a}{L} \times \left\{ V_{DS} - \frac{2}{3}(V_{bi} - V_T) \left[\left(\frac{V_{DS} + V_{bi} - V_{GS}}{V_{bi} - V_T} \right)^{3/2} - \left(\frac{V_{bi} - V_{GS}}{V_{bi} - V_T} \right)^{3/2} \right] \right\} \quad (8.45)$$

$V_{DS} \leq V_{DS \text{ sat}}$

In current saturation, we know the slope $\partial I_D / \partial V_{DS} = 0$, the same as for a MOSFET. The onset of saturation occurs when $V_{DS} = V_{DS\text{sat}}$. Taking the derivative of Equation (8.45) and setting it to zero gives

$$V_{DS} = V_{DS \text{ sat}} = V_{GS} - V_T \quad (8.46)$$

Substituting this back into Equation (8.45), at (and above) the point of saturation we find

$$I_{D\text{sat}} = \frac{qW\mu_{lf}N'_D a}{L} \times \left\{ V_{GS} - V_T - \frac{2}{3}(V_{bi} - V_T) \left[1 - \left(\frac{V_{bi} - V_{GS}}{V_{bi} - V_T} \right)^{3/2} \right] \right\} \quad (8.47)$$

$V_{DS} \geq V_{DS \text{ sat}}$

Velocity Saturation Model for the JFET For the MOSFET, the simple long-channel model was poor for short-channel devices, and the same is true for the JFET. We now correct this JFET model to include velocity saturation. As for a MOSFET, the mobility μ decreases at high longitudinal field \mathcal{E}_L . Again using the relation

$$\mu = \frac{\mu_{\text{lf}}}{1 + \frac{\mu_{\text{lf}} \mathcal{E}_L}{v_{\text{sat}}}} \quad (8.48)$$

and using the relation $|\mathcal{E}_L| = dV_{\text{ch}}/dy$, we substitute into Equation (8.48). Integrating Equation (8.44), we obtain for $V_D \leq V_{D\text{sat}}$

$$I_D = \frac{qW\mu_{\text{lf}}N'_D a}{L\left(1 + \frac{\mu_{\text{lf}}V_{DS}}{v_{\text{sat}}L}\right)} \times \left\{ V_{DS} - \frac{2}{3}(V_{bi} - V_T) \left[\left(\frac{V_{DS} + V_{bi} - V_{GS}}{V_{bi} - V_T}\right)^{3/2} - \left(\frac{V_{bi} - V_{GS}}{V_{bi} - V_T}\right)^{3/2} \right] \right\} \quad (8.49)$$

$$V_{DS} \leq V_{DS\text{ sat}}$$

and in saturation:

$$I_{D\text{sat}} = \frac{qW\mu_{\text{lf}}N'_D a}{L\left(1 + \frac{\mu_{\text{lf}}V_{DS\text{sat}}}{v_{\text{sat}}L}\right)} \times \left\{ V_{DS\text{sat}} - \frac{2}{3}(V_{bi} - V_T) \left[\left(\frac{V_{DS\text{sat}} + V_{bi} - V_{GS}}{V_{bi} - V_T}\right)^{3/2} - \left(\frac{V_{bi} - V_{GS}}{V_{bi} - V_T}\right)^{3/2} \right] \right\}$$

$$V_{DS} \geq V_{DS\text{ sat}} \quad (8.50)$$

We observe that the preceding equations can be expressed as for a MOSFET:

$$I_D = \frac{I_D(\text{model neglecting velocity saturation})}{1 + \frac{\mu_{\text{lf}} V_{DS}}{L v_{\text{sat}}}} \quad (8.51)$$

$$I_{D\text{sat}} = \frac{I_{D\text{sat}}(\text{model neglecting velocity saturation})}{1 + \frac{\mu_{\text{lf}} V_{DS\text{sat}}}{L v_{\text{sat}}}} \quad (8.52)$$

As for a MOSFET, for short channels, the velocity saturation effect reduces the current from that predicted with velocity saturation neglected.

An expression for $V_{DS\text{sat}}$ can be found by setting $\partial I_D / \partial V_{DS} = 0$ as before, but in this case the resulting expressions are difficult to solve analytically.

I_D-V_{DS} Characteristics of MESFETs The I_D - V_{DS} characteristics for Si-based MESFETs are the same as those for JFETs. However, since many MESFETs are made from GaAs rather than silicon, we should point out that for n-channel GaAs-based FETs, Equation (8.48) is a poor approximation for electron mobility and the dependence of mobility on \mathcal{E}_L is appreciably more complicated than considered here. This is because of the peak in the velocity-field relation, as discussed in Chapter 3 (Figure 3.9).

8.8 SUMMARY

In Chapter 7 the basic operation of MOSFETs was discussed using Si-based planar n-channel devices as an example. In this chapter a procedure to experimentally determine the threshold voltage (V_T) and low-field mobility (μ_{lf}) was described. The discussion was expanded to include Si-based devices using both n-channel and p-channel MOSFETs with a CMOS digital inverter circuit as an example. The switching characteristics of such an inverter were investigated. The propagation (gate) delay during switching is

$$t_d = \left[\frac{C_L V_{DD}}{2I_{D\text{satn}}} + \frac{C_L V_{DD}}{2I_{D\text{satp}}} \right]$$

We also examined the power consumption of MOSFETs. Although we usually assume that $I_D \approx 0$ for $V_{GS} < V_T$, in reality some current does flow. We must choose V_T such that in the **off** state, when $V_{GS} = 0$, negligible current flows. This value of V_T fixes the supply voltage ($V_{DD} \approx 5V_T$), which in turn influences the dynamic power dissipation associated with switching. For a simple CMOS inverter,

$$P_{\text{dynamic}} = C_L V_{DD}^2 f$$

The devices discussed are volatile, in that they lose their state when power is removed. A nonvolatile MOSFET is one that remains in its current state with the removal of power. A floating gate nonvolatile MOSFET was briefly described.

Then variations of the Si-based MOSFET were examined, including silicon on insulator (SOI) devices and 3D FinFETs in which the MOSFETs are vertical extensions of the horizontal substrates.

Other FET structures were then examined. These include heterojunction field-effect transistors (HFETs), metal oxide field-effect transistors (MESFETs), and junction field-effect transistors (JFETs).

In the devices discussed so far, the current is limited by the carriers that can surmount a potential barrier in the channel, and the minimal subthreshold swing S achievable is

$$S = 2.3 \frac{kT}{q} \text{ mV per decade current}$$

or 60 mV per decade current at room temperature. To overcome this limitation, and thus permit the use of lower operating voltages resulting in less dissipated power, we discussed tunnel field-effect transistors (TFETs), using as an example a gated-channel p^+in^+ device.

8.9 REFERENCES

1. See for example Y. Taur, G. J. Hu, R. H. Dennard, L. M. Terman, C. Y. Ting, and K. E. Petrillo, “A self-aligned 1 μm channel CMOS technology with retrograde n well and thin epitaxy,” *IEEE Trans. Electron Devices*, ED-32, pp. 203–209, 1985.
2. Richard C. Jaeger, *Microelectronic Circuit Design*, Section 7.7, McGraw-Hill, New York, 1997.
3. M. Yoshimi, H. Hazama, M. Takahashi, S. Kambayashi, J. Wada, and H. Tango, “Two-dimensional simulations and measurement of high-performance MOSFETs made on very thin SOI films,” *IEEE Trans. Electron Devices*, ED-36, pp. 493–503, 1989.
4. Mark Lundstrom, *ECE Nanoscale Transistors* (Fall 2008), <https://nanohub.org/resources/5328>, Lecture 25.
5. A. Chini, R. Coffie, G. Meneghesso, E. Zanoni, D. Buttari, S. Heikman, S. Keller, and U. K. Mishra, “2.1 A/mm current density AlGaN/GaN HEMT,” *Electronics Letters*, 39, no. 7, pp. 625–629, 2003.
6. R. B. Fair, and H. W. Wivell, “Zener and avalanche breakdown in As-implanted low voltage Si n-p junctions,” *IEEE Trans. Electron Devices* ED-23, pp. 512–518, 2002.
7. R. Ghandi, Z. Chen, N. Singh, K. Bannerjee, and S. Lee, “CMOS compatible vertical silicon nanowire gate-all-around p-type tunneling FETs with ≤ 50 mV/decade subthreshold swing,” *IEEE Transactions on Electron Devices*, 58, no. 11, pp. 1504–1506, 2011.

8. Márcio D. V. Martino, Filepe S. Neves, Paula G. D. Agopelan, João A. Martino, Rita Rooyackers, and Cor Claeys, Nanowire tunnel field transistors at high temperatures, *Journal of Integrated Circuits and Systems*, 8, no. 2, pp. 2110–2115, 2013.

8.10 REVIEW QUESTIONS

1. Explain why the I_D - V_{GS} characteristic is not a straight line for large V_{GS} .
2. Why is it important to reduce the subthreshold leakage current? What must be traded off against this in the design process?
3. Explain the operation of the inverter circuit of Figure 8.5.
4. Why should NFETs and PFETs have different dimensions when used in CMOS circuits?
5. What is pass-through current?
6. Comparing Figure 8.12c and Figure 8.2, explain how a silicon-on-insulator design decreases the junction capacitances C_{JS} and C_{JD} below those of conventional CMOS circuits.
7. Considering the MOSFET, HFET, MESFET, and JFET, identify for each the region in which the controlling field occurs. For each, what is the mechanism used to control it (e.g., field across an oxide, changing the voltage across a depletion region, etc.)
8. What is the use of an asymmetrical double-gate SOI MOSFET with only one conducting channel?
9. For an ultra-thin body (UTB) symmetrical two-gate SOI MOSFET, why does the threshold voltage depend on body width?
10. Figure 8.22a shows ten FinFETs in parallel. Why not use a single FinFet?
11. What is the purpose of a nonvolatile MOSFET?
12. Why is the channel electron mobility of a GaN HFET degraded little from its bulk value, unlike that for a Si MOSFET?
13. Why is the minimum swing of a MOSFET 60 mV/decade current at room temperature?
14. Why can a TFET have a smaller swing than a MOSFET?

8.11 PROBLEMS

- 8.1 For the device whose I_D - V_{GS} characteristics are shown in Figure P8.1, find the value of μ_0 , V_T , and θ . The measurements were taken at

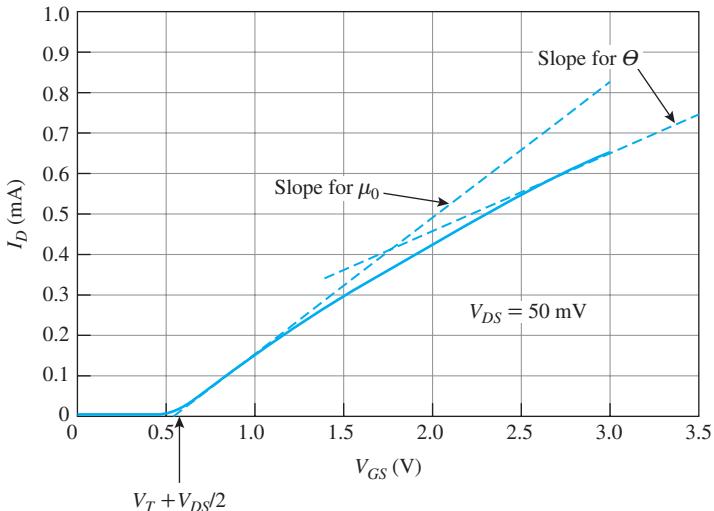


Figure P8.1

$V_{DS} = 50$ mV and $W = 10$ mm, $L = 0.5$ mm, and $t_{ox} = 5$ nm. The oxide is SiO_2 .

- 8.2** A particular MOSFET process produces $C'_B = 10^{-7}$ F/cm² and $I_0 = 4 \times 10^{-20}$ A, and a threshold voltage of $V_T = 0.5$ V. For gate oxide thicknesses of 2 nm and 4 nm, find n and S in the subthreshold region. Which device is better, and why?
- 8.3**
- a. Find W_p/W_n needed to match $I_{D\text{sat}}$ for CMOS transistors if $\mu_{l\text{fn}} = 250$ cm²/V · s, $\mu_{l\text{fp}} = 100$ cm²/V · s, $L = 0.5$ μm , and $V_G - V_T = 1$ V. Assume $v_{\text{sat}} = 10^7$ cm/s.
 - b. Find $V_{DS\text{sat}}$ for the NMOS and the PMOS.
 - c. Adjust the length of the NFET to equalize the $V_{DS\text{sat}}$'s. What should the new W_p/W_n be to keep the $I_{DS\text{sat}}$'s equal?
- 8.4** A CMOS inverter drives a load that consists of the gate of another FET. The gate area is 0.2×5 μm . Find the dynamic power dissipation if the clock frequency is 350 MHz, the supply voltage is $V_{DD} = 2.5$ V, and the oxide (SiO_2) thickness is 2 nm. If a medium-scale-integration (MSI) circuit has 5000 transistors, what is the power consumption due just to switching? (Neglect feedthrough current.)
- 8.5** A CMOS inverter has $W_p/W_n = 1.5$. The channel lengths are $L = 1$ μm , and $W_n = 10$ μm . Find the propagation delay time if the load capacitance is 1 pF. Let $t_{ox} = 4$ nm (SiO_2), $V_{DD} = 2.5$ V, and $V_{GS} - V_T = 2$ V.

- 8.6** SOI devices are more rad-hard (insensitive to radiation environments such as those in outer space or nuclear reactors) than conventional CMOS as discussed in the text. Examples of ionizing radiation are bombardment with high kinetic energy neutrons, electrons, protons, or gamma rays (which are photons with energies of a few MeV). The high-energy particles can disrupt the silicon crystal in conventional CMOS, while in SOI the amount of crystalline Si in the device is much smaller. Furthermore, charges that accumulate in the buried SiO_2 layer from the bombardment will generally not affect device operation. Explain why SOI devices will also be less affected by the absorption of gamma rays.
- 8.7**
- Draw the energy band diagram for the HFET, perpendicular to the gate, for the case when the transistor is strongly on. Indicate the polarity of the gate voltage.
 - Repeat for the depletion case (no channel). Now what is the polarity of the gate voltage?
 - Draw the energy band diagram along the channel for the equilibrium case and when the channel is conducting.
- 8.8** There are two primary reasons that GaAs-based HFETs are faster than silicon MOSFETs. One reason, given in the text, is that the electrons are traveling in a lightly doped channel. Explain how this increases their velocity. What do you think is the other reason?
- 8.9** The channel in the HFET is often referred to as supporting a two-dimensional electron gas. The electron gas part of this name refers to the sea of electrons that exists in the channel when the transistor is conducting. Why is it called two-dimensional?
- 8.10** Draw the energy band diagram at the source and the channel ends of the JFET under saturation (e.g., the same type of figure as Figure 8.36 for the MESFET). Point out the similarities and differences in operation of the two different types of devices.
- 8.11** An n-channel JFET has a channel doping of 10^{16} cm^{-3} . Let the thickness of the n layer be $1 \mu\text{m}$, the length of the channel be $4 \mu\text{m}$, and the width be $40 \mu\text{m}$. Neglect the depletion region adjacent to the substrate and neglect velocity saturation effects.
- What is the threshold voltage?
 - If the gate voltage is -2 V , at what drain voltage does the current saturate?
 - What is the saturation current under these conditions, using the simple long-channel model?
- 8.12** Figure 8.19 shows schematics of two FinFET versions. For similar geometries, which version will permit greater **on** current? Explain your answer.

- 8.13** Floating gate MOSFETs are more difficult to fabricate and are less reliable than conventional MOSFETs. Why bother with them?
- 8.14** Discuss the choice of band gap energy for the semiconductor used in a TFET.
- 8.15** Compare the **on** current density for direct gap and indirect gap semiconductors for use in TFETs.

Supplement to Part 3: Additional Consideration for MOSFETs

S3.1 INTRODUCTION

Chapters 7 and 8 were dedicated to describing the basic principles of FET operation. In this supplement, some of the more specialized topics associated with MOSFETs and MOS-based capacitance structures are briefly described.

S3.2 DEPENDENCE OF THE CHANNEL CHARGE Q_{ch} ON THE LONGITUDINAL FIELD \mathcal{E}_L

In the previous model (Chapter 7) for the I_D - V_{DS} characteristics of a MOSFET, we let the channel charge Q_{ch} approach zero when the device was in saturation. Recall that in that model, the carrier velocity v approached infinity such that the $Q_{ch}v$ product was finite and thus I_D remained at its constant value of I_{Dsat} . For high longitudinal fields, however, the carrier velocity saturates, meaning the channel charge Q_{ch} will reach some minimum value to keep the $Q_{ch}v$ product constant. In this section, the Q_{ch} dependence on v_{sat} is discussed.

Recall from Equation (III.3) that $I_D = -WQ_{ch}v$. Thus,

$$Q_{ch}(y) = -\frac{I_D}{Wv(y)} \quad (\text{S3.1})$$

where Q_{ch} and the carrier velocity v are evaluated at a given position y . At small longitudinal field strengths \mathcal{E}_L , such that $v \ll v_{sat}$, we have

$$Q_{ch} = Q_{chl} = -C'_{ox}[V_{GS} - V_T - V_{ch}(y)] \quad (\text{S3.2})$$

where Q_{chl} is the channel charge at small longitudinal fields and is independent of \mathcal{E}_L .

Since I_D is constant everywhere in the channel, the product $Q_{ch}v$ is constant. When the longitudinal field is large enough that the carrier velocity saturates, the channel charge reaches some minimum value Q_{chmin} :

$$Q_{chmin} = -\frac{I_D}{Wv_{sat}} \quad (S3.3)$$

We define a critical longitudinal field \mathcal{E}_{Lc} to be that field at which Q_{chlf} , given by Equation (S3.2), is equal to Q_{chmin} , given by Equation (S3.3). Then, for $\mathcal{E}_L \geq \mathcal{E}_{Lc}$

$$Q_{chlf} = Q_{chmin} \quad \mathcal{E}_L \geq \mathcal{E}_{Lc} \quad (S3.4)$$

Figure S3.1 shows a plot of Q_{ch} for this model (solid line) as a function of field strength \mathcal{E}_L for an n-channel MOSFET with $V_{GS} = 2$ V and assuming $v_{sat} = 5.5 \times 10^6$ cm/s. [1] This model assumes that the field dependence of the carrier velocity changes abruptly at \mathcal{E}_{Lc} from its low field value $v = \mu_{lf}\mathcal{E}_L$ to v_{sat} at the critical field \mathcal{E}_{Lc} . This underestimates Q_{ch} where \mathcal{E}_L is close to \mathcal{E}_{Lc} .

An empirical model for Q_{ch} has been proposed to eliminate this discrepancy [1]:

$$Q_{ch} = Q_{lf} + \frac{I_D}{Wv_{sat}} \left(1 - \frac{I_D}{WQ_{lf}\mu_{lf}\mathcal{E}_L} \right) \quad (S3.5)$$

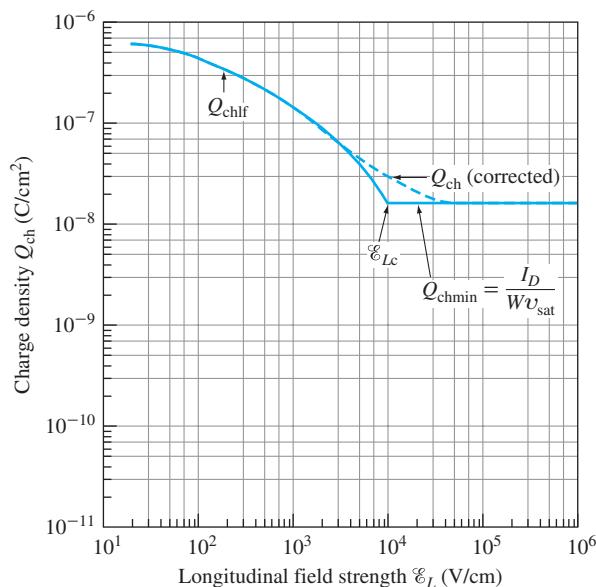


Figure S3.1 Model for the channel charge Q_{ch} (solid line) for $Q_{ch} = Q_{chlf}$ below the critical field \mathcal{E}_{Lc} and $Q_{ch} = Q_{chmin} = I_D/Wv_{sat}$ for $\mathcal{E}_L \geq \mathcal{E}_{Lc}$. Also shown (dashed line) is a correction to Q_{ch} in the vicinity of \mathcal{E}_{Lc} .

While this model is somewhat arbitrary, it does reduce to Q_{lf} at small \mathcal{E}_L as required, and at large \mathcal{E}_L where Q_{lf} is small, it reduces to Equation (S3.3). The result from Equation (S3.5) is shown by the dashed line in Figure S3.1.

S3.3 THRESHOLD VOLTAGE FOR MOSFETS

We consider the threshold voltage V_T to be the gate voltage required to just form a good conducting channel. While this parameter can be determined experimentally once the device is made, it is important to know its dependence on the various material properties of the device so that its value can be controlled during fabrication.

The threshold voltage is commonly defined as the gate voltage required to produce a surface potential of $\phi_s = 2\phi_f$ at the source end of the channel. It is the gate voltage that produces a carrier concentration at the surface at the source end of the channel equal to, but opposite in sign to, that in the substrate. This surface potential is dependent on the work function difference between the gate and substrate, as well as the oxide thickness. In addition to these previously discussed causes of the surface potential, there can be charges in the material below the gate, as shown in Figure S3.2.

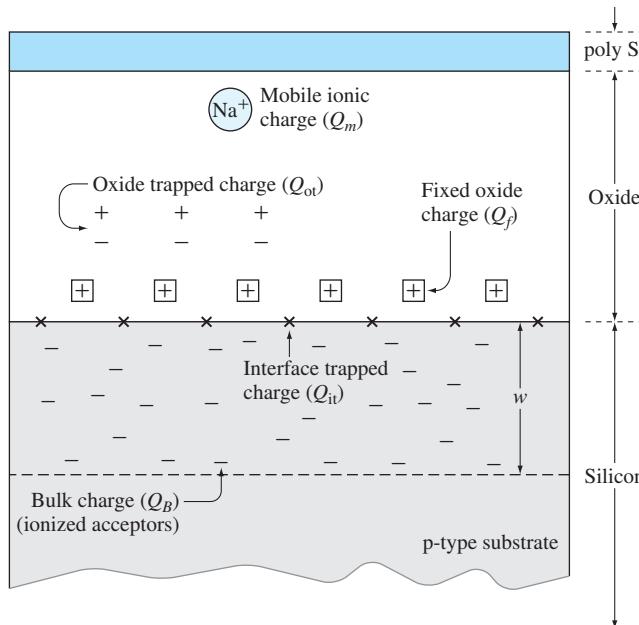


Figure S3.2 The charges in a MOSFET affect the threshold voltage. These include the trapped charge Q_{ot} distributed through the oxide, fixed charge Q_f in the oxide near the Si/oxide interface, bulk charge Q_B in the substrate depletion region (w), mobile ionic charge Q_m in the oxide, and charge Q_{it} trapped in the interface states at the Si/oxide interface. Not to scale.

The figure shows an n-channel enhancement MOSFET; the substrate is p type. There is charge at the oxide-substrate interface, along with the usual bulk (depletion) charge in the substrate adjacent to the channel. (Note that this is an enhancement mode device, so at equilibrium there is no conducting channel, and thus no mobile charge carriers are shown in the channel.)

The concentrations of these charges, as well as their locations, influence the threshold voltage. These charges need to be controlled during the fabrication process, or, if they cannot be controlled, they should be accounted for and compensated for in the design of the device.

The various types of charges (coulombs per unit area) that influence the value of the threshold voltage are:

Q_{ot} , oxide trapped charges, which are fixed in position and distributed throughout the oxide.

Q_f , fixed oxide charge, the charge associated with the positive Si^+ ions in the oxide near (within about 2 nm) the oxide/Si interface. These also are fixed in position.

Q_m , mobile ion charge in the oxide (e.g., Na^+ ions), which are quite mobile in the oxide and can move under the influence of the electric field in the oxide.

Q_{it} , interface trapped charge, the charge associated with interface states at the oxide/Si interface (these states are due to crystal defects at the interface, e.g., dangling bonds). This charge is dependent on the surface potential, which is dependent on the gate voltage.

Q_B , depletion region or bulk charge, the charge associated with the ionized acceptors and donors in the Si depletion region at the source end of the channel.

In addition, if a channel exists:

Q_{ch} , the mobile electron charge that is responsible for channel current (i.e., the electrons in the channel)

The charges Q_{it} and Q_B depend on the surface potential, and for $V_{DS} \neq 0$ they depend on position y in the channel since the surface potential is dependent on y . Since the threshold voltage depends on their values at the source end of the channel, however, they are evaluated at $y = 0$.

Of these, the concentrations of oxide trapped charge (Q_{ot}) and the oxide mobile ion charge (Q_m) have been problems in the past for SiO_2 , but modern processing methods have reduced them in SiO_2 to levels small enough that they have a small effect on the device characteristics. As a result, we will not consider them further. [With the trend toward using oxides with higher dielectric constants than SiO_2 (e.g., HfO_2), these charges can be a problem.] The other charges remain important and we will consider their effects next.

S3.3.1 FIXED CHARGE

The fixed oxide charge Q_f in the oxide near the interface is a natural result of the thermal oxidation process used to produce the oxide insulating layer. During thermal oxidation, oxygen diffuses through the existing oxide layer and reacts

with the Si in the substrate, thus increasing the oxide thickness (and eating into the substrate). This oxidation process is not uniform, however, and there exists a region about 2 nm thick between the Si and the oxide that contains some partially oxidized Si, which is positively charged. As the oxidizing process continues, these Si ions become oxidized (bond with oxygen), but some positive Si ions are incorporated into the newly formed oxide. With current processing methods the concentration of these Si ions can be reduced to the low level below about 10^{10} cm^{-2} .

S3.3.2 INTERFACE TRAPPED CHARGE

The interface trapped charge Q_{it} results from dangling bonds at the Si/oxide interface. While the bonding of the oxide to the Si surface reduces the number of dangling bonds and thus the number of interface states, some danglers remain. One method for reducing the number of remaining interface states is to introduce some hydrogen into the silicon dioxide. Since hydrogen, being a small atom, is relatively mobile in oxide, at elevated temperatures during processing some of the H atoms diffuse to the oxide/Si interface, where they attach to the dangling bonds. The dangling bonds are said to be *passivated*, meaning rendered inert. However, since hydrogen can diffuse in the oxide, even at room temperature, the degree of passivation can change with time, with a resultant change in the electrical characteristics of the transistor. Alternatively, deuterium, an isotope of hydrogen, can be used instead of hydrogen for passivation. Deuterium, being larger than hydrogen, has a smaller diffusion coefficient and results in a more stable device. [2] The addition of nitrogen to SiO_2 (SiON) helps to passivate the interface. Since nitrogen has one fewer electrons in its outer shell than oxygen, it can bond with the Si interface dangling bonds. The addition of nitrogen to SiO_2 also reduces the ionic diffusion in the oxide.

The surface potential ϕ_s , and therefore the threshold voltage, depends on the charge in these interface states. Figure S3.3 shows the energy band diagram at equilibrium. Note that the interface trapped charge Q_{it} is the charge in *interface* states. In principle Q_{it} can be either positive or negative. For Si MOSFETs, however, the net Q_{it} is normally positive at equilibrium.

S3.3.3 BULK CHARGE

The threshold voltage also depends on the bulk charge Q_B , which consists of fixed ionized acceptors (in p-channel devices they would be ionized donors) in the depletion region at the source end of the channel. The bulk charge is a function of the net doping level in the substrate and also depends on the width $w(y)$ of the depletion region.

$$Q_B(y) = -qN'_A w(y) \quad (\text{S3.6})$$

where Q_B is the bulk charge per unit area, and

$$w(y) = \left[\frac{2\epsilon_s}{qN'_A} \phi_s(y) \right]^{1/2} \quad (\text{S3.7})$$

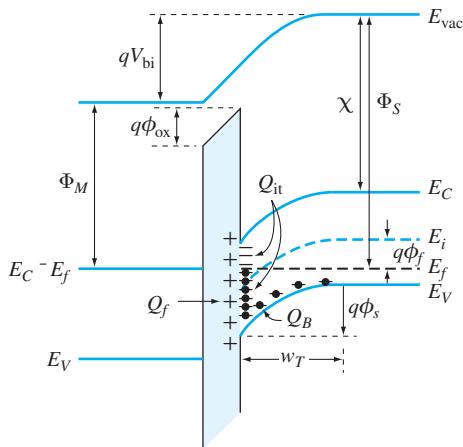


Figure S3.3 Equilibrium energy band diagram normal to the gate. The fixed charge Q_f , the bulk charge Q_B , and the interface trapped charge Q_{it} are indicated. The interface states are largely filled below the Fermi level and empty above it.

Since the threshold voltage depends on the source-channel barrier (at $y = 0$), and since at threshold $\phi_s(0) = 2\phi_f$, then at threshold at the source end of the channel

$$Q_{B(\text{threshold})}(y=0) = -[2qN'_A\epsilon_s(2\phi_f)]^{1/2} \quad (\text{S3.8})$$

S3.3.4 EFFECT OF CHARGES ON THE THRESHOLD VOLTAGE

Consider a MOSFET with source, substrate (body), drain, and gate connected together (equilibrium). Since we are using the source as a reference, $V_{GS} = V_{DS} = V_{BS} = 0$, where V_{BS} is the voltage between substrate (body) and source. The built-in potential energy qV_{bi} between gate and substrate is the difference between the work functions Φ_M for the gate and Φ_S for the semiconductor, or $V_{bi} = |\Phi_{MS}|$ where $\Phi_{MS} = (\Phi_M - \Phi_S)$ (see Figure S3.3).¹ The built-in potential is dropped partly across the oxide and partly across the substrate as indicated in Figure S3.3. If now a voltage $V_{GB} = V_T$ is applied to the gate, the electrostatic potential between gate and substrate is at threshold

$$V_{GB} + V_{bi} = V_T - \frac{\Phi_{MS}}{q} = \phi_{ox}(V_{GB}) + \phi_s(V_{GB}) \quad (\text{S3.9})$$

¹The subscript M is for “metal,” although in modern FETs, the gate is often degenerately doped polysilicon.

where $\phi_{ox}(V_{GB})$ and $\phi_s(V_{GB})$ represent respectively the electrostatic potential across the oxide and the surface potential for a gate-substrate voltage V_{GB} . The threshold voltage is then

$$V_T = \frac{\Phi_{MS}}{q} + \phi_{ox}(V_T) + \phi_s(V_T) \quad (\text{S3.10})$$

At threshold, $\phi_s = \phi_s(V_T) = 2\phi_f$ and

$$\phi_{ox}(V_T) = -\left[\frac{Q_f + Q_{it}(2\phi_f) + Q_B(2\phi_f)}{C'_{ox}} \right] \quad (\text{S3.11})$$

where Q_{it} and Q_B are evaluated at $\phi_s = 2\phi_f$.

The threshold voltage can then be expressed as

$$V_T = \frac{\Phi_{MS}}{q} + 2\phi_f - \frac{Q_f}{C'_{ox}} - \frac{Q_{it}(2\phi_f)}{C'_{ox}} - \frac{Q_B(2\phi_f)}{C'_{ox}} \quad (\text{S3.12})$$

S3.3.5 FLAT BAND VOLTAGE

As the gate voltage is varied, the band bending in Figure S3.3 also varies. It is possible for the applied voltage to cause the energy bands in the semiconductor to bend up instead of down. Figure S3.4a shows an enhancement-type NMOS at equilibrium. In this case, applying a negative gate voltage will reduce the band bending. At some value of applied gate voltage the bands in the Si are flat, as shown in Figure S3.4b. The value of V_{GS} required for the substrate band edges to be flat, or $\phi_s = 0$, is referred to as the flat band voltage V_{FB} :

$$V_{FB} = \frac{\Phi_{MS}}{q} + \phi_{ox}(0) = \frac{\Phi_{MS}}{q} - \frac{(Q_f + Q_{it}(0))}{C'_{ox}} \quad (\text{S3.13})$$

where $Q_{it}(0)$ and $\phi_{ox}(0)$ are evaluated at the flat band condition ($\phi_s = 0$). Figure S3.4c illustrates that a more negative gate voltage can reverse the direction of band bending.

The energy band diagram at threshold is shown in Figure S3.4d. Here the applied gate voltage is positive. Note that the amount of charge trapped at the interface is different from the flat band condition, or $Q_{it}(0) \neq Q_{it}(2\phi_f)$.

The threshold voltage can be expressed in terms of the flat band voltage,

$$V_T = V_{FB} + 2\phi_f - \frac{Q_B(2\phi_f)}{C'_{ox}} - \frac{Q_{it}(2\phi_f) - Q_{it}(0)}{C'_{ox}} \quad (\text{S3.14})$$

It is often assumed that

$$Q_{it}(2\phi_f) \approx Q_{it}(0) \quad (\text{S3.15})$$

and the last term in Equation (S3.14) is neglected.

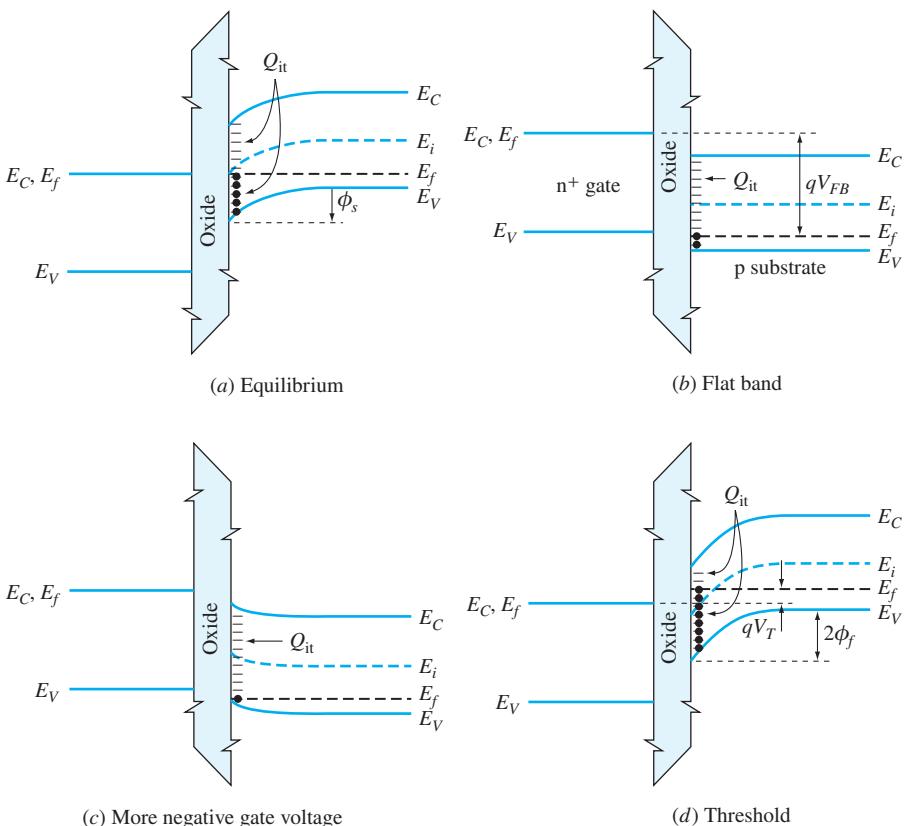


Figure S3.4 Occupation of surface states (traps). (a) Equilibrium. (b) Flat band condition. (c) Under increasingly negative gate voltage, the bands can actually bend up. (d) Threshold. For donor traps, those below the Fermi level are occupied and thus neutral. Traps above the Fermi level are vacant and thus positive. It can be seen that Q_{it} is greater (more positive) at flat band than at threshold.

EXAMPLE S3.1

Find the threshold voltage for an n-channel MOSFET with an n⁺ poly Si gate, with $N'_A = 10^{16} \text{ cm}^{-3}$ in the substrate, and $Q_f = Q_{it}(2\phi_f) = q \times 10^{10} \text{ C/cm}^2$, an oxide (SiON with dielectric constant = 5) thickness of 2 nm, and a poly Si gate doped to $2 \times 10^{19} \text{ cm}^{-3}$.

Solution

We evaluate the contribution of each term in Equation (S3.12). First, to find Φ_{MS}/q , we use

$$\Phi_{MS} = \Phi_M - \Phi_S$$

The polysilicon is degenerately doped. From Figure 2.25 we see that for $N'_D = 2 \times 10^{19} \text{ cm}^{-3}$, the apparent band gap narrowing is 0.07 eV. Assume that the Fermi level coincides with the (reduced) conduction band edge or E_f is 0.07 eV below the unperturbed conduction band edge (at the apparent conduction band edge). Since at equilibrium $E_{fM} = E_{fS}$,

$$\Phi_M = \chi_s + 0.07 \text{ eV}$$

and

$$\Phi_S = \chi_s + E_G - (E_f - E_V)$$

Then

$$\Phi_{MS} = 0.07 - [E_G - (E_f - E_V)]$$

To locate the Fermi level ($E_f - E_V$) in the substrate under the gate, we use the relation (since the p region is not degenerately doped)

$$p_0 = N'_A = N_V e^{-(E_f - E_V)/kT}$$

or

$$E_f - E_V = kT \ln \frac{N_V}{N'_A} = 0.026 \ln \frac{3.1 \times 10^{19}}{10^{16}} = 0.21 \text{ eV}$$

Since the band gap $E_g = 1.12 \text{ eV}$, we combine these values and get

$$\Phi_{MS} = 0.07 - 1.12 + 0.21 = -0.84 \text{ eV}$$

The next term in Equation (S3.12) is found from the relation

$$\begin{aligned} p_0 &= N'_A = n_i e^{q\phi_f/kT} \\ \phi_f &= \frac{kT}{q} \ln \frac{N'_A}{n_i} = 0.026 \left(\ln \frac{10^{16}}{1.08 \times 10^{10}} \right) = 0.36 \text{ V} \\ 2\phi_f &= 0.72 \text{ V} \end{aligned}$$

We will need the oxide capacitance per unit area for the last three terms:

$$C'_\text{ox} = \frac{\epsilon_\text{ox}}{t_0} = \frac{5 \times 8.85 \times 10^{-14} \text{ F/cm}}{2 \times 10^{-7} \text{ cm}} = 2.21 \times 10^{-6} \text{ F/cm}^2$$

Now it remains to find the charges. It is given that

$$Q_f = Q_{it} = q \times 10^{10} \text{ C/cm}^2 = (1.6 \times 10^{-19} \text{ C})(10^{10} \text{ charges/cm}^2)$$

from which we find the quantity: Q_f

$$\frac{Q_f}{C'_\text{ox}} = \frac{Q_{it}(2\phi_f)}{C'_\text{ox}} = 0.00072 \text{ V}$$

For the bulk charge term:

$$\begin{aligned} Q_B &= -\sqrt{2\epsilon_s q N'_A (2\phi_f)} \\ &= -\sqrt{2 \cdot (11.8 \times 8.85 \times 10^{-14})(1.6 \times 10^{-19})(10^{16})(2)(0.36)} \\ &= -4.9 \times 10^{-8} \text{ C/cm}^2 \\ \frac{Q_B}{C'_\text{ox}} &= -0.022 \text{ V} \end{aligned}$$

Substituting all of these into Equation (S3.12), we find the value of the threshold voltage to be, for this n-channel device,

$$V_T = -0.84 + 0.72 - 0.00072 - 0.00072 + 0.022 = -0.10 \text{ V}$$

We can see from Example S3.1 example that the contribution of the interface charge is small and Equation (S3.15) is a reasonable approximation.

In the preceding calculation, we considered the impurity-induced band-gap narrowing for the degenerate gate. If we had neglected this effect and taken the Fermi level to coincide with the unperturbed conduction band edge (as is often done for degenerate semiconductors), the calculated value of V_T would be 0.07 eV less than calculated above; i.e., $V_T \approx -0.17$ V.

EXAMPLE S3.2

Find the flat band voltage associated with the device of Example S3.1. Assume $Q_{it}(2\phi_f) = Q_{it}(0)$.

■ **Solution**

From Equation (S3.13),

$$V_{FB} = \frac{\Phi_{MS}}{q} - \frac{(Q_f + Q_{it})}{C'_\text{ox}} = -0.84 - \frac{1.6 \times 10^{-19}(10^{10} + 10^{10})}{2.21 \times 10^{-6}} = -0.84 \text{ V}$$

S3.3.6 THRESHOLD VOLTAGE CONTROL

In Example S3.1, a threshold voltage of $V_T = -0.10$ V was obtained for the n-channel MOSFET. This makes it a depletion mode device—a voltage more positive than this (including $V_G = 0$) results in a conducting channel. In digital circuits, however, enhancement MOSFETs are normally preferred, such that a voltage is required to turn them on rather than to turn them off. A typical design rule is to aim for threshold voltages of about 20 percent of the supply voltage.² Thus, for a device expected to operate with a supply voltage of +2.5 V, the threshold voltage should be $V_T \approx +0.5$ V.

This threshold voltage adjustment can be accomplished by ion implanting a p-type impurity (e.g., boron) into the Si substrate in the channel region. Figure S3.5 shows how this implant is done. In ion implantation, dopant atoms are ionized and accelerated electrostatically toward the sample, where they strike with sufficient force that they become implanted. The implant is done through a thin sacrificial oxide layer to prevent the ions from going too deep. After the implant, the oxide is removed, and the gate oxide is grown in a later processing step.

If the implant is very shallow, the approximation can be made that all of the impurities are at the oxide/Si interface. Since these acceptors are ionized, they contribute an extra charge layer (in this case negative charge), assumed to be at the interface. This localization of the impurities then adjusts the threshold voltage by the value

$$\Delta V_T = -\frac{Q_{ii}}{C'_\text{ox}} = +\frac{qN_{ii}}{C'_\text{ox}} \quad (\text{S3.16})$$

²For high-performance logic devices, however, $V_T \approx 0.1 V_{DD}$ and for circuits (e.g., memories), in which low standby power is desired, $V_T \approx 0.5 V_{DD}$.

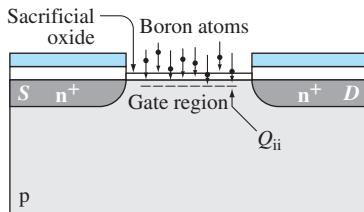


Figure S3.5 Ion implantation of (in this case) boron atoms leaves excess negative charges near the silicon surface. The gate oxide is grown later. The excess charges are used to adjust the threshold voltage.

where Q_{ii} is the implanted charge per unit area and N_{ii} is the number of implanted boron atoms per unit area. In this case, the change in the threshold voltage is positive since the added charges are negative. The threshold voltage can also be controlled by nonuniform doping in the substrate adjacent and normal to the channel.

EXAMPLE S3.3

To change the NMOSFET of the example in the previous section to an enhancement device following the design guidelines, we would want to correct the fabrication process to raise the threshold voltage by 0.60 V.

■ **Solution**

Using the ion-implantation technique just described, we would have to implant:

$$N_{ii} = \frac{22.1 \times 10^{-7} \times 0.60}{1.6 \times 10^{-19}} = 8.3 \times 10^{12} \text{ boron atoms/cm}^2$$

Ion implantation is a technique for adjusting the threshold voltage during device fabrication. There is a way, however, to adjust the threshold voltage of a MOSFET after it has already been made—that is, to change the bias on the substrate with respect to the source. The voltage applied to the substrate changes the width of the depletion region between the channel and the bulk substrate below it, and thus changes the bulk charge Q_B at the source end of the channel to the value

$$Q_B = -[2\epsilon_s q N'_A (2\phi_s - V_{BS})]^{1/2} \quad (\text{S3.17})$$

where V_{BS} is the substrate-to-source voltage.

If we use this adjusted value of the bulk charge in Equation (S3.12), we have an adjustment to the threshold voltage of

$$\Delta V_T = \frac{\sqrt{2\epsilon_s q N'_A}}{C'_{ox}} [(2\phi_f - V_{BS})^{1/2} - (2\phi_f)^{1/2}] \quad (\text{S3.18})$$

Let

$$\gamma = \frac{\sqrt{2\epsilon_s q N'_A}}{C'_{\text{ox}}} = \frac{t_{\text{ox}}}{\epsilon_{\text{ox}}} \sqrt{2\epsilon_s q N'_A} \quad (\text{S3.19})$$

where γ is referred to as the *body effect coefficient*. Equation (S3.18) can then be expressed as

$$\Delta V_T = \gamma [(2\phi_f - V_{BS})^{1/2} - (2\phi_f)^{1/2}] \quad (\text{S3.20})$$

An alternate expression is

$$\Delta V_T = \gamma \sqrt{2\phi_f} \left[\left(1 - \frac{V_{BS}}{2\phi_f} \right)^{1/2} - 1 \right] \quad (\text{S3.21})$$

EXAMPLE S3.4

Consider an n-channel MOSFET with $N'_A = 10^{18} \text{ cm}^{-3}$, and a SiON gate oxide with $\epsilon_r = 5$ and $t_{\text{ox}} = 2 \text{ nm}$. Find the change in threshold voltage for $V_{DS} = -1 \text{ V}$.

Solution

From the relation $\phi_f = \frac{kT}{q} \ln \frac{N'_A}{n_i}$, $\phi_f = 0.026 \ln \frac{10^{18}}{1.05 \times 10^{10}} = 0.48 \text{ V}$

Then from Equation S3.19,

$$\begin{aligned} \gamma &= \frac{2 \times 10^{-7} \text{ cm}}{5(8.854 \times 10^{-14} \text{ F/cm})} \sqrt{2(11.8 \times 8.854 \times 10^{-14} \text{ F/cm})(1.6 \times 10^{-19} \text{ C})(10^{18} \text{ cm}^{-3})} \\ &= 0.26 \text{ V}^{1/2} \end{aligned}$$

From Equation (S3.21),

$$\Delta V_T = 0.26 \text{ V}^{1/2} \times \sqrt{2 \times 0.48 \text{ V}} \left[\left(1 + \frac{1 \text{ V}}{2 \times 0.48 \text{ V}} \right)^{1/2} - 1 \right] = 0.11 \text{ V}$$

S3.3.7 CHANNEL QUANTUM EFFECTS

Up to now, we have considered the channel in a MOSFET classically. That is, the density of states function $S(E)$ has the value:

$$S(E) = \frac{1}{2\pi^2} \left(\frac{2m_{ds}^*}{\hbar^2} \right)^{3/2} \sqrt{E - E_C}$$

This is a continuous function, and an electron could be at any energy in the band.

In the channel of a MOSFET, however, the carriers are confined in the narrow potential well at the Si/oxide interface. At high electric fields ($\mathcal{E}_T > 10^5 \text{ V/cm}$), the width of the well is small enough that the carriers must be treated quantum mechanically. You will recall from the one-dimensional potential well problem that the electron is restricted to a finite number of discrete states—not every energy is allowed. This is shown in Figure S3.6a on the left. There the potential energy is constant in the well. In a MOSFET channel, however, shown on the right, the

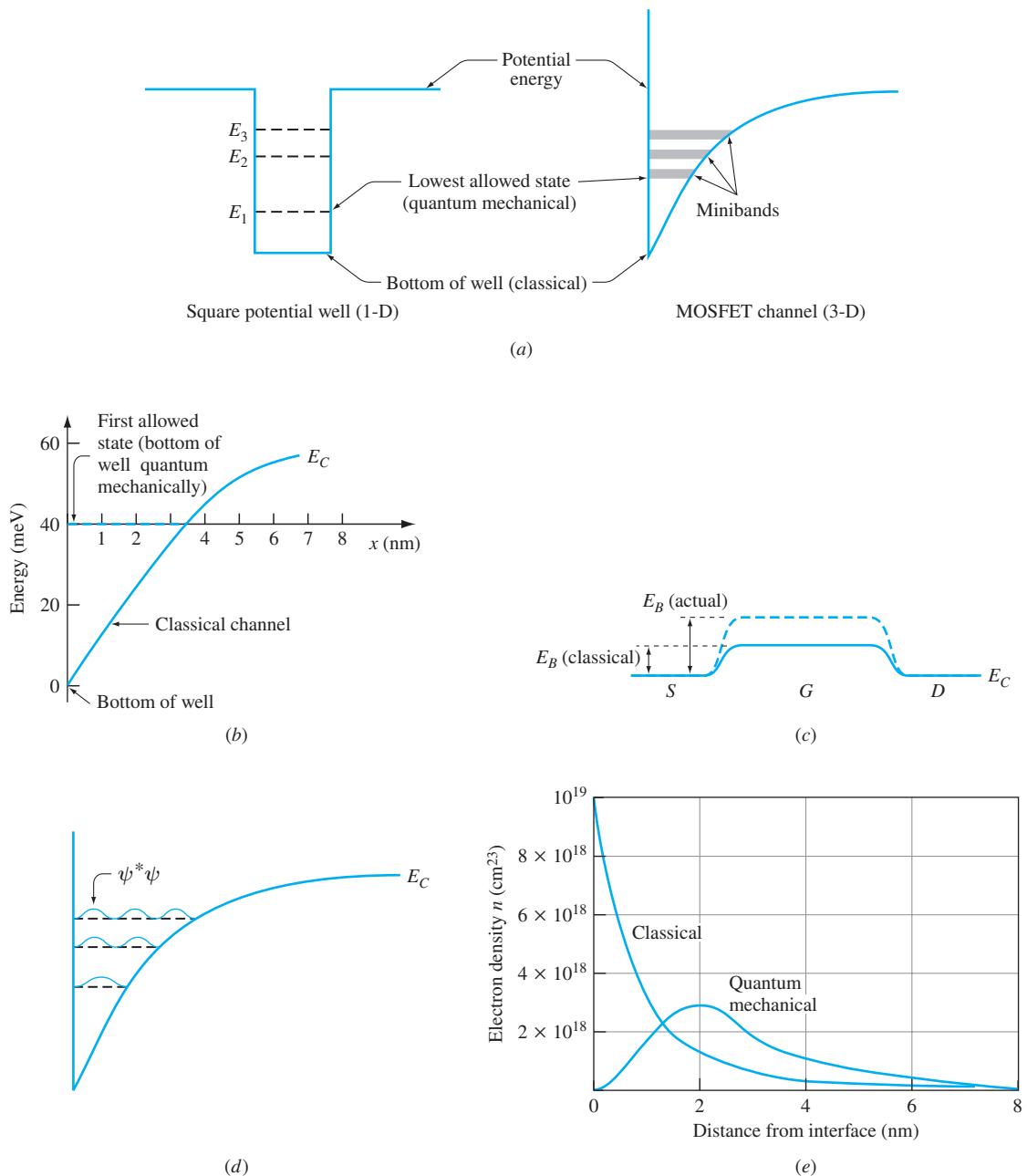


Figure S3.6 The channel in a MOSFET can be narrow enough to behave like a potential well. (a) Comparison of square well to MOSFET channel well; (b) comparison of classical channel depth to quantum-mechanical lowest energy level; (c) the minimum state above the classical bottom of the well results in a higher potential barrier for electrons entering the well; (d) the wave functions indicate that the charge is concentrated somewhere in the interior of the well rather than at the interface; (e) electron concentration normal to the gate for the classical case and the quantum case. In both cases, $Q_{ch}/q = 10^{12} \text{ cm}^{-3}$. Adapted from References 3 and 4.

potential varies with position in the well. The principles are the same, in that there are discrete states and the number and actual energies vary with well depth and width, but the quantitative results are somewhat different since the well shape is different. The potential well in a MOSFET, however, is three-dimensional and is narrow only in the x direction. For this case each discrete state in the figure is actually the lowest energy of a “miniband.” The bottom of the lowest energy band is somewhat above the classical band minimum at the interface. [3]

The existence of the quantization of states perpendicular to the gate has two effects. We consider an n-channel MOSFET.

1. Because the lowest energy states are above the classical bottom of the well at the surface, it requires a larger gate voltage to bend the bands sufficiently to permit electrons to enter the channel. This then increases the threshold voltage. Figure S3.6b indicates this for a particular case of 10^{12} electrons/cm² in the well. [4] The lowest allowed state is 40 meV above the bottom of the well, increasing the potential barrier E_B for electrons entering the channel, Figure S3.6c.
2. Recall that the electron in a potential well has a wave function that oscillates in the well and goes to zero outside the well. This forms a standing wave whose maxima are in the interior of the well. Because of the standing electron wave in the x direction in the well, $\psi^*\psi$, and thus the maximum charge concentration for a given electron, is removed from the Si/oxide interface as indicated in Figure S3.6d for the device of (a). [4] Figure S3.6e shows a plot of charge density as a function of position for the classical case and for the actual (quantum-mechanical) case. Here it can be seen that, from classical considerations, the channel charge should be concentrated near the interface, but when the quantization is considered, the maximum charge is on the order of 1 or 2 nm from the interface. Since the charge is located deeper into the semiconductor, this effect is analogous to having an increased oxide thickness. This results in a reduced transconductance.

For this case the threshold voltage has increased somewhat more than the 40 meV difference between the lowest occupied state and the classical bottom of the well. Because the oxide is in effect thicker, and because any applied gate voltage is dropped partly across the “oxide” and partly in the semiconductor, to produce the required extra band bending of 40 meV, approximately 60 meV must be applied to the gate.

Since these quantum mechanical effects increase with increasing transverse field, they are more pronounced for heavier substrate doping and for increased channel charge.

S3.4 MOSFET ANALOG EQUIVALENT CIRCUIT

In addition to the fundamental principles of MOSFET operation that have already been discussed in this and the preceding two chapters, there are a number of parasitic elements that must be taken into account. These are indicated in Figure S3.7.

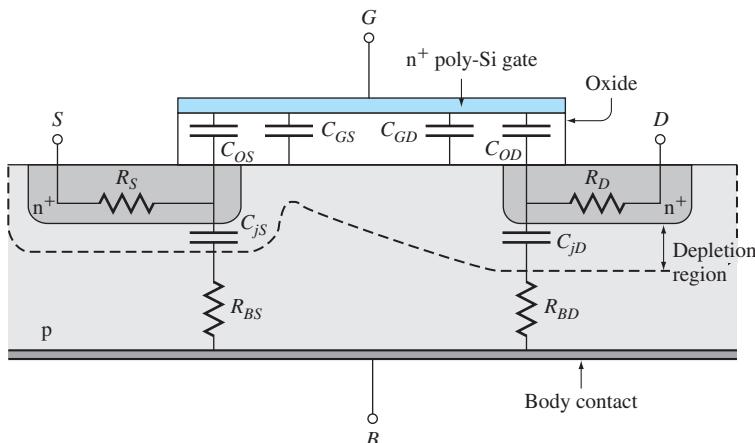


Figure S3.7 Schematic diagram of an n-channel MOSFET showing the various resistances and capacitances.

The source and drain resistances R_S and R_D are the resistances from the source and drain contacts to the edges of the channel. The source-to-substrate and drain-to-substrate resistances, R_{BS} and R_{BD} , are the resistances from the edges of the source-substrate and drain-substrate depletion regions to the substrate (body) contact.

The gate oxide prevents direct current from flowing into the gate. (Although in modern devices with thin oxides, some small tunnel current is tolerated). There is, however, some capacitance between the gate and the n⁺ source, C_{OS} , and between the gate and the n⁺ drain, C_{OD} . These are called *overlap capacitances* since the gate overlaps the source and drain regions. In addition, there is capacitance between the gate and the channel. While strictly speaking this is a distributed capacitance, it is normally modeled as two lumped capacitances—between gate and source, C_{GS} , and between gate and drain, C_{GD} . There are also junction capacitances, C_{JS} and C_{JD} , between the n⁺ source and drain and the p-type substrate. Note that although the C'_ox that we have used in the past is a capacitance per unit area, the capacitances C_{OS} , C_{OD} , C_{GS} , C_{GD} , C_{JS} , and C_{JD} are just capacitances (not per unit area).

The overlap capacitance can be reduced by the use of a self-aligned gate structure in which the gate is used as a mask for the source and drain implantation process. Since the overlap capacitances can therefore be made small, we will neglect them in the next section, where we cover the small-signal equivalent circuit of a MOSFET.

S3.4.1 SMALL-SIGNAL EQUIVALENT CIRCUIT

We have discussed the CMOS inverter as a representative example of a digital circuit. MOSFETs are also used in analog circuits, and to analyze such circuits it is useful to have a device characterized by its small-signal equivalent circuit.

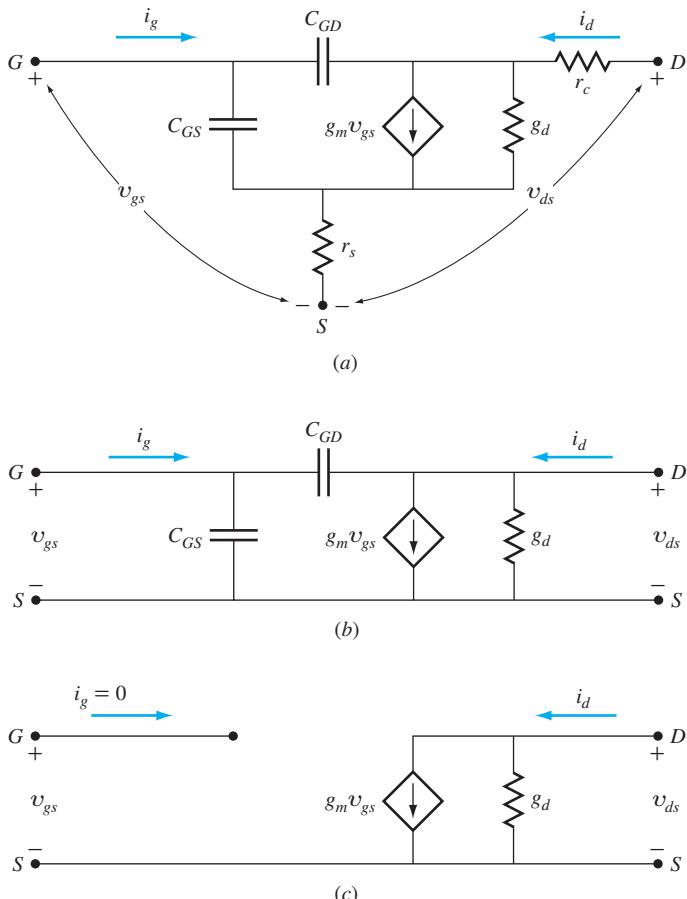


Figure S3.8 Small-signal equivalent circuit for a MOSFET (a) at high frequencies; (b) neglecting the source and drain resistances; (c) at low frequencies.

The small-signal equivalent circuit of Figure S3.7 is shown in Figure S3.8a. The current generator is $g_m v_{gs}$ where g_m is the transconductance, as will be discussed later. For simplicity, in (b) the resistances have been neglected. An output conductance g_d has also been added. The input small signal voltage³ is v_{gs} , and the output small signal voltage is v_{ds} . The input and output small signal currents are i_g and i_d .

Figure S3.8c shows an equivalent circuit for a MOSFET at low frequencies. By low frequencies, we mean signals that are varying slowly enough that the capacitances have negligible effect on the signals.

³We use the conventional notation that capital letters are used for steady-state or dc values (V_{GS} , V_{DS} , I_D , etc.) and small variations around those steady-state points are represented with small letters (v_{gs} , v_{ds} , i_d , etc.).

At both high and low frequencies, the output current is characterized by a current generator $g_m v_{gs}$ in parallel with an output conductance g_d . It is important to notice that the input signal is a voltage, but the output quantity that we generally consider with these transistors is a current. The transfer function $i_{\text{out}}/v_{\text{in}}$ has units of conductance. The transconductance g_m of the device, then, is

$$g_m \equiv \frac{i_d}{v_{gs}} = \frac{\partial I_D}{\partial V_{GS}} \quad (\text{S3.22})$$

where V_{DS} is considered constant, or v_{ds} is zero.

The output conductance results from the slope of the I_D - V_{DS} characteristics with V_{GS} constant.

$$g_d \equiv \frac{i_d}{v_{ds}} = \frac{\partial I_D}{\partial V_{DS}} \quad (\text{S3.23})$$

EXAMPLE S3.5

Consider an NFET with $t_{\text{ox}} = 4.7 \text{ nm}$, $L = 0.27 \mu\text{m}$, $W = 8.6 \mu\text{m}$, and $V_T = 0.3 \text{ V}$, whose I_D - V_{DS} characteristics are shown in Figure S3.9. Find the transconductance in saturation, g_{msat} , at $V_{DS} = 1.5 \text{ V}$, and the output conductance in saturation g_{dsat} at $V_{GS} = 1.8 \text{ V}$.

Solution

To find the transconductance at $V_{GS} = 1.5 \text{ V}$, we choose the points A and B at a constant $V_{DS} = 1.5 \text{ V}$. From Equation (S3.22), we have, in saturation,

$$g_{msat} = \frac{\partial I_D}{\partial V_{GS}} \approx \frac{\Delta I_D}{\Delta V_{DS}} = \frac{4.3 \text{ mA} - 2.2 \text{ mA}}{1.8 \text{ V} - 1.2 \text{ V}} = 3.5 \frac{\text{mA}}{\text{V}} = 3.5 \text{ mS}$$

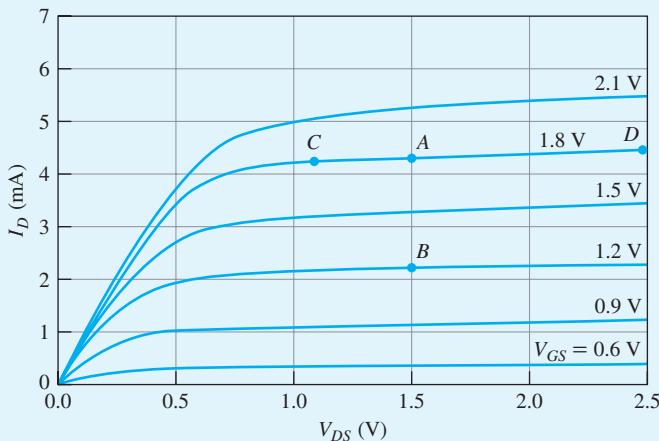


Figure S3.9 How to find g_m and g_d for a MOSFET from electrical measurements (Example S3.5).

The output conductance is found from points *C* and *D*, using Equation (S3.23), and applying it in the saturation region:

$$g_{dsat} = \left. \frac{\partial I_D}{\partial V_{DS}} \right|_{sat} \approx \frac{4.50 \text{ mA} - 4.25 \text{ mA}}{2.5 \text{ V} - 1.2 \text{ V}} = 0.178 \text{ mS}$$

This is equivalent to $(0.178 \times 10^{-3})^{-1} = 5.3 \text{ k}\Omega$ of output resistance.

We will now present some approximate equations for the transconductance g_m and the output conductance g_d . To do this, some parasitic effects are ignored. These are the source and drain resistances and the influence of the transverse field \mathcal{E}_T on the mobility μ_{lf} , both covered in Chapter 7. We will also ignore the influence of V_{DS} on V_T and on the effective channel length L . We will however, consider velocity saturation here since it has a major impact on the I_D - V_{DS} characteristics, and we will use the velocity saturation model for the I_D - V_{DS} characteristics as the starting point for the derivation.

We see from Equations (S3.22) and (S3.23) that the two conductances can be found from the I_D - V_{DS} characteristics. The I_D - V_{DS} relation was given by Equation (7.66) and is repeated here:

$$I_D = \frac{WC'_{ox} \mu_{lf} \left(V_{GS} - V_T - \frac{V_{DS}}{2} \right) V_{DS}}{L \left(1 + \frac{\mu_{lf} V_{DS}}{L v_{sat}} \right)} \quad V_{DS} \leq V_{DSsat} \quad (\text{S3.24})$$

This is from the model taking velocity saturation into account. The corresponding saturation current is [Equation (7.67)]

$$I_D = I_{Dsat} = \frac{WC'_{ox} \mu_{lf} \left(V_{GS} - V_T - \frac{V_{DSsat}}{2} \right) V_{DSsat}}{L \left(1 + \frac{\mu_{lf} V_{DSsat}}{L v_{sat}} \right)} \quad V_{DS} \geq V_{DSsat} \quad (\text{S3.25})$$

We start with the transconductance. Taking the partial derivative as prescribed by Equation (S3.22), we obtain for $V_D \leq V_{DSsat}$

$$g_m = \frac{WC'_{ox} \mu_{lf} V_{DS}}{L \left(1 + \frac{\mu_{lf} V_{DS}}{L v_{sat}} \right)} \quad V_{DS} \leq V_{DSsat} \quad (\text{S3.26})$$

This is the result below saturation. Above saturation, the transconductance g_{msat} can be determined from Equations (S3.22) and (S3.25). This is more complicated than it looks, however, because V_{DSsat} is a function of V_{GS} . Let us do as was done before in Chapter 7. There we used Equation (7.66) [the same as Equation (S3.24)], and found the value of I_{Dsat} by setting the slope $\partial I_D / \partial V_{DS}$ to zero.

For small L such that $\mu V_{DSsat} / v_{sat} L \gg 1$, Equation (S3.25) becomes

$$I_{Dsat} \approx WC'_{ox} v_{sat} \left(V_{GS} - V_T - \frac{V_{DSsat}}{2} \right) \quad (\text{S3.27})$$

Inserting this into Equation (S3.22) and taking the partial derivative with respect to V_{GS} gives

$$g_{msat} = Wv_{sat} C'_{ox} \left(1 - \frac{\partial V_{DSsat}}{\partial V_{GS}} \right) \quad (\text{S3.28})$$

Using V_{DSsat} as expressed by Equation (7.68), we obtain

$$g_{msat} = Wv_{sat} C'_{ox} \left\{ 1 - \frac{1}{2} \left[1 + \frac{2\mu_{lf}(V_{GS} - V_T)}{v_{sat}L} \right]^{-1/2} \right\} \quad (\text{S3.29})$$

Now let us examine the output conductance g_d . We use the definition of g_d [Equation (S3.23)] as the starting point. Differentiating Equations (S3.24) and (S3.25) below and above saturation, we obtain

$$g_d = \frac{WC'_{ox} \mu_{lf}}{L} \left(\frac{(V_{GS} - V_T - V_{DS}) - \frac{\mu_{lf} V_{DS}^2}{2L v_{sat}}}{\left(1 + \frac{\mu_{lf} V_{DS}}{L v_{sat}} \right)^2} \right) \quad V_{DS} \leq V_{DSsat} \quad (\text{S3.30})$$

and

$$g_{dsat} = 0 \quad V_{DS} \geq V_{DSsat} \quad (\text{S3.31})$$

Of course, the saturation output conductance is not really zero because of some of the effects we neglected in this derivation. One of these is channel-length modulation, discussed in Chapter 7, in which the effective channel length varies with drain-source voltage V_{DS} . The second effect is the dependence of the threshold voltage on V_{DS} .

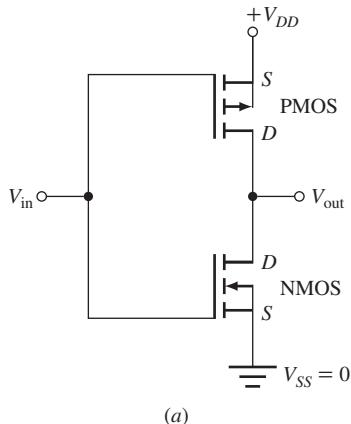
S3.4.2 CMOS AMPLIFIERS

CMOS inverters have been discussed in relation to digital circuits. They also are useful as small signal amplifiers. Consider the CMOS inverter of Figure 8.3 repeated here as Figure S3.10. The transfer characteristics (V_{out} - V_{in}) of such a device are shown in Figure S3.10b. For $V_{in} = V_{DD}/2 + v_{ac}$, where v_{ac} is a small-signal voltage, it can be seen that for a small change in V_{in} (e.g., v_{ac}) the output voltage changes by a large amount or the ac voltage gain is large.

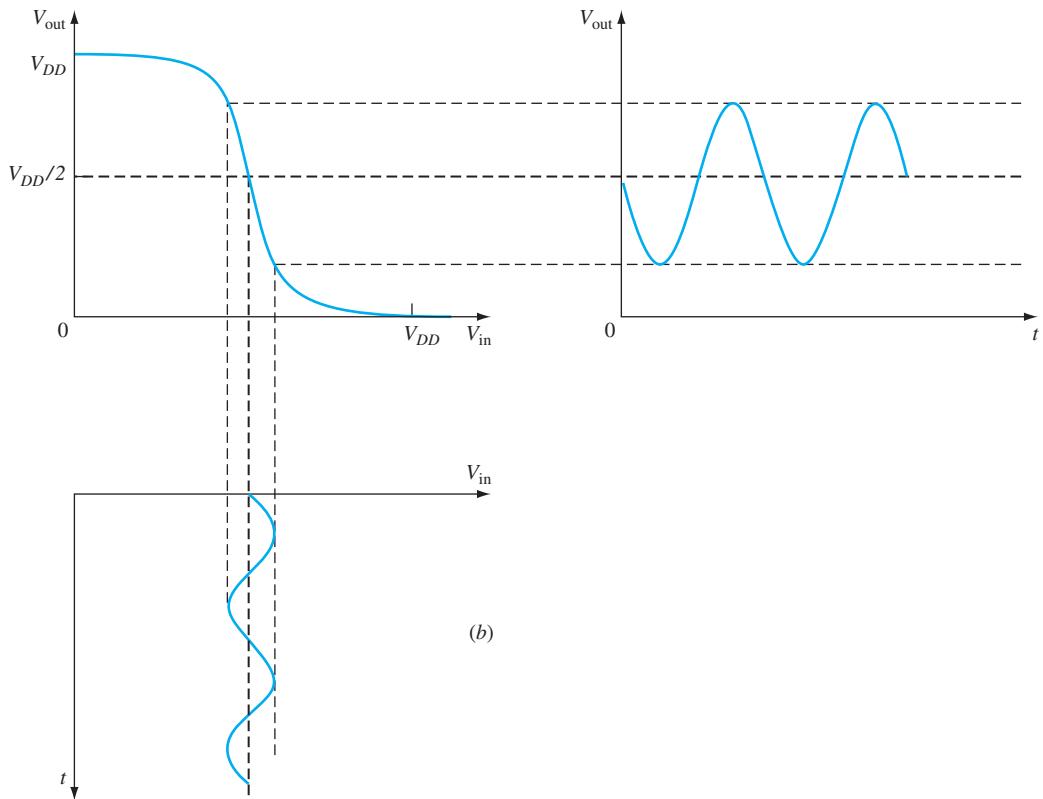
S3.5 UNITY CURRENT GAIN CUTOFF FREQUENCY f_T

At low frequencies, the gate capacitance blocks any gate current (excluding a very small leakage through the oxide), so the input current is zero. Since the current gain of the transistor is i_d/i_g , then at low frequencies the current gain is infinite.

At high frequencies, however, there will be (displacement) current flowing in the gate oxide, which will tend to decrease the current gain. The current gain cutoff frequency f_T is defined as the frequency at which the magnitude of the current gain is reduced to unity, with the ac output of the circuit of Figure S3.8



(a)

**Figure S3.10** (a) CMOS inverter. (b) The input voltage is amplified at the output.

short circuited.⁴ In the high-frequency model of Figure S3.8a, there are two capacitances shown explicitly. For the case when the output is short circuited, these two capacitors are in parallel, so their capacitances add. Ignoring the

⁴The ac output is short circuited to ground via the load capacitance. The dc output is not short circuited, of course, since to operate, $V_{DS} > 0$.

overlap capacitances, the input current is related to the input voltage v_g by the admittance:

$$i_g = j2\pi f(C_{GS} + C_{GD})v_{gs} \approx j2\pi f(C'_{ox}WL)v_{gs} \quad (\text{S3.32})$$

where $(C_{GS} + C_{GD}) \approx C'_{ox}WL$. (Recall that C_{GS} and C_{GD} are capacitances while C'_{ox} is capacitance per unit area.)

At the drain, the short-circuit output current is

$$i_d = g_m v_{ds} \quad (\text{S3.33})$$

Setting the current gain magnitude to unity, then, we make $|i_g| = |i_d|$. Solving for $f = f_T$, we find the current gain cutoff frequency from Equations (S3.32) and (S3.33):

$$f_T = \frac{g_m}{2\pi C'_{ox}WL} \quad \text{or} \quad \omega_T = 2\pi f_T = \frac{g_m}{C_{gate}} \quad (\text{S3.34})$$

which is a figure of merit for a MOSFET.

If the MOSFET is operating in the current saturation region, we substitute the expression for g_{msat} from Equation (S3.29) to obtain

$$f_T = \frac{v_{sat}}{2\pi L} \left\{ 1 - \left[1 + \left(\frac{2\mu_f(V_{GS} - V_T)}{v_{sat}L} \right)^{-1/2} \right] \right\} \quad (\text{S3.35})$$

Figure S3.12 shows a plot of the current cutoff frequency f_T as a function of channel length for the same transistor as in Figure S3.10 [i.e., $v_{sat} = 4 \times 10^6$ cm/s, $\mu_f = 500$ cm²/V · s, and $(V_{GS} - V_T) = 2.6$ V]. Here we see that f_T decreases from 56 GHz for the short-channel NFET ($L = 0.1$ μm) to 430 MHz for a longer-channel NFET ($L = 5$ μm), and from 50 GHz to 234 MHz for the equivalent PFET. Clearly short channels make for significantly higher frequency devices.

In these plots (Figures S3.10 and S3.12), we considered only the velocity saturation effect on the given parameters. These are optimistic results. Series resistance and short-channel effects reduce g_m and f_T from the values presented here.

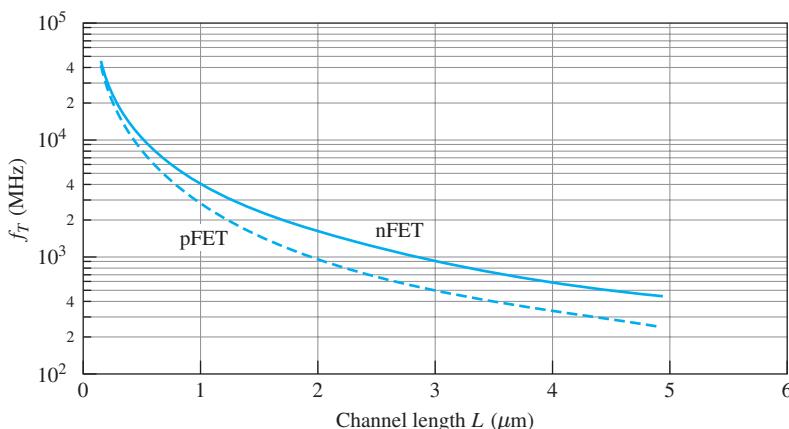


Figure S3.11 Current cutoff frequency as a function of channel length for n-channel and p-channel MOSFETs.

S3.6 MOS CAPACITORS

We noted in Chapter 7 that the gate structure of a field-effect transistor looks like a capacitor. In fact, the same fabrication processes can be used to put capacitors as circuit elements into integrated circuits. Further, the electrical characteristics of a MOS capacitor can be used as a diagnostic tool to determine various material properties resulting from the fabrication processes. Thus, we will examine MOS capacitors in some detail in this section.

The structure of a MOS capacitor is shown in Figure S3.12. It is similar to a MOSFET except that there are no source or drain regions. As is the case for MOSFETs, the “metal” gate is often degenerately doped Si. We will consider the case of a metal electrode, and the insulating region to be SiO_2 , so we have a metal/ SiO_2 /Si MOS capacitor, and we’ll take the Si to be p type.

S3.6.1 IDEAL MOS CAPACITANCE

The capacitance between the gate and the substrate depends on the applied voltage, the metal-semiconductor work function difference, and all the charges involved in the determination of MOSFET threshold voltage. In this section we idealize the problem by assuming that Φ_{MS} , Q_f , and Q_{it} are all equal to zero. (We will include these terms in the following section.) We will consider only the influence of the depletion region width in the bulk substrate on the bulk charge Q_B , and the mobile charge in the Si at the Si/oxide interface, which we call Q_i . Note that in MOSFETs, this mobile charge was referred to as Q_{ch} , the channel charge per unit area. There is no channel in the capacitor since there is no source or drain.

Figure S3.13 shows the energy band diagram for various values of gate-substrate voltage V_G for a metal/oxide/p-type semiconductor and for a metal/oxide/n-type semiconductor. In (a), a gate voltage of the polarity shown is applied to bend the bands such that majority carriers accumulate near the surface of the

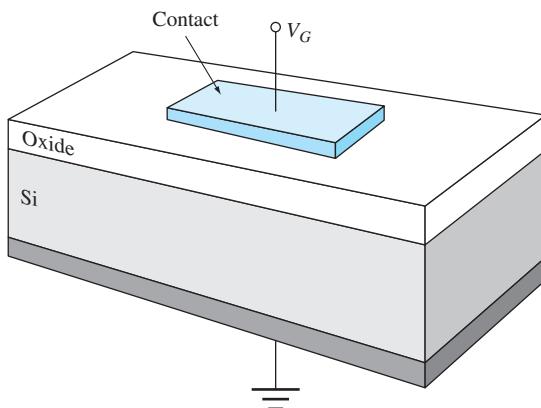
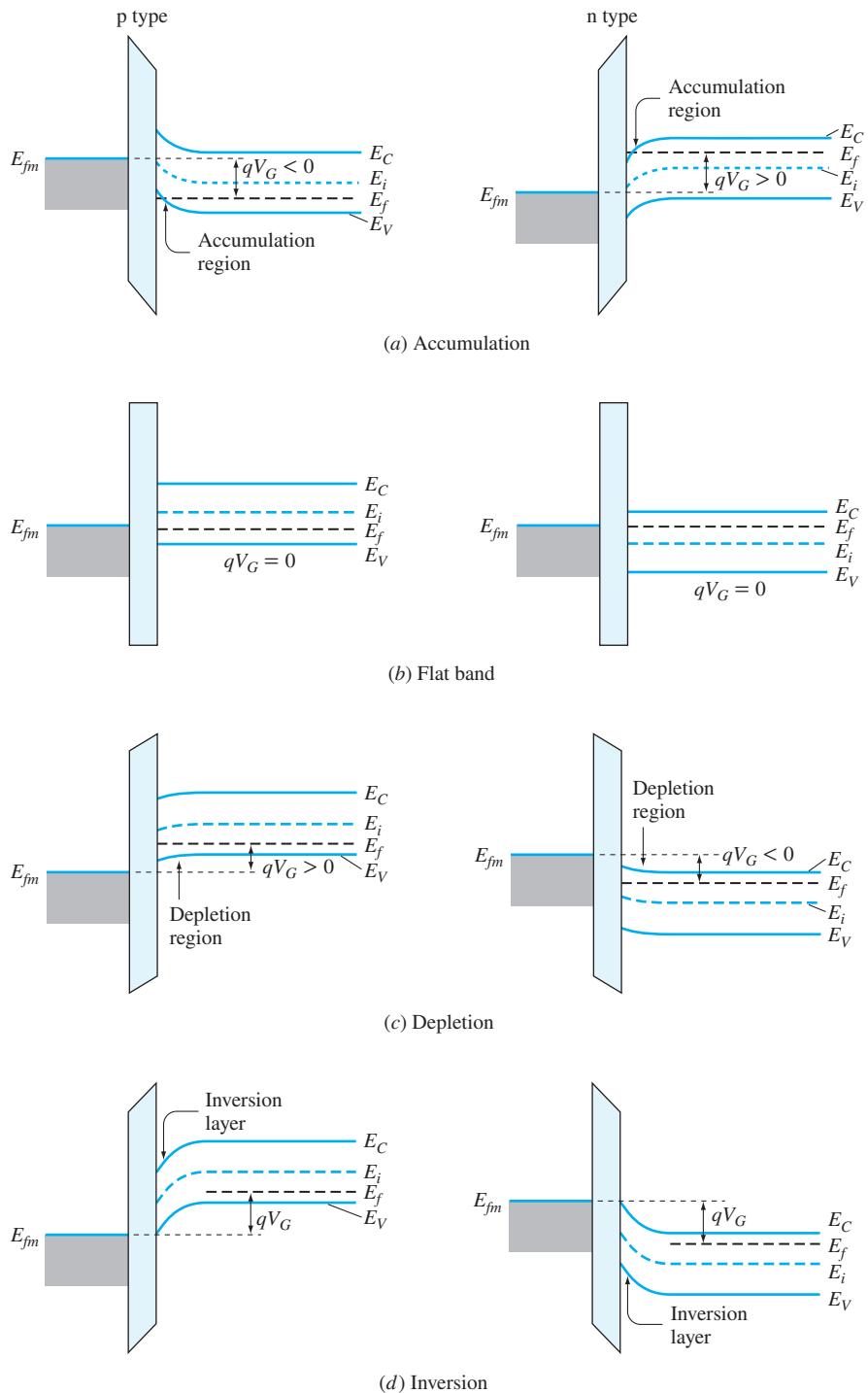


Figure S3.12 Schematic of a MOS capacitor.

**Figure S3.13**

Energy band diagram for ideal MOS structures using a metal gate and p-type silicon substrate (left) and n-type substrate (right). In (a) the bands bend in such a way that the surface of the silicon is accumulated (with majority carriers). For this special case where $\Phi_M = \Phi_S$, the flat-band case (b) is also the equilibrium case. In (c) the polarity of V_G is such that a depletion region is established in the silicon. In (d), V_G is sufficient to produce an inversion layer.

semiconductor. Since in the simplified ideal case being considered, $\Phi_{MS} = 0$, the flat band case (b) is also the equilibrium case.

In (c), a gate voltage is applied that creates a depletion region in the Si, and in (d), the gate voltage is large enough that the Si is inverted near the interface.

Let us consider the metal/SiO₂/p-type semiconductor. The charges associated with each of the bias conditions of Figure S3.13 are indicated schematically in Figure S3.14.

- In accumulation (a), holes in the valence band exist at the interface, creating a positive mobile interface charge Q_i . An equal and opposite electron charge exists on the metal gate.
- In the flat band condition (b), neutrality exists everywhere in the silicon.
- In depletion (c), the charge in the semiconductor results from the ionized acceptors Q_B in the depletion region. There are no mobile charges, so $Q_i \approx 0$.
- In inversion (d), in addition to the bulk charge Q_B resulting from the ionized acceptors in the depletion region, a negative Q_i exists resulting from electrons at the interface.

The gate-substrate capacitance C' per unit area as a function of V_G is indicated in Figure S3.15 for this ideal case at low frequencies ($f \sim 10$ Hz) and at high frequency ($f \sim 1$ MHz). Here V_G varies slowly, so that above threshold ($\phi_S = 2\phi_f$), the variation of V_G causes a change in Q_i , and Q_B is constant. The C' - V_G plot can be explained as follows:

- For $V_G < 0$ V, the Si surface is in accumulation as indicated in Figure S3.13a. Since $C' = |dQ_m/dV_G|$, when there is a change dV_G , then the charge on the metal changes by $dQ_m = -dQ_i$. This structure looks like and behaves as a parallel plate capacitor, with SiO₂ as the dielectric and with capacitance per unit area of

$$C' = C'_{ox} \quad \text{accumulation} \quad (\text{S3.36})$$

Point *a* on the C - V_G plot indicates accumulation.

- At flat band (point *b* in Figure S3.15), there is still some mobile charge at the interface. If the gate voltage is varied by a small amount from V_{FB} , the bands will bend up or down slightly. Thus, the effect of dV_G penetrates some small distance into the substrate. This distance is known as the Debye length L_D . This adds, in effect, a second capacitor in series with the oxide capacitor. Thus, at $V_G = V_{FB}$,

$$\frac{1}{C'} = \frac{1}{C'_{ox}} + \frac{L_D}{\epsilon_{Si}} = \frac{t_{ox}}{\epsilon_{SiO_2}} + \frac{L_D}{\epsilon_{Si}} \quad \text{near flat band} \quad (\text{S3.37})$$

- Note that in accumulation, the number of mobile charges is large and they are all close to the interface, such that the second term in Equation (S3.37) is negligible. Equation (S3.37) then reduces to Equation (S3.36).
- For $V_G > V_{FB}$ (remember $V_{FB} = 0$ for this ideal capacitor), a depletion region forms in the substrate. For point *c* on the plot, there is no mobile charge, and

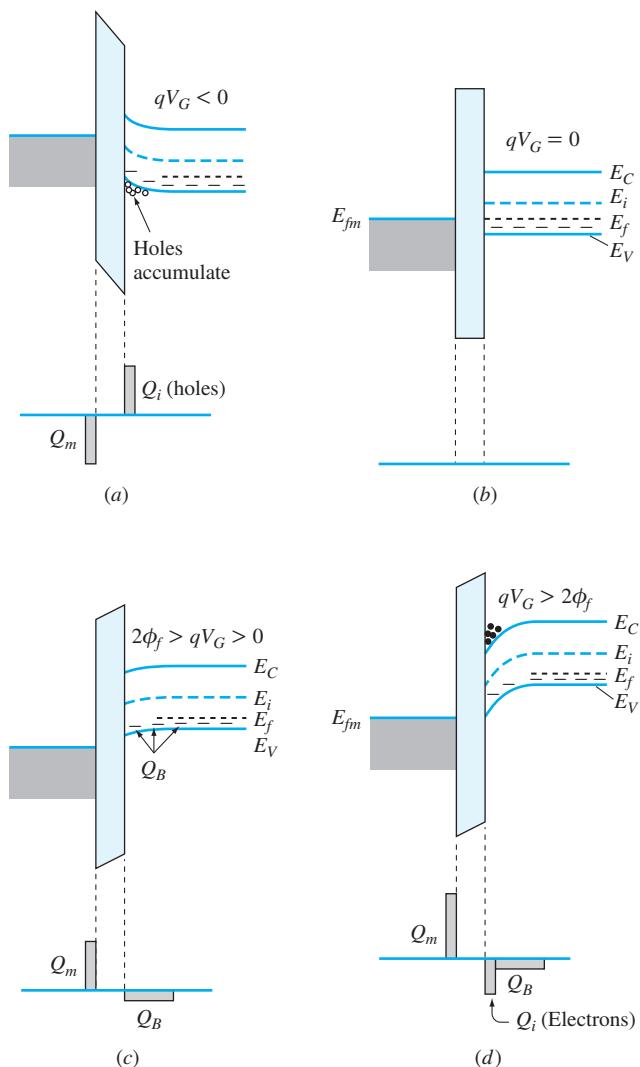


Figure S3.14 Charges associated with the ideal MOS p-Si capacitor of Figure S3.12. In (a) holes accumulate near the surface, attracting an equal and opposite negative charge on the gate. In (b) flat band is also the case for equilibrium (for this case in which $\Phi_M = \Phi_S$). (c) For small positive V_G , a depletion region is formed in the silicon, with negatively ionized acceptors near the surface not neutralized by holes. An equal and opposite positive charge appears on the gate. (d) For V_G such that $\phi_s > 2\phi_f$, the bulk charge Q_B remains, and in addition there are mobile electrons occupying states in the silicon potential well.

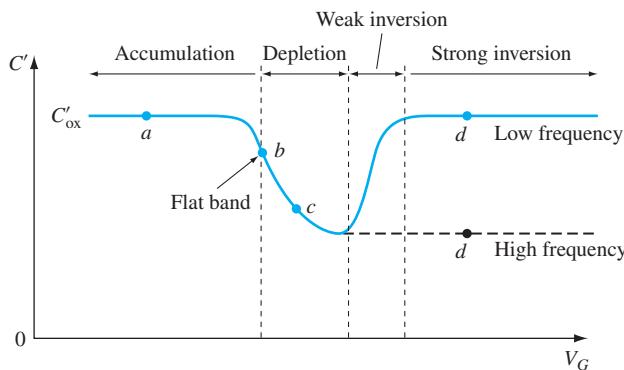


Figure S3.15 Capacitance-voltage characteristic for the ideal MOS capacitor at low and high frequencies. (a) Accumulation; (b) flat band; (c) depletion; (d) strong inversion.

$Q_i = 0$. The capacitance is a series combination of the oxide capacitance C'_ox and the (bulk) junction capacitance C'_B

$$\frac{1}{C'} = \frac{1}{C'_\text{ox}} + \frac{1}{C'_B} \quad (\text{S3.38})$$

where C'_B is the capacitance per unit area associated with the depletion region:

$$C'_B = \left(\frac{\epsilon_{\text{Si}} q N_A'}{2 \phi_s} \right)^{1/2} \quad (\text{S3.39})$$

and w_B , the depletion region depth, is

$$w_B = \left(\frac{2 \epsilon_{\text{Si}} \phi_s}{q N_A'} \right)^{1/2} \quad (\text{S3.40})$$

- In the inversion region (point d), in addition to the depletion charge Q_B , there exists an interface charge Q_i . The capacitance in this case depends on whether the measurements are made at high frequencies or at low frequencies.

At high frequencies (typically 1 to 10 MHz), $f \gg 1/\tau$, where τ is the lifetime associated with the carrier generation and recombination in the depletion region. In this case, the change in V_G is rapid enough that negligible generation or recombination can occur during one cycle, meaning that Q_i is not affected. Thus $|dQ_m/dV_G| = |dQ_B/dV_G|$, and the capacitance is the series combination of C'_ox and C'_B

$$\frac{1}{C'} = \frac{1}{C'_\text{ox}} + \frac{1}{C'_B} \quad (\text{S3.41})$$

This value is independent of the value of V_G , as shown by the dashed line in Figure S3.15.

At low frequencies, however (typically 1 to 100 Hz), for $f \ll 1/\tau$, the generation and recombination of minority carriers in the depletion region can follow the change in V_G . Thus, Q_i can vary in response to the changing gate voltage. Recall from “More About Threshold” in Section 7.2.4, however, that above threshold, ϕ_s is nearly constant and that any additional change in gate voltage is dropped across the oxide. That implies that Q_B is constant. Thus, in this case $|dQ_m/dV_G| = |dQ_i/dV_G|$. As a result, in this regime C' is essentially equal to C'_{ox} and is independent of the dc value of V_G .

S3.6.2 THE C-V_G CHARACTERISTICS OF REAL MOS CAPACITORS

In the above, Φ_{MS} , Q_{it} , and Q_f were assumed to equal zero. In a real MOS capacitor, however, they must be considered. We will take the effect of each in turn.

The Effect of Φ_{MS} In the ideal MOS capacitor, flat band occurred at $V_G = 0$. As discussed in Section S3.3.5, and expressed in Equation (S3.13), for a real MOS capacitor the flat band voltage becomes

$$V_{FB} = \frac{\Phi_{MS}}{q} - \frac{[Q_f + Q_{it}(0)]}{C'_{ox}} \quad (\text{S3.42})$$

The effect, then, is to shift the C-V plot of Figure S3.16 to the right or left on the voltage scale, depending on the sign of Φ_{MS} .

Interface States It was indicated earlier that the charge Q_{it} trapped in interface states depends on the band bending in the semiconductor ϕ_s . At a given value of ϕ_s , the interface trap states below the Fermi level in the semiconductor are

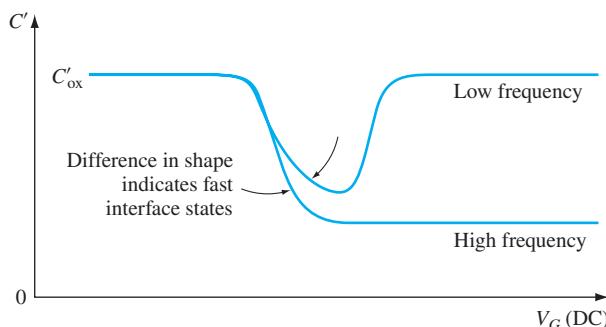


Figure S3.16 C-V characteristics of a real MOS capacitor at low and high frequency. The difference in the general shape from that of the ideal MOS capacitor in Figure S3.15 results from the presence of fast interface states. Since the occupancy of these states can follow the low-frequency variation in V_G but not the high-frequency variation, the difference between the two plots is a measure of the concentration of fast states.

predominantly occupied while those above the Fermi level are mostly vacant. As V_G and thus the band bending changes, the value of Q_{it} also changes. The interface states, also referred to as *fast interface states*, however, respond to V_G at frequencies below about 1 kHz but not at frequencies on the order of 1 MHz. Thus, their concentration can be determined from a comparison of high-frequency capacitance with the low-frequency capacitance (Figure S3.16).

S3.6.3 MOSFET PARAMETER ANALYSES FROM C-V_G MEASUREMENTS

We have seen how some of the physical parameters of the MOSFET can affect the $C'-V_G$ curves. Conversely, examination and interpretation of these curves can reveal the values of some parameters, including the oxide thickness, the bulk or depletion capacitance C'_B , the depletion region width w_B , and the substrate doping level.

In the accumulation mode, and at low frequencies in the inversion mode (see Figure S3.15), the oxide thickness can be determined if the oxide permittivity is known, from

$$C' = C'_{ox} = \frac{\epsilon_{\text{oxide}}}{t_{\text{ox}}} \quad (\text{S3.43})$$

From the high-frequency $C-V_G$ characteristics in the inversion mode, and using the measured value of C'_{ox} , one can find the bulk or depletion region capacitance:

$$C'_{\text{inversion}} = \frac{C'_{ox} C'_B}{C'_{ox} + C'_B} = \frac{C'_{ox}}{\frac{C'_{ox}}{C'_B} + 1} \quad (\text{S3.44})$$

This can be solved for C'_B . Furthermore, since

$$C'_B = \frac{\epsilon_{\text{Si}}}{w_B}$$

the maximum depletion width w_B is now known. From

$$w_B = \left[\frac{2\epsilon_{\text{Si}}(2\phi_f)}{qN_A} \right]^{1/2} \quad (\text{S3.45})$$

we can find the doping concentration in the substrate using

$$\phi_f = \frac{kT}{q} \ln \frac{N'_A}{n_i} \quad (\text{S3.46})$$

As discussed in Supplement 2, N'_A can also be determined from the $C-V$ characteristics of a Schottky barrier obtained by depositing a metal onto the Si surface with the oxide removed.

S3.7 DYNAMIC RANDOM-ACCESS MEMORIES (DRAMS)

MOS technology is also used to make memory cells, for example random-access memory (RAM). [5] An attractive feature of a RAM is that each memory cell can be written to and read from individually. In RAM memory, the information stored in each cell remains there until the cell is rewritten or until the power to the circuit is turned off (volatile memory).

The most common type of RAM is the dynamic random-access memory (DRAM, pronounced dee-ram). Each cell is composed of a MOS capacitor and an n^+ p junction separated by a transfer gate. A cross section of one such “one-transistor memory cell” is shown schematically in Figure S3.17a. The capacitor is connected to a constant voltage, 2.5 V in this illustration. The transfer gate of the DRAM cell, when it is at zero volts, isolates the capacitor from the diode as in any other enhancement field-effect device. When $V_G = 2.5$ V, a conducting channel connects capacitor well and diode.

Figure S3.17b shows the hybrid diagram for this memory cell for $V_G = 0$, when the capacitor and diode are isolated from each other. We define a zero (**0**) to be stored in the capacitor well if it is empty and a one (**1**) to be stored if it is filled. The dashed lines represent the case for the source well empty of charge (logical **0**) and for the diode biased at 2.5 V. The solid lines represent the state for a logical **1**, in which case the potential well under the capacitor contains charge. Note that the state stored depends only on the charge in the capacitor well, not on the diode voltage V_D . We illustrate two different values of V_D , however, because we will use both conditions in what follows.

We now examine the process of writing a **1** or a **0** into the capacitor well. For the write operation, the capacitor well is initially empty (logical zero). To

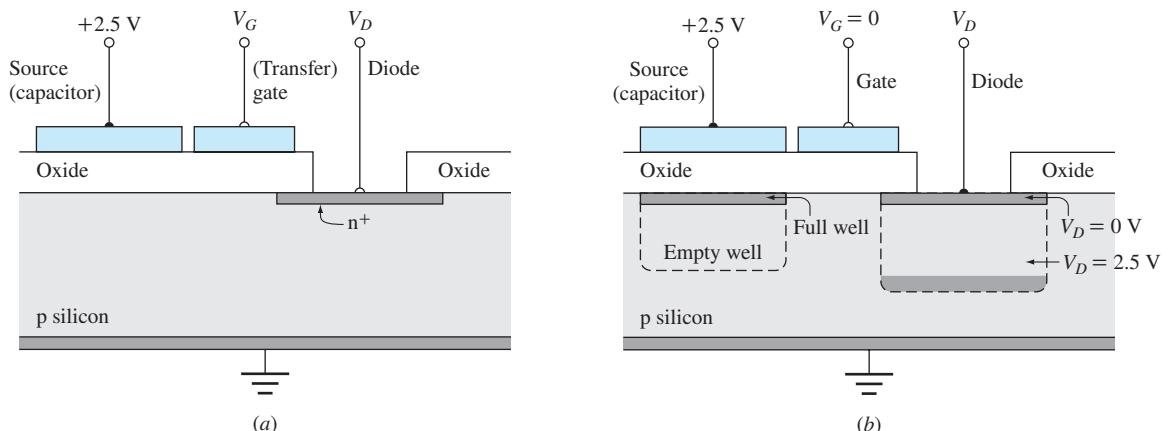


Figure S3.17 (a) The cross section of a DRAM and (b) the hybrid diagram. This one-transistor memory cell consists of a diode and a MOS capacitor, separated by a transfer gate. (b) When $V_G = 0$, the source and drain are isolated.

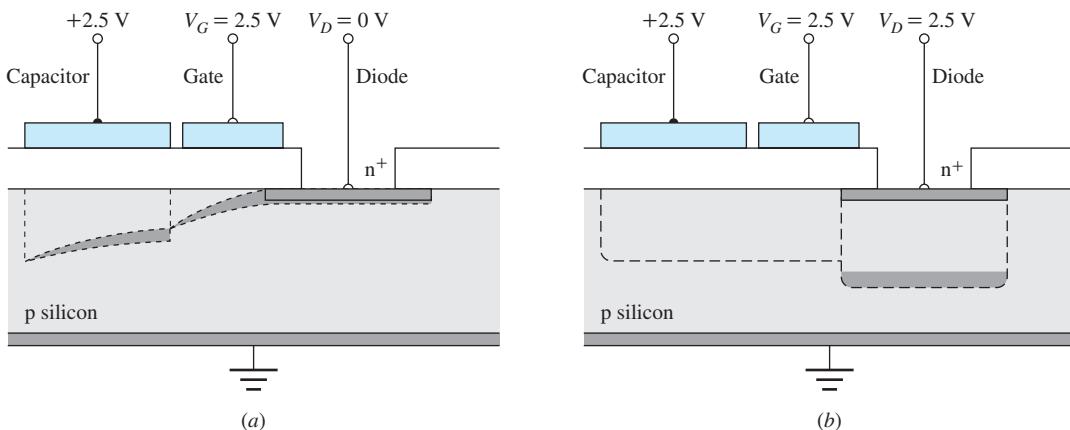


Figure S3.18 (a) Illustration of writing a **1** into a memory cell. With $V_D = 0$ and $V_G = 2.5$ V, charge flows from diode to capacitor to fill the potential well of the source capacitor. (b) To write a **0**, V_D is made positive when the gate is opened. In this case no charge flows from diode to capacitor.

write a **1**, V_D is set to zero volts at the same time that the gate is pulsed on. While $V_G = 2.5$ V, a channel exists. Charge then flows from diode to capacitor as indicated in Figure S3.18a. For this operation, the diode plays the role of the source of a MOSFET, while the capacitor acts as a drain.

Note that charge flow is by drift as well as diffusion. There is more charge at the right-hand end of each well during the transfer process, since the charge starts from the diode and diffuses to the left. Because the depth of the wells under the capacitor and the gate depends on the electron concentration at any position, the increased charge on the right of each well creates an electric field

$$\left(\mathcal{E} = \frac{1}{q} \frac{dE_C}{dy} \right)$$

which accelerates the electrons to the left. The drift and diffusion continue until the capacitance well is filled or contains a **1**. The gate pulse is removed, the channel disappears, and the charge remains in the capacitor potential well.

To write a **0**, we set $V_D = 2.5$ V and pulse the gate on. Although a channel exists (the channel is not shown in the figure), we can see from Figure S3.19b that no charge transfers because in this case the diode is at the same potential as the capacitor. When the gate is turned off ($V_G = 0$) the source well is still empty, i.e., uncharged.

Thus, to write a **1**, we set the diode voltage to 0 V and pulse the gate on. This charges the capacitor. To write a **0**, we set the diode voltage to 2.5 V and pulse the gate on. In either case, the capacitor voltage remains at 2.5 V.

To read what is stored in the well, V_D is set to 2.5 V and the gate turned on. If the capacitor contains a **1**, the charge is transferred to the diode as indicated in Figure S3.19a. The current resulting from this charge transfer is interpreted as

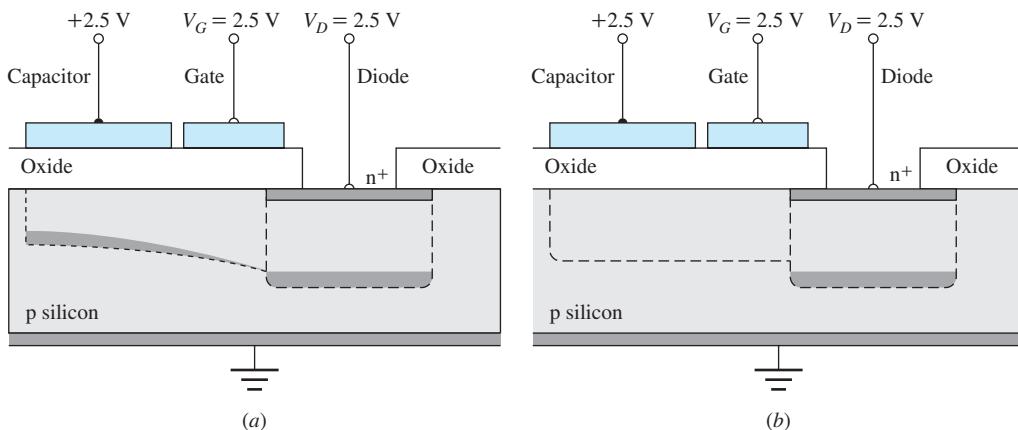


Figure S3.19 To read the contents of a DRAM memory cell, V_D is made positive when the gate is opened ($V_G = 2.5$ V). (a) If the cell contains a **1**, charge flows from capacitor to diode. (b) If the cell contains a **0**, no charge flows.

having read a stored **1**. For this read operation, the capacitor acts as a MOSFET source while the diode acts as a MOSFET drain. If the capacitor well is empty (uncharged), no charge flows during the read cycle and the absence of diode current is interpreted as reading a stored **0**, as indicated in Figure S3.19b.

Thermal generation in the region of the capacitor well, however, will trap electrons in the well. With time, an empty well will fill; that is, a **0** will become a **1**. For this reason the stored charge must be “refreshed” periodically (every few milliseconds), i.e., read out and rewritten.

S3.8 MOSFET SCALING [6]

MOSFET scaling involves the reduction in size of a MOSFET chip by systematically reducing the size of the MOSFETs and interconnects. Smaller devices permit more chips of a given functionality on a wafer. Since the cost of fabricating a wafer is relatively constant, the cost per chip is approximately inversely proportional to the number of chips per wafer. Smaller MOSFETs (shorter channel length) also means shorter source-to-drain transit time and thus faster switching (digital circuits) and higher frequency response (analog circuits).

There are two major approaches to MOSFET scaling: these are *constant field scaling* and *constant voltage scaling* as indicated in Table S3.1. In both approaches the device dimensions L and W are reduced by the same factor, k . The gate oxide capacitance per unit area (C'_ox) is also increased by the same factor. Assuming no change in the gate oxide material, the oxide thickness t_ox is also reduced by the factor k .⁵

⁵To avoid confusion with the symbol k often used for the dielectric constant, in some of the semiconductor literature, the inverse scaling factor S is used where $S = 1/k$.

Table S3.1 MOSFET scaling guidelines. The dimensional scaling factor k is typically 0.7.

Parameter	Symbol	Constant field scaling	Constant voltage scaling
Field	\mathcal{E}	1	$1/k$
Voltage	V	k	1
Channel width	W	k	k
Channel length	L	k	k
Area/device	A	k^2	k^2
Oxide thickness	t_{ox}	k	k
Gate capacitance per unit area	C'_{ox}	$1/k$	$1/k$
Gate capacitance	C_{ox}	k	k
Substrate doping	N'	$1/k$	$1/k^2$
On current	I_D	k	1
Power per device	P/D	k^2	1
Power density	P	1	$1/k^2$

In constant field scaling, all fields in a MOSFET remain the same, and thus supply voltage V_{DD} and substrate doping concentration must be changed as indicated in Table S3.1. The scaling factor k is on the order of 0.7 for each MOSFET generation, on the order of two years.

Because of the reluctance to change the supply voltage used in earlier circuits, the constant voltage scaling scheme is normally used for several generations of MOSFETs before the supply voltage is reduced with constant field scaling. A third scaling procedure called *general scaling* is often used. It is a modification of the constant field scaling scheme except that the voltage is reduced by a factor other than k .

There are several problems associated with reduction of the device dimensions. One is the drain-induced barrier lowering (DIBL) effect discussed in Chapter 7, which increases as the channel length decreases. Also, for very short channels the source and drain depletion regions overlap, thus reducing the source-to-channel barrier (punch-through), increasing the channel current. This can cause appreciable current for V_{GS} equal to zero, since the device is always **on**. Both of these problems can be overcome by reducing the supply voltage.⁶ The problem of punch-through can also be reduced by increasing the substrate doping. For substrate doping greater than about 10^{19} cm^{-3} , however, band-to-band tunneling of the reverse biased drain-substrate junction increases the drain-substrate current.

The punch-through problem was reduced by a very shallow ultra-heavily doped n^{++} region near the drain (drain extension), effectively moving the drain-substrate depletion region further from the channel (and the source). This separates the drain depletion region from the source depletion region and avoids

⁶The minimum supply voltage, however, is limited to that required to turn **on** the MOSFET, about 1 V for S_i .

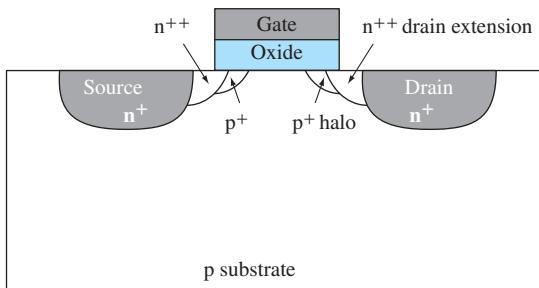


Figure S3.20 Cross section of MOSFET structure to reduce short-channel effects.

punch-through. A highly doped region (halo) of the same type as the substrate, p⁺ in the n-channel device of Figure S3.20, is added adjacent to the drain to suppress the drain depletion region on the source side.

In the constant field approach to scaling, the gate channel capacitance (C'_{ox}) is reduced by the same factor k as in constant field scaling. When the gate insulator is SiO₂, however, where t_{ox} is less than about 2 nm, tunneling from gate to channel is excessive. This can be alleviated by increasing the dielectric constant of the insulator, thus permitting a thicker insulating layer t_{ox} to obtain a given C'_{ox} . Silicon oxynitride (SiON) with a dielectric constant of about 5 compared to 3.9 for SiO₂ does improve this somewhat. The tunneling probability is exponentially dependent on band gap and the oxide thickness as well.⁷ Table S3.2 gives the dielectric constant and band gap for some of the metal oxides considered for gate insulators.

Note that in Figure S3.20, because of symmetry of the fabrication process, the drain extension and the halo structures are also present on the source side. While reducing V_{DD} , and thus the drain depletion region, reduces the short-channel effects, it also results in a smaller threshold voltage V_T . Since subthreshold current increases exponentially with gate-to-source voltage, V_{GS} , a smaller V_T results in a large increase in off current ($V_{GS} = 0$). In a chip with billions of MOSFETs, this increase can be substantial, resulting in an unacceptable power dissipation. This limits the number of devices, and thus the functionality per chip.

Table S3.2 Dielectric constant and band gap for some insulators considered for gate dielectric

	SiO ₂	SiON	HfO ₂	Al ₂ O ₃	ZrO ₂	TiO ₂	Ta ₂ O ₅
Dielectric constant	3.9	~5	20	9.5	22	80	25
Band gap (eV)	9	~6.5	4.5	8.8	4	3	5

⁷Also, replacing SiO₂ with SiON reduces the boron diffusion in the oxide.

*S3.9 DEVICE AND INTERCONNECT DEGRADATION

In this section, we examine a number of failure mechanisms in MOSFETs. Some of these are thermally activated, meaning that the degradation mechanism operates on thermal energy. That in turn implies that degradation will be faster at higher temperatures. Examples would be (1) oxide breakdown, in which thermally excited charges lodge in the oxide and eventually cause it to be conductive; (2) ionic diffusion in the oxide, which can cause changes in the threshold voltage over time; and (3) electromigration [7].

Electromigration occurs in the metal lines used to interconnect the various devices on the chip. As electrons flow through these lines, they can collide with the metal atoms, giving them enough momentum to move. With time, this atomic movement can cause voids in the line or shorts between adjacent lines. Electromigration is thermally activated because at higher temperatures, the atoms have more vibrational kinetic energy and thus are more easily dislodged from their equilibrium positions.

For a thermally activated mechanism, the mean time to failure (MTTF) is given by

$$\text{MTTF} \propto e^{E_a/kT} \quad (\text{S3.47})$$

where E_a is the activation energy. Typical values for E_a are 0.3 eV for oxide breakdown, 0.7 eV for ionic diffusion in the oxide, and 0.7 eV for electromigration. These mechanisms, however, can be reduced by proper device design.⁸

Another failure mechanism, known as *hot-carrier-induced degradation*, is not thermally activated but does depend on the kinetic energy of the channel carriers. It can cause severe degradation. To illustrate the origin of this phenomenon, we consider an n-channel MOSFET biased in the current saturation region. The energy diagram along the channel is shown in Figure S3.21a. In the region near the drain where the longitudinal field \mathcal{E}_L is high, the electrons can attain a high kinetic energy between collisions. They are called *hot carriers* by analogy to electrons that have high kinetic energy due to high temperatures.

There is also a transverse field \mathcal{E}_T , which accelerates these fast-moving carriers toward the oxide as indicated in Figure S3.21b. Those electrons that strike the oxide with sufficient kinetic energy can create interface states that can trap electrons, or they can penetrate a short distance into the oxide and become trapped. This trapped negative charge shifts the conduction band edge upward in the region adjacent to the drain since \mathcal{E}_L is maximum there.

Figure S3.22 shows the energy band diagram for $V_{DS} = 0$ for a virgin device (a) and for a device that has been subjected to hot-carrier stress (b). For the case shown, the potential energy “hump” prevents the conducting channel from connecting source to drain.

⁸For example, electromigration can be reduced by the use of copper conducting lines rather than the much lighter aluminum lines.

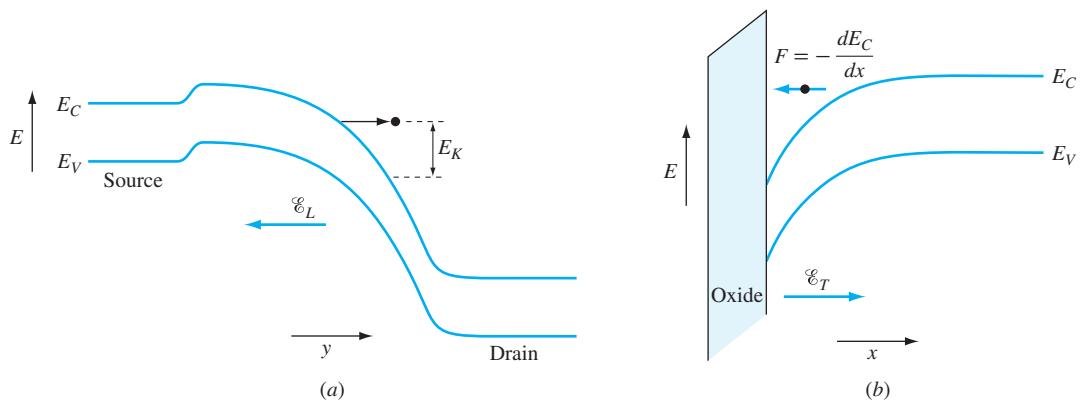


Figure S3.21 (a) The energy band diagram along the channel. The high lateral field means that between collisions, the electron can gain appreciable kinetic energy. (b) Because of the transverse field, the hot electrons are accelerated toward the oxide, where they can create interface states or become embedded in the oxide, thus causing a negative charge buildup.

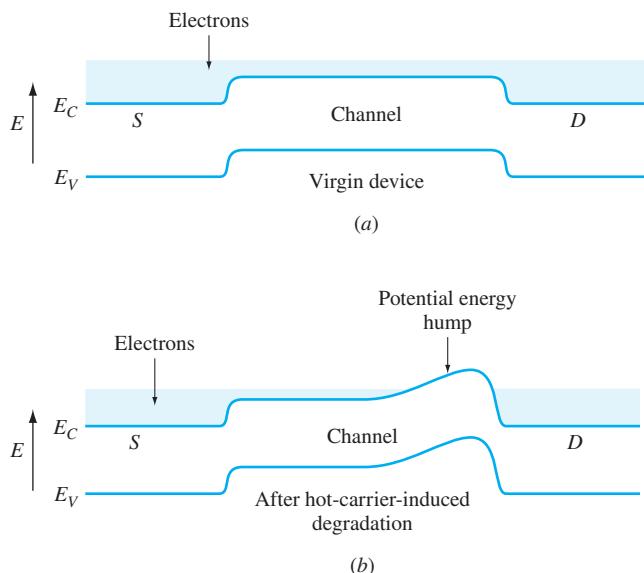


Figure S3.22 Effect of hot-carrier stress on a MOSFET energy band diagram along the channel of a MOSFET for $V_{DS} = 0$ and V_{GS} large enough to create a channel. (a) The virgin device just off the shelf, and (b) the device after long operation in the forward mode. For the stressed device, no channel exists near the drain.

Figure S3.23 indicates what happens in the stressed device as the drain voltage V_{DS} is varied. The energy band diagram is drawn for a small value of V_{DS} (a) and for a larger V_{DS} (b). At small V_{DS} the hump prevents current flow. In (b), the drain voltage is large enough that the hump has a negligible effect on the current, so that at large V_{DS} there is not a noticeable aging effect. For intermediate values of V_{DS} , however, the presence of the hump reduces the current from that of a virgin device.

Recall that MOSFETs are normally fabricated such that the source and drain are interchangeable. On a virgin device, therefore, the electrical characteristics are the same for forward and inverse operation. The hot-carrier-induced potential hump in the channel near the drain, however, causes an asymmetry in the I_D-V_{DS} characteristics. The energy band diagram for inverse operation is shown in Figure S3.24. Here the polarity of the drain voltage is reversed from the forward operation of Figure S3.23. Because the hump is near the acting source (S), the positive voltage on the acting drain (D) has little effect on the size of the barrier. This means that the threshold voltage for inverse operation is increased.

Figure S3.25 shows the I_D-V_{DS} characteristics for forward and inverse operation of a MOSFET in its virgin state and after stress. The forward characteristics are plotted in the first quadrant while the inverse characteristics are in the third quadrant. The solid lines indicate the characteristics of the virgin device, while the dashed lines show the characteristics after prolonged operation in the forward mode. The I_D-V_{DS} characteristics for the virgin device are symmetric in forward and inverse operation, but that is not true for the stressed device.

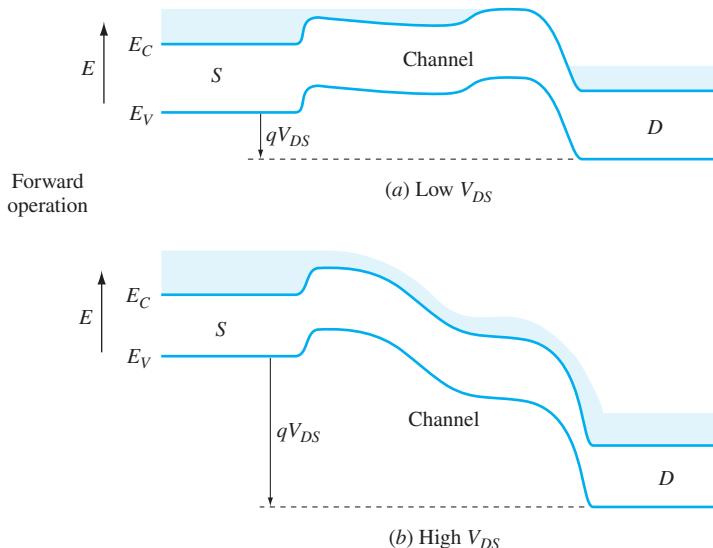


Figure S3.23 Energy band diagrams for stressed device operating in the forward mode. (a) For small V_{DS} , I_D is reduced from that of a virgin MOSFET. (b) At high V_{DS} , the hot-carrier-induced hump has little effect on current flow.

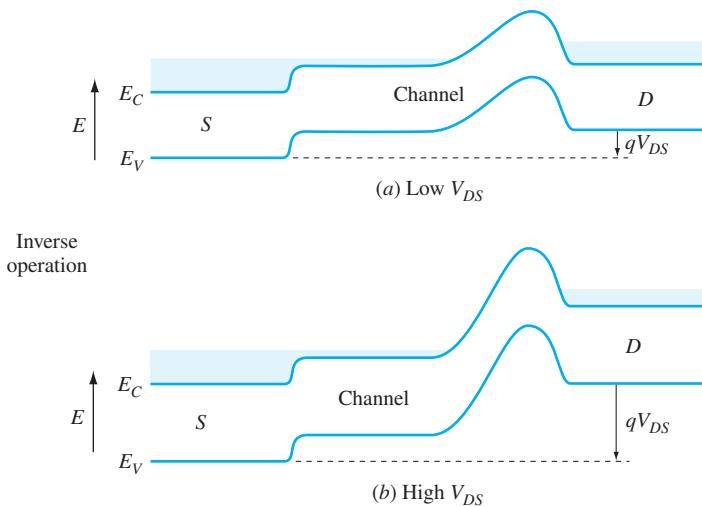


Figure S3.24 Energy band diagrams for inverse operation (drain voltage polarity is inverted) for a stressed device. In this mode of operation the actual drain is effectively the source and the actual source is effectively the drain. Since the potential energy hump is near the acting source, the voltage on the acting drain is not enough to create a conducting channel. In effect, the threshold voltage V_T has increased.

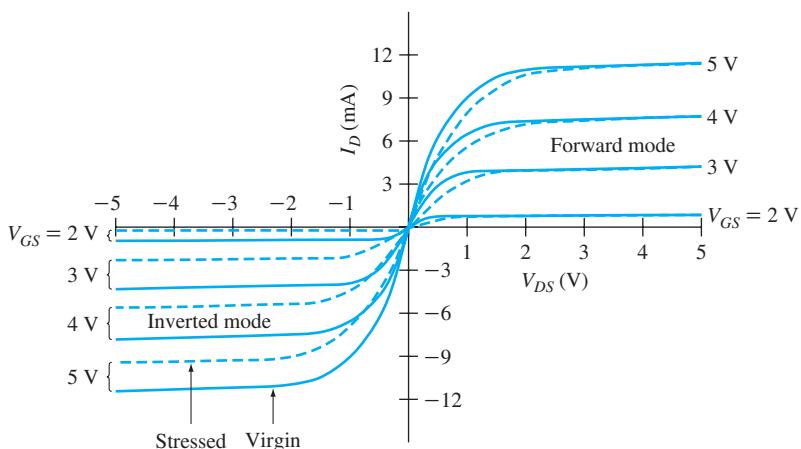


Figure S3.25 The I_D - V_{DS} characteristics of a virgin device (solid lines) and a hot-carrier-stressed device (dashed lines) for forward and inverse operation. The reduction in V_T and thus I_D is apparent for inverse operation.

The characteristics change for both forward and inverse operation. For forward operation with small V_{DS} , the drain current I_D is reduced somewhat from its virgin value. At large enough V_{DS} , in the saturation region, the current is equal to that before stress, and the device appears to operate normally.

Next, compare the I_D-V_{DS} characteristic at $V_{GS} = 2$ V in Figure S3.25. In the forward mode, the saturation current is the same for the virgin and stressed device. In the inverse mode, however, even with a large V_{DS} , no conducting channel exists. To initiate conduction a larger V_{GS} is required. Put another way, V_T is increased, which results in reduced current at all drain-to-source voltages.

There are some circuits in which the transistors operate in both forward and inverse modes. These circuits (e.g., bilateral transmission gates) are more susceptible to the effects of hot-carrier-induced degradation.

Hot-carrier-induced degradation is increasingly pronounced with decreasing L . This is because, in a short channel, the high longitudinal field region at the drain end reaches closer to the source. The potential energy hump extends over a larger fraction of the channel.

As expected, hot-carrier-induced degradation is reduced by the use of smaller drain voltages.

*LIGHTLY DOPED DRAIN (LDD) MOSFETS

The hot-carrier degradation in MOSFETs discussed in the previous section can be reduced by reducing the maximum longitudinal field \mathcal{E}_L . One way to reduce the field is to use smaller drain voltages, as mentioned earlier. Another method is to use a structure known as a lightly doped drain (LDD).

In the LDD MOSFET, there is a more lightly doped n region between the n⁺ drain and the channel, as shown schematically in Figure S3.26. The more lightly doped region is also more resistive, and part of the drain voltage is dropped across it. This reduces the field in the channel near the drain. Because it is easier from a fabrication standpoint to make the transistors symmetric, a similar lightly doped region is present between the n⁺ source and channel. This region, however, serves no useful purpose.

There is a trade-off. These lightly doped regions introduce extra series resistance in the source and drain, which degrades the I_D-V_{DS} characteristics as discussed in Section 7.3.3. In modern MOSFETs with reduced voltages, LDDs are not necessary, and the device of Figure S3.26 is replaced by that of Figure S3.20.

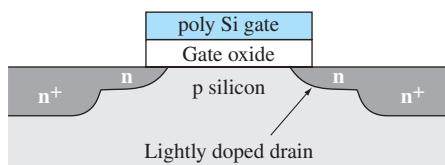


Figure S3.26 Cross section of the lightly doped drain (LDD) MOSFET.

S3.9.1 MOSFET INTEGRATED CIRCUIT RELIABILITY

With scaling, as devices get smaller, individual defects become increasingly important in characteristics variation and reliability degradation. Both devices and interconnects contribute to failure. Early failures are due to defects introduced during processing. Devices made inoperable by these early failures can be identified and eliminated by a process called *burn in*, in which the chips are subjected to temperature, humidity, and bias cycling for several hours. After the early failures, the rate of failures is small and approximately constant for a time on the order of 10 to 20 years, after which the materials wear out. This is indicated in Figure S3.27, which shows the variation of failure rate with time on a logarithmic scale. This distribution is referred to as the *bathtub curve*.

MOSFET Variability and Degradation Mechanisms Most of the variability and degradation mechanisms in MOSFETs can be traced to the gate structures and are caused by trapping and de-trapping of carriers (electrons and holes), either in the states in the forbidden gap of the oxide layer or in defect states at the oxide-silicon interface. Some of these states result from the fabrication processes. Because the various materials have differing coefficients of thermal expansion, mechanical stresses occur as the chip is cooled from the high temperatures of fabrication to room temperature. Still other charges are created by injection of energetic carriers into the gate oxide.

Random Telegraph Noise (RTN) With the thin gate oxides of modern MOSFETs, the presence or absence of electrons in the oxide or interface states results in a change in the channel potential near the defect and thus affects the channel current. For n-channel MOSFETs, the current is decreased when an electron is trapped in a state, and the current rises again when the electron is de-trapped. This fluctuating current creates noise in analog circuits (random telegraph noise or RTN) and in digital circuits it affects the minimum operating voltage. This is especially important in SRAM circuits.

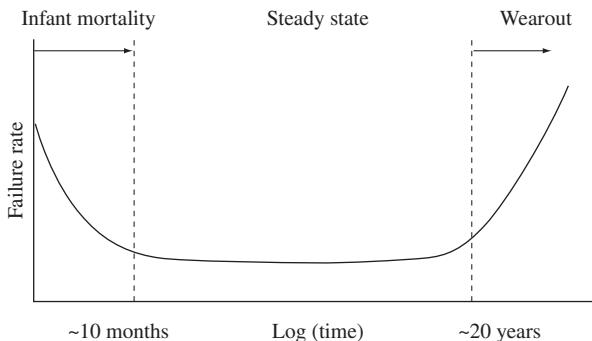


Figure S3.27 Bathtub curve for IC failure rates

Gate Oxide Reliability In addition to RTN, the process of trapping and detrapping can create defects in the oxide that can permanently trap carriers. With a sufficient concentration of traps, the threshold voltage is altered. Traps can also be created by electrons (or holes) tunneling into the oxide in the high-field channel region near the drain, or they can be produced by “soft errors” caused by energetic particles from radioactive decay of the package materials (e.g., metals) or from cosmic rays. This is referred to as *time-dependent dielectric breakdown (TDDB)*.

Creation of interface state traps at the oxide-silicon interface can cause “interface drain current” by electrons jumping from trap to trap, thus from source to drain.

Interconnect Degradation As with the degradation of the gate oxide dielectrics, the insulating dielectrics between metallic interconnect layers can break down with time (TDDB). Thin-film metallic interconnects carry high current densities, on the order of 10^6 A/cm^2 . The electrons flowing in the conducting films collide with the metal film atoms, thus transferring momentum to the atoms and moving some atoms, primarily at grain boundaries, in the direction of electron flow. This atomic movement (electromigration) can cause the metal width to decrease. This thinner region causes locally increased current density, which in turn increases atomic movement. The eventual result is an open circuit or void. Originally the interconnect metal was aluminum. To reduce this electromigration, the aluminum films were doped with copper, which tended to segregate at the grain boundaries. Because copper atoms are much heavier than aluminum atoms, they are less affected by the electron collisions. More recently, copper has replaced aluminum interconnects. Since the copper atoms are less affected by the electron collisions than aluminum, copper interconnects permit a greater current density. To further reduce electromigration, the copper lines are often covered by refractory cladding layers (e.g., TiN) to reduce the surface atomic movements.

S3.10 SUMMARY

This supplement amplifies some of the topics presented in Chapters 7 and 8. In those chapters, we presented various models from which to find the current in a MOSFET. In the Supplement to Part 3, we refined the formulation for the channel charge Q_{ch} to include the effects of the variations of depletion region charge along the channel and the dependence of Q_{ch} on the longitudinal field \mathcal{E}_L .

In the earlier chapters, the threshold voltage was treated as an experimentally measured quantity whose value for a given process was assumed to be known. In this supplement, the threshold voltage was related to physical parameters. Its dependence on the work function difference between the gate and the substrate, Φ_{MS} , was investigated, as well as the effects of various charges, including charges in the oxide, charges at the oxide-semiconductor interface, and charges within the semiconductor depletion region. Adjustment of the threshold voltage to its desired value is achieved by ion implantation of impurities in the semiconductor near its interface with the oxide.

An equivalent circuit for MOSFETs in analog circuits was presented. With decreasing channel length, the frequency response is increased.

Next, we examined the MOS capacitor, a useful circuit element and valuable diagnostic tool. The capacitance-voltage characteristics in particular are useful to measure material properties resulting from the fabrication processes, including the concentration of interface states, the depletion width, the oxide thickness, and the substrate doping. In the category of capacitance-related devices, dynamic random-access memories (DRAMs) were briefly described.

Methods for reducing the size (scaling) of MOSFETs were discussed. Reducing device size permits more chips of a given functionality on a wafer, thus reducing the cost per chip. Smaller devices also reduce carrier transit time from source to drain, resulting in reduced switching time and higher frequency response. Problems associated with scaling were discussed along with solutions to minimize these problems.

Mechanisms responsible for device degradation were also briefly examined. These include thermally activated mechanisms such as oxide breakdown and electromigration, as well as hot-carrier-induced degradation resulting from operation at high lateral channel fields.

S3.11 REFERENCES

1. Dennis Hoyniak, Edward Nowak, and Richard L. Anderson, “Channel electron mobility dependence on lateral electric field in field-effect transistors,” *J. Appl. Phys.*, 87, pp. 876–881, 2000.
2. Kangguo Cheng, Jinju Lee, Karl Hess, Joseph W. Lyding, Young-Kwang Kim, Young-Wug Kim, and Kwang-Pyuk Suh, “Improved hot-carrier reliability of SOI transistors by deuterium passivation of defects at oxide/silicon interfaces,” *IEEE Trans. Electron Devices*, ED-49, pp. 529–531, 2002.
3. F. Stern and W. E. Howard, “Properties of semiconductor surface inversion layers in the quantum limit,” *Phys. Rev.*, 163, p. 816, 1967.
4. F. Stern, “Quantum properties of surface space charge layers,” *CRC Crit. Rev. Solid-State Sci.*, 4, p. 499, 1974.
5. R. H. Dennard, “Scaling challenges for DRAM and microprocessors in the 21st century,” *Electrochemical Society Proceedings*, 97-3, pp. 519–532, 1997.
6. Cor Claeys, et. al. “Advanced semiconductor devices for future CMOS technologies,” *Electrochemical Society Transactions*, 66, no. 5, pp. 49–60, 2015.
7. Anthony S. Oates, Richard C. Blish, Gennadi Bersuker, and L. Kasprzak, “Reliability of electron devices, interconnects and circuits,” Chapter 9 in *Guide to State-of-the-Art Electron Devices*, Joachim N. Burghartz, ed., Wiley-IEEE Press, Chichester, West Sussex, 2013.

S3.12 REVIEW QUESTIONS

1. Why should we realistically expect the depletion region charge adjacent to the channel to vary with position along the channel? When this is taken into account, what effect does it have on the current-voltage characteristics of a MOSFET?
2. Explain how the presence of fixed oxide charge can influence the value of the threshold voltage. What other charges must also be considered?
3. What is meant by flat band voltage?
4. How is the threshold voltage of MOSFETs controlled after fabrication?
5. When the channel is very narrow, quantum mechanical effects come into play. What is the result on the allowed energies for electrons in the channel?
6. Explain qualitatively how the I_D - V_{GS} relationship can be used to find the low-field mobility and its dependence on the gate voltage.
7. Explain how C - V measurements can be used to experimentally find various parameters of the process. Which parameters can be obtained with this technique?
8. Explain the operation of a DRAM cell.
9. What is meant by hot-carrier-induced degradation? What effect does it have on the MOSFET operation? How can it be mitigated?
10. What is meant by transconductance?
11. What is meant by the current gain cutoff frequency f_T of a transistor?
12. What is constant field scaling in MOSFETs? What parameters are involved?
13. Why does the frequency response of a MOSFET increase as the channel length is decreased?
14. To increase the threshold voltage of an n-channel MOSFET, would donors or acceptors be implanted? Why?
15. What is the purpose of the n^{++} drain extension region in an n-channel MOSFET?
16. What is the purpose of the “halo” in the device of Figure S3.20?

S3.13 PROBLEMS

- S3.1** A Si NMOS with $t_{ox} = 2 \text{ nm}$ has a threshold voltage of 0.2 V. What is the required implant dose (atoms/cm^2) to increase the threshold voltage to 0.5 V?
- S3.2** Adjacent MOSFETs are isolated from each other by a thick *field oxide* region as indicated in Figure PS3.1. The gate oxide thickness is 5 nm and the field oxide region is $0.5 \mu\text{m}$ thick. Both gate and interconnect are of the same metal with $\Phi_{MS} = -0.4 \text{ eV}$. The Si net doping concentration

is 10^{16} cm^{-3} . The threshold voltage for the NMOS is 0.5 V. What is the threshold voltage for the parasitic NMOS between the active devices?

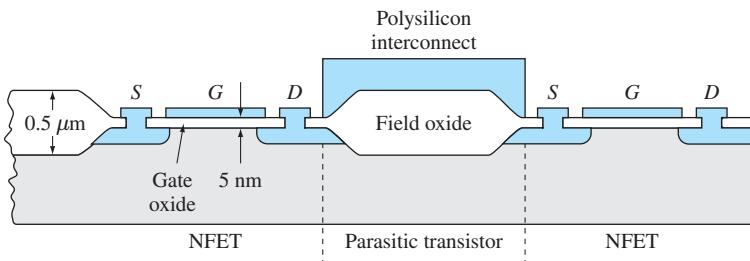


Figure PS3.1

S3.3 If the amount of positive charge Q_f in Figure S3.3 is increased, what will be the effect on V_T ? Explain the physics of your answer.

S3.4 Find V_T for a silicon n-channel FET with the following parameters:

$$N_{D\text{gate}} = 10^{20} \text{ cm}^{-3}$$

$$N_A = 10^{17} \text{ cm}^{-3} \text{ in the channel}$$

$$t_{\text{ox}} = 4 \text{ nm}$$

$$Q_f = q \times 5 \times 10^{10} \text{ C/cm}^2$$

$$Q_I = q \times 10^{10} \text{ C/cm}^2$$

- Is this an enhancement or depletion device? Assume that the emitter Fermi level is at the apparent conduction band edge.
- What is the result if bandgap narrowing is neglected?

S3.5 For the device of Example S3.1, what is the minimum value the combined fixed charge and interface charge can be and still ensure an enhancement device? What is the corresponding density of charges?

S3.6 A MOSFET process produces

$$\Phi_{MS} = -0.5 \text{ eV}$$

$$C'_{\text{ox}} = 6.9 \times 10^{-7} \text{ F/cm}^2$$

$$Q_f = q \times 10^{10} \text{ C/cm}^2$$

$$Q_{\text{it}}(2\phi_f) = q \times 10^{10} \text{ C/cm}^2$$

$$Q_B(2\phi_f) = -5 \times 10^{-8} \text{ C/cm}^2$$

$$\phi_f = 0.4 \text{ eV}$$

What should the ion implantation be to make this an enhancement mode device intended to operate in a 3.3 V operating circuit?

S3.7 Consider a MOS capacitor where the gate is degenerate n⁺ Si and the substrate is n-type Si with $N'_D = 10^{16} \text{ cm}^{-3}$. Except for the difference in work functions of gate and substrate, the capacitor can be considered to be ideal. Neglect band-gap narrowing in the degenerate gate.

- What is the built-in voltage of the device?
- Sketch the equilibrium energy band normal to the gate.
- What is the flat band voltage?
- Sketch the energy band diagram at flat band.
- Sketch the low-frequency and the high-frequency $C-V_G$ characteristics of the device.
- Sketch the charge distribution as a function of position for $V_G = 0$, $V_G = +5$ V, and $V_G = -5$ V.

S3.8 The $C-V_G$ characteristics of an (ideal) MOS capacitor of area 10^{-3} cm 2 is shown in Figure PS3.2.

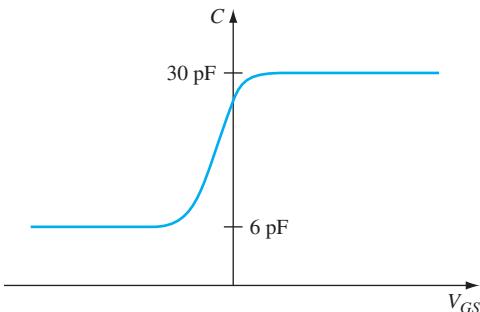
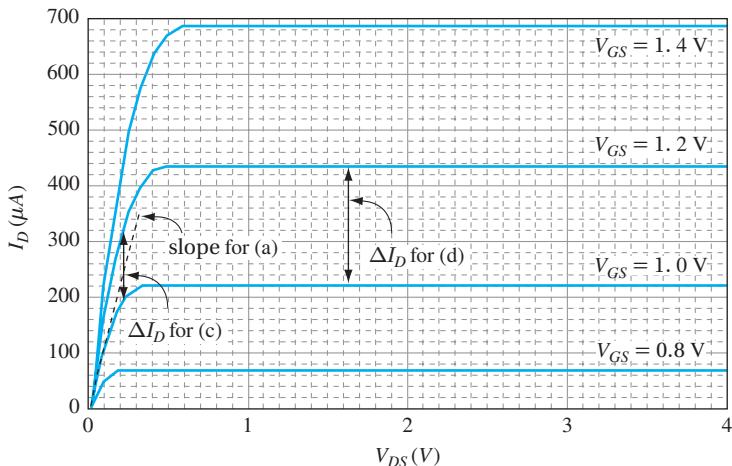


Figure PS3.2

- Is the measurement made at high frequency or low frequency?
 - Is the semiconductor n type or p type?
 - What is the oxide thickness?
 - What is the net doping concentration of the semiconductor?
- S3.9** Explain how a logic 1 and logic 0 can be written into a DRAM cell if the initial state of the cell is a 1 instead of 0 as assumed in the text.
- S3.10** A CMOS circuit must operate at 500 MHz. What is the maximum channel length that can be tolerated, neglecting short-channel effects?
- S3.11** Practice using circuit models. Suppose a FET was found to have a finite input resistance, R_{in} . How would you include this in the model of Figure S3.8a? How would you include the channel resistance?
- S3.12** For the transistor whose I_D-V_{DS} characteristics are shown in Figure PS3.3, find the small signal parameters:
- g_d for $V_{GS} = 1.0$ V, $V_{DS} < V_{DSSat}$
 - g_d for $V_{DS} > V_{DSSat}$, $V_{GS} = 1.1$ V
 - g_m for $V_{DS} = 0.2$ V, $V_{GS} = 1.1$ V
 - g_{msat} for $V_{DS} > V_{DSSat}$, $V_{GS} = 1.1$ V

**Figure PS3.3**

- S3.13** For thermally activated failure mechanisms, the mean time to failure is given by Equation (S3.47) $MTTF \propto e^{\frac{E_a}{kT}}$, where E_a is the activation energy. Find the ratio of the MTTF for electromigration ($E_a = 0.7$ eV) at 50°C and 100°C compared to that at room temperature (27°C).
- S3.14** Why is constant field scaling impractical as devices get very small?
- S3.15** Obtain expressions for transconductance g_{msat} for three MOSFET models: the long-channel drift model, the short-channel drift model as $L \rightarrow 0$, and the ballistic model.
- S3.16** Compare the transconductance per unit gate width, $\frac{g_{msat}}{W}$, for the drift model with $L = 5 \mu\text{m}$, for the drift model as $L \rightarrow \infty$, and for the ballistic model. Assume that $\mu_{lf} = 400 \text{ cm}^2/\text{V} \cdot \text{s}$ for the device with $L = 5 \text{ nm}$, $250 \text{ cm}^2/\text{V} \cdot \text{s}$ for the other two devices. Let $v_{sat} = 10^7 \text{ cm/s}$ and $v_T = 1.05 \times 10^7 \text{ cm/s}$. In all cases the gate oxide is SiON with $\epsilon_r = 5$ and oxide thickness of 2 nm.

Bipolar Junction Transistors

The preceding chapters dealt with field-effect transistors, which are unipolar devices, i.e., devices in which only one type of carrier, electrons or holes, contributes to current flow in the device. In this section, we discuss bipolar devices, in which both electrons and holes must be considered.

Like a FET, a bipolar junction transistor (BJT) is a two-junction, three-terminal device in which one terminal controls the current flow between the other two terminals. The terminology is different, however. In Table IV.1, the analogous regions of a BJT and a FET are compared.

In a FET, the source acts as a source of carriers, which are emitted into the channel where they flow into the drain. In a BJT, the emitter acts as the carrier source and emits carriers into the base where they flow to the collector and are collected. In both cases, the voltage at the control electrode (gate or base), determines the number of carriers available to flow from source to drain or from emitter to collector. The physics of the control mechanism differs between a FET and a BJT, however. In a FET, the control (gate) voltage is capacitively coupled to the channel. In a BJT, a direct connection is made to the base. Another distinction is that in a FET, carrier flow from source to drain is primarily by drift. In a BJT, carrier flow from emitter to collector is at least in part by diffusion.

Figure IV.1 indicates the (planar) structure of an integrated circuit (IC) npn (emitter-base-collector) BJT in which all contacts, the emitter (*E*), base (*B*) and collector (*C*) are made from the top.¹ In (a) the top view is indicated and in (b) the cross-sectional schematic of the device is shown. The structure of Figure IV.1 is representative of BJTs of the mid 1970s. Because of its relative simplicity, it will be used to develop the basic operating principles of bipolar junction transistors. The structure and operation of more modern devices is discussed later. Note that this integrated circuit transistor exists in an n-type *well* in a p-type substrate.

¹This is in contrast to a discrete BJT in which the collector contact is usually made from the bottom.

Table IV.1 Comparison of terminology of a BJT with that of a FET

FET	BJT
Source	Emitter
Drain	Collector
Channel (gate electrode)	Base

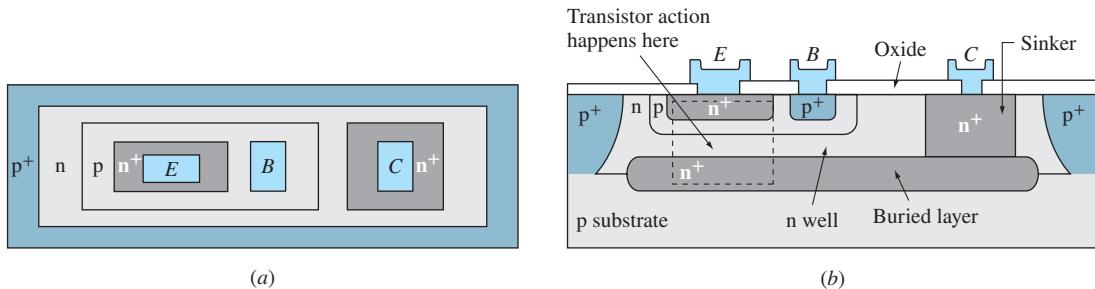


Figure IV.1 Schematic diagram of an npn bipolar transistor used in integrated circuits. (a) Top view and (b) cross-sectional view.

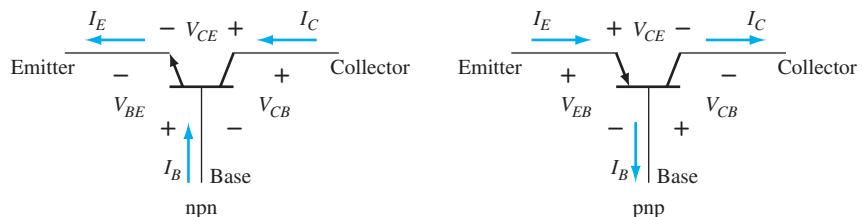


Figure IV.2 Circuit symbols for npn and pnp BJTs. The arrow in the emitter indicates the direction of current flow. The polarities of the voltages are those for bias in the active mode.

The p-type substrate is connected to the most negative voltage of the circuit such that the junctions between the n wells and the p substrate are never forward biased. In this way, the various transistors on a chip are isolated from each other. The transistor action, which consists of electrons being injected from the emitter into the base and ending up in the collector, occurs in the region indicated in the figure. The buried n⁺ layer provides a low-resistance path from the collector region underneath the emitter to the collector contact on the right.

The circuit symbols for npn and pnp devices are indicated in Figure IV.2, where the various voltage differences and currents are indicated. From Kirchhoff's voltage law,

$$V_{CE} = V_{CB} + V_{BE} \quad (\text{IV.1})$$

and from Kirchhoff's current law,

$$I_E = I_C + I_B \quad (\text{IV.2})$$

The arrow in the emitter of the circuit symbol represents the direction of current flow in the active mode, thus indicating whether it is npn or pnp.

Figure IV.3 indicates two common circuit configurations for an npn BJT, along with the input and output voltages and currents. In (a), the base is common to both input and output and so is referred to as the *common-base configuration*. In (b), the *common-emitter configuration*, the emitter is common to both input and output.

Figure IV.4a and b show the I_C - V_{CE} characteristics for the npn and pnp transistors operating in the common-emitter configuration. Qualitatively, they are similar to those of the FET. One key difference, however, is that the different curves correspond to different values of base *current*, whereas in the FET the varying parameter was the gate *voltage*. We will see why later.

With two junctions, each of which can be forward or reverse biased, there are four possible modes of operation, as indicated in Table IV.2. The analogous bias regimes for the FET are shown in the last column. The forward and saturation bias regimes are also shown in Figure IV.4.

These biasing modes are used in circuits in various ways, as shown in Figure IV.4c and d. In digital circuits, for example, a BJT is operated in two regions:

Table IV.2 Biasing modes for a BJT

Biasing polarity and typical values for Si BJTs			
Biasing mode	E-B junction	C-B junction	FET analog
Active	Forward (0.7 V)	Reverse (5 V)	Current saturation
Voltage saturation	Forward (0.7 V)	Forward (0.5 V)	Sublinear
Inverted	Reverse (5 V)	Forward (0.7 V)	Inverted
Cutoff	Reverse, zero, or weakly forward (0 V)	Reverse (5 V)	Cutoff

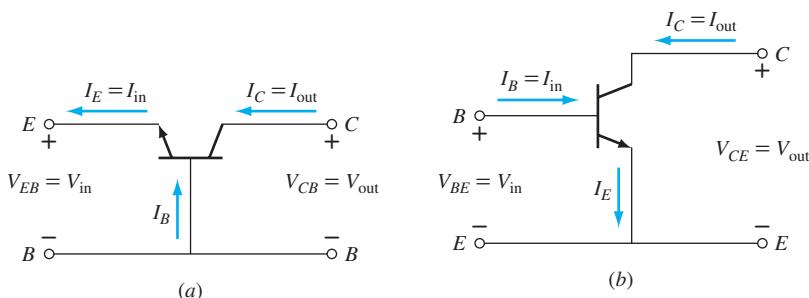


Figure IV.3 (a) Common-base and (b) common-emitter configurations for an npn BJT.

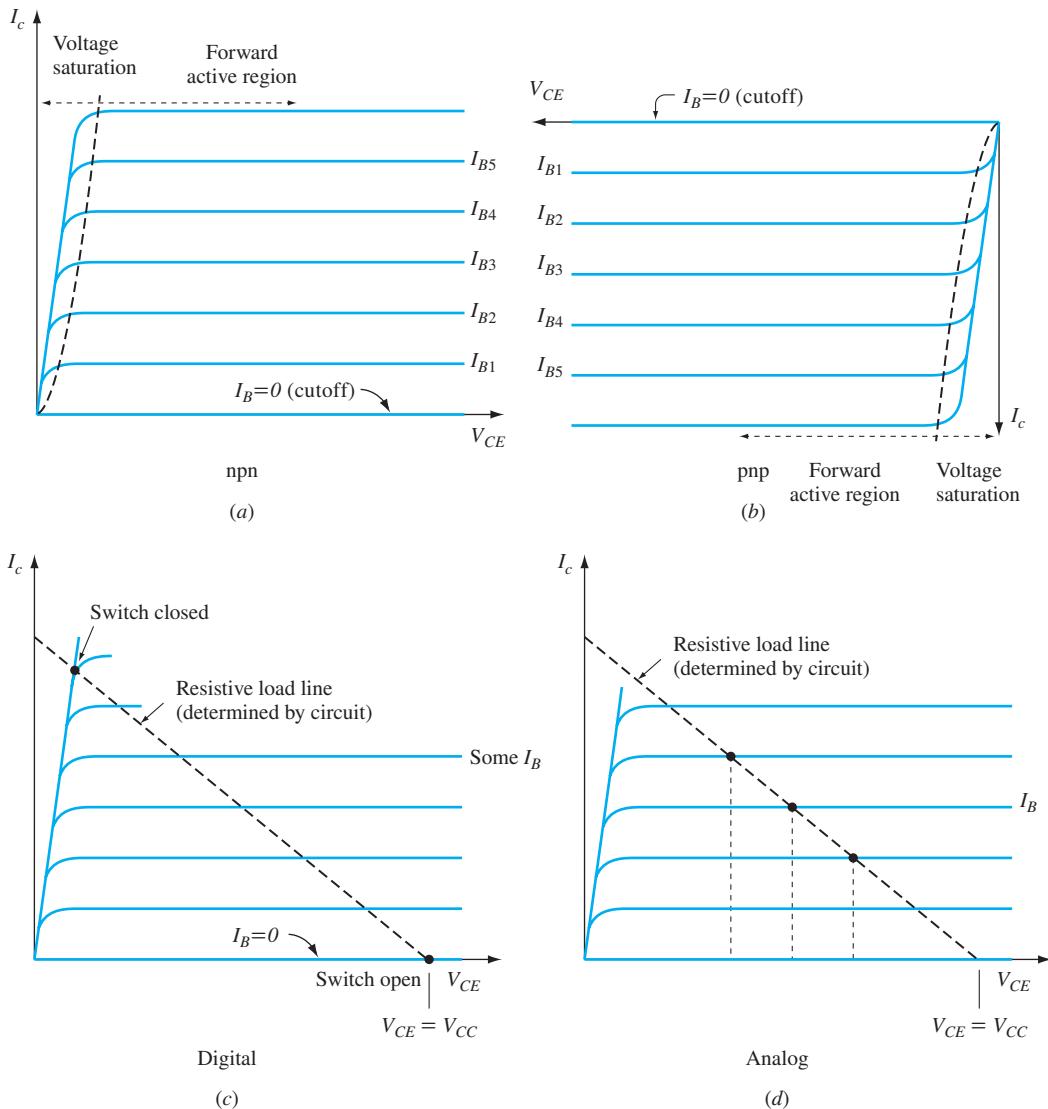


Figure IV.4 Typical I_c - V_{CE} curves for (a) an npn transistor and (b) a pnp transistor in the common-emitter configuration. An npn BJT can be used in a digital circuit (c) or an analog circuit (d).

the low-voltage, high-current *saturation* mode, corresponding to the **on** state, and in the *cutoff* mode, which is a low-current, high-voltage **off** state. In the **on** state, both the emitter-base and base-collector junctions are forward biased, and in cutoff they are both reverse biased. The switching behavior of BJTs is discussed in Chapter 10. In analog circuits, the BJT is operated in the *forward active* mode. In this mode, the emitter-base junction is forward biased and the base-collector junction is reverse biased.

The electrical characteristics of a BJT are obtained by solving the continuity equations with appropriate boundary and initial conditions in each region of the device. These equations are for zero optical generation (i.e., in the dark).

$$\frac{\partial n}{\partial t} = \frac{\partial \Delta n}{\partial t} = \frac{1}{q} \frac{\partial J_n}{\partial x} - \frac{\Delta n}{\tau_n} \quad (\text{IV.3})$$

$$\frac{\partial p}{\partial t} = \frac{\partial \Delta p}{\partial t} = -\frac{1}{q} \frac{\partial J_p}{\partial x} - \frac{\Delta p}{\tau_p} \quad (\text{IV.4})$$

where, in general, J_n or J_p is the sum of the respective drift and diffusion current densities.

$$\begin{aligned} J_n &= J_{n\text{drift}} + J_{n\text{diff}} \\ J_p &= J_{p\text{drift}} + J_{p\text{diff}} \end{aligned}$$

To obtain the steady-state (dc) characteristics, the time-dependent terms are set to zero. For transient analysis they must be considered.

In Chapter 9, the dc model of BJT operation is developed. First, the ideal (prototype) BJT is considered, in which the doping is uniform in each region—emitter, base, and collector. Later, the effects of nonideal parameters are discussed, and the graded-doping base transistor is introduced along with heterojunction bipolar transistors, including the SiGe-base graded-base transistor. In Chapter 10, we examine the ac operation of a BJT. We introduce the Ebers-Moll model and the hybrid pi small-signal model. We examine BJT capacitance and transient behavior. Chapter 10 also describes some specific BJTs, including the double-poly (polysilicon) self-aligned bipolar junction transistor, BJT switching transistors, and BiMOS, a hybrid technology that involves both BJTs and FETs. This technology combines the advantages of FETs with those of BJTs. ■

9

CHAPTER

Bipolar Junction Transistors: Statics

9.1 INTRODUCTION

In the introduction to Part 4, we described the basic bipolar junction transistor (BJT). We saw that the current-voltage characteristics are similar to those of a field-effect transistor, in that the current flowing between two terminals is controlled by a control signal applied to the third terminal. In a FET the control signal is the voltage applied to the gate, but in a BJT the control signal is the current applied to the base. They are similar, however, in the sense that in both cases carriers have to surmount a potential barrier for current to flow. In the FET, the voltage applied to the gate changes the source-channel barrier. In the BJT, the base-emitter voltage controls the barrier. Although V_{BE} controls the barrier height, it is directly related to the base current. It is convenient to consider the base current as the control signal.

Recall that in a pn junction under forward bias, minority carriers are injected across the junction, and the number of carriers injected depends on the bias voltage. In an npn BJT in the active mode, the (forward-biased) base-emitter current controls the electron flux F_n injected from the emitter into the base. This flux is indicated in Figure 9.1. The electron flux is the number of electrons crossing an area per unit time. Current is equal to the product of the carrier charge and the flux. The electron current flows in the direction opposite to the electron flux, but hole current flows in the same direction as hole flux.

In realistic transistors, the base is currently made quite thin (tenths of a micrometer), while the emitter area dimensions are on the order of micrometers. A thin, or short, base is essential to the efficient operation of a BJT, as will be seen later.

Recall from Figure IV.1 that the transistor action occurs between the emitter, base, and collector under the emitter region. The actual transistor is much larger to permit electrical contact to the emitter, base, and collector regions. For clarity in discussing the ideal transistor, the “box” diagram is used as indicated in Figure 9.2a, where the base thickness is exaggerated for clarity.

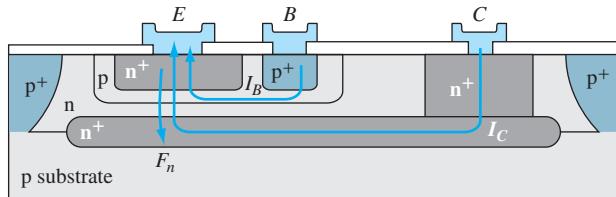


Figure 9.1 Illustration of electron flux F_n and currents of an npn BJT.

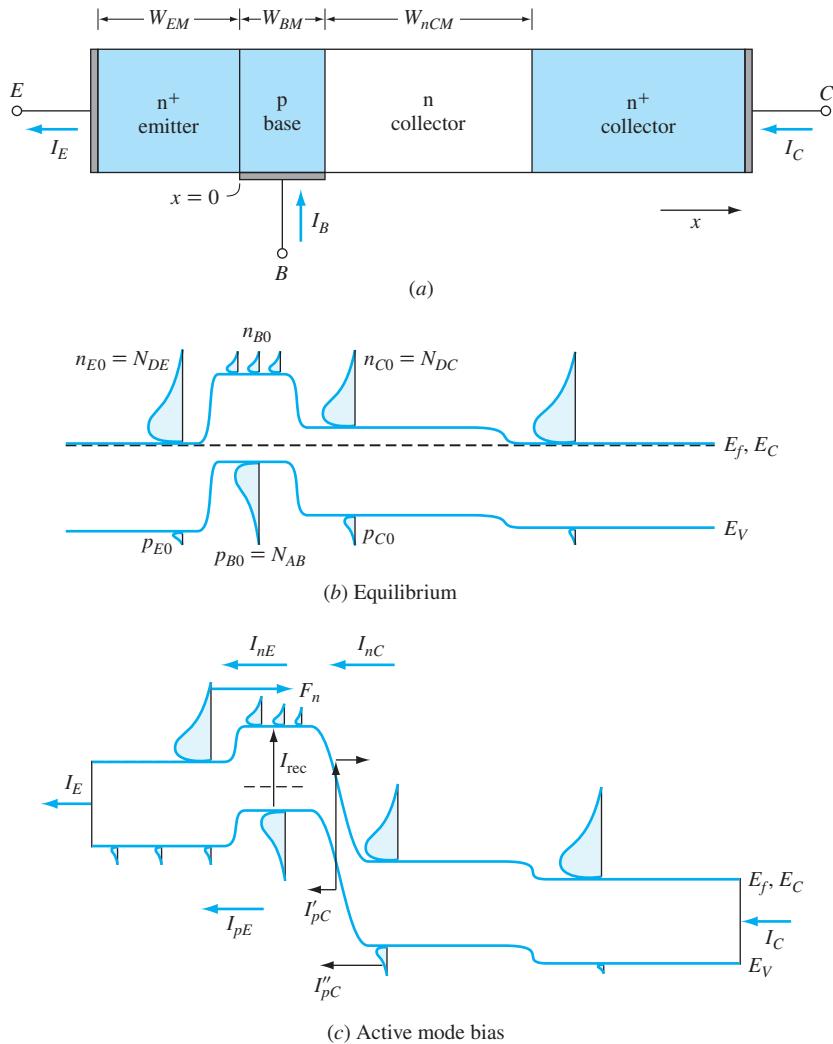


Figure 9.2 (a) A cut along the current path from emitter to collector for the npn device of Figure 9.1. (b) The equilibrium energy band diagram for a BJT with uniform doping in each region. (c) The energy band diagram for operation in the active mode.

The primary current flow in an npn BJT results from electrons from the emitter contact flowing through the emitter, over the emitter-base potential barrier, then through the base, then into the n and n⁺ regions of the collector and out the collector contact. The width of the emitter region (between the metallurgical junctions) is W_{EM} , that of the base region is W_{BM} , and that of the n collector region is W_{nCM} .

The best way to understand the physics of any semiconductor device is by use of the energy band diagram. The equilibrium energy band diagram is shown in Figure 9.2b. To construct this energy band diagram, we make the following assumptions:

1. The doping in the n⁺ emitter is degenerate and the emitter is uniformly doped.
2. The doping in the p-type base is nondegenerate and uniform (except for the p⁺ region under the base contact, which is not in this energy band diagram).
3. The n collector region is nondegenerate and uniformly doped.
4. The degenerate n⁺ collector region is uniformly doped.
5. Impurity-induced band-gap narrowing is neglected.
6. The Fermi level coincides with the conduction band edge in n⁺ emitter and n⁺ collector.

We refer to a BJT with uniform doping in each region as a *prototype* transistor. The energy band diagram of this prototype BJT under normal (analog) operating conditions (active mode) is shown in Figure 9.2c. The emitter-base junction is forward biased, which permits electrons to be injected from the emitter into the base and holes to be injected from the base into the emitter. The base-collector junction is reverse biased. For n_E electrons injected into the base from the emitter, a small fraction recombine in the base and contribute to base current I_B . Most of the electrons injected from the emitter into the base reach the collector-base junction and are collected and contribute to collector current I_C . Since $I_C < I_E$, the current gain $\alpha = I_C/I_E$ from emitter to collector is less than unity. However, most circuits use the common-emitter configuration in which the input current is I_B and the output current is I_C . The current gain then is $\beta = I_C/I_B$, which is typically on the order of 100. Note that the electrons, which are minority carriers, in this prototype BJT flow across the base only by diffusion since there is no electric field in the base, or

$$\mathcal{E} = \frac{1}{q} \frac{dE_{V_{AC}}}{dx} = \frac{1}{q} \frac{dE_C}{dx} = 0$$

In Figure 9.2c the emitter current I_E consists of the current resulting from electron injection from emitter to base, I_{nE} , plus the hole current injected from base to emitter, I_{pE} :

$$I_E = I_{nE} + I_{pE} \quad (9.1)$$

Note that for this example, the recombination current within the E-B junction is neglected. This recombination current is considered later.

The collector current I_C is composed of an electron component and a hole component. The electron current I_{nC} is the electron current resulting from the electrons that cross the base from the emitter and reach the collector. The hole

current I_{pC} has two components: the hole current, I''_{pC} extracted from the collector into the base, and I'_{pC} , the current due to the electron-hole generation in the reverse-biased B-C junction. Thus

$$I_C = I_{nC} + I_{pC} \quad (9.2)$$

where $I_{pC} = I'_{pC} + I''_{pC}$.

The collector electron current I_{nC} is the emitter electron current injected into the base minus the current lost to recombination in the base:

$$I_{nC} = I_{nE} - I_{\text{rec}} \quad (9.3)$$

where I_{rec} is the current due to electron-hole recombination within the base.

The current into the base, I_B , is then

$$I_B = I_{pE} + I_{\text{rec}} - I_{pC} \quad (9.4)$$

Some general observations concerning the relative magnitudes of the various current components can be qualitatively determined.

Since the emitter is much more heavily doped than the base, we know that

$$I_{nE} \gg I_{pE}$$

Further, recalling that the base is made thin enough that its width is much less than an electron diffusion length, the probability of electrons recombining within the base is small, and

$$I_{\text{rec}} \ll I_{nE}$$

Thus

$$I_{nC} \approx I_{nE}$$

In the active mode, I_C is on the order of milliamperes, but the leakage current for a reverse-biased pn junction is several orders of magnitude smaller than this. Thus, I_{pC} is small and

$$I_C \approx I_{nC}$$

The relative magnitudes of the preceding currents are shown in Figure 9.3, where I_{pE} , I_{pC} , and I_{rec} are exaggerated for clarity.

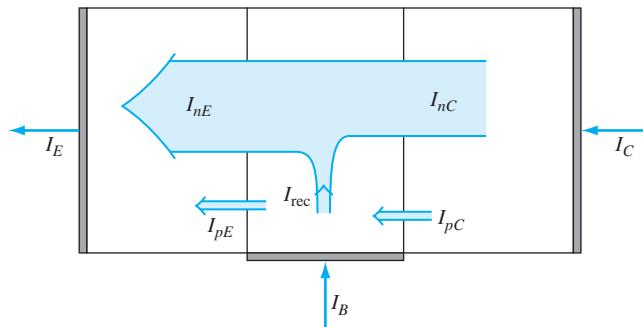


Figure 9.3 Schematic indicating the relative magnitudes (not to scale) of the various current components in an npn BJT operating in the active mode.

9.2 OUTPUT CHARACTERISTICS (QUALITATIVE)

A typical (idealized) family of curves for the output characteristics of an npn BJT operated in the *common-base configuration* with input (emitter) current as a parameter is shown in Figure 9.4a, and for the *common-emitter configuration* in Figure 9.4b with the input (base) current as a parameter. Figure 9.4c shows the output characteristics for the common-emitter configuration with input voltage (V_{BE}) as a parameter.

The output characteristics for the common-base configuration of Figure 9.4a are easily explained. First take the case of $I_E = 0$. With no emitter current, $I_C = -I_B$, and the current from the collector flows through the collector-base junction and out the base contact. This is then a simple diode and the I_C-V_{CB} characteristic of an np junction is obtained. The diode curve in this figure looks inverted compared with the way it is usually drawn. When the collector-base junction is forward biased, V_{CB} and I_C are negative and when it is under reverse bias, they are positive, according to convention.

Now let us consider what happens when the emitter-base junction is forward biased so that I_E is positive. Since the emitter is n type in this case, electrons are injected from the emitter into the base. Note that while electrons are majority carriers in the emitter, once they reach the base they become minority carriers. Normally those electrons would recombine in the p-type base region in an average time (lifetime) τ_n , or within a diffusion length L_n . The minority carrier lifetime is on the order of $\tau_n = 1 \mu\text{s}$ in Si, corresponding to L_n on the order of tens of micrometers. The base region is thin, however, and before the injected carriers recombine, most of them diffuse to the collector edge of the base junction. If the C-B junction is reverse biased, the field of this junction accelerates them into the collector. Thus, most of the emitter current contributes to collector current. The effect on the I_C-V_{CE} characteristics of Figure 9.4a is that the entire diode curve translates upward by an amount αI_E , where α is the fraction of I_E that contributes to I_C . This fraction is typically on the order of 99 percent.

For the common-emitter configuration, the physics is the same but the I-V characteristics look different because now I_C is plotted versus V_{CE} with the varying parameter I_B . If the p base is made positive with respect to the emitter, thus forward biasing the base-emitter junction, electrons will be injected from the emitter into the base and holes from the base into the emitter. The injected holes recombine in the emitter or at the emitter contact and flow out the common-emitter contact, but a small fraction of the electrons injected into the base will recombine and flow out the base contact. Most of the electrons diffuse across the thin base to be collected by the collector.

As can be seen from Figure 9.3, the base input current I_B (that is supplied by the base) has three components. It consists of hole current injected into the emitter I_{pE} , plus the electron-hole recombination current I_{rec} , minus the collector-base leakage current I_{pC} . Each of these is small. Thus, I_B is on the order of 1 percent of the injected electron current (supplied through the emitter contact). Most of the emitter electron current ends up at the collector, and we take the collector current to be the output. The output current I_C is thus on the order of 100 times the input current I_B . This is shown in Figure 9.4b.

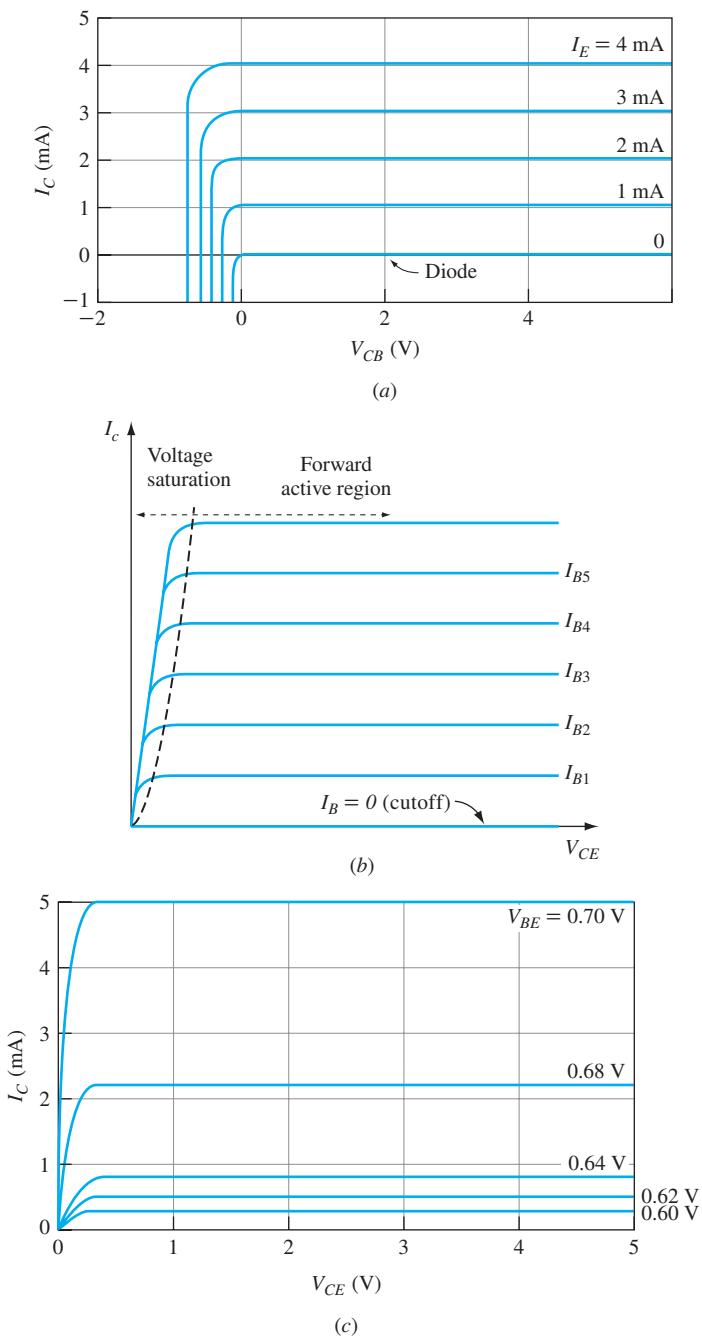


Figure 9.4 Idealized output characteristics of an npn transistor with input current as a parameter operating in (a) the common-base configuration and (b) the common-emitter configuration. (c) The output characteristics of the common-emitter configuration with input voltage as a parameter.

From Equation (IV.1), we see that for a given collector-emitter voltage (V_{CE}), applying a base-emitter voltage (V_{BE}) tends to reduce the collector-base voltage (V_{CB}), since their sum is a constant. At low V_{CE} , the base-emitter voltage may be larger than V_{CE} , which would make V_{CB} negative. This means that both base-emitter and base-collector junctions are forward biased and the device is said to operate in the *saturation region*; i.e., the collector-emitter voltage (V_{CE}) is saturated at a small value (about 0.2 to 0.3 V). At larger V_{CE} , the *current* saturates for a given value of I_B as indicated. This current saturation corresponds to the saturation region for FETs, but in BJTs the term *saturation* refers to the region indicated in Figure 9.4b. In Figure 9.4a and b, where input current is the varying parameter between curves, the output currents are approximately proportional to the input currents. In (c) however, the output current increases exponentially with input voltage. Because of this nonlinearity, this latter representation is seldom used and will not be considered further here.

9.3 CURRENT GAIN

As indicated earlier, two common modes of operation of a BJT are the common-base and common emitter configurations shown in Figure IV.3. In the common-base circuit, the emitter current is taken as the input and the collector current is the output. The current gain for common-base operation is

$$\alpha = \frac{I_{\text{out}}}{I_{\text{in}}} = \frac{I_C}{I_E} \quad (9.5)$$

This ratio is close to unity. A small fraction of the electrons injected into the base from the emitter are lost by recombination, and a small hole current is injected from the base to the emitter. In addition, as is discussed later, the hole current from collector to base is negligible. The collector current is close to but smaller than the emitter current, so $I_C < I_E$ and $\alpha < 1$, typically about 0.99.

In the common-emitter circuit, the base current is the input and the output is the collector current, so the common-emitter current gain is

$$\beta = \frac{I_{\text{out}}}{I_{\text{in}}} = \frac{I_C}{I_B} \quad (9.6)$$

Since $I_E = I_C + I_B$, the common-emitter current gain β can be expressed as

$$\beta = \frac{I_C}{I_B} = \frac{I_C}{I_E - I_C} = \frac{\frac{I_C}{I_E}}{1 - \frac{I_C}{I_E}} = \frac{\alpha}{1 - \alpha} \quad (9.7)$$

or

$$\alpha = \frac{\beta}{\beta + 1} \quad (9.8)$$

However, since $\alpha < 1$, for large β it is desired to have α approach unity.

To improve our understanding of these current gains, we can break up the various currents into their electron and hole components. First, let us consider α . Multiplying and dividing Equation (9.5) by I_{nE} and I_{nC} and rearranging gives

$$\alpha = \frac{I_{nE} I_{nC} I_C}{I_E I_{nE} I_{nC}} \quad (9.9)$$

where I_{nE} is the electron current produced by the electron flux injected from the emitter into the base and I_{nC} is the current that results from electrons reaching the collector, as shown in Figure 9.2c.

We can look at each of the terms in Equation (9.9) individually. The first term is called the injection efficiency γ , which is defined as the fraction of the emitter current that is due to electron injection into the base.

$$\gamma = \frac{I_{nE}}{I_E} = \frac{I_{nE}}{I_{nE} + I_{pE}} = \frac{1}{1 + \frac{I_{pE}}{I_{nE}}} \quad (9.10)$$

where I_{pE} is the hole current injected from base to emitter.

Ideally, γ is close to unity, which is accomplished by making $I_{pE} \ll I_{nE}$.

The base transport efficiency α_T (often called the current transport factor) is the fraction of the electrons injected from the emitter into the base that reach the collector.

$$\alpha_T = \frac{I_{nC}}{I_{nE}} = \frac{I_{nE} - I_{\text{rec}}}{I_{nE}} = 1 - \frac{I_{\text{rec}}}{I_{nE}} \quad (9.11)$$

where I_{rec} is the electron recombination current in the base. The value of α_T approaches unity for $I_{\text{rec}} \ll I_{nE}$.

The last term is the collection efficiency M :

$$M = \frac{I_C}{I_{nC}} = \frac{I_{nC} + I_{pC}}{I_{nC}} = 1 + \frac{I_{pC}}{I_{nC}} \quad (9.12)$$

where I_{pC} is the hole leakage current from collector to base in the reverse-biased B-C junction. This collection efficiency is often called the *collection multiplication factor* since it can be greater than unity for carrier multiplication under high base-collector voltages. From Equation (9.9) then,

$$\alpha = \gamma \alpha_T M \quad (9.13)$$

To determine values for α and β then, we must relate the quantities γ , α_T , and M to the currents I_{pE} , I_{nE} , I_{rec} , I_{nC} , and I_{pC} . To do this analytically requires a model for the device.

9.4 MODEL OF A PROTOTYPE BJT

Now we will go through the operation of the BJT in the active mode more quantitatively, starting with a simple but not entirely realistic model. Once we understand the basic operating principles, we will add refinements.

We take as our example the Si npn BJT considered before in Figure IV.1 and in Figure 9.1. It consists of a degenerately doped n⁺ emitter (*E*), a nondegenerate p base (*B*), and a collector region (*C*) that includes both nondegenerate n and degenerate n⁺ regions. In this prototype transistor, all regions are assumed to be uniformly doped.

Notice that the metal terminal contacts are made to degenerate n⁺ emitter and n⁺ collector regions and to a degenerate p⁺ region in the base. These metal-degenerate semiconductor contacts form Schottky barriers thin enough to permit tunneling and thus ensure low-resistance contacts.

The n collector region is relatively lightly doped to reduce the collector-base capacitance and to ensure that the breakdown voltage of the collector-base junction is high enough that breakdown doesn't occur (recall that the base-collector junction is reverse biased in forward active bias). The n⁺ collector region (buried layer) is used to provide a low-resistance path in the collector. The n⁺ *sinker* between the collector contact and the n⁺ collector region also reduces the collector resistance.

To obtain quantitative results for the currents in each region, the continuity equations must be solved with the appropriate boundary conditions. To obtain the steady-state (dc) currents, the continuity equations [Equations (IV.3) and (IV.4)] become

$$\frac{\partial J_n}{\partial x} = q \frac{\Delta n}{\tau_n} \quad (9.14)$$

$$\frac{\partial J_p}{\partial x} = -q \frac{\Delta p}{\tau_p} \quad (9.15)$$

For the prototype transistor being considered, since the emitter, base, and collector are each uniformly doped, at equilibrium there are no electric fields in these regions. Under forward bias only minuscule fields exist, and minority carrier flow is primarily by diffusion. In the p-type base, the influx of minority carriers, electrons in this case, increases the minority carrier concentration near the junction. To maintain neutrality, an equal number of holes are drawn into the base via the base contact. For the low-injection condition considered here, the excess hole concentration is negligible compared with the equilibrium hole concentration resulting from ionized acceptors. Similarly, the holes injected into the emitter have negligible effect on the electron concentration there. The currents of primary interest, then, are the minority carrier diffusion currents. In general, to find these currents we must solve the continuity equation in each region using the appropriate boundary conditions, not always an easy task. However, since the doping in each region is uniform, $\mathcal{E} \approx 0$, only diffusion current needs to be considered for minority carriers.

Figure 9.5 repeats Figure 9.2a with the emitter-base and base-collector depletion regions exaggerated. Because, for an npn BJT, positive current is from collector to emitter, in the negative *x* direction, the currents obtained from solutions of the continuity equations will result in negative quantities, i.e., in the negative *x*

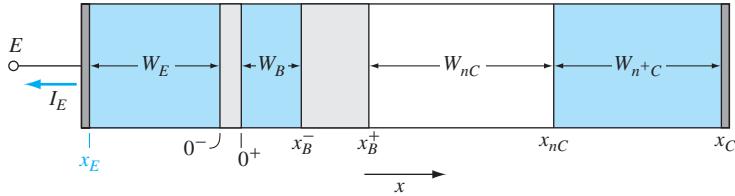


Figure 9.5 Widths and coordinates for simple prototype model. The depletion region widths have been exaggerated.

direction. The emitter edge of the $E-B$ junction is designated $x = 0^-$ and the base edge is 0^+ . Similarly the base edge of the $B-C$ junction is designated x_B^- and the collector edge is x_B^+ . The width of the quasi-neutral region in the emitter is W_E , of the base, W_B , and of the n collector, W_{nC} .

In the emitter the diffusion current density is

$$J_{pE} = -q D_{pE} \frac{d\Delta p_E}{dx} \quad (9.16)$$

and Equation (9.15) becomes

$$D_{pE} = \frac{d^2 \Delta p_E}{dx^2} = \frac{\Delta p_E}{\tau_{pE}} \quad (9.17)$$

with the boundary conditions

$$\begin{aligned} \Delta p_E(x = 0^-) &= p_{E0}(e^{qV_{BE}/kT} - 1) \\ \Delta p_E(x = x_E) &= 0 \end{aligned}$$

where p_{E0} is the equilibrium hole concentration in the emitter. In the base, the electron diffusion current density is

$$J_{nB} = q D_{nB} \frac{d\Delta n_B}{dx} \quad (9.18)$$

and Equation (9.14) becomes

$$D_{nB} \frac{d^2 \Delta n_B}{dx^2} = \frac{\Delta n_B}{\tau_{nB}} \quad (9.19)$$

with the boundary conditions

$$\begin{aligned} \Delta n_B(0^+) &= n_{B0}(e^{qV_{BE}/kT} - 1) \\ \Delta n_B(x_B^-) &= n_{B0}(e^{qV_{BC}/kT} - 1) \end{aligned}$$

In the collector, the hole diffusion current density is

$$J_{pC} = -q D_{pC} \frac{d\Delta p_C}{dx} \quad (9.20)$$

and

$$D_{pC} \frac{d^2 \Delta p_C}{dx^2} = \frac{\Delta p_C}{\tau_{pC}} \quad (9.21)$$

with the boundary conditions

$$\begin{aligned}\Delta p(x_B^+) &= p_{C0}(e^{qV_{BC}/kT} - 1) \\ \Delta p(x_{nC}) &\approx 0\end{aligned}$$

The last boundary condition assumes that $W_{nC} \gg L_{pC}$, and therefore all the excess minority carriers have recombined by the time they reach the end of the collector region.

To calculate the diffusion current in each region, Δp_E , Δn_B , and Δp_C are solved with appropriate boundary conditions and used in Equations (9.16), (9.18), and (9.20) to determine the minority carrier diffusion currents in each region. The results in general involve hyperbolic functions. [1, 2] The task can be simplified, however, with the preceding assumptions and for operation in the active mode.

To find the current gain parameters $\alpha = \gamma\alpha_T M$ and $\beta = \alpha/(1 - \alpha)$ we must solve for I_{nB} , I_{pE} , I_{nC} , I_{rec} , and I_{pC} . To obtain a quantitative result, we make the following additional assumptions appropriate for most modern BJTs:

1. The width of the base region W_B , measured between the edges of the depletion regions, is much smaller than the diffusion length L_{nB} for electrons in the base, or $W_B \ll L_{nB}$.
2. The width of the emitter region $W_E \ll L_{pE}$, where L_{pE} is the diffusion length for holes in the emitter.
3. The concentration of electrons injected into the base is small enough that, everywhere in the base, the electron concentration is always much less than the hole concentration, or $n_B \ll p_B$. This is the “low-injection” condition. We will consider high injection in the Supplement to Part 4.

The energy band diagram of Figure 9.2c for operation in the active mode is redrawn in Figure 9.6, where the $E-B$ and $B-C$ transition regions are indicated. For simplicity, the n^+ collector region is not shown. Note that the minority carrier concentration in the base goes to zero at the collector edge (where the minority carriers are extracted because of the reverse bias on that junction), and the excess minority carrier concentration injected into the emitter decays to zero at the emitter edge because of the presence of the contact (not shown).

9.4.1 COLLECTION EFFICIENCY M

From Equation (9.12), the collection efficiency M is

$$M = \frac{I_{nC} + I_{pC}}{I_{nC}} = 1 + \frac{I_{pC}}{I_{nC}} \quad (9.22)$$

where I_{pC} consists of the hole current diffusing from collector to base plus the hole generation current in the collector-base junction. Electron-hole pairs are

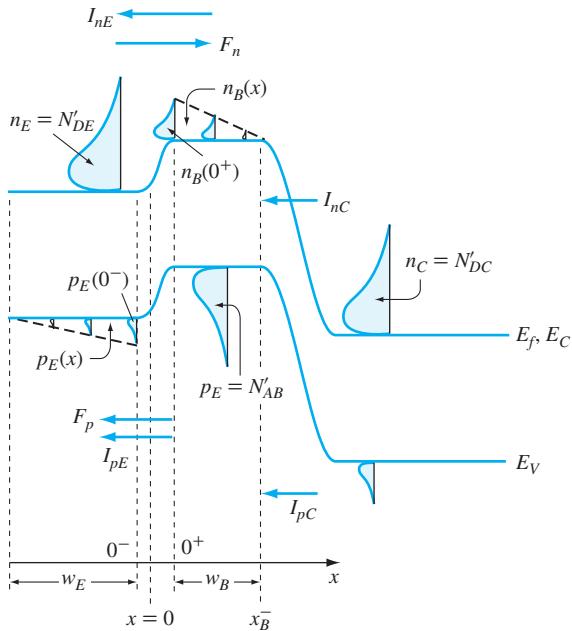


Figure 9.6 Active mode energy band diagram of a prototype npn BJT showing currents used to determine transport properties.

continually being generated thermally in the junction, and the electric field separates them and drives them away from the junction (not shown). Except when the reverse bias is large enough that significant carrier multiplication occurs (approaching breakdown, a situation to be avoided), I_{pC} is several orders of magnitude less than I_{nC} . Thus, for normal operation

$$M \approx 1 \quad (9.23)$$

and

$$\alpha \approx \gamma \alpha_T \quad (9.24)$$

Finally, for a reverse-biased junction, the generation current is much greater than the diffusion current. Thus, for the base-collector junction, Equation (9.20) need not be solved.

9.4.2 INJECTION EFFICIENCY γ

From Equation (9.10)

$$\gamma = \frac{1}{1 + \frac{I_{pE}}{I_{nE}}} \quad (9.10)$$

However, I_{nE} is the injected electron current across the forward-biased E - B junction, and thus can be expressed as

$$I_{nE} = qA_E D_{nB} \frac{dn_B}{dx} \Big|_{x=0^+} \quad (9.25)$$

where A_E is the area of the emitter-base junction, D_{nB} is the diffusion constant for electrons in the base, and $x = 0^+$ represents the position of the base edge of the emitter-base transition region.

To find an expression for dn_B/dx , we note that excess carriers are injected into the base at the base-emitter junction and $n_B(0^+)$ is the electron concentration at the emitter edge of the base ($x = 0^+$), but at the collector end of the base, the carriers are extracted. Since the base is much shorter than a minority carrier diffusion length ($W_B \ll L_{nB}$), the electron-hole recombination in the base is small, and under these conditions, to first approximation, the excess carrier concentration in the base varies linearly from $n_B(0^+)$ to $n_B(x_B^-) \approx 0$. Thus

$$n_B(x) = n_B(0^+) \left[1 - \frac{x - 0^+}{W_B} \right]$$

as indicated by the straight dashed line in the base region conduction band in Figure 9.6. Therefore

$$\frac{dn_B}{dx} \Big|_{x=0^+} = -\frac{n_B(0^+)}{W_B}$$

and Equation (9.25) becomes

$$I_{nE} = -\frac{qA_E D_{nB} n_B(0^+)}{W_B} \quad (9.26)$$

Similarly, since the emitter is also thin, such that $W_E \ll L_{pE}$, we can write at the emitter end of the E - B junction

$$I_{pE} = I_{pB} = -qA_E D_{pE} \frac{dp}{dx} \Big|_{x=0^-} \approx -\frac{qA_E D_{pE} p_E(0^-)}{W_E} \quad (9.27)$$

where D_{pE} is the diffusion constant for holes in the emitter. Then substituting these expressions into Equation (9.10) and simplifying gives

$$\gamma \approx \frac{1}{1 + \frac{p_E(0^-) D_{pE} W_B}{n_B(0^+) D_{nB} W_E}} \quad (9.28)$$

To maximize the injection efficiency, from Equation (9.28) one should minimize the second term in the denominator. One has little control of D_{pE} or D_{nB} , but one can minimize $p_E(0^-)/n_B(0^+)$, the ratio of the minority carrier concentrations. In other words, one should dope the emitter more heavily than the base. In addition, W_B/W_E should also be made small, or the base should be made thinner than the emitter.

9.4.3 BASE TRANSPORT EFFICIENCY α_T

The electron collector current I_{nC} is equal to the emitter electron injection current, less any that is lost to recombination in the base. Recalling that the recombination current is the ratio of the electron charge in the base $qA_E\langle n_B \rangle W_B$ to the minority carrier lifetime τ_{nB} , we have

$$I_{\text{rec}} = \frac{qA_E\langle n_B \rangle W_B}{\tau_{nB}} \quad (9.29)$$

where $\langle n_B \rangle$ is the average electron concentration in the base. Also

$$\langle n_B \rangle \approx \frac{n_B(0^+)}{2} \quad (9.30)$$

With the aid of Equations (9.26), (9.29), and (9.30), Equation (9.11) becomes

$$\alpha_T \approx 1 - \frac{W_B^2}{2D_{nB}\tau_{nB}} = 1 - \frac{W_B^2}{2L_{nB}^2} \quad (9.31)$$

where we have used $D_{nB}\tau_{nB} = L_{nB}^2$. Therefore, to maximize the current transport factor α_T , the base width W_B should be small. Note that Equation (9.31) is for an npn transistor, and that L_{nB} is the minority carrier diffusion length *in the base*. For a pnp transistor, the equation is the same except that L_{pB} would be used.

It is of interest to calculate β for an npn transistor. We will do two examples below. In Example 9.1, a nondegenerate emitter will be considered. In Example 9.2, these results will be adapted to that of a somewhat more realistic BJT with a degenerately doped emitter. In both cases it is assumed that each region is uniformly doped such that the diffusion current predominates and relations for γ [Equation (9.28)] can be used.

EXAMPLE 9.1

Find β for a bipolar junction transistor with a nondegenerate emitter. Assume that the emitter, base, and collector are noncompensated and that

$$\begin{aligned} N'_{DE} &= N_{DE} = 2 \times 10^{18} \text{ cm}^{-3} \\ N'_{AB} &= N_{AB} = 10^{16} \text{ cm}^{-3} \\ N'_{DC} &= N_{DC} = 10^{15} \text{ cm}^{-3} \\ W_E &= 0.2 \mu\text{m} \\ W_B &= 0.1 \mu\text{m} \end{aligned}$$

Solution

To find β we must first determine γ , α_T , and α . From Equation (9.28)

$$\gamma \approx \frac{1}{1 + \frac{p_E(0^-) D_{pE} W_B}{n_B(0^+) D_{nB} W_E}} \quad (9.28)$$

To find the first factor in the denominator, we recall that for a given value of V_{BE} ,

$$p_E(0^-) = p_{E0} e^{qV_{BE}/kT} \quad (9.32)$$

$$n_B(0^+) = n_{B0} e^{qV_{BE}/kT} \quad (9.33)$$

where p_{E0} is the equilibrium hole concentration in the emitter and n_{B0} is the equilibrium concentration of electrons in the base. Taking the ratio, we obtain

$$\frac{p_E(0^-)}{n_B(0^+)} = \frac{p_{E0}}{n_{B0}} \quad (9.34)$$

Since both regions are considered to be nondegenerate

$$p_{E0} n_{E0} = n_i^2 \quad (9.35)$$

$$n_{B0} p_{B0} = n_i^2 \quad (9.36)$$

Substituting $n_{E0} = N'_{DE}$ and $p_{B0} = N'_{AB}$, we have

$$p_{E0} = \frac{n_i^2}{N'_{DE}}$$

$$n_{B0} = \frac{n_i^2}{N'_{AB}}$$

and

$$\frac{p_E(0^-)}{n_B(0^+)} = \frac{N'_{AB}}{N'_{DE}} = \frac{10^{16}}{2 \times 10^{18}} = \frac{1}{200}$$

Next, we determine D_{pE}/D_{nB} . For the doping levels given, the minority carrier diffusion coefficients are

$$D_{pE} = 5.5 \text{ cm}^2/\text{s}$$

$$D_{nB} = 32 \text{ cm}^2/\text{s}$$

and thus $D_{pE}/D_{nB} = 0.17$.

The base and emitter widths W_B and W_E are given, so

$$\gamma = \frac{1}{1 + \frac{p_E(0^-) D_{pE} W_B}{n_B(0^+) D_{nB} W_E}} = \frac{1}{1 + \frac{1}{200} \times 0.17 \times \frac{0.1}{0.2}} = 0.9996$$

Next, the transport efficiency α_T can be found from Equation (9.31). For $N_{AB} = 10^{16} \text{ cm}^{-3}$, $L_{nB} = 300 \mu\text{m}$ (remember, it is the minority carrier diffusion length in the base) and

$$\alpha_T = 1 - \frac{1}{2} \left(\frac{0.1}{300} \right)^2 = 0.9999999 \approx 1$$

As an aside, we observe that our assumption that the base is much shorter than a minority carrier diffusion length is met, since $W_B = 0.1 \mu\text{m}$ and $L_{nB} = 300 \mu\text{m}$. Similarly, the

minority carrier diffusion length in the emitter is $L_{pE} = 16 \mu\text{m}$, and the emitter length of $0.2 \mu\text{m}$ is short compared with this, and so the assumption that $W_E \ll L_{pE}$ is reasonably valid. Continuing with our example, from Equation (9.24), we have for $\alpha = \gamma\alpha_T \approx \gamma$. Then, from Equation (9.7)

$$\beta = \frac{\alpha}{1 - \alpha} \approx \frac{\gamma}{1 - \gamma} = \frac{0.9996}{1 - 0.9996} = 2500$$

This result is clearly much larger than typical values of β , which are on the order of 100 to 200 for a Si npn BJT. One reason is that in practical transistors, the emitter and base regions are more heavily doped than in the previous example. In the degenerately doped n-type emitter, the conduction band edge is lowered by the band-gap narrowing effect discussed in Chapter 2. This is shown in Figure 9.7. Thus, electrons being injected into the base from the emitter face a slightly larger energy barrier than was accounted for in Example 9.1. The barrier for holes is essentially unaffected, however, since neither the emitter nor the base have degenerate p-type doping in this example.¹ The increased barrier for electrons reduces the injection of electrons, but the hole injection remains the same. This, in turn, reduces the injection efficiency γ [Equation (9.10)]. Since the injection varies exponentially with the barrier height, the impact is significant. In Example 9.1, the injection efficiency was the most important factor in the value of β . Note that the conduction band edge in the p-type base is unaffected.

For bipolar junction transistors the energy of the Fermi level (or quasi-Fermi energy for electrons) in the degenerate emitter is not well known, and rather than band gap narrowing, an apparent band gap narrowing ΔE_g^* is of interest. It is given by the relation [3]

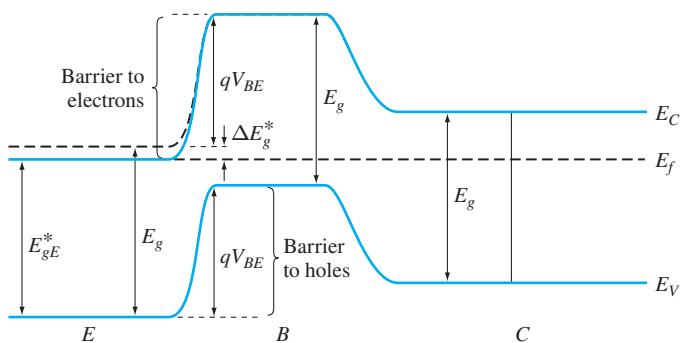


Figure 9.7 Apparent band gap narrowing for n- and p-type silicon as a function of majority carrier doping.

¹It is often the case in modern transistors that the base is doped heavily enough that its band gap is also affected.

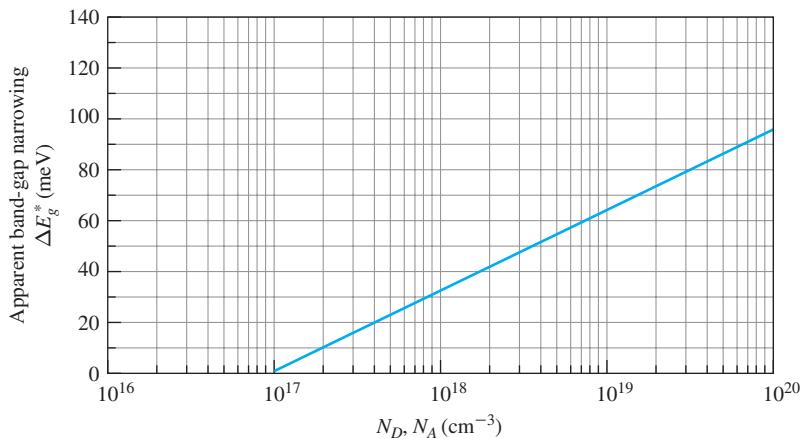


Figure 9.8 The apparent band-gap narrowing in the heavily doped emitter results in a higher barrier to electrons than to the holes at the emitter-base junction.

$$\Delta E_g^* = 6.92 \left[\ln \frac{N}{1.3 \times 10^{17}} + \sqrt{\left(\ln \frac{N}{1.3 \times 10^{17}} \right)^2 + 0.5} \right] \text{meV} \quad (9.37)$$

It is plotted in Figure 9.8 for both n- and p-type silicon. Here, N represents N_D in n-type Si and N_A in p-type Si.

Let us now consider a somewhat more realistic example and include the effects of band-gap narrowing.

EXAMPLE 9.2

We consider a degenerate emitter. Assume again an npn prototype transistor of the same geometry as Example 9.1 in which each region is uniformly doped, but now the emitter is degenerate. Let $N'_{DE} = 5 \times 10^{19} \text{ cm}^{-3}$, and to keep the N'_{DE}/N'_{AB} ratio the same as in the previous example, let $N'_{AB} = 2.5 \times 10^{17} \text{ cm}^{-3}$.

Solution

The injection efficiency is still given by Equation (9.28):

$$\gamma \approx \frac{1}{1 + \frac{p_E(0^-) D_{pE} W_B}{n_B(0^+) D_{nB} W_E}}$$

To determine γ we must again find D_{pE}/D_{nB} and $p_E(0^-)/n_B(0^+)$, but in this case, to find the $p_E(0^-)/n_B(0^+)$ ratio we must consider the case of the degenerate emitter.

As in the previous case,

$$p_E(0^-) = p_{E0} e^{qV_{BE}/kT}$$

$$n_B(0^+) = n_{B0} e^{qV_{BE}/kT}$$

$$\frac{p_E(0^-)}{n_B(0^+)} = \frac{p_{E0}}{n_{B0}}$$

and in the nondegenerate base, as before, at equilibrium

$$n_{B0} = \frac{n_i^2}{N'_{AB}}$$

But, in the degenerate emitter, from Equation (2.102),

$$n_0 p_0 = n_i^2 e^{\frac{\Delta E_g^*}{kT}} \quad (9.38)$$

where ΔE_g^* is the impurity-induced apparent band-gap narrowing.

$$\frac{p_E(0^-)}{n_B(0^+)} = \frac{N'_{AB}}{N'_{DE}} e^{\Delta E_g^*/kT}$$

The injection efficiency in this case is

$$\gamma \approx \frac{1}{1 + \frac{p_E(0^-) D_{pE} W_B}{n_B(0^+) D_{nB} W_E}} = \frac{1}{1 + \frac{N'_{AB}}{N'_{DE}} e^{\Delta E_g^*/kT} \frac{D_{pE} W_B}{D_{nB} W_E}} \quad \text{emitter degenerate} \quad (9.39)$$

Since $N'_{DE} = 5 \times 10^{19} \text{ cm}^{-3}$ and, from Figure 9.7 or Equation (9.37), $\Delta E_g^* = 0.083 \text{ eV}$, we have, using $D_{pE} = 3.3 \text{ cm}^2/\text{s}$ and $D_{nB} = 15 \text{ cm}^2/\text{s}$ from Figure 3.11,

$$\gamma = \frac{1}{1 + \frac{2.5 \times 10^{17}}{5 \times 10^{19}} e^{0.083/0.026} \times \frac{3.3}{15} \times \frac{1}{2}} = 0.987$$

The value of α_T can be obtained from Equation (9.31) as before, but for this base doping, $L_n = 40 \mu\text{m}$ (note that the assumption that $W_B \ll L_{nB}$ is still valid), and

$$\alpha_T = 1 - \frac{1}{2} \left(\frac{0.1}{40} \right)^2 = 0.999999$$

Again $\alpha_T \approx 1$ and

$$\alpha \approx \gamma$$

Then

$$\beta = \frac{\alpha}{1 - \alpha} = \frac{\gamma}{1 - \gamma} = 75$$

which is on the order of the expected value.² Note that, although we changed the doping in this example, we kept the ratio of N'_{DE}/N'_{AB} the same. If band-gap narrowing had been neglected with this new doping, we would have obtained a result similar to that in Example 9.1, modified only by a slightly different value of D_{pE}/D_{nB} .

²It is somewhat low compared with realistic values on the order of 100 to 200. Sections 9.5 and 9.6 show how the nonuniform doping and graded composition in real BJTs tends to raise the value of β .

The main factor in reducing the calculated value of β in Example 9.2 from that calculated in the previous example is the consideration of the impurity-induced apparent band-gap narrowing, ΔE_{gB}^* . This effect cannot be ignored in practical transistors. In Example 9.2, the effect of band-gap narrowing in the base is neglected since the base is nondegenerate and the effect is small. If this effect is taken into account, Equation (9.39) becomes

$$\gamma = \frac{1}{1 + \frac{N'_{AB}}{N'_{DE}} e^{\Delta E_{gBE}^*/kT} \frac{D_{pE}}{D_{nB}} \frac{W_B}{W_E}} \quad \text{base and emitter degenerate} \quad (9.40)$$

where

$$\Delta E_{gBE}^* = E_{gB}^* - E_{gE}^* \quad (9.41)$$

and E_{gB}^* is the apparent band gap of the base and E_{gE}^* is that of the emitter. Thus ΔE_{gBE}^* is the difference in the (apparent) band gaps of base and emitter.

EXAMPLE 9.3

Show that for a prototype transistor, β can be approximated by

$$\beta \approx \frac{n_B(0^+)}{p_E(0^-)} \frac{D_{nB}}{D_{pE}} \frac{W_E}{W_B}$$

Solution

Since $\beta \approx \gamma/(1 - \gamma)$, $\gamma = \beta/(1 + \beta)$. From Equation (9.28), letting

$$Z = \frac{p_E(0^-)}{n_B(0^+)} \frac{D_{pE}}{D_{nB}} \frac{W_B}{W_E}$$

we can write $\gamma = 1/(1 + Z)$, and

$$\beta \approx \frac{\frac{1}{1+Z}}{1 - \frac{1}{1+Z}} = \frac{\frac{1}{1+Z}}{\frac{Z}{1+Z}} = \frac{1}{Z}$$

and

$$\beta \approx \frac{n_B(0^+)}{p_E(0^-)} \frac{D_{nB}}{D_{pE}} \frac{W_E}{W_B}$$

Note that, for a nondegenerate emitter,

$$\beta \approx \frac{N'_{DE}}{N'_{AB}} \frac{D_{nB}}{D_{pE}} \frac{W_E}{W_B} \quad (9.42)$$

and, for a degenerate emitter,

$$\beta \approx \frac{N'_{DE}}{N'_{AB}} \frac{D_{nB}}{D_{pE}} \frac{W_E}{W_B} e^{-\Delta E_{gBE}^*/kT} \quad (9.43)$$

Note that the apparent bandgap narrowing can be determined from the variation of β with emitter doping level and with temperature.

It should be emphasized that the model for a prototype BJT is overly simplified. In most BJTs, the doping levels in each region are not constant. As we will see in the next section, the base impurity grading tends to increase the value of β over that calculated from the prototype model in which band-gap narrowing is considered.

9.5 DOPING GRADIENTS IN BJTS

The prototype transistor we considered had uniform doping in each region. We now consider a more realistic bipolar junction transistor. We take as an example an npn BJT fabricated in the *bipolar-CMOS* (BiCMOS) technology.³ Its doping profile is shown in Figure 9.9a. The profile, with x positive from left to right and $x = 0$ at the emitter surface, was measured by a technique known as *secondary-ion mass spectroscopy* (SIMS). The impurity concentration is plotted vertically on a log scale, and the horizontal axis is distance into the material. The base region is doped with boron. The n-type dopants (phosphorus, antimony, and arsenic) are used to dope the emitter and collector. Because the ion implants used to create this profile were made through a surface oxide layer, some dopants remain in the oxide region ($x < 0$).

This profile is replotted on an expanded scale in Figure 9.9b for the region of the active transistor. Where the donor concentration exceeds the acceptor concentration, the region is n type. The emitter-base and base-collector metallurgical junctions are at the locations where $N_A = N_D$. The metallurgical emitter width is $W_{EM} = 0.13 \mu\text{m}$ and the metallurgical base width is $W_{BM} = 0.14 \mu\text{m}$.

It can be seen from Figure 9.9b that the doping profiles in the emitter, base, and collector are all nonuniform, which complicates the analysis. Because of the impurity gradients, electric fields exist in each region, and the current flow is by a combination of drift and diffusion. These added complications are not amenable to hand calculations, and device simulation programs are required for analysis.⁴ We will, however, discuss the physics of effects of the graded doping, and present some results.

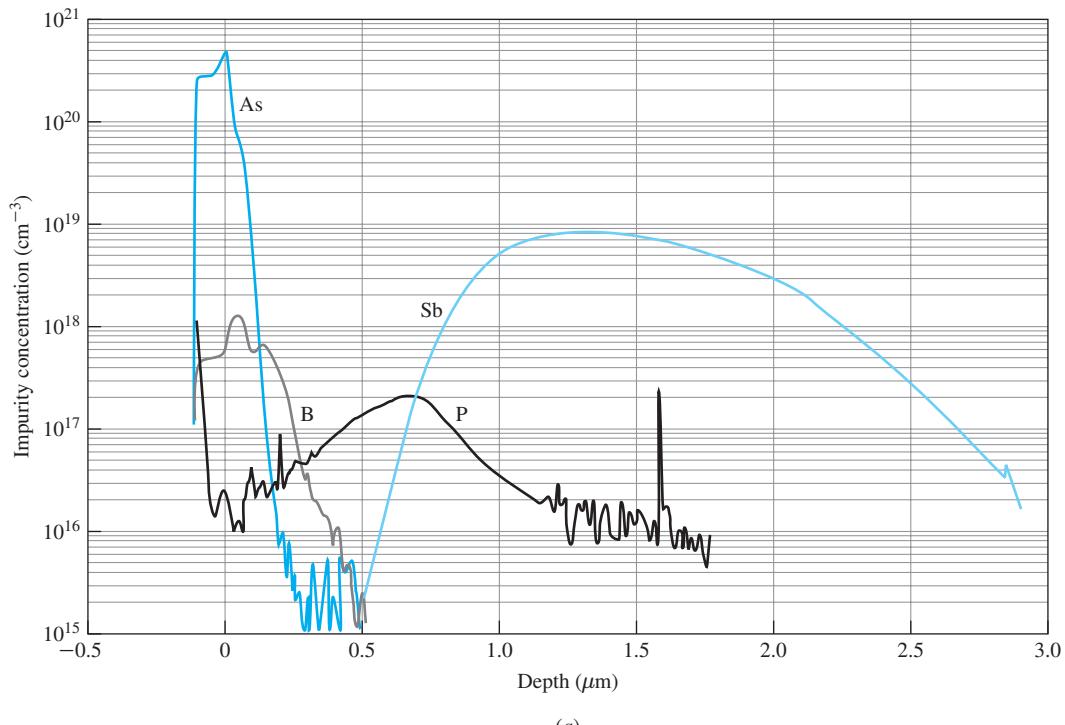
The doping gradients in the emitter, base, and collector cause the conduction band and valence band edges to bend, producing internal electric fields. These fields must be considered. An energy band diagram of this device is shown in Figure 9.10a, in which the variation in doping is taken into account. This figure is for operation in the active mode. The emitter-base junction is forward biased by an amount V_{BE} . The collector-base junction is reverse biased.

From the energy band diagram we can see:

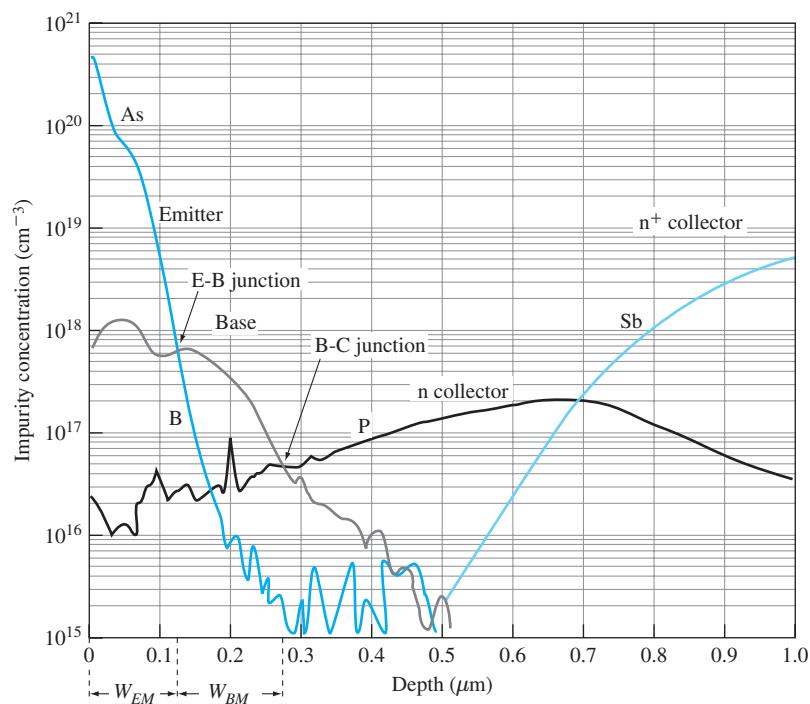
- Because of the doping dependence of the band-gap narrowing in the emitter (primarily a decrease in E_C), an effective field exists for electrons,

³In this technology, n-channel and p-channel MOSFETs and npn and pnp BJTs are fabricated in the same chip.

⁴The nominal common-emitter gain for this device is $\beta = 90$.



(a)



(b)

Figure 9.9 SIMS data. (a) Measured impurity concentration profile for an npn BJT. (b) Diagram expanded in the x direction in the active region.

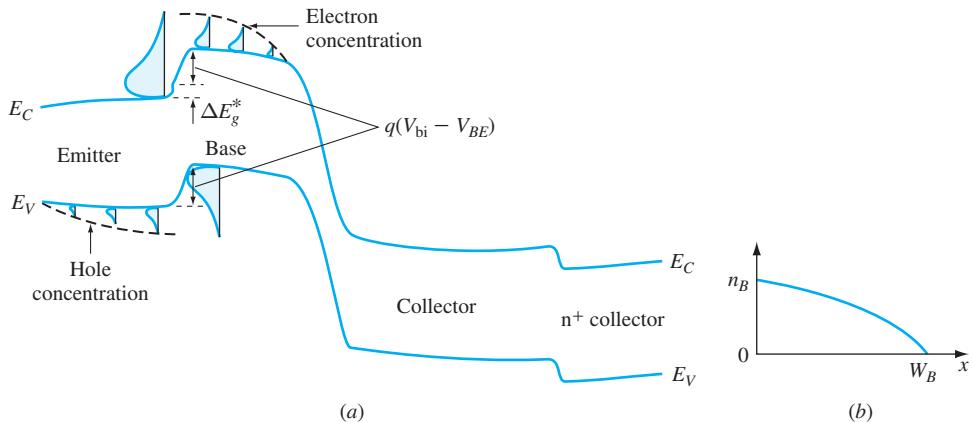


Figure 9.10 (a) Approximate active mode energy band diagram for the BJT of Figure 9.9 considering band-gap shrinkage. (b) The resultant electron distribution in the base.

$\mathcal{E}_e^* = (1/q)/(dE_C/dx)$. The conduction band edge bends closer to the valence band edge where the doping is heaviest, near the surface.

2. The effective field for holes in the emitter, $\mathcal{E}_h^* = (1/q)/(dE_V/dx)$, is in the opposite direction of that for electrons.
3. In the base, an electric field exists that accelerates electrons toward the collector.
4. Because of the relatively small doping in the base and the corresponding small band-gap narrowing, here the slopes of the conduction and valence band edges are equal ($dE_C/dx = dE_V/dx$). In other words, the effective fields for electrons and holes are equal ($\mathcal{E}_e^* \approx \mathcal{E}_h^*$).
5. The barrier for electrons at the emitter-base junction is greater than that for holes by the amount ΔE_g^* . This is the amount of apparent band-gap narrowing in the heavily n-doped emitter.

This device is referred to as a *graded-base transistor*. Because of the effective field for electrons in the base, the injected electrons are accelerated toward the collector. Thus, the electron concentration gradient is not constant as was the case for uniform base doping. The electron concentration in the base for a graded-base BJT is shown in Figure 9.10b. The presence of the field in the base creates a drift component to the current that increases I_{nB} above its value for diffusion. The electron distribution in the base of Figure 9.10b and the way in which the base field increases β are treated in the next section.

9.5.1 THE GRADED-BASE TRANSISTOR

It was seen in the previous section that, by grading the base doping, an electric field could be created in the base region that helps accelerate electrons across the base. This tends to improve the injection efficiency. Now we will look at this more closely.

Consider an npn BJT with some arbitrary base doping profile, $N'_{AB}(x)$. The hole (majority) current density in the base is

$$J_{pB} = q \mu_p p \mathcal{E} - q D_p \frac{dp}{dx} \quad (9.44)$$

From the Einstein relation, $D/\mu = kT/q$, Equation (9.44) can be rewritten as

$$J_{pB} = q D_p \left[\frac{q}{kT} p \mathcal{E} - \frac{dp}{dx} \right] \quad (9.45)$$

At equilibrium, there is no net current, so $J_{pB} = 0$ and

$$\frac{q}{kT} p \mathcal{E} = \frac{dp}{dx} \quad (9.46)$$

Considering that all acceptors are ionized, we have $p(x) = N'_{AB}(x)$, and the field in the base is

$$\mathcal{E} = \frac{kT}{q N'_{AB}(x)} \frac{dN'_{AB}(x)}{dx} \quad (9.47)$$

Equation (9.47) is a general relation for nondegenerate semiconductors.

For the special case where the doping profile in the base varies exponentially with position, a condition that is a reasonable approximation in many BJTs,

$$N'_{AB}(x) = N'_{AB}(x_0) e^{-(x-x_0)/\lambda} \quad (9.48)$$

where $N'_{AB}(x_0)$ is the doping level at some position $x = x_0$ in the base and λ is a characteristic length that is a measure of the doping gradient. An exponential doping gradient results in a constant electric field:

$$\mathcal{E} = -\frac{kT}{q\lambda} \quad (9.49)$$

From Poisson's equation,

$$\frac{d\mathcal{E}}{dx} = \frac{Q_V}{\epsilon} \quad (9.50)$$

where Q_V is the charge per unit volume. Since \mathcal{E} is constant, however, $Q_V = 0$, meaning the base is electrically neutral. If N_{AB} is not exponential, a small net charge exists and the interior of the base is referred to as a *quasi-neutral* region.

Let us take an example. We consider the impurity concentration of the graded-base transistor of Figure 9.11. Here the net doping concentration $|N_D - N_A|$ is plotted against depth from the surface. On a semilog plot, a straight line represents an exponential distribution. Thus, from the graph, over most of the base region we can approximate the net acceptor concentration as in Equation (9.48).

It is convenient to introduce the parameter [4]

$$\eta = \frac{W_B}{\lambda} = (W_B \mathcal{E}) \left(\frac{q}{kT} \right) \quad (9.51)$$

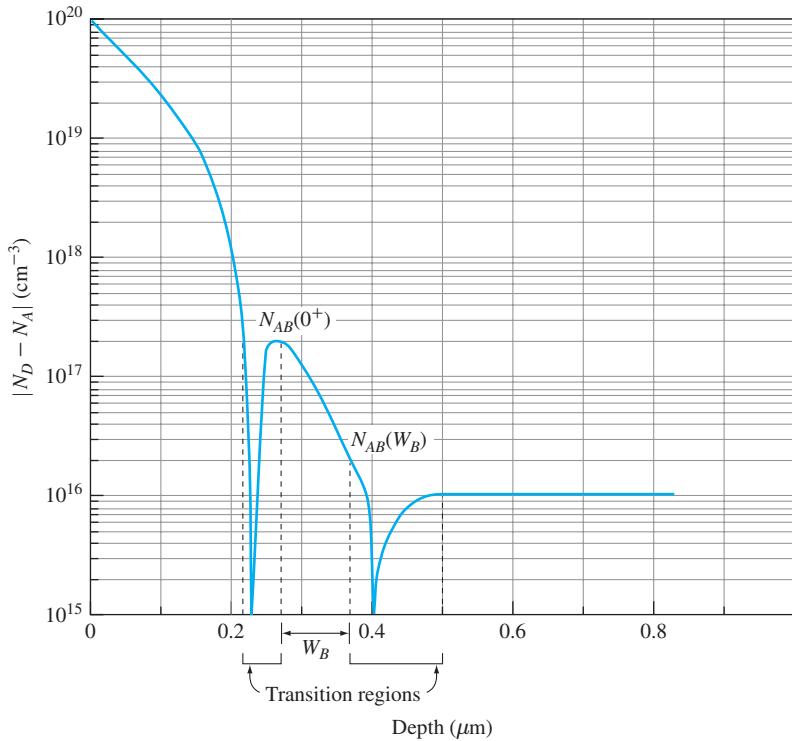


Figure 9.11 Doping profile of a graded-base transistor. The base doping profile as a function of position x is approximated as a straight line in this semilog plot, indicating an exponential impurity distribution (and thus a constant electric field) in the base.

This is, in effect, a measure of the value of the voltage drop across the base, since it varies with the parameter λ , which describes the doping profile. Substituting Equation (9.48) into Equation (9.51) and solving for η gives

$$\eta = \frac{W_B}{x} \ln \frac{N_A(0^+)}{N_A(x)} \quad (9.52)$$

Evaluating this expression at $x = W_B$ results in

$$\eta = \ln \frac{N_A(0^+)}{N_A(W_B)} \quad (9.53)$$

The electric field strength can then be written

$$\mathcal{E} = \frac{kT}{q} \frac{1}{\lambda} = \frac{kT}{q} \frac{\eta}{W_B} \quad (9.54)$$

Let us apply these results.

EXAMPLE 9.4

Find η and \mathcal{E} in the base for the transistor of Figure 9.11.

Solution

Since N_{AB} approximates a straight line on this semilog plot, it can be considered an exponential function, and so we can use the analysis above. From Figure 9.11, the base width is $W_B = 0.10 \mu\text{m}$. The doping at each end of the base is

$$N_{AB}(0^+) = 2 \times 10^{17} \text{ cm}^{-3} \quad N_{AB}(W_B) = 2 \times 10^{16} \text{ cm}^{-3}$$

From Equation (9.53), we can find the factor

$$\eta = \ln \frac{2 \times 10^{17}}{2 \times 10^{16}} = 2.3$$

From Equation (9.54) the field strength is

$$\mathcal{E} = 0.026 \times \frac{2.3}{0.10} = 0.6 \text{ V}/\mu\text{m} = 6 \text{ kV}/\text{cm}$$

Note that for this high a field, electrons in the base are traveling near their saturation velocity.

Also note that this constant built-in field is established by the exponential *majority* carrier (acceptor) distribution. However, this field affects the *minority* carrier electron current in the base (and the rate at which the electrons cross the base to arrive at the collector). We can analyze this situation as follows. The electron current density in the base is the sum of the diffusion and drift currents:

$$J_{nB} = q\mu_n n_B \mathcal{E} + qD_n \frac{dn_B}{dx} \quad (9.55)$$

Again using the relation $D/\mu = kT/q$, and solving Equation (9.55) for dn_B/dx , we obtain

$$\frac{dn_B}{dx} = \frac{J_{nB}}{qD_n} - \frac{qn_B \mathcal{E}}{kT} \quad (9.56)$$

We can substitute the expression for \mathcal{E} from Equation (9.54) into Equation (9.56). We also set the electron concentration at the collector end of the base to $n(W_B) = 0$, since the collector is reverse biased. This expression for the electron distribution in the base, $n_B(x)$, then becomes

$$n_B(x) = \frac{J_{nB} W_B}{q D_n} \left[\frac{1 - e^{\eta(x/W_B - 1)}}{\eta} \right] \quad \text{graded-doping base} \quad (9.57)$$

where D_n is evaluated at the electric field associated with η . (Recall that at large \mathcal{E} , μ and D decrease with increasing \mathcal{E} .)

The term outside the brackets is a constant for a given transistor, for a given bias level. The bias level determines J_{nB} . We will therefore plot the normalized electron concentration distribution in the base,

$$n_{B(\text{norm})} = \frac{n_B(x)}{\frac{J_{nB} W_B}{q D_n}}$$

as a function of normalized distance x/W_B , Figure 9.12. This normalization is useful because it means that the plot can be applied to any exponentially doped region.⁵ The plot shows the term in brackets, with η as a parameter. The case of $\eta = 0$ is the case for uniform base doping. With increasing field (increasing η) the carrier distribution with distance becomes flatter, meaning that the diffusion current ($J_{n(\text{diff})} = q D_n d n_B / dx$) becomes increasingly less important. The current becomes more dominated by drift. Note that for a given J_{nB} , the total electron charge in the base decreases with increasing η .

The current J_{nB} in each curve in Figure 9.12 is the same, so with increasing η , less J_{pE} is required to achieve a given J_{nB} due to the increased drift. Thus J_{nE} (and J_C) increases for the same $n_B(0^+)$.

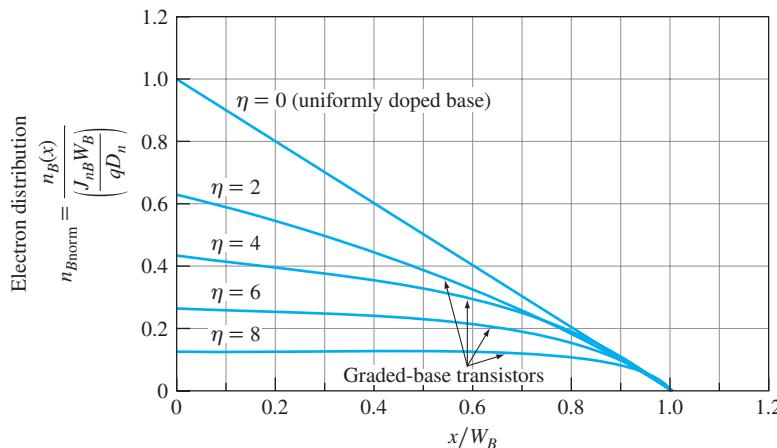


Figure 9.12 The distribution of excess carriers across the base of a graded-base npn transistor for a given I_C . The electron distribution is normalized as indicated, and the horizontal axis is a function of distance across the base. The distributions are plotted for various values of the grading parameter η ; a uniformly doped transistor has an η of 0. As η and thus \mathcal{E} increases, the slope of the concentration decreases, particularly at the emitter end, meaning the current is carried less by diffusion and more by drift. This plot assumes that D_n is independent of \mathcal{E} and thus η .

⁵This formulation is most useful for \mathcal{E} small enough that electrons are not traveling at their saturation velocity and the diffusion coefficient is independent of \mathcal{E} , e.g., the low-field diffusion coefficient.

As an example, consider a transistor having the fairly typical value of $\eta = 4$. We see from Figure 9.12 that for $\eta = 4$, the electron concentration is reasonably constant over most of the base. Since the diffusion current is $J_{nB} = qD_n dn/dx \approx 0$, the diffusion current is negligible and drift predominates. With increasing x , toward the collector end of the base region, diffusion current becomes more pronounced.

9.5.2 EFFECT OF BASE FIELD ON β

The presence of a field in the base alters the mechanism by which the injected carriers cross the base, and thus will have an effect on β . We recall that

$$\beta = \frac{I_C}{I_B} \approx \frac{J_{nB}}{J_{pE}} \quad (9.58)$$

We use Equation (9.57) for a graded-base transistor and solve for J_{nB} at $x = 0^+$:

$$J_{nB} = \frac{qD_n n_B(0^+)}{W_B} \left(\frac{\eta}{1 - e^{-\eta}} \right) \quad (9.59)$$

For a given doping at the base edge $N'_{AB}(0^+)$ and emitter-base voltage, the minority carrier concentration $n_B(0^+)$ and back injection J_{pE} are constant. Thus, the only thing different between this transistor and a uniformly doped one is the factor in brackets. To see the effect of η on the current gain, we again normalize, this time normalizing the value of β for a graded-doping $\beta(\eta)$ to the uniformly doped transistor's $\beta(0)$ (η is zero). We therefore write for the normalized β

$$\frac{\beta(\eta)}{\beta(0)} \approx \frac{\eta}{1 - e^{-\eta}} \quad (9.60)$$

This function is plotted in Figure 9.13. We can see that β increases with increasing η , or increasing base field. At $\eta = 4$, for example, β is 4 times larger than for the uniform base device.⁶ We also note that for negative η (N'_{AB} increasing with x), β is quite small.

9.6 HETEROJUNCTION BIPOLAR TRANSISTORS (HBTS)

We have seen that, in a homojunction BJT, to increase β we wanted to maximize the injection efficiency γ , so the emitter was doped much more heavily than the base. The heavy emitter doping, however, produced impurity-induced band-gap narrowing in the emitter, which resulted in the potential energy barrier for electrons being greater than that for holes for an npn transistor, Figure 9.14a.

⁶These results are somewhat optimistic, since the analysis neglects saturation velocity at high base fields.

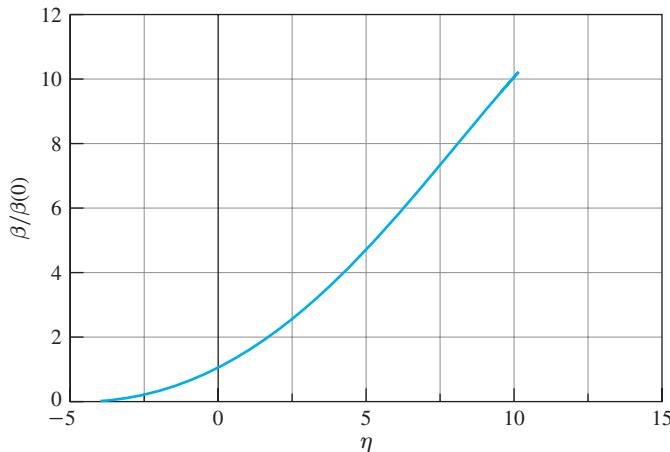


Figure 9.13 The effect of the grading parameter η on the current gain β of a BJT. If the grading goes in the wrong direction ($\eta < 0$), β actually becomes smaller.

This lowered the electron injection efficiency while the back-injection of holes remained the same, reducing β . Recall from Example 9.2 that the result is that $\beta \approx \gamma/(1 - \gamma)$, which is reduced by the factor $e^{-\Delta E_{gBE}^*/kT}$, where

$$\Delta E_{gBE}^* = E_g^*(\text{base}) - E_g^*(\text{emitter}) \quad (9.61)$$

In effect, the heavy doping produced an extra barrier that worked against us. We would like it to be the other way around—we'd like to *increase* the barrier to back-injected holes while we *decrease* the barrier for the injected electrons.

This can be done by choosing an emitter material with a larger band gap than that of the base, Figure 9.14b. Such a structure is referred to as a heterojunction bipolar transistor, or HBT.

9.6.1 UNIFORMLY DOPED HBT

Consider the uniformly doped $\text{Al}_x\text{Ga}_{1-x}\text{As}:\text{GaAs}$ (AlGaAs:GaAs) HBT, where the base and collector are GaAs and whose equilibrium energy band diagram is shown in Figure 9.14b. Here we assume that each region (emitter, base, and collector) is uniformly doped and nondegenerate. The emitter consists of the wider band-gap material $\text{Al}_x\text{Ga}_{1-x}\text{As}$, and the base and collector are made of GaAs. There is a discontinuity ΔE_C in the conduction band edge and a discontinuity ΔE_V in the valence band edge. Normally, during fabrication the Al content of the emitter is graded near the base contact such that the potential energy spike is not present, as indicated in Figure 9.14c. We can see that the barrier for electrons (E_{Bn}) is appreciably smaller than that for holes (E_{Bp}).

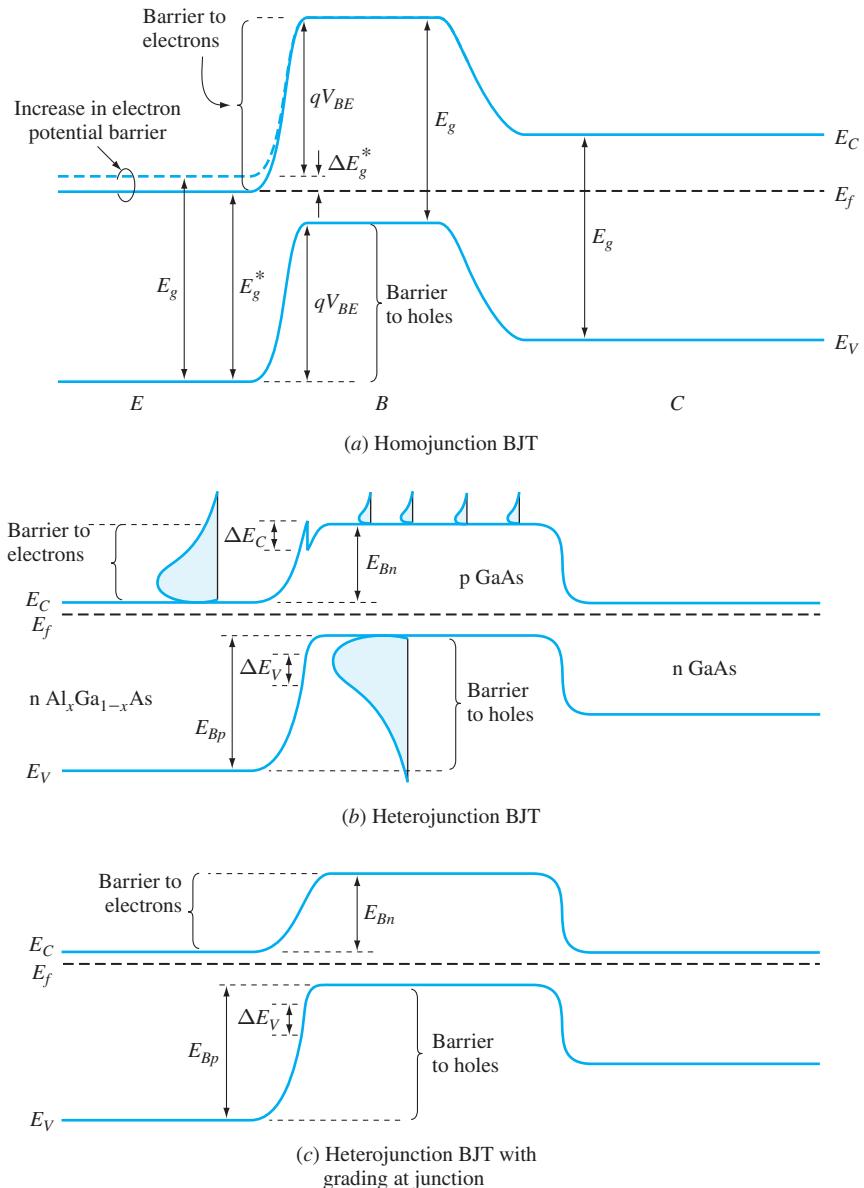


Figure 9.14 Equilibrium energy band diagrams of BJTs. (a) Homojunction npn transistor: The degenerately doped emitter causes band-gap narrowing, which increases the barrier to injection of electrons and reduces β . (b) The heterojunction bipolar transistor (HBT) uses a wide-band-gap emitter to create a large barrier to holes to reduce back-injection. (c) The wide-gap material is graded near the emitter-base junction to eliminate the potential energy spike in the conduction band.

Figure 9.15 shows the energy band diagram for this HBT operating in the active mode. The injected electron current is several orders of magnitude greater than the injected hole current. In practical devices, however, unless the emitter and base have virtually identical lattice constants, β is smaller than predicted. This is because if the lattices are not well matched, there is some base-emitter current that flows by recombination via interface states at the heterojunction, which reduces the injection efficiency. Still, for this structure, β can be made appreciably greater than for homojunction transistors.

There are other advantages to using an HBT structure apart from increased current gain. We saw earlier that in a homojunction transistor, to get adequate γ and thus β , the emitter must be much more heavily doped than the base. In a homojunction BJT, however, γ (and thus β) is degraded by the ΔE_{gBE}^* resulting from the heavy emitter doping. In a heterojunction, γ is controlled by the differences in the band gaps, removing these doping restrictions. Thus, the base can be heavily doped to reduce the base resistance, while the emitter can be lightly doped to reduce the emitter-base junction capacitance.

A plot of β as a function of collector current I_C for an npn $\text{Al}_x\text{Ga}_{1-x}\text{As}$: GaAs HBT is shown in Figure 9.16 for three temperatures. [5] For this device the aluminum fraction is $x = 0.4$, the emitter thickness is $W_E = 1.0 \mu\text{m}$, the base width $W_B = 0.4 \mu\text{m}$, and the doping concentrations are $N'_{DE} = 3 \times 10^{17} \text{ cm}^{-3}$,

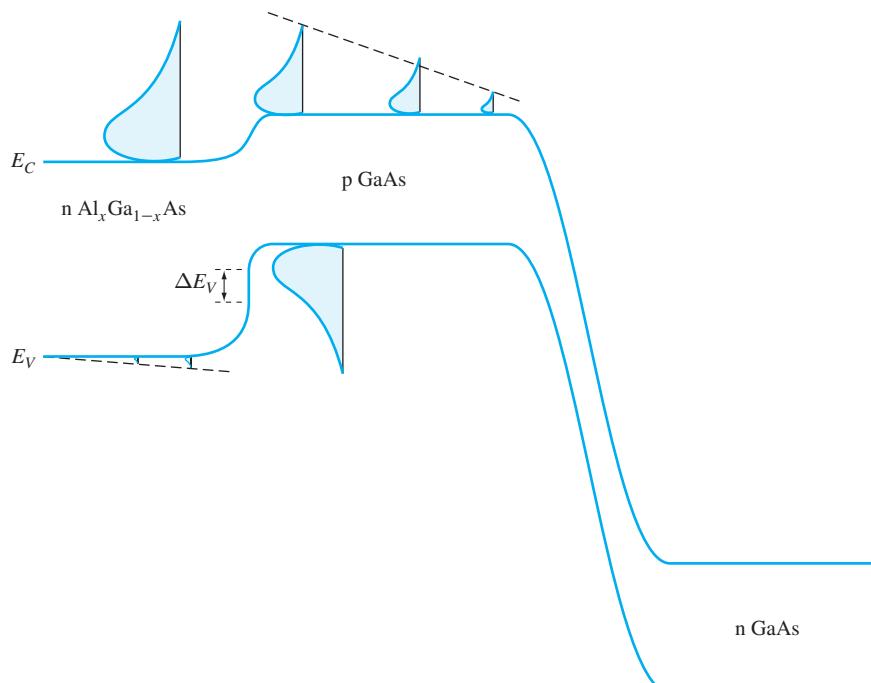


Figure 9.15 The HBT under forward active bias.

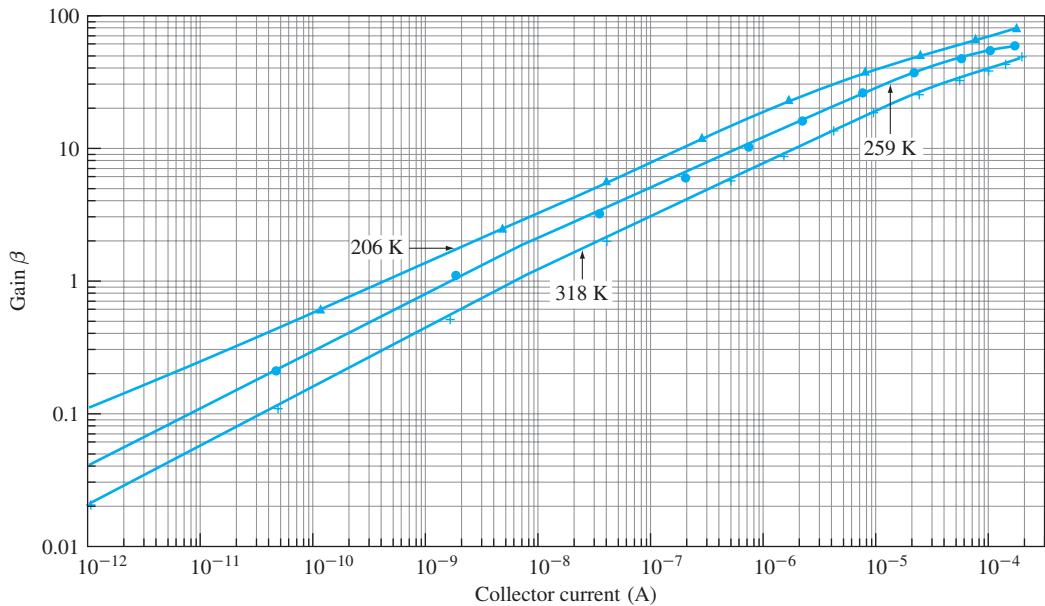


Figure 9.16 Gain versus collector current for an AlGaAs:GaAs:GaAs HBT at three temperatures.

$N'_{AB} = 10^{18} \text{ cm}^{-3}$, and $N'_{DC} = 10^{15} \text{ cm}^{-3}$. [6] The current gain is low at small I_C , since here the parasitic (recombination) currents predominate. The increase in β with increasing I_C results from the greater dependence of I_C on V_{BE} ($I_C \propto e^{qV_{BE}/kT}$) compared with the dependence of the recombination current on V_{BE} ($I_B \propto e^{qV_{BE}/nKT}$, where $n \approx 2$).

It is seen in Figure 9.16 that β increases with decreasing temperature. This is the opposite behavior of homojunction transistors, where β decreases with decreasing temperature. This can be explained as follows: For an npn homojunction transistor with degenerate emitter and nondegenerate uniformly doped base, from Equation (9.43),

$$\beta \approx \frac{N'_{DE}}{N'_{AB}} \frac{D_{nB}}{D_{pE}} \frac{W_E}{W_B} e^{-\Delta E_{gBE}^*/kT} \quad (9.62)$$

The temperature dependence of the diffusion constants is small compared with that of the term $e^{-\Delta E_{gBE}^*/kT}$, which is the impurity-induced reduction of the apparent band gap in the emitter. Thus, to first approximation,

$$\beta \propto e^{-\Delta E_{gBE}^*/kT} \quad (9.63)$$

and β decreases with decreasing temperature. The value of ΔE_{gBE}^* (which is positive) can then be estimated from the slope ($\Delta E_{gBE}^*/k$) of a plot of $\ln \beta$ as a function of $1/T$.

In a heterojunction transistor, however, ΔE_{gBE}^* is normally negative since the emitter band gap is larger than that of the base, and β can increase with decreasing temperature. While the current gain of the Si homojunction transistor decreases appreciably with decreasing temperature, that of the HBT increases. The HBT therefore appears to have promise for operation at cryogenic temperatures.

9.6.2 GRADED-COMPOSITION HBT: (Si: SiGe-BASE: Si HBT)

The results of Figure 9.16 are for an HBT with uniform base doping. Just as for a homojunction, performance can be improved by appropriately grading the base doping to create a built-in field in the base, which accelerates the minority carriers to the collector. [7]

Another way to create a built-in field in the base is to grade the composition of the material, so that the band gap changes with position across the base. Such a grading is shown schematically in Figure 9.17 for an npn Si:Si_xGe_{1-x}:Si HBT with a uniformly doped base. [8] Here both the emitter and collector are silicon, and the base is a SiGe alloy. The Ge content is increased from zero slightly inside the Si emitter to the final Ge content somewhat inside the collector. The grading continues into the collector to avoid potential energy spikes in the conduction band. Since the band gap E_g decreases with increasing Ge content, the E-B junction is a heterojunction, graded in this case to avoid spikes in the band edges. The decreasing E_g with position induces a built-in field for electrons in the base.

The base field can be further increased by combining a graded composition with graded doping in the base, as shown in the SIMS plot of Figure 9.18. [9]

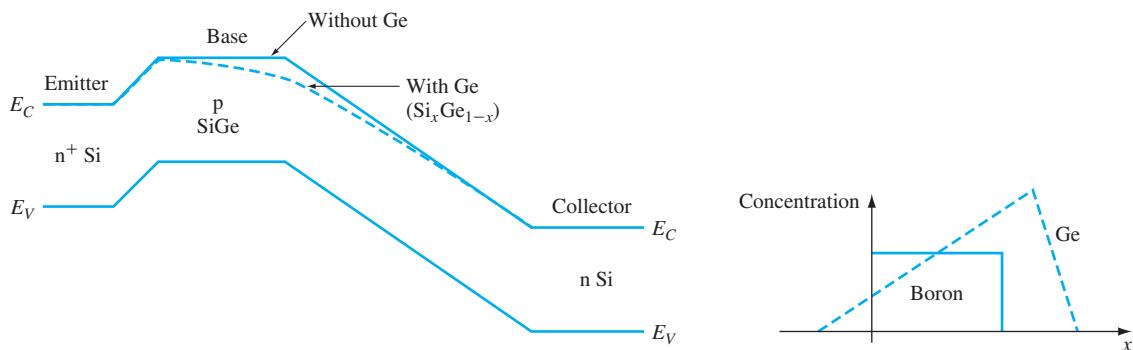


Figure 9.17 The Si:Si_xGe_{1-x}:Si HBT energy band diagram (left) and plot of germanium concentration (right). The addition of Ge starts slightly into the emitter and extends a short distance into the collector. (Source: Adapted from John D. Cressler and Katsuyoshi Washio, "Bipolar Transistors," in *Guide to State-of-the-Art Electron Devices*, Wiley/IEEE Press, John Wiley and Sons Ltd, pp. 3–20, 2013.)

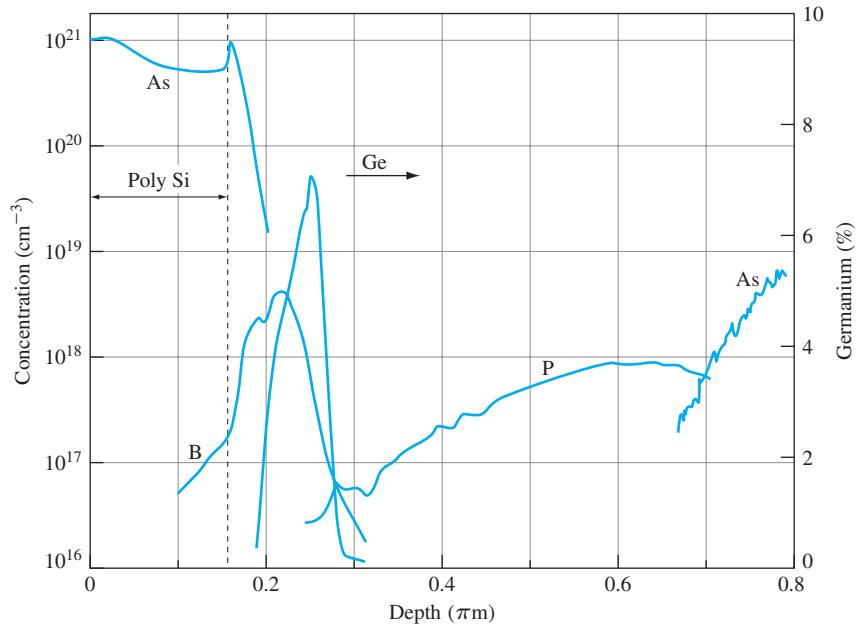


Figure 9.18 SIMS data showing the concentrations of the dopants and the Ge in the $\text{Si:Si}_x\text{Ge}_{1-x}\text{:Si}$ heterojunction bipolar transistor. (Source: Adapted from G. Niu, J. D. Cressler, S. Zhang, W. E. Ansley, C. S. Webster, and D. L. Harame, “A unified approach to RF and microwave noise parameter modeling in bipolar transistors,” *IEEE Transactions on Electron Devices*, 48, no. 11, pp. 2568–2574, 2001.)

Here the increasing Ge content and the decreasing acceptor doping in the base combine to create a high base field (on the order of 10^3 to 10^4 V/cm). This field reduces the base transit time, which increases the frequency response. Values for the unity gain frequency f_T well in excess of 300 GHz have been measured, compared to about 50 GHz for Si BJTs. [10]

In SiGe alloys, however, there is a difference in lattice constant (4 percent) between the Si and the Ge. Therefore we would expect dangling bonds, and thus interface states, to be present at the emitter-base junction and within the base because of the variation of germanium concentration with position. These states would allow recombination at the interface and within the base and degrade the current gain β of the transistor. The lattice-matching defects can be avoided, however, if the Ge content is kept relatively small (less than about 20 percent), and the alloy layer is thin. In this case, the $\text{Si}_x\text{Ge}_{1-x}$ alloy on one side of the interface bonds with the Si on the other side atom for atom, without dangling bonds. Because this produces strain in the alloy, it is referred to as a *strained layer*. This strain in the SiGe base is compressive, which reduces the base band gap, and that further increases the injection efficiency. If the layer is too thick or the Ge

content too high, the strain will be relieved with the creation of interband states. These will trap carriers and thus reduce β .

We make another point about HBTs. In Section 9.6.1 we saw that β would be expected to increase with decreasing temperature, and this is experimentally observed. Figure 9.19 shows plots of β normalized to its room temperature value [$\beta(T)/\beta(300 \text{ K})$] as a function of inverse temperature $1/T$. The figure compares a Si homojunction transistor to a $\text{Si}:\text{Si}_x\text{Ge}_{1-x}:\text{Si}$ heterojunction transistor.

As indicated earlier, the graded doping in a BJT increases the Early voltage and reduces the base transit time as well as increasing β . Figure 9.20 shows the relative improvements in these quantities as a function of maximum bandgap difference across the base of a $\text{Si}:\text{Si}_x\text{Ge}_{1-x}:\text{Si}$ BJT relative to that of a similar device with a Si base. [11]

9.6.3 DOUBLE HETEROJUNCTION BIPOLAR TRANSISTOR (DHBT)

So far, we have discussed single heterojunction bipolar transistors, in which the emitter bandgap is larger than that in the base and collector, shown schematically in Figure 9.21a. This type is called a “wide-bandgap-emitter HBT.”

In a double heterojunction bipolar transistor the emitter and collector have wider bandgaps than the base as indicated in Figure 9.21b. Here the emitter-base and base-collector junctions are composition-graded to eliminate their spikes.

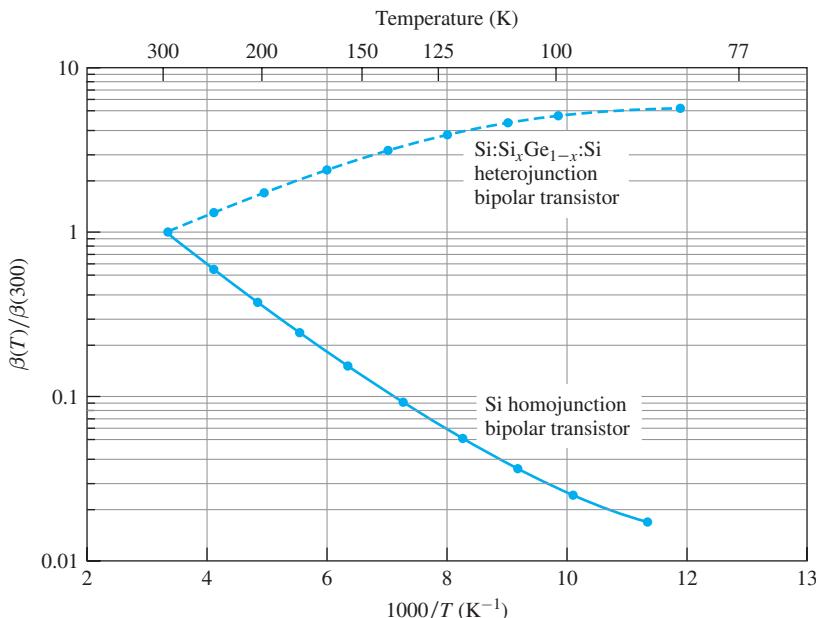


Figure 9.19 Normalized variation of β with temperature for a Si bipolar transistor (solid line) and a $\text{Si}:\text{Si}_x\text{Ge}_{1-x}:\text{Si}$ HBT (dashed line).

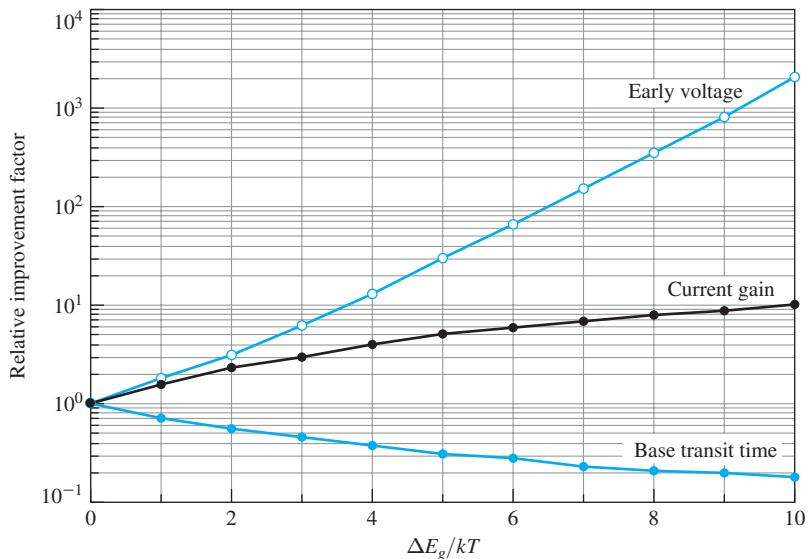


Figure 9.20 Improvement in Early voltage, current gain, and base transit time for a Si:Si_xGe_{1-x}:Si BJT compared with a similar device with a silicon base. (Source: Adapted from T. Ning, “History and Future Perspectives of the Modern Silicon Bipolar Transistor,” *IEEE Trans. Electron Devices*, 48, pp. 2485–2491, © IEEE 2001.)

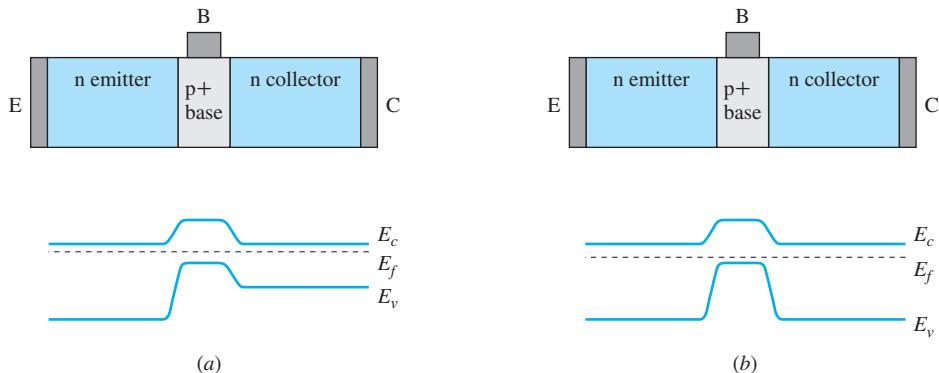


Figure 9.21 Cross section of a wide-bandgap-emitter HBT (a) and a double heterojunction bipolar transistor (b) with equilibrium energy band diagrams.

A double heterojunction transistor has advantages over a wide-bandgap-emitter HBT. Because the doping in the collector is less than that in the base, under reverse C-B bias, most of the voltage is dropped in the collector. Also, since the collector has a larger bandgap, the breakdown voltage is increased. Another advantage of the larger collector bandgap is that under forward bias,

a negligible number of holes (minority carriers) are injected into the collector. Recall that in switching from forward to reverse bias, these minority carriers must be removed by diffusion (a relatively slow process), so in the DHBT the switching time is greatly reduced. Another advantage is that a DHBT can be used in symmetrical operation, i.e., the collector can be used as emitter and vice versa.

9.7 COMPARISON OF Si-BASE, SiGe-BASE, AND GaAs-BASE HBTS

Device parameters of Si (Si-base) BJTs, SiGe-base, and GaAs-base HBTs are presented for devices similar in structure and at the same collector current density.

- The base transit time of a SiGe-base transistor is less than for a Si-base device since the graded composition in the base creates a quasi-electric field for electrons, which accelerates them to the collector.
- The base transit time for a GaAs-base HBT is smaller than for a Si-base BJT because of the higher low-field electron mobility of GaAs.
- The base transit time of a (uniformly doped) GaAs-base HBT is comparable to that of a SiGe-base device.⁷
- The base collector junction transit times t_{tBC} are comparable because for the high fields in the base-collector depletion region, the electron velocities are at their saturation values, which are comparable.
- The emitter-base junction capacitances of the SiGe-base and Si-base heterojunction BJTs are comparable because of the comparable emitter and base dopings. For the GaAs-base HBT, however, the emitter-base junction capacitance is much reduced since the emitter can be lightly doped (on the order of 10^{17} cm^{-3} compared with a doping on the order of 10^{19} cm^{-3} for the Si-base and SiGe-base transistors).
- The base resistance of a GaAs-base HBT is about an order of magnitude smaller than that of the Si-base and SiGe-base devices since the GaAs base can be heavily doped (on the order of 10^{19} cm^{-3}) without degrading the current gain.
- A major advantage of the SiGe-base device over the GaAs-base HBT is that SiGe is compatible with Si integrated circuit processing and can be used to fabricate BiCMOS circuits.

9.8 THE BASIC EBERS-MOLL DC MODEL

Until now, we have considered only BJT operation in the active mode, in which the emitter-base junction is forward biased and the base-collector junction is reverse biased. This mode of operation is used for analog or linear circuits. In digital circuits, however, operation in all four of the modes that were listed in

⁷The base transit time of a GaAs-base HBT can be decreased by grading the base doping and/or composition.

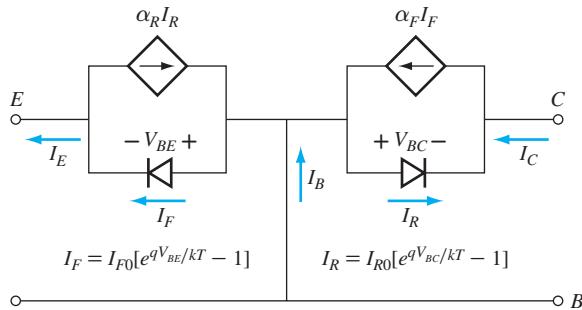


Figure 9.22 The Ebers-Moll model represents a bipolar junction transistor as two interacting back-to-back diodes. This is the common-base configuration.

Table IV.2 is possible. To model transistors operating in digital switching circuits, the Ebers-Moll model [12] is commonly used.

In the Ebers-Moll model, an npn BJT is represented by two back-to-back diodes with interaction between the two diodes, as shown in Figure 9.22. The figure shows the common-base configuration. Each junction is represented by a diode in the model. Each junction also has a dependent current source whose current depends on the current in the *other* junction.

Let I_F be the forward-biased emitter-base current. Because of the thin base region, a fraction of the carriers injected into the base from the forward-biased emitter-base junction will flow into the collector. The factor α_F is the forward common-base current gain. Then $\alpha_F I_F$ represents the emitter-base current that is collected by the reverse-biased collector. This current is modeled by the dependent current source belonging to the collector-base junction. Thus far, we can model the active mode of operation.

For operation in the inverse mode, the collector-base junction is forward biased while the base-emitter junction is reverse biased. For this model, the collector-base current is represented by I_R , while $\alpha_R I_R$ is that current collected by the reverse-biased base-emitter junction.⁸ This current is represented by the dependent current source corresponding to the emitter-base junction. Here α_R is called the inverse-mode common-base current gain.

Normally transistors work poorly in the inverse mode, and $\alpha_R \ll \alpha_F$ (except for the DHBT where they can be comparable). Recall that the current transfer ratio $\alpha = I_C/I_E \approx \gamma \alpha_T$. We showed in Section 9.4 that to maximize the injection ratio γ , the emitter should be much more heavily doped than the base. In the inverse mode, the collector is acting as an emitter, but because the collector

⁸The subscripts F and R may be confusing; they refer to the bias of the respective junctions under forward active bias conditions. In the inverse mode, the emitter-base junction is reverse biased and the current I_F is actually a reverse-bias current. The model will still work, however; α_F is still the fraction of emitter-base current arriving at the collector and α_R is the fraction of the base-collector current arriving at the emitter.

doping is less than the base doping, the injection efficiency γ for inverse operation is small. Further, the emitter area is normally much smaller than the collector area. Thus, few electrons injected from collector to base find their way to the emitter, resulting in a small (inverse-mode) transport efficiency, α_T .

The currents I_F and I_R of the diodes in the model are the usual diode currents and can be expressed

$$\begin{aligned} I_F &= I_{F0}(e^{qV_{BE}/kT} - 1) \\ I_R &= I_{R0}(e^{qV_{BC}/kT} - 1) \end{aligned} \quad (9.64)$$

where I_{F0} and I_{R0} are the saturation currents for the emitter-base and collector-base diodes respectively. The emitter and collector terminal currents are

$$\begin{aligned} I_E &= I_F - \alpha_R I_R \\ I_C &= \alpha_F I_F - I_R \end{aligned} \quad (9.65)$$

or, from Equation (9.64),

$$\begin{aligned} I_E &= I_{F0}(e^{qV_{BE}/kT} - 1) - \alpha_R I_{R0}(e^{qV_{BC}/kT} - 1) \\ I_C &= \alpha_F I_{F0}(e^{qV_{BE}/kT} - 1) - I_{R0}(e^{qV_{BC}/kT} - 1) \end{aligned} \quad (9.66)$$

The base terminal current is then

$$I_B = I_E - I_C = I_F(1 - \alpha_F) + I_R(1 - \alpha_R) \quad (9.67)$$

The forward and reverse common-base current gains are related by the expression [13]

$$\alpha_F I_{F0} = \alpha_R I_{R0} = I_S \quad (9.68)$$

where the term I_S is introduced since it is a parameter used in SPICE.

For circuit analysis, it is normally more convenient to use the common-emitter representation of the Ebers-Moll model shown in Figure 9.23, where

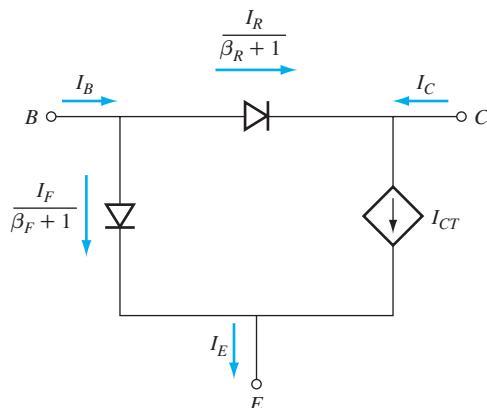


Figure 9.23 The common-emitter representation of the Ebers-Moll model.

$$\begin{aligned}\beta_F &= \frac{\alpha_F}{1 - \alpha_F} & \alpha_F &= \frac{\beta_F}{\beta_F + 1} \\ \beta_R &= \frac{\alpha_R}{1 - \alpha_R} & \alpha_R &= \frac{\beta_R}{\beta_R + 1} \\ I_{CT} &= (\alpha_F I_F - \alpha_R I_R) = \frac{\beta_F I_F}{\beta_F + 1} - \frac{\beta_R I_R}{\beta_R + 1}\end{aligned}\quad (9.69)$$

The terminal currents are then

$$\begin{aligned}I_E &= I_{CT} + \frac{I_F}{\beta_F + 1} = \left(1 + \frac{1}{\beta_F}\right) I_S (e^{qV_{BE}/kT} - 1) - I_S (e^{qV_{BC}/kT} - 1) \\ I_C &= I_{CT} - \frac{I_R}{\beta_R + 1} = I_S (e^{qV_{BE}/kT} - 1) - \left(1 + \frac{1}{\beta_R}\right) I_S (e^{qV_{BC}/kT} - 1) \\ I_B &= \frac{I_F}{\beta_F + 1} + \frac{I_R}{\beta_R + 1} = \frac{I_S}{\beta_F} (e^{qV_{BE}/kT} - 1) + \frac{I_S}{\beta_R} (e^{qV_{BC}/kT} - 1)\end{aligned}\quad (9.70)$$

where the right-hand side of Equation (9.70) uses SPICE variable names. However in SPICE, I_E is defined as positive in the opposite direction to the conventional direction of I_E . Thus in SPICE, I_E will have the opposite sign from that indicated above.

EXAMPLE 9.5

Find the terminal currents for an npn BJT operating in the active mode.

■ Solution

In the active mode $e^{qV_{BE}/kT} \gg 1$ and $e^{qV_{BC}/kT} \ll 1$. Then from Equation (9.70),

$$\begin{aligned}I_E &= \left(1 + \frac{1}{\beta_F}\right) I_S e^{qV_{BE}/kT} = \frac{I_C}{\alpha_F} \\ I_C &= I_S e^{qV_{BC}/kT} \\ I_B &= \frac{I_S}{\beta_F} e^{qV_{BC}/kT} = \frac{I_C}{\beta_F}\end{aligned}$$

The Ebers-Moll equations, or variations of them (e.g., the Gummel-Poon equations) [14] are often used in circuit analysis programs such as SPICE.⁹

⁹The Gummel-Poon equations include second-order effects not present in the Ebers-Moll model, e.g., nonconstant β .

9.9 SUMMARY

In this chapter, we have introduced some concepts fundamental to the operation of bipolar junction transistors. In BJTs, carriers are injected into the base from the emitter. Ideally, all of the injected carriers cross the base and become collected. The amount of injection depends strongly on the barrier height at the emitter-base junction, so that small changes in base voltage create large changes in transistor current.

BJTs are characterized by the currents in the various modes of operation. A figure of merit for BJTs operating in the active mode used for analog circuits is

$$\beta = \frac{I_C}{I_B} = \frac{\gamma \alpha_T M}{1 - \gamma \alpha_T M}$$

where, for a prototype nondegenerate emitter with $W_B \ll L_{nB}$ and $W_E \ll L_{pE}$,

$$\gamma = \frac{1}{1 + \frac{N'_{AB} D_{pE}}{N'_{DE} D_{nB}} \frac{W_B}{W_E}} \quad \text{uniformly doped npn}$$

$$\gamma = \frac{1}{1 + \frac{N'_{DB} D_{nE}}{N'_{AE} D_{pB}} \frac{W_B}{W_E}} \quad \text{uniformly doped pnp}$$

and

$$\alpha_T = 1 - \frac{W_B^2}{2 L_{nB}^2} \quad \text{npn}$$

$$\alpha_T = 1 - \frac{W_B^2}{2 L_{pB}^2} \quad \text{pnp}$$

But since α_T and M are close to unity,

$$\beta \approx \frac{\gamma}{1 - \gamma}$$

For both emitter and base nondegenerate (not a common situation),

$$\beta_{npn} \approx \frac{I_{nE}}{I_{pE}} \approx \frac{N'_{DE} D_{nB}}{N'_{AB} D_{pE}} \frac{W_E}{W_B} \quad \text{nondegenerate emitter and base}$$

$$\beta_{pnp} \approx \frac{I_{pE}}{I_{nE}} \approx \frac{N'_{AE} D_{pB}}{N'_{DB} D_{nE}} \frac{W_E}{W_B} \quad \text{nondegenerate emitter and base}$$

The normal situation is for the emitter to be degenerately doped. The heavy emitter doping causes band-gap narrowing in the emitter, which tends to increase the barrier for electron injection and significantly reduce the current gain β . For degenerate emitter and nondegenerate uniform base,

$$\beta_{nnp} \approx \frac{N'_{DE} D_{nB}}{N'_{AB} D_{pE}} \frac{W_E}{W_B} e^{-\Delta E_{gBE}^*/kT} \quad \text{n-p-n degenerate emitter, uniform base}$$

$$\beta_{pnp} \approx \frac{N'_{AE} D_{pB}}{N'_{DB} D_{nE}} \frac{W_E}{W_B} e^{-\Delta E_{gBE}^*/kT} \quad \text{p-n-p degenerate emitter, uniform base}$$

where $\Delta E_{gBE}^* = E_g^*(\text{base}) - E_g^*(\text{emitter})$ is the difference in the apparent band-gap narrowing in base and emitter.

If the impurity gradient and the resultant field in the base are considered, β is increased somewhat from the preceding values. The field in the base increases the fraction of injected electrons reaching the collector, and thus increases the collector current.

Heterojunction bipolar transistors have some inherent advantages over homojunction BJTs. These include increased injection efficiency, thus resulting in higher current gain β and higher operating frequencies. As in a homojunction BJT, the performance can be increased by appropriately grading the base doping, i.e., decreasing the acceptor doping with distance from the emitter edge of the base. The performance can be further increased, as in SiGe base BJTs, by grading the base composition such that its band gap decreases with distance from the emitter edge of the base. The use of a double heterojunction decreases the minority carrier injected into the collector under forward bias, thus increasing the switching speed.

9.10 REFERENCES

1. Yuan Taur and Tak H. Ning, *Fundamentals of Modern VLSI Devices*, 2nd ed. Cambridge University Press, Cambridge, U.K., 2009.
2. S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, 3rd ed, John Wiley & Sons, Inc. Hoboken, 2007.
3. D. B. M. Klassen, J. W. Slotboom and H. C. deGraaff, “Unified apparent bandgap narrowing in n- and p-type silicon,” *Solid State Electron*, 35, 125–129, 1992.
4. Joseph Lindmayer and Charles Wrigley, *Fundamentals of Semiconductor Devices*, Chap. 4, Van Nostrand, Princeton, 1965.
5. R. J. Ferro, “Base and collector current mechanisms in bipolar NPN Gallium aluminum arsenide/gallium arsenide heterojunction transistors,” Ph.D. dissertation, University of Vermont, 1985.
6. A. Marty, G. Rey, and J. P. Bailbe, “Electrical behavior of an NPN AlGaAs/GaAs heterojunction transistor,” *Solid State Electronics*, 22, pp. 549–557, 1979.
7. H. Kroemer, “Theory of wide-gap emitter for transistors,” *Proc. IRE*, 45, pp. 1535–1539, 1957.
8. J. D. Cressler (ed.), *Silicon Heterojunction Handbook: Devices, Circuits, and Applications of SiGe and Si Strained-Layer Epitaxy*, Boca Raton FL, CRC Press. 2006.

9. Guofu Niu, John D. Cressler, Shiming Zhang, William E. Ansley, Charles S. Webster, and David L. Harame, “A unified approach to RF and microwave noise parameter modeling in bipolar transistors,” *IEEE Trans. Electron Devices*, ED48, pp. 2568–2574, 2001.
10. John D. Cressler and Katsuyoshi Washio, “Bipolar transistors,” in *Guide to State-of-the Art Electron Devices*, John N. Burghartz (ed.), John Wiley and Sons, pp. 3–20, 2013.
11. Tak H. Ning, “History and future perspectives of the modern silicon bipolar transistor,” *IEEE Trans. Electron Devices*, 48, pp. 2485–2491, 2001.
12. J. J. Ebers and J. L. Moll, “Large-signal behavior of junction transistors,” *Proc. IRE*, 42, pp. 1761–1772, 1954.
13. R. S. Muller and T. I. Kamins, *Device Electronics for Integrated Circuits*, John Wiley & Sons, New York, 1977.
14. H. K. Gummel and H. C. Poon, “An integral charge control model for bipolar transistors,” *Bell Syst. Tech. J.*, 49, pp. 827–852, 1970.

9.11 REVIEW QUESTIONS

1. What is the purpose of the n^+ sinker in Figure IV.1?
2. Equation (9.1) describes the prototype BJT under forward active bias. Why is it reasonable to neglect the recombination current in the emitter-base junction?
3. Explain why it is advantageous to make the base thin in a bipolar junction transistor.
4. Explain the reasons for doping the emitter heavily compared with the base, and for doping the collector lightly.
5. Why should we reduce back-injection (from the base back into the emitter)? What can be done to control it?
6. Explain how band-gap narrowing in the degenerate emitter affects the current gain β . What is the physics behind this?
7. From the SIMS plot of Figure 9.8, it is clear that the doping is graded in all three regions—the emitter, base, and collector. Why, then, is this device referred to as a *graded-base transistor*?
8. Explain in words how grading the base doping can increase the common-emitter current gain of a transistor.
9. Explain why, in Figure 9.9, E_C bends downward and E_V bends upward toward the surface of the semiconductor.
10. Explain why an increasing V_{CE} causes the collector current to increase, even in the active region.
11. Explain why in an HBT having the emitter band gap greater than that of the base increases β .

12. Why is the base of a BJT doped more lightly than the emitter but in an HBT the base is more heavily doped?
13. What is the purpose of having the collector band gap of a DHBT greater than that of the base?

9.12 PROBLEMS

- 9.1 For each of the transistors in Figure P9.1, indicate the mode of operation (forward active, cutoff, saturation, etc.)

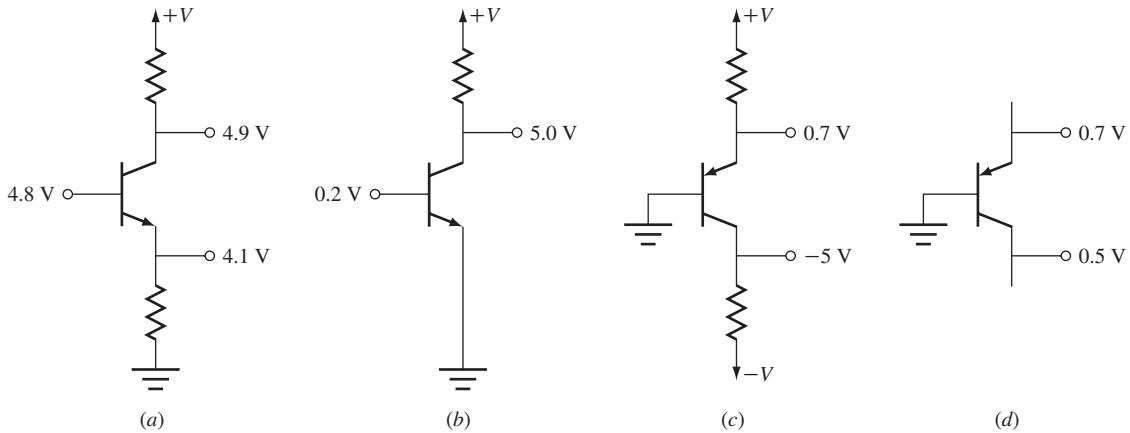


Figure P9.1

- 9.2 For each area in Figure P9.2, identify how the region should be doped to make a good pnp transistor (n^+ , n, intrinsic, p, p^+), and indicate the reason(s). Indicate where the active pnp transistor occurs.

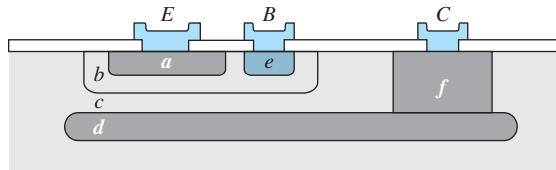


Figure P9.2

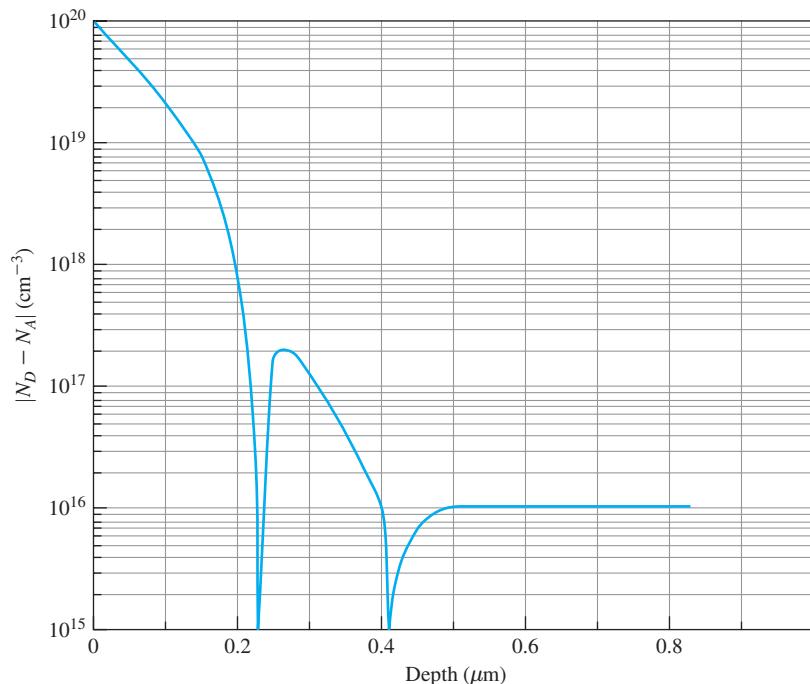
- 9.3 Draw the energy band diagram for a pnp transistor at equilibrium and under forward active bias.
- 9.4 From Equation (9.13), $\alpha = \gamma\alpha_T M$, where M is the carrier multiplication factor in the base-collector junction. For small base-collector voltage, $M = 1$ and $\alpha = \gamma\alpha_T$ and $\beta = \alpha/(1 - \alpha)$. Show that for avalanche breakdown, $M = 1 + 1/\beta$.

- 9.5** Using the prototype model (i.e., ignore the apparent emitter band-gap narrowing reduction on the injection efficiency), find α and β for an npn BJT with $N'_{DE} = 10^{19} \text{ cm}^{-3}$, $N'_{AB} = 2 \times 10^{17} \text{ cm}^{-3}$, and $N'_{DC} = 10^{17} \text{ cm}^{-3}$. Indicate clearly all your steps. Let $V_{BE} = 0.8 \text{ V}$ and $V_{CB} = 2.0 \text{ V}$. The metallurgical widths are $W_{EM} = 0.2 \mu\text{m}$ and $W_{BM} = 0.2 \mu\text{m}$.
- 9.6** For the device of Example 9.1, estimate the emitter-base junction width, the width appearing on the emitter side, and that appearing on the base side. Assume that the junction is forward biased with $V_j = V_{bi} - V_a = 0.2 \text{ V}$. Draw the emitter and base to scale and indicate the locations of the edges of the transition region.
- 9.7** Derive the simple-model expression for β for a pnp transistor with a non-degenerate emitter [the pnp equivalent of Equation (9.42)].
- 9.8** A prototype npn transistor has its emitter doped to $N'_{DE} = 5 \times 10^{19} \text{ cm}^{-3}$. The base is doped with $N_A = 2 \times 10^{17} \text{ cm}^{-3}$. Find ΔE_g^* , accounting for narrowing of the emitter conduction band. Find β if $W_E = 0.12 \mu\text{m}$ and $W_B = 0.07 \mu\text{m}$.
- 9.9** In Example 9.2, we neglected the band-gap narrowing in the base. Find the current gain β , this time taking ΔE_{gB}^* into account. Draw the equilibrium energy band diagram, and indicate the true band edges and apparent band edges in both the emitter and the base. What is the effective barrier height for electrons? Holes?
- 9.10** Suppose that a degenerate Si layer of $N_{DE} = 10^{20} \text{ cm}^{-3}$ and $N_A = 0$ is epitaxially deposited onto the emitter of the transistor of Section 9.4. Explain how such a layer could increase β slightly. Both emitter layers together are still much shorter than the diffusion length of either electrons or holes in the emitter. [Hint: There is a small amount of band-gap narrowing due to the acceptors in the base (and the compensated acceptors in the emitter)].
- 9.11** Research and learn about SIMS, and write up a short explanation of what it does and how it works.
- 9.12** If a pnp transistor is made with the same dimensions and doping concentrations as an npn, explain why the pnp will have a lower β . Neglect band-gap narrowing effects.
- 9.13** If an npn transistor is considered to have all step junctions, with the properties in the table below, find:

$N'_{DE} = 5 \times 10^{19} \text{ cm}^{-3}$	$N'_{AB} = 4 \times 10^{17} \text{ cm}^{-3}$	$N'_{DC} = 10^{17} \text{ cm}^{-3}$
$W_{EM} = 0.13 \mu\text{m}$	$W_{BM} = 0.20 \mu\text{m}$	
$V_{BE} = 0.8 \text{ V}$ (forward bias)	$V_{CB} = 2 \text{ V}$ (reverse bias)	

- The minority carrier lifetimes
- The corresponding diffusion coefficients and diffusion lengths
- The built-in voltage of the collector-base junction

- d. The built-in voltage of the emitter-base junction (neglecting band-gap narrowing)
e. W_B
f. γ , neglecting band-gap narrowing
g. α_T
h. α , assuming $M = 1$,
i. β
- 9.14** Repeat Problem 9.13, but this time account for band-gap narrowing in the emitter.
- 9.15** Find β for an npn transistor with a degenerately doped emitter at $N'_{DE} = 10^{20} \text{ cm}^{-3}$ and $N'_{AB} = 5 \times 10^{17} \text{ cm}^{-3}$. Let $W_E = 0.15 \mu\text{m}$ and $W_B = 0.1 \mu\text{m}$.
- 9.16** Consider the graded-base transistor whose SIMS profile is shown in Figure P9.3.
- Find η .
 - Find the built-in electric field in the base.
 - Estimate β .

**Figure P9.3**

- 9.17** A graded-base transistor has its emitter degenerately doped to 10^{19} cm^{-3} and its base has a doping concentration of $5 \times 10^{17} \text{ cm}^{-3}$ at the emitter

edge and a grading parameter $\eta = 3$. The emitter width is $W_E = 0.2 \mu\text{m}$ and base width is $W_B = 0.15 \mu\text{m}$. Find β .

- 9.18** An npn BJT is operating in the forward active region. Assuming the default SPICE values $I_S = 10^{-16} \text{ A}$, $\beta_F = 100$, find the terminal currents I_C , I_B , and I_E for $V_{BE} = 0.7 \text{ V}$.
- 9.19** Typical forward β 's are on the order of 100, while typical reverse β 's are on the order of 0.1. Find the corresponding α_F and α_R .
- 9.20** Figure P9.4 shows some experimental data for the current gain (normalized to that at 300 K) of a bipolar junction transistor as a function of temperature. Also included is a straight-line approximation to the linear region. From the data, estimate the value of ΔE_g^* .

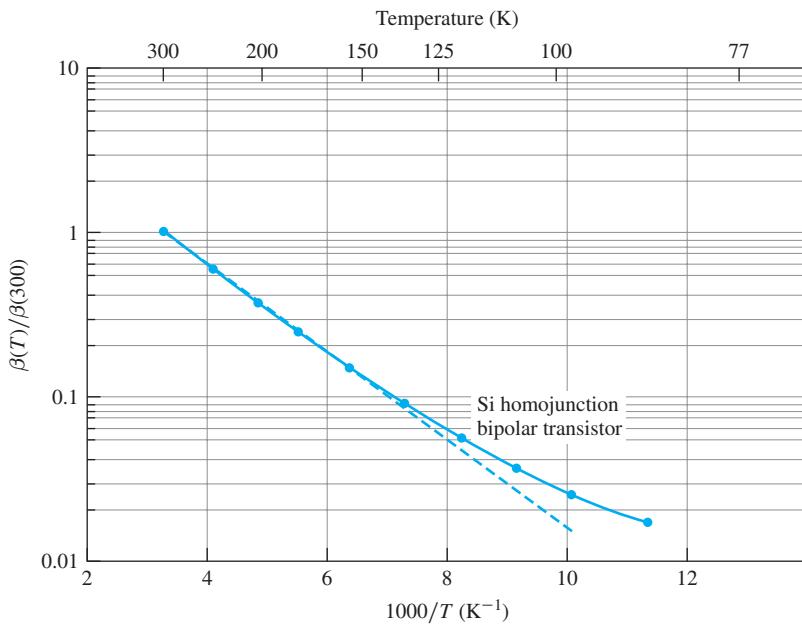


Figure P9.4

- 9.21** Equation (9.28) states that

$$\gamma = \frac{1}{1 + \frac{p_E(0^-) D_{pE}}{n_B(0^+) D_{nB}} \frac{W_B}{W_E}}$$

Discuss the effect of the heterojunction on the term $p_E(0^-)/n_B(0^+)$, even for equal doping. What effect does that have on α and β ?

- 9.22** Consider a nondegenerately doped HBT. Adapt Equation (9.42) to find an expression for the β of this transistor, as a function of the band gaps of the emitter and base.

10

CHAPTER

Time-Dependent Analysis of BJTs

10.1 INTRODUCTION

In Chapter 9 the dc characteristics of bipolar junction transistors were discussed to illustrate the physics of operation. However, in many systems high frequency operation in analog systems and fast switching speed in digital circuits are of great importance. In this chapter, the time-dependent characteristics of BJTs are examined. First, small-signal ac models are examined for use in analog circuits. Then switching transients are investigated for digital circuits. The advantages and disadvantages of BJTs compared with MOSFETs are examined and the use of both types of transistors on a chip (BiMOS) is discussed.

10.2 EBERS-MOLL AC MODEL

In Chapter 9, Section 9.8, the Ebers-Moll dc model of a BJT was presented. To model the ac behavior, the parasitic resistances and capacitances must be considered. Figure 10.1 shows the Ebers-Moll ac common emitter equivalent circuit for a BJT. It is the dc equivalent circuit of Figure 9.23 with the parasitic capacitances added. Since the capacitances are more important than the parasitic resistances in determining the time-varying behavior, for simplicity, the parasitic resistances are omitted (we will attend to them later, in the hybrid-pi model). In Figure 10.1, C_{BE} and C_{BC} represent the base-emitter and base-collector junction capacitances respectively, while C_{scBE} and C_{scBC} represent the stored-charge capacitances associated with the forward-biased base-emitter and base-collector junctions. The terminal currents from Section 9.8 are repeated here.

$$I_E = I_{CT} + \frac{I_F}{\beta_F + 1}$$

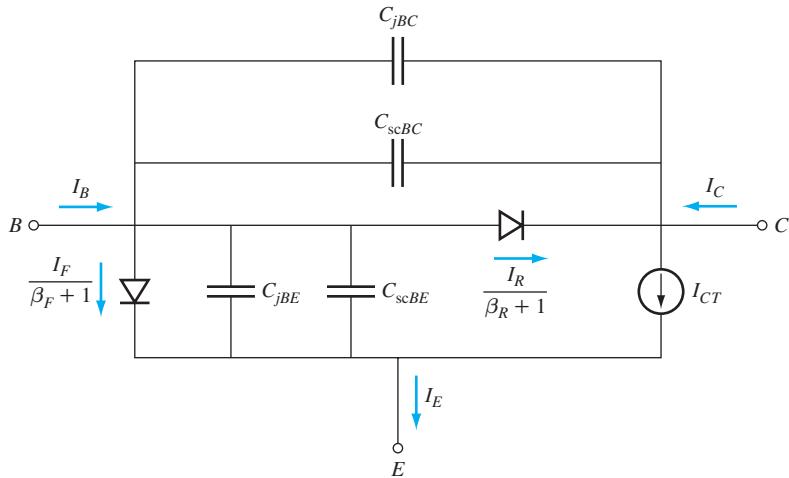


Figure 10.1 Ebers-Moll common emitter ac model for a BJT.

$$I_C = I_{CT} - \frac{I_R}{\beta_R + 1}$$

$$I_B = \frac{I_F}{\beta_F + 1} + \frac{I_R}{\beta_R + 1}$$

where

$$I_{CT} = \frac{\beta_F I_F}{\beta_F + 1} - \frac{\beta_R I_R}{\beta_R + 1}$$

The Ebers-Moll equivalent circuit operating in the active mode is shown in Figure 10.2. Since the base-collector junction is reverse biased, I_R and C_{scBC} are neglected.

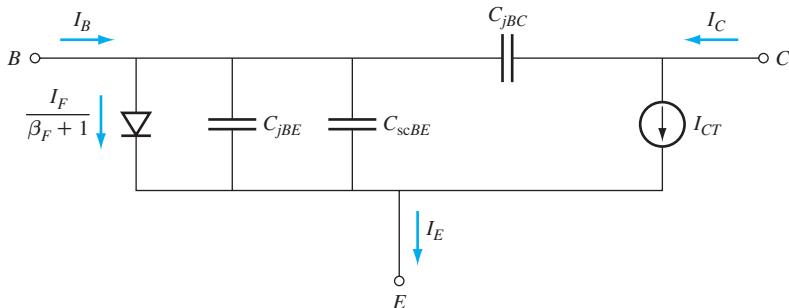
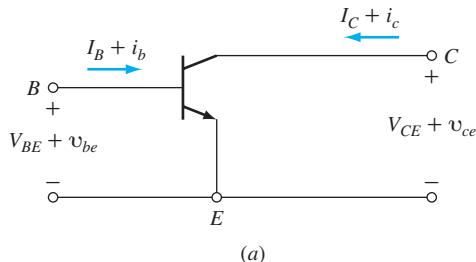


Figure 10.2 Ebers-Moll common emitter model for a BJT operating in the forward active mode. The collector-base junction is reverse biased, so C_{scBC} and I_R are neglected.

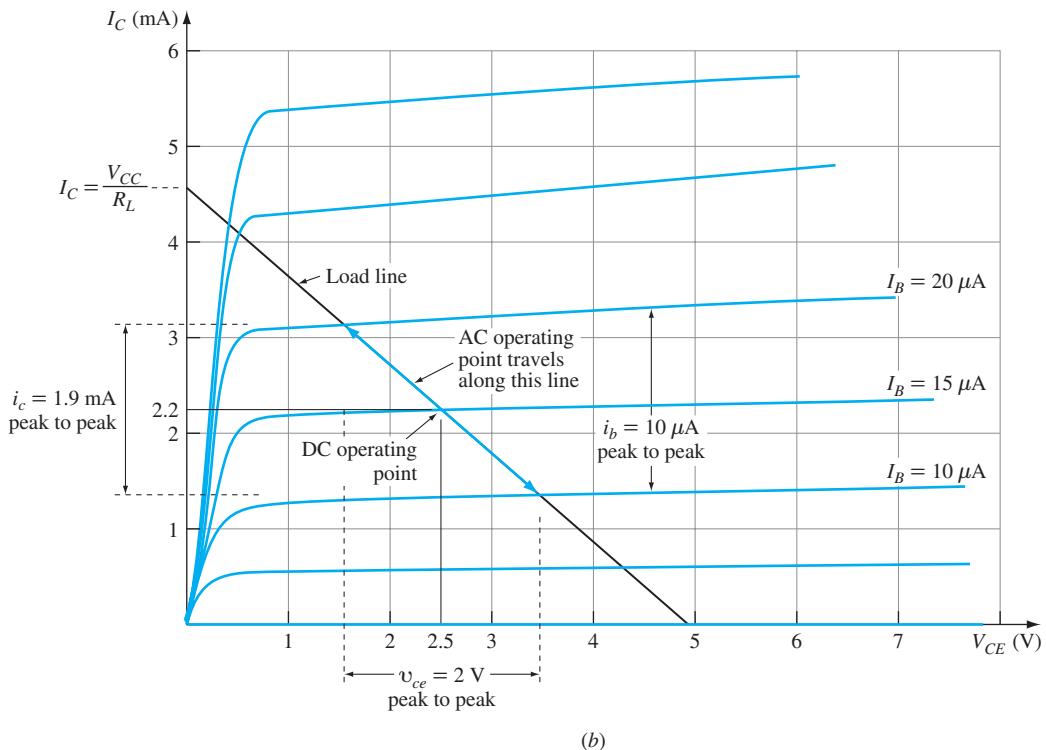
The Ebers-Moll model is useful for analyzing both switching and analog circuits, and in the next section the use of Ebers-Moll for small-signal analysis is discussed.

10.3 SMALL-SIGNAL EQUIVALENT CIRCUITS

Many analog circuits use BJTs operating in the active mode to amplify small-signal voltages and currents. We will consider an amplifier operating in the common emitter configuration as indicated in Figure 10.3. In (a) the input and output



(a)



(b)

Figure 10.3a–b (a) The common emitter circuit with the DC and small-signal, or ac, quantities indicated; (b) illustration of operation along the load line. For $I_B = 15 \mu\text{A}$, the dc operating point is at $I_C = 2.2 \text{ mA}$ and $V_{CE} = 2.5 \text{ V}$.

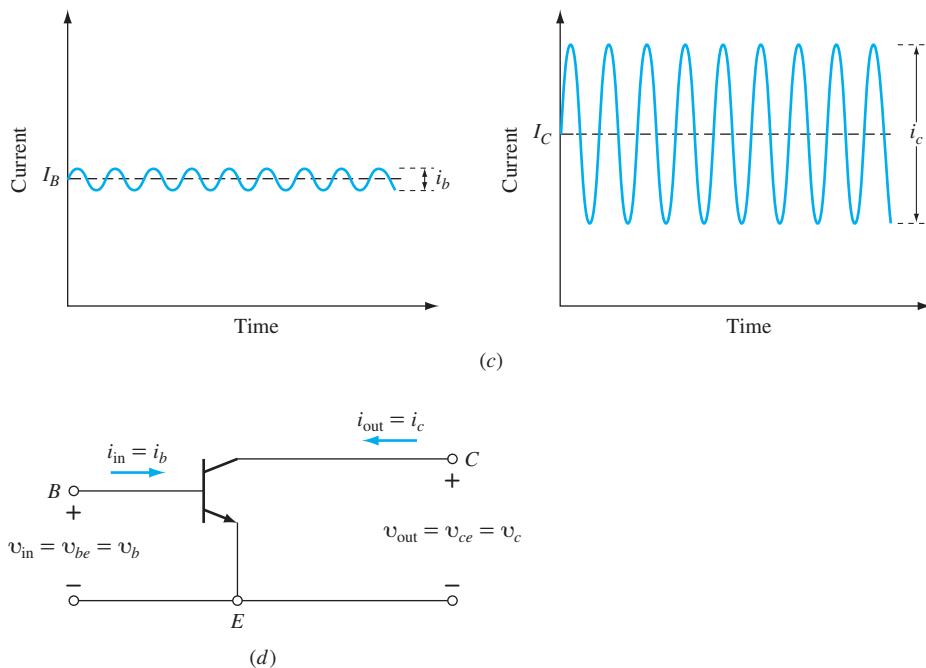


Figure 10.3c-d (c) The total input and output signals have a dc component (I_B and I_C respectively) and an ac component (i_b and i_c respectively); (d) the common emitter circuit, considering only the ac quantities.

voltages and currents are indicated. We use the convention that uppercase symbols such as I_B , I_C , V_{BE} , and V_{CE} represent dc quantities, and lowercase symbols i_b , i_c , v_{be} , and v_{ce} represent ac quantities. Each node will have a dc bias point and a dc current associated with that bias point, and the analog signal is taken to be a small variation around that bias point. For example, in Figure 10.3b, the dc bias point is at $I_B = 15 \mu\text{A}$, $V_{CE} = 2.5 \text{ V}$, and $I_C = 2.2 \text{ mA}$. The base current varies around the dc point by $10 \mu\text{A}$ peak-to-peak. For a load resistance R_L and a supply voltage V_{CC} , the dc collector current is

$$I_C = \frac{V_{CC} - V_{CE}}{R_L}$$

which is represented by the load line in (b). Since $I_C = \beta I_B$, the operating point corresponds to the intersection of the load line with the I_C - V_{CE} characteristics for a given value of I_B . The operating point moves along the load line indicated in the figure. If I_B increases by a small amount, I_C increases by β times that amount. The output voltage V_{CE} drops proportionally.

If the dc bias point is fixed and small ac changes are made, then small variations in i_b cause large variations in i_C , as shown in Figure 10.3c. The corresponding circuit for small-signal analysis is shown in Figure 10.3d. Since the common

emitter is used as reference, the ac input current, output current, input voltage, and output voltage are respectively i_b , i_c , v_b , and v_c .

There are a number of possible small-signal equivalent circuits obtainable from Figure 10.3d, but the most common is the so-called hybrid-pi equivalent circuit, so we will consider this next.

10.3.1 HYBRID-PI MODELS

To model the small-signal behavior, we consider the integrated circuit BJT of Figure 10.4 (the BJT of the previous chapter). The small-signal currents and voltages are indicated. Also shown are the series base resistance r_b , the collector resistance r_c , and the emitter resistance r_e . These arise from the finite conductivities of the semiconductor regions. These are the actual, physical resistances in the device, and include the contact resistances.

The equivalent circuit model for an ideal transistor in the common emitter configuration is shown in Figure 10.5a. This is known as the *hybrid-pi model*.¹ The resistances are shown, along with several parasitic capacitances. This model can be analyzed just like an actual circuit, and it will yield the same (small-signal) behavior as an ideal transistor. In an ideal transistor, the resistances would be zero, and the hybrid-pi model would reduce to the Ebers-Moll model of Figure 10.2.

In the hybrid-pi model, the capacitance C_μ between collector and base is the junction capacitance of the reverse-biased collector-base junction:

$$C_\mu = C_{jBC} \quad (10.1)$$

The capacitance between the base and the emitter is

$$C_\pi = C_{jBE} + C_{scBE} \quad (10.2)$$

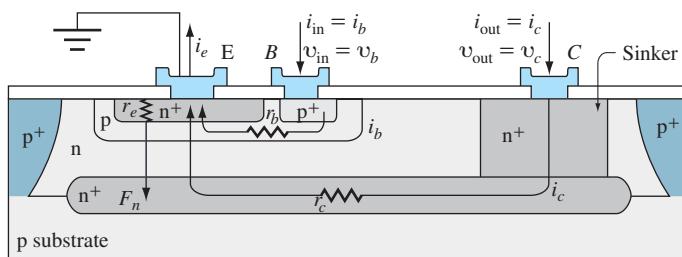


Figure 10.4 An integrated circuit bipolar junction transistor in the common emitter configuration, showing the small-signal currents.

¹It is called this because the arrangement of the resistors in the model could be imagined to form the Greek letter pi.

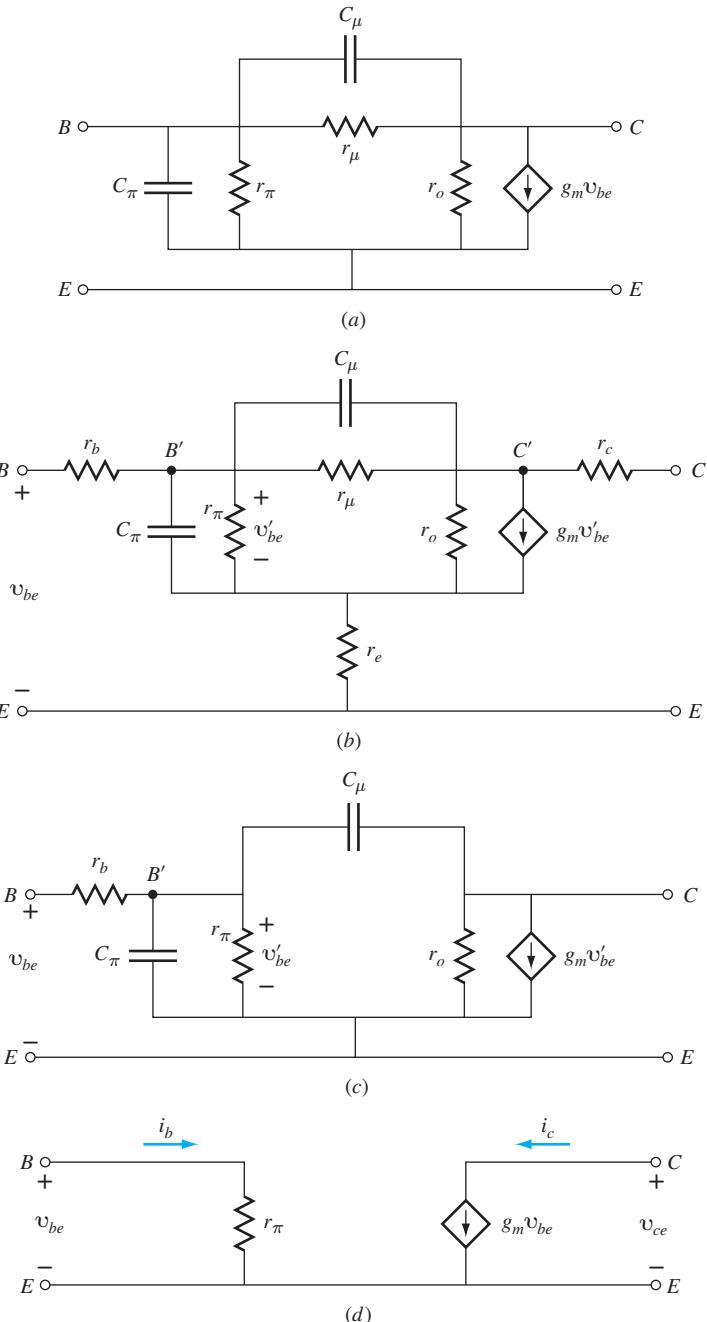


Figure 10.5 Hybrid-pi models: (a) Ideal BJT (all resistances can be neglected). (b) All resistances including contact resistances are included. In practical BJTs, r_e and r_c are often small enough to be neglected, and the feedthrough resistance r_μ is large (the resistance of the reverse-biased C-B junction). These approximations simplify the hybrid-pi model to (c). (d) Low-frequency model in which the capacitances are neglected.

This capacitance C_π is the sum of two capacitances, C_{jBE} , the junction capacitance of the forward-biased base-emitter junction, and C_{scBE} , the base-emitter stored-charge capacitance. Charge storage capacitance was discussed in connection with diodes and has the same meaning in BJTs.

The hybrid-pi model shows a differential input resistance r_π , which is a measure of how the base current changes for a differential change in base voltage. Since the I - V characteristics are not linear for a diode junction, this depends on the slope of the I - V curve at the operating point of interest:

$$r_\pi \approx \frac{1}{\left. \frac{\partial I_B}{\partial V_{BE}} \right|_{V_{CE}}} = \frac{v_{be}}{i_b} \quad (10.3)$$

We know that the relationship between I_B and V_{BE} is $I_B \approx I_{BO} (e^{qV_{BE}/kT} - 1)$ so

$$\left. \frac{\partial I_B}{\partial V_{BE}} \right|_{V_{CE}} = \frac{q}{kT} I_{BO} e^{qV_{BE}/kT} \approx \frac{qI_B}{kT} \quad (10.4)$$

Therefore combining Equations (10.3) and (10.4) gives

$$r_\pi \approx \frac{kT}{qI_B} \approx \frac{\beta_{DC} kT}{qI_C} \quad (10.5)$$

In this last step we also used the relationship $\beta_{DC} = I_C/I_B$, which represents the dc or low-frequency current gain.

We can express r_π in terms of the device parameters by expressing I_B as

$$I_B = I_{nB} + I_{pB} \approx I_{pB} \quad (10.6)$$

since the recombination current in the base, I_{nB} , is normally much smaller than the base-to-emitter injection current I_{pB} .

Since from Chapter 9, for a nondegenerate prototype transistor, $I_{pB} = I_{pE} = (qA_E D_{pB} p_E(0^-)/W_E)$ [Equation (9.27)]²

$$I_B = qA_E \left(\frac{D_{pE}}{W_E} p_E(0^-) \right) \quad (10.7)$$

Substituting $p_E(0^-) = (n_i^2/N'_{DE}) \exp(qV_{BE}/kT)$ yields

$$I_B = qA_E n_i^2 \left(\frac{D_{pE}}{W_E N'_{DE}} \right) e^{qV_{BE}/kT} \quad (10.8)$$

From Equation (10.5),

$$r_\pi = \frac{\frac{kT}{q^2 A_E n_i^2} e^{-qV_{BE}/kT}}{\left(\frac{D_{pE}}{W_E N'_{DE}} \right)} = \frac{kT W_E N'_{DE}}{q^2 A_E n_i^2 D_{pE}} e^{-qV_{BE}/kT} \quad (10.9)$$

²Here the currents are taken as being positive in the negative x direction.

The feedthrough resistance r_μ between the collector and the base is

$$r_\mu = \frac{1}{\left. \frac{\partial I_C}{\partial V_{CB}} \right|_{V_{BE}}} \quad (10.10)$$

This is the differential resistance of the reverse-biased collector-base junction (the reciprocal slope of the common base output characteristics). This is normally large enough that it can be ignored.

The output resistance r_o is

$$r_o \approx \frac{1}{\left. \frac{\partial I_C}{\partial V_{CE}} \right|_{V_{CB}}} = \frac{V_A}{I_C} \quad (10.11)$$

which is the reciprocal slope of the common emitter output characteristics. The quantity V_A is the Early voltage.

The hybrid-pi model's dependent current generator produces a current proportional to the base-emitter voltage v_{be} . The value of that current is $g_m v_{be}$, where g_m is the device transconductance,

$$g_m = \left. \frac{\partial I_C}{\partial V_{BE}} \right|_{V_{CE}} = \frac{i_c}{v_{be}} \quad (10.12)$$

The transconductance is a measure of how the input voltage controls the output current in a transistor.

In a transistor operated in the forward active region the collector current is essentially equal to the emitter current, and the emitter current depends on the base-emitter voltage, giving

$$I_C \approx I_E \approx I_{E0} e^{qV_{BE}/kT} \quad (10.13)$$

where I_{E0} is the base-emitter junction leakage current. Thus, from Equation (10.12),

$$g_m = \frac{q I_C}{kT} \quad (10.14)$$

But, from Equation (10.5)

$$\frac{1}{r_\pi} = \frac{q I_B}{kT}$$

Combining the last two equations gives $g_m / (1/r_\pi) = I_C/I_B = \beta_{DC}$, or

$$g_m = \frac{\beta_{DC}}{r_\pi} \quad (10.15)$$

Now we have expressions for the parameters of the hybrid-pi model, but so far we have discussed this model only for the ideal transistor. If the series

resistances are included, the equivalent circuit of Figure 10.5b results. For example, now some voltage is dropped across the input resistor r_b . Here, then, the dependent current generator has the value $g_m v'_{be}$, where v'_{be} is the voltage across r_π ; i.e.,

$$v'_{be} = v_{be} - i_b r_b - i_e r_e \quad (10.16)$$

Because the emitter is heavily doped and thin, r_e is small and can often be neglected.³

The value of the collector resistance r_c is reduced by the use of a degenerate n⁺ collector or buried layer and the n⁺ sinker that were indicated in Figure 10.4. In a well-designed BJT, r_c is small enough that it also can often be neglected. Further, the resistance r_μ between the base and collector is very large because that junction is reverse biased. Thus, this feedthrough resistance can be approximated as being infinite. With these approximations, the equivalent circuit of Figure 10.5c results.

Since the parasitic resistances (r_b , r_e , and r_c) reduce the voltage gain of the circuit, they should be minimized. Designers can decrease the base resistance by making the base region more heavily doped and increasing the base width. There are trade-offs here, however, since increased base doping increases C_{jBE} , while increasing the base thickness decreases β_{DC} and thus g_m .

For a well-designed transistor, then, r_e , r_b , and r_c are often small enough to be neglected. At low frequencies, such that the capacitances can be ignored, one arrives at the simplified equivalent circuit of Figure 10.5d. This circuit, because of its simplicity, is used extensively for first-order analysis.

For higher frequencies, the hybrid-pi model of Figure 10.5b or c is used. In the next section, we examine the effect of the capacitances on the high-speed behavior of BJTs. The model of Figure 10.5b, however, is valid only up to frequencies on the order of a few gigahertz. Above this, the accuracy decreases with increasing frequency because it does not take into account the carrier delay (transit time) from emitter to collector.

10.4 STORED-CHARGE CAPACITANCE IN BJTS

We discussed the stored-charge capacitance of a pn junction in Chapter 5. Here we re-examine it in terms of the BJT with the aid of Figure 10.6. This figure shows a transistor having uniform base and emitter dopings. In part (a) of the figure we show the energy band diagram for an npn BJT operating in the active mode for a given emitter-base voltage. Also shown are the minority carrier distributions in the base and the emitter. Excess minority carriers (electrons) are injected into the base from the emitter and extracted at the base-collector junction. Similarly excess holes are injected from base to emitter and extracted at the emitter contact.

³In modern BJTs designed for high-frequency operation, r_e is not always negligible. We discuss this further in Section 10.6.

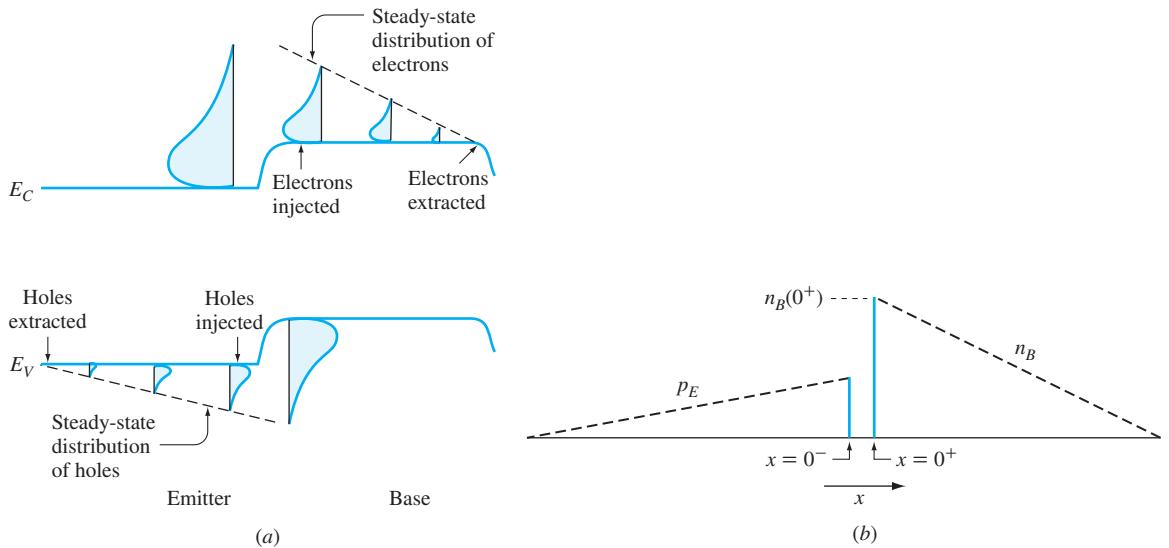


Figure 10.6 The injected charge on each side of the base-emitter junction acts as a capacitance. (a) The energy band diagram; (b) the charge distributions (not to scale).

The distribution of charge in the base, if uniformly doped, is very close to a straight line, as shown in Figure 10.6b. The total stored charge in the base is

$$Q_B = -\frac{qA_E n_B(0^+) W_B}{2} \quad (10.17)$$

where $n_B(0^+)$ is the electron density at the emitter edge of the base, and A_E is the area of the emitter. But for a uniformly doped base the electron current results from diffusion, so, we have

$$I_n = -qA_E D_n \frac{dn(x)}{dx}$$

where the minus sign arises because current is taken as positive in the negative x direction in our present coordinate system.

The slope of the electron concentration is $-n_B(0^+)/W_B$, and

$$I_{nB} = \frac{qA_E n_B(0^+) D_n}{W_B} \quad (10.18)$$

Thus, combining Equations (10.17) and (10.18) gives

$$Q_B = -\frac{W_B^2 I_{nB}}{2 D_n} \quad (10.19)$$

Similarly, the charge stored in the emitter is

$$Q_E = -\frac{W_E^2 I_{pE}}{2 D_p} \quad (10.20)$$

In a well-designed BJT, however, the hole current in the emitter is much smaller than the electron current in the base ($I_{pE} \ll I_{nB}$), and so we can often neglect the effect of stored carrier charge in the emitter. This especially true in heterojunction bipolar transistors.

Figure 10.7a shows the distribution of carriers in a uniformly doped base at some time $t = 0$. When the emitter-base voltage is decreased by an amount dV_{BE} ,

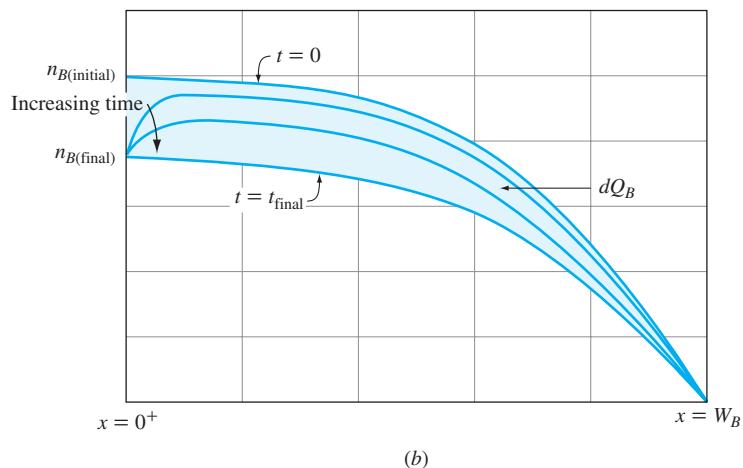
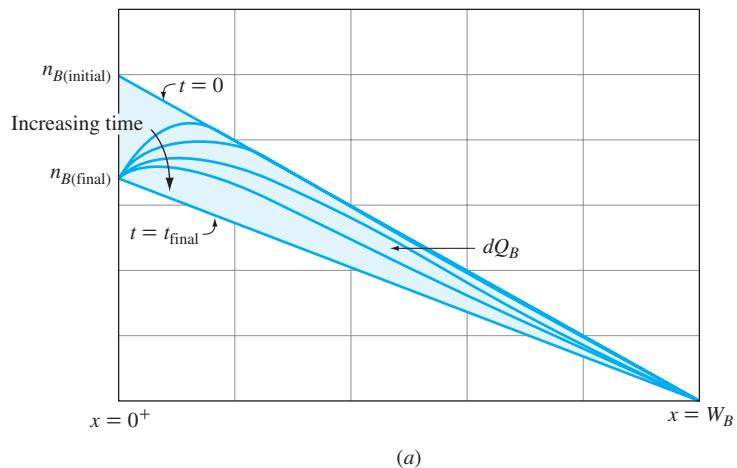


Figure 10.7 The change in the charge distribution when the injection is changed in a uniformly doped junction. It takes time to get rid of the excess charge, which is equivalent to discharging a capacitor.
(a) Uniformly doped base; (b) graded base.

it reduces the number of carriers being injected across the $x = 0$ plane (in other words, $n_B(0^+)$ changes). The steady-state stored base charge is decreased by an amount equal to the shaded area:

$$dQ_B = \frac{W_B^2}{2D_n} \frac{dI_{nB}}{dV_{BE}} dV_{BE} \quad (10.21)$$

The evolution of the charge distribution in a uniform base is also shown in Figure 10.7a. Of this charge dQ_B that must be gotten rid of, some diffuses back to the emitter, and a portion diffuses to the collector. Only the recoverable fraction δ of dQ_B that is recovered by the emitter flows in the external circuit and contributes to the stored-charge capacitance. We define the recoverable charge dQ_{Br} by

$$dQ_{Br} = \delta dQ_B \quad (10.22)$$

For the case of the uniformly doped base, δ is about $\frac{2}{3}$, or two-thirds of dQ_B is recovered by the emitter.

For the case of the graded-base transistor of Figure 10.7b, the field in the base accelerates the base electrons to the collector and reduces the fraction that returns to the emitter. That reduces the reclaimable charge, and so correspondingly reduces the stored-charge capacitance.⁴ For a typical base gradient, the recovered fraction of dQ_B is on the order of $\delta = 0.2$ to 0.3 .

Determination of this reclaimable charge is beyond the scope of this book, so we simply state without proof that the stored charge in the base, Q_B , is given by [1]

$$Q_B = \frac{I_C W_B^2}{\langle D_{nB} \rangle} \left[\frac{\eta - 1 - e^{-\eta}}{\eta^2} \right] \quad (10.23)$$

where η is the doping parameter, $\langle D_{nB} \rangle$ is the average diffusion coefficient in the base, and the charge that is reclaimable, Q_{Br} , is

$$Q_{Br} = \frac{I_C W_B^2}{\langle D_{nB} \rangle} \left[\frac{1 - e^{-\eta}}{\eta^2} \right] \left[\frac{\sinh \eta - \eta}{\cosh \eta - 1} \right] \quad (10.24)$$

Then the fraction δ of reclaimable charge for a change in voltage dV_{BE} is

$$\delta = \frac{dQ_{Br}}{dQ_B} = \frac{(1 - e^{-\eta}) \left(\frac{\sinh \eta - \eta}{\cosh \eta - 1} \right)}{\eta - 1 + e^{-\eta}} \quad (10.25)$$

⁴For uniform base doping and uniform composition, diffusion is responsible for the recovered base charge, and the resultant capacitance is referred to as *diffusion capacitance*. Since in most BJTs the recovered charge is influenced by drift as well as diffusion, we refer to the resultant capacitance as *stored-charge capacitance*.

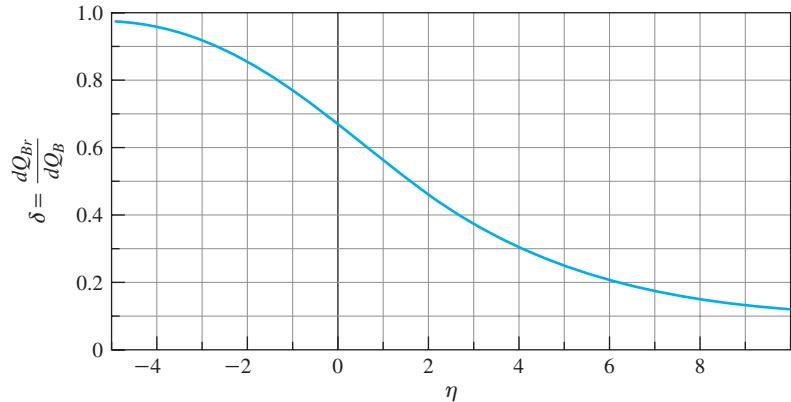


Figure 10.8 Fraction of reclaimable charge in the base as a function of grading parameter η in the base. As the fraction of reclaimable charge decreases, the stored-charge capacitance also decreases, and the device response time decreases (the device operates faster).

which is plotted in Figure 10.8. We can see that the fraction of the reclaimable charge reduces with increasing η , from $\frac{2}{3}$ for $\eta = 0$ (uniformly doped, uniform composition case), to 0.13 for $\eta = 10$. Note that this corresponds to a decrease in stored-charge capacitance, so that the graded-base and/or graded composition transistor will have a faster operating speed than a uniformly doped one. Negative values of η indicate a field that accelerates injected charge back toward the emitter. Thus for $\eta = -3$, about 90 percent of the injected charge is reclaimable, which would increase the capacitance considerably and slow the device response.

We define the stored-charge capacitance as

$$C_{scBE} \equiv \delta \left| \frac{dQ_B}{\partial V_{BE}} \right| \quad (10.26)$$

Then from Equation (10.21),

$$C_{scBE} = \frac{\delta W_B^2}{2\langle D_{nB} \rangle} \frac{\partial I_{nB}}{\partial V_{BE}} \quad (10.27)$$

Since the collector current consists primarily of the electrons coming from the base, we have

$$\frac{\partial I_{nB}}{\partial V_{BE}} \approx \frac{\partial I_C}{\partial V_{BE}} \approx \beta_{DC} \frac{\partial I_B}{\partial V_{BE}} = \frac{\beta_{DC}}{r_\pi} \quad (10.28)$$

where we have used $I_C = \beta_{DC} I_B$ and Equation (10.3). Combining Equations (10.27) and (10.28) gives

$$C_{scBE} = \frac{\delta W_B^2 \beta_{DC}}{2\langle D_{nB} \rangle r_\pi} = \frac{\delta W_B^2 q I_C}{2\langle D_{nB} \rangle kT} \quad (10.29)$$

For uniform base doping, $\delta = \frac{2}{3}$ and

$$C_{scBE} = \frac{W_B^2 \beta_{DC}}{3D_{nB} r_\pi} \quad \text{uniformly doped base} \quad (10.30)$$

For the graded base considered above, $\delta \approx 0.2$ and

$$C_{scBE} \approx \frac{W_B^2 \beta_{DC}}{10\langle D_{nB} \rangle r_\pi} \quad \text{graded-base with } \eta = 6 \quad (10.31)$$

We can see that the stored-charge capacitance decreases with decreasing base width, and it also decreases through the recoverable fraction δ , which decreases with increasing field in the base.

10.5 FREQUENCY RESPONSE

At high frequencies, the capacitance reduces the current gain of a BJT. In this section, we investigate the frequency response of the short-circuit current gain, using a BJT in the common emitter configuration. Figure 10.9 shows the hybrid-pi circuit of Figure 10.5c with r_b and r_c neglected and the output short-circuited. The output current is

$$i_c = (g_m - j\omega C_\mu) v_{BE} \quad (10.32)$$

while the base current is

$$i_b = \left(\frac{1}{r_\pi} + j\omega C_\pi + j\omega C_\mu \right) v_{BE} \quad (10.33)$$

The short-circuit current gain is then

$$\beta(\omega) \frac{i_c}{i_b} = \frac{g_m - j\omega C_\mu}{\frac{1}{r_\pi} + j\omega(C_\pi + C_\mu)} = \frac{r_\pi(g_m - j\omega C_\mu)}{1 + j\omega r_\pi(C_\pi + C_\mu)} \quad (10.34)$$

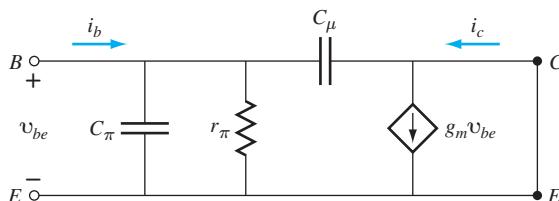


Figure 10.9 The hybrid-pi model for high-frequency short-circuit current gain.

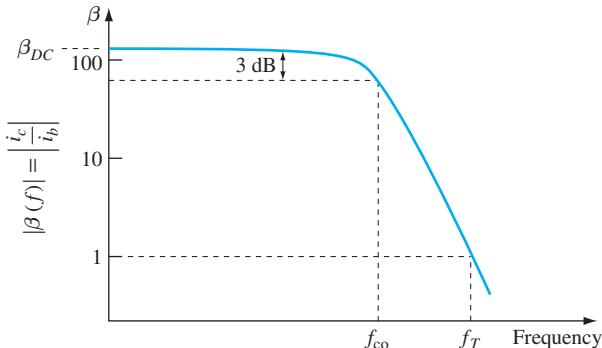


Figure 10.10 The frequency response of a BJT, showing the cutoff frequency f_{co} and the unity gain frequency f_T .

At normal operating frequencies and currents, $g_m \gg \omega C_\mu$ and so

$$\beta(\omega) \approx \frac{r_\pi g_m}{1 + j\omega r_\pi(C_\pi + C_\mu)} = \frac{\beta_{DC}}{1 + j\omega r_\pi(C_\pi + C_\mu)} \quad (10.35)$$

The magnitude of the current gain as a function of frequency is

$$\beta(f) = \frac{\beta_{DC}}{\sqrt{1 + \left(\frac{f}{f_{co}}\right)^2}} \quad (10.36)$$

where the cutoff frequency for the frequency response of β , f_{co} is

$$f_{co} = \frac{1}{\sqrt{2\pi r_\pi(C_\pi + C_\mu)}} \quad (10.37)$$

This function is plotted in Figure 10.10.

10.5.1 UNITY CURRENT GAIN FREQUENCY f_T

Figure 10.10 indicates the frequency f_T at which current gain is unity. This figure of merit is known as the *unity current gain frequency* or *unity gain frequency*. From Equation 10.36, $\beta(f) = \beta(f_T) = 1$, or $\beta_{DC} = \sqrt{1 + (f_T/f_{co})^2}$. However, for $\beta = 1$, $f \gg f_{co}$ and

$$f_T = \beta_{DC} f_{co} \quad (10.38)$$

Note that f_T is equal to the current gain-bandwidth product ($\beta_{DC} f_{co}$).

EXAMPLE 10.1

Estimate f_T for a prototype BJT with a base doping of 10^{17} cm^{-3} and a base width of $W_B = 0.1 \mu\text{m}$.

Solution

To obtain a rough estimate of f_T , we use the equivalent circuit of Figure 10.9, but neglect C_μ since it is usually small. The base current i_b results from the voltage v_{be} across r_π and C_π in parallel:

$$i_b = v_{be} \left(\frac{1}{r_\pi} + j\omega C_\pi \right)$$

The collector current is $i_c = g_m v_{be}$. Then the current gain $\beta(f)$ is

$$\beta(f) = \frac{i_c}{i_b} = \frac{g_m v_{be}}{v_{be} \left(\frac{1}{r_\pi} + j\omega C_\pi \right)} = \frac{g_m}{\left(\frac{1}{r_\pi} + j\omega C_\pi \right)}$$

which can be expressed

$$\beta(f) = \frac{g_m r_\pi}{1 + j\omega r_\pi C_\pi} = \frac{\beta_{DC}}{1 + j\omega r_\pi C_\pi}$$

This reduces to β_{DC} for $\omega = 0$, as expected.

For $|\beta(f)| = 1$, $\omega r_\pi C_\pi \gg 1$ and

$$|\beta(f)| = 1 = \frac{\beta_{DC}}{2\pi f_T r_\pi C_\pi} = \frac{\beta_{DC}}{2\pi f_T r_\pi (C_{jBE} + C_{scBE})}$$

Except at low currents, the storage capacitance is appreciably larger than the junction capacitance and thus $C_\pi \approx C_{scBE}$. Solving for f_T and using Equation (10.29) gives

$$f_T \approx \frac{\beta_{DC}}{\frac{2\pi\delta r_\pi W_B^2 \beta_{DC}}{2D_n r_\pi}} = \frac{2D_n}{2\pi\delta W_B^2}$$

For a prototype transistor, $\delta = \frac{2}{3}$, and for a base doping of 10^{17} cm^{-3} , $D_n \approx 20 \text{ cm}^2/\text{s}$, from Figure 3.11. For a base width of $0.1 \mu\text{m} = 10^{-5} \text{ cm}$, we have

$$f_T = \frac{2 \times 20 \text{ cm}^2/\text{s}}{2\pi \times \frac{2}{3} \times (10^{-5} \text{ cm})^2} = 9.5 \times 10^{10} \text{ Hz} = 95 \text{ GHz}$$

This result is approximate, however, since the presence of parasitic resistances will lower this value.

In Example 10.1, we estimated the value of f_T as $f_T = 1/2\pi r_\pi C_{scBE}$. Another approach is to estimate f_T by the time delay for carriers to cross the base from emitter to collector, discussed next.

10.5.2 BASE TRANSIT TIME t_T

For an npn transistor, the electron current in the base is

$$I_{nB} = qA_E \Delta n(x) v_n(x) \quad (10.39)$$

where $v(x)$ is the average velocity at position x in the base. Neglecting the small recombination current, the overall base current I_B is independent of x , and the $\Delta n(x)v_n(x)$ product is constant. The time an electron requires to traverse a distance dx is

$$dt = \frac{1}{v_n(x)} dx \quad (10.40)$$

The time required to cross the base, then, is

$$t_T = \int_0^{t_{IB}} dt = \int_0^{W_B} \frac{1}{v_n(x)} dx \quad (10.41)$$

The velocity can be obtained from Equation (10.39) and

$$t_T = \int_0^{W_B} \frac{qA_E \Delta n}{I_{nB}} dx = \frac{1}{I_{nB}} \int_0^{W_B} qA_E \Delta n dx \quad (10.42)$$

The integral on the right-hand side (not containing I_{nB}) is the total stored minority carrier charge in the base, Q_B . The base transit time is then

$$t_T = \frac{Q_B}{I_{nB}} = \frac{Q_B}{I_C} \quad (10.43)$$

Equation (10.43) is valid for an arbitrary base doping profile. For uniform doping in the base, the electron distribution is linear, decreasing from emitter to base:

$$\Delta n(x) = \Delta n(0^+) \left(1 - \frac{x}{W_B}\right) \quad (10.44)$$

and the stored charge is

$$Q_B = \frac{qA_E \Delta n(0^+) W_B}{2} \quad (10.45)$$

Since the collector current is almost entirely due to electrons arriving from the base,

$$I_C = I_{nB} = qA_E D_n \frac{d\Delta n}{dx} \quad (10.46)$$

Thus, with the aid of Equation (10.44),

$$I_C = \frac{qA_E D_n \Delta n(0^+)}{W_B} \quad (10.47)$$

and from Equation (10.43),

$$t_T = \frac{W_B^2}{2D_n} \quad (10.48)$$

For this approximation,

$$f_T = \frac{1}{2\pi t_T} \quad (10.49)$$

EXAMPLE 10.2

Find the base transit time for electrons in the npn prototype transistor of Example 10.1 [base doping level of 10^{17} cm^{-3} and base width of $W_B = 0.1 \mu\text{m}$ (10^{-5} cm)].

Solution

From Equation (10.48),

$$t_T = \frac{W_B^2}{2D_n} = \frac{(10^{-5})^2}{2 \times 20} = 2.5 \times 10^{-12} \text{ s} = 2.5 \text{ ps}$$

where $D_n \approx 20 \text{ cm}^2/\text{s}$. This result is more than five orders of magnitude smaller than the electron lifetime in the base, indicating that on the order of 10^{-5} of the base electrons recombine. That is, only one electron in about 100,000 is lost to recombination in the base, and $\alpha_T \approx 1$.

In this approximation,

$$f_T = \frac{1}{2\pi t_T} = \frac{1}{2\pi \times 2.5 \times 10^{-12} \text{ s}} = 63.6 \text{ GHz}$$

It is reduced from that of Example 10.1 by the factor $1/\delta$.

In a graded-base or graded-composition transistor, the preceding value of t_T is reduced and thus f_T is increased because the field in the base accelerates the electrons toward the collector (i.e., the transport is by drift as well as diffusion.) For a base field greater than about 5 kV/cm, the electrons traverse the base at the saturation velocity. In this case,

$$t_T \approx \frac{W_B}{v_{\text{sat}}} = \frac{0.1 \mu\text{m}}{10^7 \text{ cm/s}} = 1 \text{ ps}$$

for Example 10.2.

10.5.3 BASE-COLLECTOR TRANSIT TIME t_{BC}

In addition to the delay due to the transit time across the base, there is a delay due to the time t_{BC} required for a carrier to traverse the depletion region between base and collector. Because in this region the field is large, the carrier velocity over most of the depletion region is equal to its saturation velocity:

$$t_{BC} = \frac{w_{BC}}{v_{\text{sat}}} \quad (10.50)$$

where w_{BC} is the width of the base-collector depletion region and depends on the doping levels and collector-base voltage.

10.5.4 MAXIMUM OSCILLATION FREQUENCY f_{MAX}

While the unity current gain frequency f_T is a convenient figure of merit for BJTs, particularly at low current levels, it does not consider the effects of parasitic resistances. Another common figure of merit is the *maximum oscillation frequency* f_{max} , the frequency at which the power gain of the device is unity when the base resistance is considered. This can be expressed [2]

$$f_{\text{max}} = \left(\frac{f_T}{8\pi r_b C_{jBC}} \right)^{1/2} \quad (10.51)$$

where r_b is the base resistance and C_{jBC} is the base-collector junction capacitance. Note that, depending on the values of r_b and C_{jBC} , f_{max} may be either larger or smaller than f_T . In modern high-performance transistors, usually $f_{\text{max}} > f_T$.

10.6 HIGH-FREQUENCY TRANSISTORS

In Chapter 9, the dc characteristics of conventional transistors were discussed, and so the parasitic capacitances were not important. For high-frequency operation, however, it is necessary to use structures that minimize the parasitic resistances and capacitances. In the next section, we discuss one such structure, the double poly Si self-aligned structure with a polysilicon emitter. Following that we will look at the effects of the base transit time on the high-speed operation of BJTs. (Note that the Si:Si_xGe_{1-x}:Si HBT discussed in Chapter 9, Section 9.6.2 is, in effect, a double heterojunction BJT. Values of f_T and f_{max} in excess of 300 GHz have been reported.)

10.6.1 DOUBLE POLY SI SELF-ALIGNED TRANSISTOR

The cross section of the double poly Si self-aligned transistor⁵ is shown in Figure 10.11. [3] In this high-speed design, there are two base regions, or more accurately the base region is divided so that half of it appears on either side of the emitter.⁶ To form the “extrinsic” base, p⁺ polysilicon is deposited in the region indicated, and at elevated temperatures, acceptors diffuse to form a low-resistance p⁺ base. A more lightly doped active base is formed earlier by ion implantation. To form the emitter, an n⁺ polysilicon layer is deposited, and, under heat treatment, donors diffuse to form a shallow (on the order of 30 nm) n⁺ emitter in the single crystalline Si. The emitter of this device consists of an n⁺ single-crystal region and an n⁺ polysilicon region in series.

The advantages of this structure over that of a conventional BJT are

1. The base-collector junction area can be made smaller, thus reducing the associated junction capacitance.

⁵The term *self-aligned* means that, as each layer is grown, the structures previously created on the layer below are used as the mask, instead of using multiple photomasks, each of which would have to be independently aligned.

⁶In Figure 10.11 it appears that there is only one base contact. However, both base regions are active since they are connected internally.

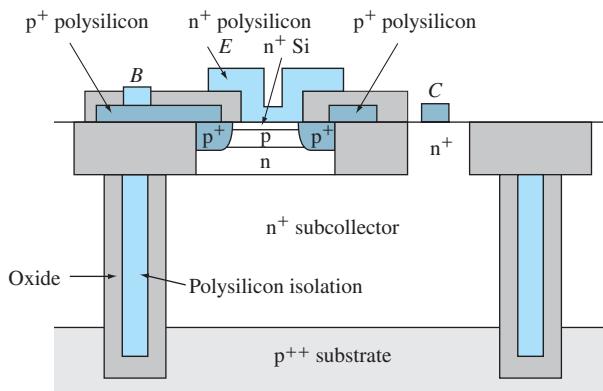


Figure 10.11 Schematic cross section of a double poly self-aligned BJT. The emitter consists of a shallow n⁺ single-crystal region and an n⁺ polysilicon region. The extrinsic base consists of a p⁺ polysilicon region and a p⁺ single-crystal region obtained from diffusion from the polysilicon. (Source: Adapted from T. Ning, “History and Future Perspectives of the Modern Silicon Bipolar Transistor,” *IEEE Transactions on Electron Devices*, 48, pp. 2485–2491, 2001.)

2. The extrinsic base regions can be made short and heavily doped, reducing the base resistance.
3. The intrinsic base resistance is reduced by a factor of 4 by the use of a double base.
4. The active emitter area can be made small, reducing its junction capacitance.
5. The polysilicon emitter can extend over an appreciably larger area than the active emitter, reducing the emitter resistance.
6. As a result of the self-alignment process, the device area can be reduced compared with conventional devices, where photolithographic alignment tolerances require larger separation between base and emitter contacts.
7. The presence of the polycrystalline portion of the emitter increases the current gain, as discussed below.

These transistors are designed for high speed, but they also exhibit higher-than-expected β 's. The increase in β that comes from the use of the polysilicon emitter structure is not well understood, and several models to explain this effect exist. [4] One of these models is based on the effects of band-gap narrowing being less in polysilicon than in monocrystalline emitters. This is explained with the aid of Figure 10.12a. The base is doped p type, and with $N'_{AB} \approx 10^{18} \text{ cm}^{-3}$, there is some band-gap narrowing in the base, raising E_V . The n⁺ monocrystalline

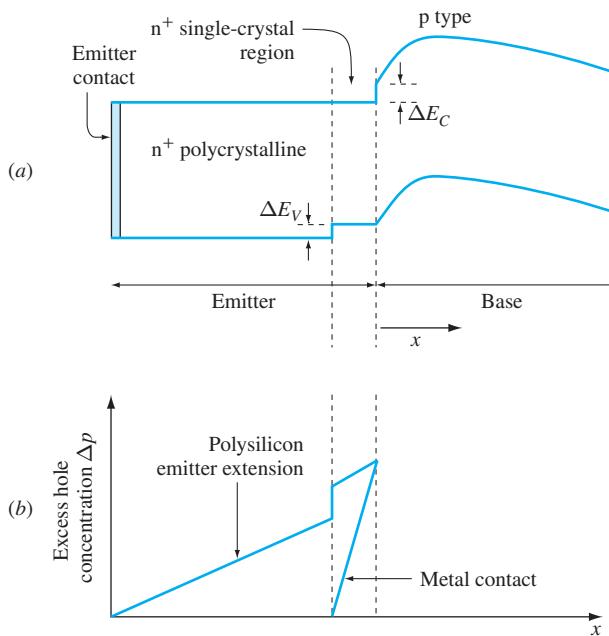


Figure 10.12 The use of a polysilicon n^+ emitter extension layer. (a) The energy band diagram. The discontinuity in the valence band acts as a barrier to holes. (b) The excess hole concentration of the polysilicon emitter compared with that of a metal contact directly on the monocrystalline emitter. The smaller gradient of Δp results in a reduced hole current and an increased β .

emitter is diffused into the p base region and thus, in addition to the increase in E_V already present, there is an additional decrease in E_C due to donors in the emitter. The n^+ polysilicon, however, is grown on top of these structures, and so contains a negligible concentration of acceptors. Thus in the n^+ polysilicon, there is only the band-gap narrowing due to the donors and E_V is not affected. As a result, the single-crystal emitter will have its valence band at a higher energy than in the polysilicon emitter. This discontinuity in E_V acts as a barrier to hole flow, reducing the base-emitter hole current and increasing the injection efficiency and thus β .

A comparison of the hole concentration profile in a polysilicon emitter device with that of a metal contact to the single-crystal emitter is shown schematically in Figure 10.12b. The gradient in the hole concentration for the polysilicon emitter extension layer case is smaller than for a metal contact applied directly to the single-crystalline emitter layer, thus reducing the hole diffusion current. There is also a discontinuity in Δp . Those holes ‘lost’ in the discontinuity do not make it to the emitter contact and do not contribute to base-emitter hole current. This reduction

of I_{pE} increases the injection efficiency γ , which in turn increases the current gain β . Another explanation for the increased β of this emitter structure is that a thin SiO₂ layer naturally forms at the polysilicon-silicon interface which acts as a barrier to reduce hole injection from base to emitter, thus increasing γ (and β).

10.7 BJT SWITCHING TRANSISTOR

In Section 10.4, it was shown that the injected minority charge in a forward-biased junction leads to stored-charge capacitance. Here, we will examine the effect of that capacitance on the switching time.

As mentioned in the discussion on field-effect transistors, the time to switch a circuit between **off** and **on** depends on the time required to charge and discharge the circuit capacitances. This is true for bipolar junction transistors also. In the BJT case, however, both minority carriers and majority carriers are involved. In addition to the majority carriers in the collector and emitter flowing into the circuit to discharge the circuit capacitances, the time associated with injecting and removing minority carrier stored charge further increases the turn-on and turn-off times.

We illustrate the effects of stored minority carriers with the circuit of Figure 10.13, in which R_C is the load resistor and C_L is the load capacitance. The base current is supplied via an input voltage V_{in} , through a base resistor R_B . In the circuit shown, the input voltage is switched from V_R to V_F . For V_R negative, the base-to-emitter junction is reverse biased and the BJT is **off**. The output voltage $V_{out} = V_{CC}$, where V_{CC} is the supply voltage. For $V_{in} = V_F$, the base-emitter is forward biased and the transistor is conducting.

For operation as a switch, V_F , R_B , and R_C are chosen such that for $V_{in} = V_F$, the BJT is operating in saturation, or the transistor is **on**, and when $V_{in} = V_R$, the transistor is in cutoff. These two states are illustrated in Figure 10.14, which illustrates the I_C - V_{CE} characteristics of a BJT with its associated resistive (R_C) load line. The collector current I_C is controlled by the base current I_B . For V_{in} negative

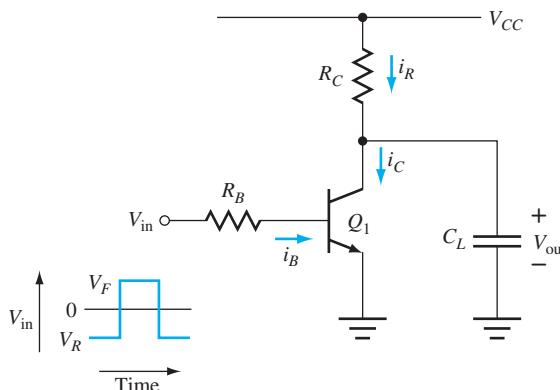


Figure 10.13 A simple bipolar transistor inverter circuit.

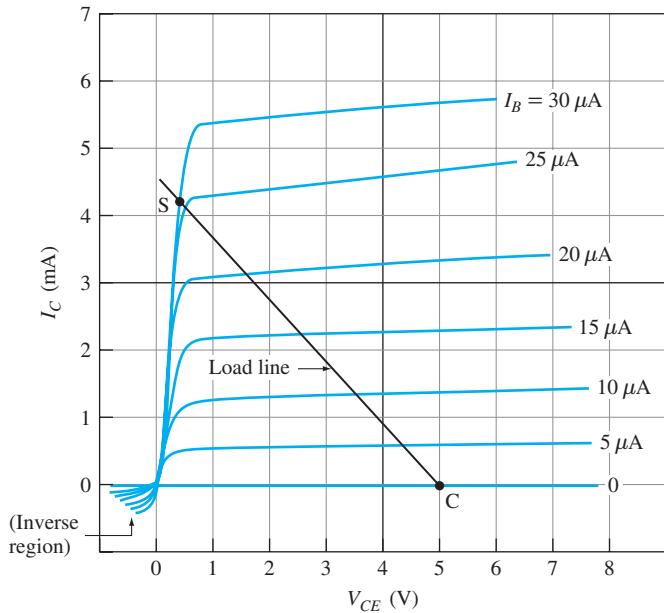


Figure 10.14 The I_C - V_{CE} characteristics of the transistor in Figure 10.13, along with the circuit load line. When the input is high, the E - B junction of transistor Q1 is forward biased, and the transistor is in saturation (point S). Current flows through the transistor and thus through R_C such that voltage is dropped across the resistor and the output goes low. When the input is low, the transistor is in cutoff (point C); the E - B junction is reverse biased. No current flows through Q1, so no voltage is dropped across R_C and the voltage output is high.

(or zero), the E - B junction is not forward biased, and $I_E \approx 0$. Therefore $I_C = I_E - I_B \approx 0$. In this case there is no voltage drop across R_C , and thus when V_{in} is negative, $V_{CE} = V_{CC}$ as indicated by point C (cutoff). For most of the region along the load line, $I_B = (V_{in} - V_{BE})/R_B \approx V_{in}/R_B$. For $V_{in} = V_F$, the E - B junction is forward biased and current flows. The voltage V_F is chosen such that a large I_C flows and most of V_{CC} is dropped across R_C , with V_{CE} remaining across the transistor (and thus the output). Thus, I_C saturates at V_{CE} , indicated by point S (saturation).

10.7.1 OUTPUT LOW-TO-HIGH TRANSITION TIME

Because it takes some time for the minority carriers injected into the base to reach the collector, the output high-to-low transition time t_{hl} is increased over the time required just to discharge the junction capacitance. More important is the effect on the output low-to-high transition time t_{lh} , or the time required to go from S to C in Figure 10.14. In saturation, both the emitter-base and collector-base

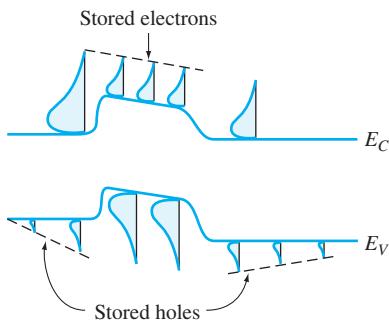


Figure 10.15 The charge “stored” in the excess carrier distributions when the transistor is in saturation. Carriers are constantly injected across both junctions, resulting in excess holes in the emitter and collector and excess electrons in the base. When the transistor goes out of saturation, it takes time for the excess carrier concentration to dissipate, and the response of the transistor is therefore slowed.

junctions are forward biased. Thus, there is injection at both junctions, and there are minority carriers stored in the emitter, base, and collector. This can be seen from the energy band diagram of Figure 10.15. When the input V_{in} is switched from V_F to V_R , the new input voltage attempts to reverse-bias the base-emitter junction. However, most of the excess electrons in the base, which were already injected from the emitter before the input was switched, continue on to the collector. This tends to keep the collector current constant for a finite time. The rest of the excess base electrons flow out the base contact. Some of the excess holes in the emitter and collector diffuse to the base and also contribute to base current. This base current is $I_B \approx V_R/R_B$ and is constant until the stored charge is small enough that the base current finally decreases. Only then can the transistor begin to change state from S (low) to C (high). This delay caused by the removal of stored minority carriers is a serious problem associated with the output low-to-high transition time of BJTs. As might be expected, this delay is dependent on the thickness of the emitter and base, and on the built-in fields of the emitter, base, and collector.

Note that t_{lh} depends on the value of the off voltage V_R as indicated in Figure 10.16. With V_R more negative, I_B is larger, and the carriers can be dissipated more quickly and the delay time is reduced.

The value of transition time is not easily calculated even if the device geometry and doping profiles are known. This is because of the dependence of the

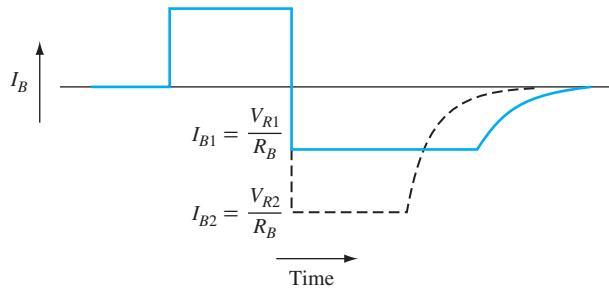


Figure 10.16 The turn-off time decreases as the voltage for the off state (V_R) increases.

junction voltage with stored charge.⁷ For the simple circuit shown (Figure 10.13), the transition time can easily approach $1\ \mu\text{s}$, a value much too long for high-performance logic circuits. Furthermore, the use of V_{in} of two polarities is inconvenient—but using $V_R = 0\ \text{V}$ results in an even larger transition time.

While the circuit of Figure 10.13 is useful to illustrate the effects of minority carrier storage, its switching time is too great to be of practical use in high-speed circuits. Next we look at two alternatives.

10.7.2 SCHOTTKY-CLAMPED TRANSISTOR

The speed of the output low-to-high transition can be increased appreciably by the use of a Schottky-clamped transistor. Such a transistor is shown schematically in Figure 10.17a. The base metallization extends over the p⁺ base region (making an ohmic contact there) and over the n-collector region (creating a Schottky barrier). In effect, the base-collector pn junction and the metal-collector Schottky barrier junction are in parallel, as shown in (b).

In the **off** state (output high), both of these junctions are reverse biased and the transistor behaves normally. In the **on** state (output low) both junctions are forward biased, and we would normally expect the transistor to operate in saturation, with a large current flowing through the base-emitter and the base-collector junctions. In the Schottky-clamped transistor, however, the built-in voltage for the Schottky barrier diode from base to collector is less than that for the base-collector pn junction. From Figure 10.18 we can see that, for a given forward voltage from base to collector, the current through the Schottky barrier is greater than that through the pn junction. Further, the Schottky current is carried by majority carriers while the pn junction current is carried by minority carriers. This implies that the minority carrier injection across the base-collector junction, and thus the stored charge, is greatly reduced (about a factor of ten) compared with that of a conventional BJT. With reduced stored charge, the time required to remove it is decreased.

⁷This problem, however, can be handled in SPICE.

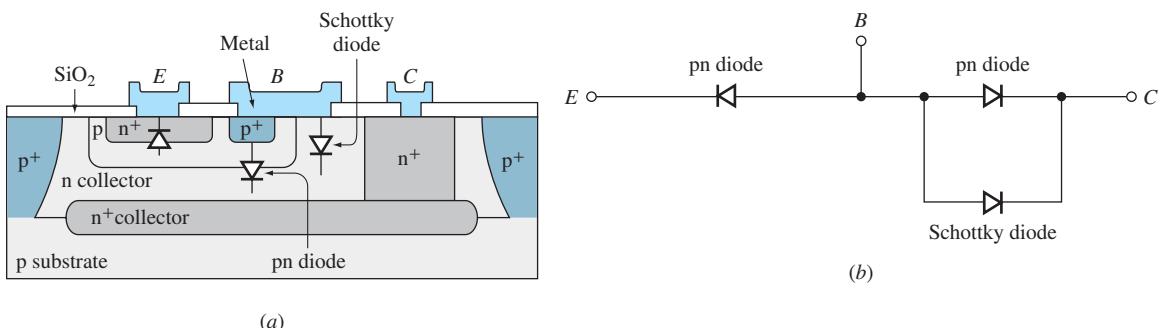


Figure 10.17 The Schottky-clamped transistor. (a) Cross-sectional view; (b) equivalent circuit schematic. The base-collector junction consists of two diodes in parallel.

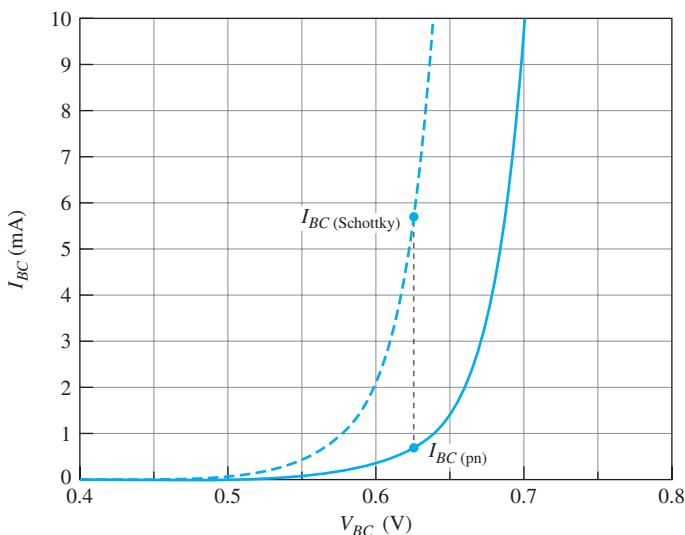


Figure 10.18 Comparison of the forward-biased $I-V_{BC}$ characteristics of the Schottky and pn diodes of the Schottky-clamped transistor. The Schottky diode carries more current than the pn junction. Since the Schottky current consists primarily of majority carriers, little minority carrier injection exists. Thus, for a given I_C , the stored-charge capacitance is less in the Schottky-clamped transistor.

10.7.3 DOUBLE HETEROJUNCTION BIPOLAR TRANSISTOR (DHBT)

An alternative method to reduce the low-to-high transition time is by the use of double heterojunction bipolar transistors. This can be accomplished in the silicon system by the use of $\text{Si}_x\text{C}_{1-x}$ in emitter and collector with a silicon base, or with a silicon emitter and collector and a $\text{Si}_x\text{Ge}_{1-x}$ base. For BJTs using III-V semiconductors, there are a number of semiconductor combinations; one possibility is to use a GaAs base with an AlGaAs emitter and collector.

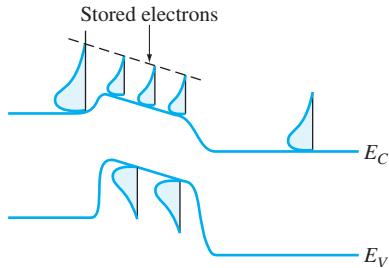


Figure 10.19 Band diagram with carrier concentrations for a DHBT in the low state. Because of the large barriers, the concentration of holes (minority carriers) in emitter and collector are negligible.

Figure 10.19 shows the energy band diagram and carrier concentrations in a DHBT with a graded composition base in the low state. It can be seen that because of the large barriers in the valence band, the concentration of holes (minority carriers) in the emitter and collector are negligible. While there are electrons stored in the p-type base, the base is narrow, and because of the electric field for electrons, they are quickly removed with the application of a high collector voltage. And since there are a negligible number of holes stored in the emitter or collector, the low-to-high transition time, t_{lh} , is markedly decreased from that of a homojunction BJT.

10.8 BJTs, MOSFETs, AND BiMOS

10.8.1 COMPARISON OF BJTs AND MOSFETs

Both BJTs and MOSFETs are used in electronic circuits. It is instructive to compare the important electrical parameters of these devices.

Input Impedance The dc input resistance is virtually infinite in a MOSFET. At high frequencies,

$$Z_{in} = \frac{1}{j\omega(C_{GS} + C_{GD})} \quad \text{FET} \quad (10.52)$$

For a BJT with a forward-biased E - B junction,

$$Z_{in} = r_\pi + \frac{1}{j\omega(C_{jBE} + C_{scBE})} \quad \text{BJT} \quad (10.53)$$

Because $(C_{jBE} + C_{scBE}) \gg (C_{GS} + C_{GD})$, the input impedance is much larger for a MOSFET than for a BJT.

Transconductance Transconductance is defined as the variation of output current with a change in input voltage

$$g_m = \left. \frac{\partial I_{\text{out}}}{\partial V_{\text{in}}} \right|_{V_{\text{out}}} = \frac{i_{\text{out}}}{v_{\text{in}}}$$

For a MOSFET, according to the simple long-channel model, in the current saturation region

$$\begin{aligned} I_{\text{out}} &= I_{D\text{sat}} = \frac{W\mu C'_{\text{ox}}(V_{GS} - V_T)^2}{2L} \\ g_m &= \frac{W\mu C'_{\text{ox}}(V_{GS} - V_T)}{L} = \frac{2I_{D\text{sat}}}{(V_{GS} - V_T)} \approx \frac{2I_{D\text{sat}}}{V_{DD}} \quad \text{FET} \end{aligned} \quad (10.54)$$

where $V_{GS} = V_{DD} \gg V_T$.

For a BJT [from Equation (10.13)]

$$\begin{aligned} I_{\text{out}} &= I_C = I_{E0} e^{qV_{BE}/kT} \\ g_m &= \frac{qI_C}{kT} \quad \text{BJT} \end{aligned} \quad (10.55)$$

The ratio of the transconductances is

$$\frac{g_m(\text{BJT})}{g_m(\text{MOSFET})} = \frac{\frac{qI_C}{kT}}{\frac{2I_{D\text{sat}}}{V_{DD}}}$$

For equal values of output current, $I_C = I_{D\text{sat}}$, and $V_{DD} = +2.5$ V,

$$\frac{g_m(\text{BJT})}{g_m(\text{MOSFET})} = \frac{2.5}{2 \times 0.026} \approx 50$$

or the transconductance of a BJT is much larger than that of a MOSFET.

The preceding comparison was for a long-channel MOSFET. In most modern MOSFETs, channel lengths are small enough that velocity saturation must be considered. When velocity saturation is taken into account, the transconductance g_m of the MOSFET is reduced with decreasing channel length, and the discrepancy is even more pronounced.

Speed The unity current gain frequencies for unloaded circuits are

$$f_T(\text{MOSFET}) \approx \frac{g_m(\text{MOSFET})}{2\pi(C_{GS} + C_{GD})}$$

$$f_T(\text{BJT}) \approx \frac{g_m(\text{BJT})}{2\pi(C_{jBE} + C_{scBE})}$$

Even though $g_m(\text{MOSFET}) \ll g_m(\text{BJT})$, it is also true that $(C_{GS} + C_{GD}) \ll (C_{jBE} + C_{sc\ BE})$, which makes the cutoff frequencies for unloaded devices comparable. For a circuit with an appreciable load capacitance C_L , however, the load capacitance predominates. Thus

$$f_T(\text{MOSFET}) \ll f_T(\text{BJT})$$

Power Dissipation The power dissipation in MOSFETs is appreciably less than for fast BJTs.

Ease of Manufacture MOSFETs require fewer processing steps than BJTs, and thus can be made at lower cost.

Summary From the above it would appear that MOSFET integrated circuits would be preferable to BJT circuits. This is the case for low- and medium-speed applications. The greatest advantage of the BJT is its much larger transconductance, which permits it to be used where the load capacitance is appreciable. The load capacitance consists of the wiring capacitance and the capacitance due to a large fan-out. The high transconductance of the BJT means a higher-output current, which can charge and discharge the capacitance more quickly.

The advantages of MOSFETs and of BJTs can be combined in a single chip using BiMOS or BiCMOS technologies, as discussed next. The use of lateral BJTs (discussed in the Supplement to Part 4), facilitates this combination.

10.8.2 BiMOS

The basic circuit diagram of a simple BiMOS analog amplifier circuit is shown in Figure 10.20. The input is to the gate of an n-channel field-effect transistor, and the output is at the collector of the bipolar transistor. Both devices are fabricated on the same chip. The input to the gate of the MOSFET (M) consists of

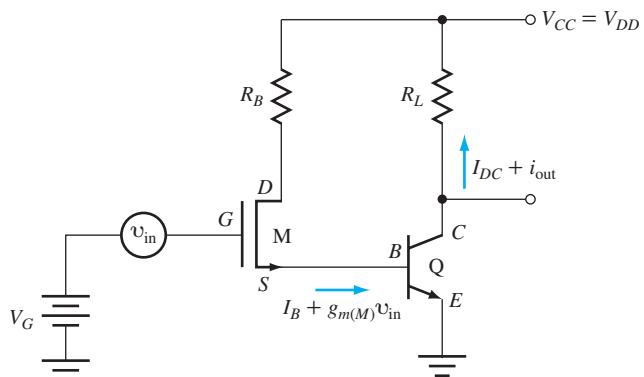


Figure 10.20 Schematic of a simple analog BiMOS amplifier. The input to the circuit is at the gate of the MOSFET (M), while the output is at the collector of the BJT (Q).

a dc voltage and a small signal ac analog voltage v_{in} . The dc gate voltage biases the MOSFET such that a drain current flows through the source and into the base of the BJT (Q). This current is sufficient to bias the BJT in the active mode. The signal voltage v_{in} produces a drain ac current $i_M = i_B = g_{m(M)}v_{in}$. This signal current is amplified by the BJT whose output signal current is

$$i_{(Q)} = i_{out} = \beta i_B = \beta g_{m(M)} v_{in}$$

The overall transconductance of this circuit is then

$$g_{m\text{total}} = \frac{i_{out}}{v_{in}} = \beta g_{m(M)}$$

This circuit has the advantage of combining the high-input impedance of the MOSFET with a high transconductance.

A simple digital inverter switch is indicated schematically in Figure 10.21. Consider the MOSFET to be an enhancement device. Then when $V_{in} = 0$, the MOSFET is **off** and current flows from V_{CC} through R_B and into the base of the BJT, turning it **on**. This causes a large I_C to flow through R_C . Since $V_C = V_{CC} - I_C R_C$, the output voltage V_{out} is low.

For a positive voltage ($V_{in} > V_T$), the MOSFET turns **on** and diverts the current from R_B through the MOSFET. The base current is then small and the BJT turns **off**. Now $V_{out} = V_{CC}$. Thus the BiMOS circuit of Figure 10.21 is an inverter. Figure 10.22 shows the cross-sectional view of this simple BiMOS circuit. The n^+ source region also serves as emitter. The BJT consists of the left n^+pn device. Note that the drain contacts overlap the p base region.

The preceding illustrations are quite simple. More advanced circuits also use both CMOS and complementary BJTs (e.g., npn and pnp) and are referred to as BiCMOS.

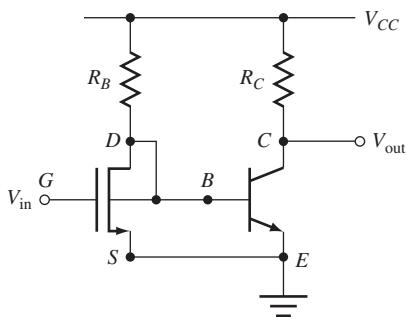


Figure 10.21 Circuit diagram of a BiMOS digital inverter. This circuit has the high input resistance and small input capacitance of the MOSFET and the high transconductance of the BJT.

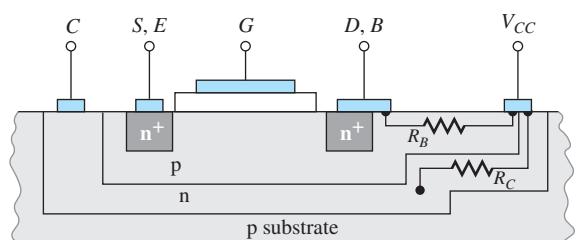


Figure 10.22 Cross-sectional schematic of the BiMOS circuit of Figure 10.20. The source and emitter are common, as are the drain and base.

10.9 SUMMARY

In this chapter, the dc models of Chapter 9 were extended to describe the time-dependent behavior of BJTs. The Ebers-Moll dc model was modified to include base-emitter and base-collector junction and the stored-charge capacitances. The small-signal hybrid-pi model was illustrated for the common emitter configuration for operation in the active mode.

The operation of BJTs as digital switches was discussed. To operate at high switching speeds it is necessary to keep the transistor out of the saturation region to minimize the base-collector stored charge. To accomplish this the base-collector should not be forward biased, or if it is, the bias should be small enough that negligible minority carrier injection exists between base and collector. Two schemes to implement this strategy were discussed: the Schottky-clamped transistor in which a Schottky diode exists in parallel with the base-collector pn junction, and the double heterojunction bipolar transistor, in which the band gaps of the emitter and collector are larger than that of the base.

Next, the electrical parameters of BJTs were compared with those of MOSFETs, to see the advantages of each technology:

- The input impedance of MOSFETs is much larger than that of BJTs. The input resistance of MOSFETs is nearly infinite for MOSFETs and is in the kilohm range for BJTs. The MOSFET input capacitance is much smaller than that for BJTs.
- The transconductance is much larger for BJTs than for MOSFETs.
- The unity current gain frequencies are comparable for two classes of devices for unloaded circuits. However, for circuits with appreciable load capacitance, the BJT is faster because of its higher transconductance.
- The advantages of each transistor type can be exploited by using MOSFETs for the input stage and BJTs for the output stage. Two examples of such BiMOS circuits were discussed.

10.10 REFERENCES

1. Joseph Lindmayer and Charles Y. Wrigley, *Fundamentals of Semiconductor Devices*, D. Van Nostrand, Princeton, NJ, Chapter 4, 1965.
2. John D. Cressler and Katsuyoshi Washio, “Bipolar transistors,” in *Guide to State-of-the Art Electron Devices*, John N. Burghartz (ed.), Wiley/IEEE Press, John Wiley and Sons, pp. 3–20, 2013.
3. Tak H. Ning, “History and future perspective of the modern silicon bipolar transistor,” *IEEE Trans. Electron Devices*, 48, pp. 2485–2491, 2001.
4. A. K. Kapoor and D. J. Roulston, eds., *Polysilicon Emitter Bipolar Transistors*, IEEE Press, New York, 1989.

10.11 REVIEW QUESTIONS

- Under what conditions can the capacitance C_{scBC} in Figure 10.1 be ignored, and why? Why is the diode representing the base-collector junction not shown in Figure 10.2?
- Explain how a small variation in I_B produces a large, proportional variation in I_C .
- How do designers minimize the parasitic base resistance in a BJT?
- What is meant by unity current gain frequency?
- In Chapter 9 we saw that making the base region thin increases the β of a transistor. Explain how a thin base also increases the cutoff frequency.
- How does the use of self-aligned structures allow one to make a higher-frequency BJT?
- Explain the operation of a Schottky-clamped transistor.
- Why is the switching speed of a DHBT greater than that of a BJT?

10.12 PROBLEMS

- 10.1** For an npn transistor with $N'_{DE} = 5 \times 10^{19} \text{ cm}^{-3}$, $N'_{AB} = 2 \times 10^{17} \text{ cm}^{-3}$, $N'_{BC} = 5 \times 10^{16} \text{ cm}^{-3}$, $W_E = 0.13 \mu\text{m}$, and $W_B = 0.15 \mu\text{m}$ under the bias conditions of $I_B = 20 \mu\text{A}$ and $V_{BC} = -2.5 \text{ V}$, find

- β
- I_C
- r_π
- g_m
- C_{BE} (C_π)
- C_μ
- f_{co}
- f_T

Note that band-gap narrowing should be accounted for, and that both sides of the $E-B$ junction are short. Let the area of the emitter junction be $A_E = 2.5 \times 10^{-7} \text{ cm}^{-2}$ and the area of the collector junction be $A_C = 8 \times 10^{-7} \text{ cm}^2$.

- 10.2** For the transistor of Problem 10.1, plot $|i_c|/|i_b|$ as a function of frequency. What is the unity current gain frequency?
- 10.3** An npn BJT with uniformly doped emitter, base, and collector has $\beta = 95$, base width W_B of $0.15 \mu\text{m}$, electron diffusion coefficient of $10 \text{ cm}^2/\text{s}$, r_π of 1000Ω , and a $C-B$ junction capacitance of 0.05 pF . What is its cutoff frequency?
- 10.4** If the base in the preceding problem is actually graded in doping, with a grading parameter $\eta = 2$,
 - What is the field in the base?
 - For the same value of I_C , what is the value of β ?
 - What is the resultant cutoff frequency?

- 10.5** Two transistors that are otherwise identical differ in that one has a uniform base and the other has a graded base with $\eta = 6$. Assume the field in the base is small enough that the diffusion constant has its low-field value.
- What is the improvement in β for the graded device?
 - What is the change in the stored-charge capacitance if both transistors are operated at the same value of I_C ?
- 10.6** Equation (10.39) indicates that the electron velocity in the base increases as $\Delta n_B(x)$ decreases (I_{nB} constant). Explain the physics of this.
- 10.7** Find the base transit time t_T for an npn prototype BJT with $W_B = 0.05 \mu\text{m}$ and base doping of $5 \times 10^{17} \text{ cm}^{-3}$.
- 10.8** Repeat problem 10.7 for a pnp prototype BJT with $W_B = 0.05 \mu\text{m}$ and base doping of $5 \times 10^{17} \text{ cm}^{-3}$.
- 10.9** Find the transit time t_{BC} across the base-collector transition region for an npn prototype BJT with base doping of $5 \times 10^{17} \text{ cm}^{-3}$, collector doping of $5 \times 10^{16} \text{ cm}^{-3}$, and collector-base voltage of 2.5 V.
- 10.10** In a particular prototype transistor fabrication process, the base width is cut in half. What is the effect on the base transit time?
- 10.11** An npn transistor's base region is doped to 10^{18} cm^{-3} . How thick would the base region need to be for the base transit time to be equal to 1/100 of the electron lifetime in the base? If such a transistor were manufactured, what would be the effect on β and the operating frequency?
- 10.12** For high-speed BJTs, the double poly self-aligned transistor is preferred to the conventional transistor discussed in Chapter 9. For similar emitter, base, and collector dopings compare the following parameters and explain your reasoning.
- Collector-base junction capacitance
 - Forward current gain β_F
 - Reverse current gain β_R
 - Base resistance
 - Early voltage
- 10.13** For a particular Schottky-clamped transistor, the leakage current density in the Schottky diode is $J_{0SD} = 10^{-5} \text{ A/cm}^2$, and the leakage current density for the base-collector junction is $J_{0BC} = 10^{-11} \text{ A/cm}^2$. For a given bias voltage across both in parallel, what is the ratio of the current flowing through the Schottky contact to the current flowing through the B-C junction? What is the impact on turn-off time? Assume that the area of the base-collector junction is 10 times that of the Schottky diode.
- 10.14** From Equations (10.43) and (10.23), find an expression for the base transit time in a graded-base transistor as a function of η . Assume that the area of the pn junction is 20 times that of the Schottky diode.
- 10.15** Why is the frequency response of a loaded MOSFET amplifier less than that of a similar BJT circuit?
- 10.16** Discuss the reason for using a BiMOS amplifier.
- 10.17** Why is the **on-off** switching speed of a DHBT much smaller than that of a homojunction BJT?

Supplement to Part 4: Bipolar Devices

S4.1 INTRODUCTION

In the previous two chapters the basic physics of operation and the electrical characteristics of bipolar junction transistors were discussed. In this Supplement some second-order effects important for operation at higher currents and voltages are considered.

S4.2 CURRENT CROWDING AND BASE RESISTANCE IN BJTs

As indicated in Section 9.3, the base region should be thin in a good bipolar transistor to increase current gain and frequency response. A thin base does have a drawback, however: a high base resistance that degrades the device performance.

Consider the discrete transistor of Figure S4.1. The emitter length L and width h are indicated. There is a base resistance, which consists of two parts. These are the resistance R'_B from base contact to the emitter edge, due primarily to the resistance of the p-type material, and R_B , the effective resistance under the emitter (often referred to as the *intrinsic base resistance*), which is normally larger because here the base is thin. In this section, we are interested only in R_B . We call R_B an “effective” resistance; it is really a distributed quantity, because I_B decreases with increasing position y , but for circuit analysis it is convenient to consider it a lumped resistance.

The base current I_B flows laterally (parallel to the junction plane) and causes a lateral IR drop in the base region. This drop along the emitter edge means that the base-emitter voltage V_{BE} is a function of position. Since the emitter current density is a strong (exponential) function of V_{BE} ,

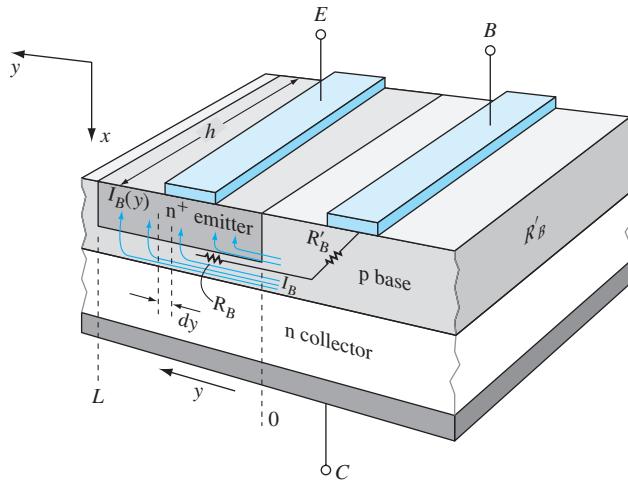


Figure S4.1 Current crowding in a BJT. There is a voltage drop across the base resistance R_B , which varies the emitter-base junction bias along the edge of the emitter. The effect is a greater bias at the end toward the base contact, and thus greater collector-emitter current flow at that end.

$$J_E(y) = J_{E0} \left(e^{qV_{BE}(y)/kT} - 1 \right) \quad (\text{S4.1})$$

a small lateral voltage drop in the base causes a large spatial dependence on emitter current density and thus collector current density. The polarity of this lateral voltage drop is such that V_{BE} is largest at the emitter edge nearest to the base contact. The emitter current and collector current then “crowd” toward the base contact.

This current crowding (often called emitter crowding) increases with increasing base current, since that increases the effect of the IR drop. That means that as I_B increases, the current crowding increases and a larger fraction of the current is flowing into the emitter from the base and from the collector at the end near the base contact than at lower currents. The implication is that the effective distance that the base current flows is reduced with increasing current, and therefore so is the equivalent lumped base resistance, R_B .

There are several ways to define a lumped resistance. [1] Here we define R_B as the magnitude of the average lateral base voltage drop divided by the terminal base current

$$R_B \equiv \frac{\langle V_B \rangle}{I_B} \quad (\text{S4.2})$$

where the average lateral base voltage drop is

$$\langle V_B \rangle = \frac{\int_0^L V_B(y) \frac{dV_B(y)}{dy} dy}{\int_0^L \frac{dV_B(y)}{dy} dy} \quad (\text{S4.3})$$

The quantity $V_B(y)$ is the base voltage at position y , and $dV_B(y)/dy$ is the base voltage distribution function with position. It can be written

$$\frac{dV_B(y)}{dy} = -\frac{R_{\square}}{h} I_B(y) \quad (\text{S4.4})$$

where $I_B(y)$ is the base current at position (y) , and R_{\square} is the sheet resistance in the base (ohms per square),

$$R_{\square} = \frac{1}{W_B} \frac{q \mu_p}{0} \int N'_{AB} dx \quad (\text{S4.5})$$

where W_B is the base width.

Solving for R_B in general is somewhat involved. However, at small base currents (low injection) such that $V_{BE}(y)$ is nearly constant, $I_B(y)$ varies linearly from I_{B0} at $y = 0$ to zero at $y = L$:

$$I_B(y) = I_B \left(1 - \frac{y}{L} \right) \quad (\text{S4.6})$$

where I_B is the base terminal current.

Combining Equations (S4.4) and (S4.6) and integrating, we find

$$V_B(y) = -\frac{R_{\square}}{h} I_B \left[y - \frac{y^2}{2L} \right] \quad (\text{S4.7})$$

Then solving for R_B with the aid of Equations (S4.2), (S4.3), (S4.4), and (S4.7) gives

$$R_B = \frac{R_{\square} L}{4h} \quad (\text{S4.8})$$

Equation (S4.8) is valid for small I_B , where current crowding is negligible. With increasing I_B , the emitter current is increasingly concentrated near the emitter edge adjacent to the base contact, which as we saw causes a reduction in R_B . Figure S4.2 shows a plot of the normalized base resistance (normalized taking into account the h/L aspect ratio of the emitter and the sheet resistance) as a function of normalized base current (normalized as indicated on the graph). There is a second line on the graph—it is an alternative normalized base equivalent

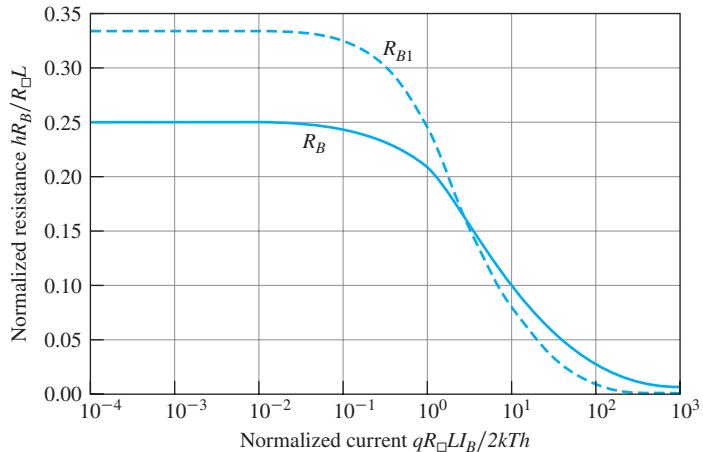


Figure S4.2 The effect of the base current on the effective resistance R_B used to represent the effect of current crowding. Two different models for lumping this distributed resistance are used. The solid line is based on the average voltage drop across the lumped resistance. The dashed line is based on the average power dissipated in the resistor.

resistance R_{B1} , arrived at using another definition of equivalency. This alternative R_{B1} is the basis of a homework problem.

While it might appear that the reduction of base resistance due to current crowding would be beneficial, this is not the case for two reasons:

1. Since the base-emitter current is concentrated at the edge of the emitter near the base contact, much of the emitter is inactive. However, the emitter-base junction capacitance associated with this inactive emitter remains, with its adverse effect on frequency response.
2. As discussed in Section S4.5, the high current density at the emitter edge reduces the current gain, β .

In *power transistors*, the currents are large enough that the current crowding causes high-injection effects that tend to reduce the current gain. In that case, two bases are often used, one on either side of the emitter as indicated in Figure S4.3a and b to reduce the crowding. (The base contact is often made to only one side.) This double base reduces the base resistance by a factor of 4. To reduce the base resistance even further, interdigitated structures, like that shown schematically in Figure S4.4a and b are often used.

In modern integrated circuit BJTs, the emitter stripe width L can be made small (in the sub-micrometer range) and for such cases, current crowding is not a major problem. Typical emitter sizes in modern devices are on the order of $0.1 \times 1 \mu\text{m}$.

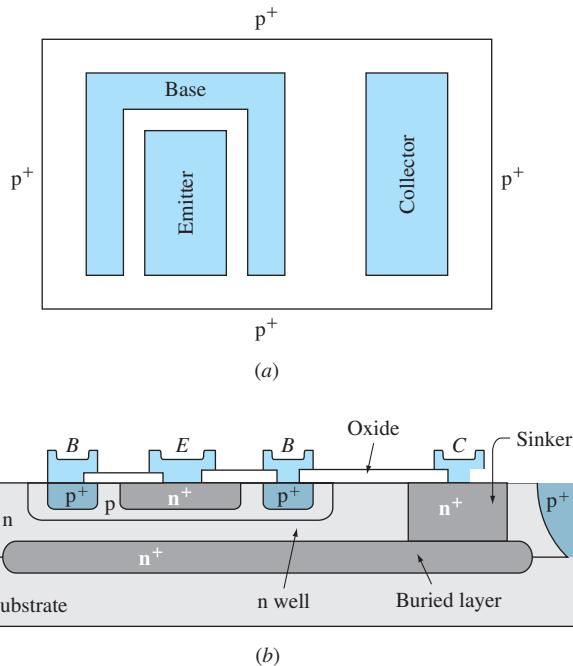


Figure S4.3 A double-base transistor. (a) Top view; (b) cross section.

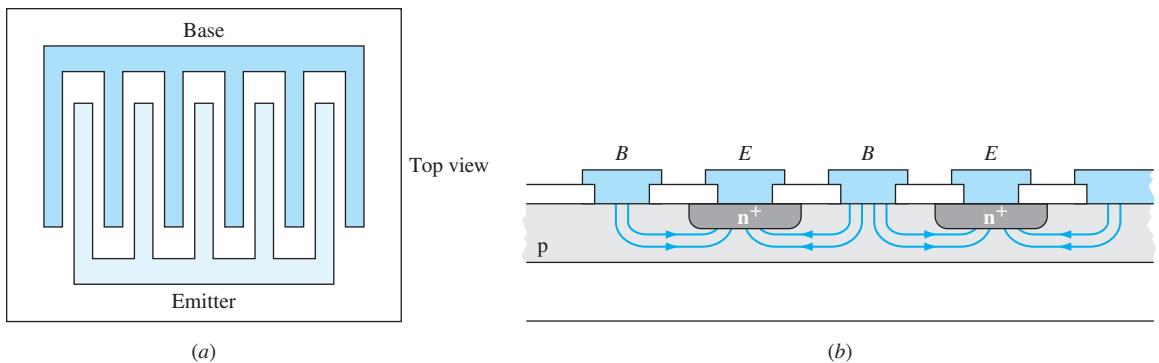


Figure S4.4 The contacts can be interdigitated, as in (a) top view and (b) cross section.

S4.3 BASE WIDTH MODULATION (EARLY EFFECT)

In Section 9.4.2, we obtained an expression for the injection efficiency γ [Equation (9.28)] that contained the quantity W_B/W_E in the denominator. The quantity W_B is the base width measured between the E - B depletion region and the B - C

depletion region edges as was shown in Figure 9.5. As the reverse bias across the collector-base junction is increased, the depletion region gets wider and the effective base width gets narrower. This has the effect of increasing γ and thus increasing the current gain β .

Figure S4.5a shows the electron distribution n_B in the base as a function of distance across the base for different values of V_{CE} for a uniformly doped base transistor, so the distribution is a straight line ($\eta = 0$). Because the reverse-biased C-B junction extracts carriers, n_B goes to zero at W_B , which shrinks with V_{CB} (and V_{CE} , assuming a constant base-emitter bias).

The effect of the shrinking base on the I_C - V_{CE} characteristics is shown in Figure S4.5b. Recall that $\beta = I_C/I_B$. For a given value of I_B , I_C increases with increasing V_{CE} , reflecting the increase in β . Extrapolations of these I_C - V_{CE} curves meet (approximately) at a voltage $-V_A$, where V_A is called the *Early voltage*. [2] Note that this is similar to the influence of drain voltage in a MOSFET on channel length, and thus on drain current. The argument is different but the effect on the I - V characteristics is the same.

Figure S4.5a is repeated in Figure S4.6 for the case of the graded-base transistor. Although a change in V_{CE} does change the effective base width, it has no effect on the field in the base and little effect on dn_B/dx at the emitter edge of the base, and thus it has little influence on saturation current. This in turn increases the magnitude of the Early voltage V_A and results in a more constant value of β . In Chapter 10 we saw that the small-signal output conductance ($g_d = dI_C/dV_{CE}$) is also reduced.

As the collector voltage gets larger, and the depletion region at the base-collector junction gets wider, it can become wide enough such that W_B becomes zero. When this happens, the depletion region of the collector actually meets the depletion region of the emitter, as shown in Figure S4.7a. The current at that point becomes very large. The electrons are swept out of the emitter to the collector and the base loses control over I_C . Transistor action disappears, as can be seen from the I_C - V_{CE} characteristics in Figure S4.7b. This effect is called *punch-through*.

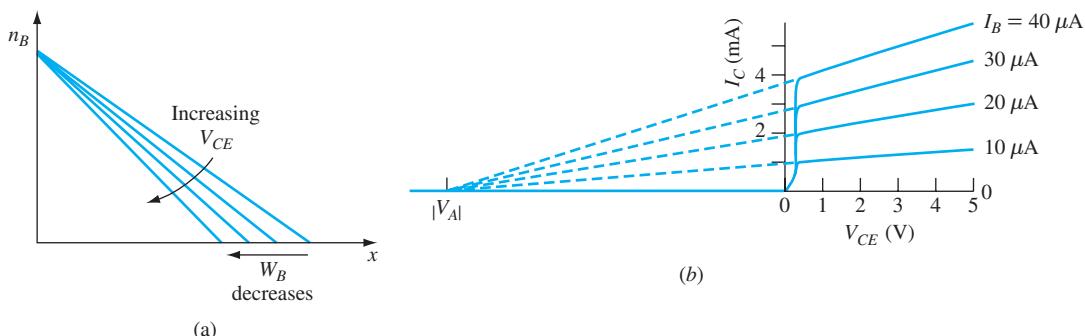


Figure S4.5 Increasing collector voltage causes a decrease in effective base width (a) resulting in increased I_C and β (b) for a uniform base transistor. The Early voltage V_A is indicated.

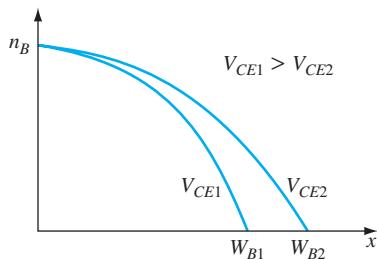


Figure S4.6 For a graded-base transistor, a change in collector voltage has little effect on n_B near the emitter, and thus little effect on I_C or β , resulting in increased V_A .

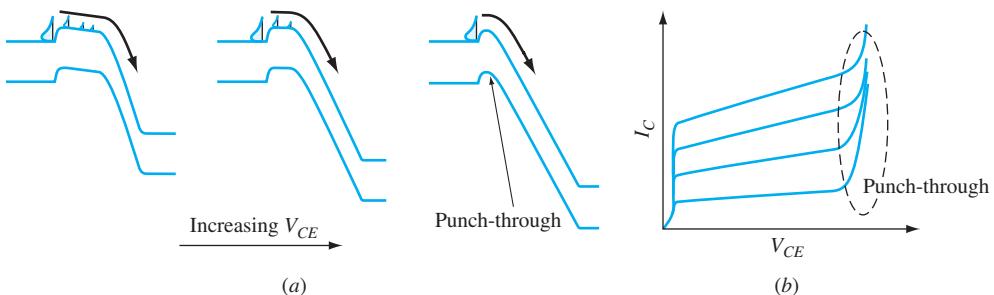


Figure S4.7 (a) As the collector-base voltage increases, the effective base width decreases. At punch-through, the two depletion regions meet, and the emitter electrons are swept directly into the collector. (b) The effect is loss of transistor action.

EXAMPLE S4.1

Consider a prototype npn BJT with emitter, base, and collector dopings:

$$\begin{aligned}N'_{DE} &= 7 \times 10^{19} \text{ cm}^{-3} \\N'_{AB} &= 4 \times 10^{17} \text{ cm}^{-3} \\N'_{DC} &= 6 \times 10^{16} \text{ cm}^{-3}\end{aligned}$$

What is the minimum value of W_{BM} , the metallurgical base width, such that the punch-through voltage $V_{CE}(PT)$ is greater than 6 V? Assume $V_{BE} = 0.75$ V.

■ Solution

Figure S4.8a shows the depletion region widths in the base for a BJT under normal bias. From the figure, we see that

$$W_B = W_{BM} - w_{pEB} - w_{pCB}$$

where W_{BM} is the metallurgical base width and w_{pEB} and w_{pBC} are the depletion widths on the p (base) sides of the E-B and B-C junctions.

At punch-through, the C-B depletion region meets the E-B depletion region, and $W_B = 0$. Then

$$W_{BM} = w_{pEB} + w_{pCB}$$

For the collector-base junction, from Equation (5.36), we have

$$w_{pCB} = \left[\frac{2\epsilon V_{jCB}}{qN'_{AB} \left(1 + \frac{N'_{AB}}{N'_{DC}} \right)} \right]^{1/2} = \left[\frac{2\epsilon(V_{biBC} + V_{CB})}{qN'_{AB} \left(1 + \frac{N'_{AB}}{N'_{DC}} \right)} \right]^{1/2}$$

where $V_{jCB} = V_{biCB} + V_{CB}$, N'_{AB} is the net doping on the p side of the junction and N'_{DC} is that on the n side.

Since the emitter-base junction can be considered to be a one-sided step junction, the depletion region width w_{pEB} is

$$w_{pEB} = \left[\frac{2\epsilon(V_{biEB} - V_{EB})}{qN'_{AB}} \right]^{1/2}$$

To find the built-in voltages we refer to Figure S4.8b, which shows the neutrality energy band diagram for this transistor. From Figure 9.8, the impurity-induced apparent band-gap narrowing in the emitter is $\Delta E_g^* \approx 0.09$ eV. We neglect the small apparent band-gap narrowing in the base and collector.

In the base the hole concentration is $p_{B0} = N'_{AB} = N_V e^{-\delta p/kT}$, so

$$\delta p = kT \ln \frac{N_V}{N'_{AB}} = 0.11 \text{ eV}$$

as indicated in the figure. The emitter-base built-in voltage then is (assuming that E_f in the emitter is at the [reduced] value of E_C),

$$V_{biEB} = \frac{1}{q} (E_g - \Delta E_g^* - \delta p) = 1.12 - 0.09 - 0.11 = 0.92 \text{ V}$$

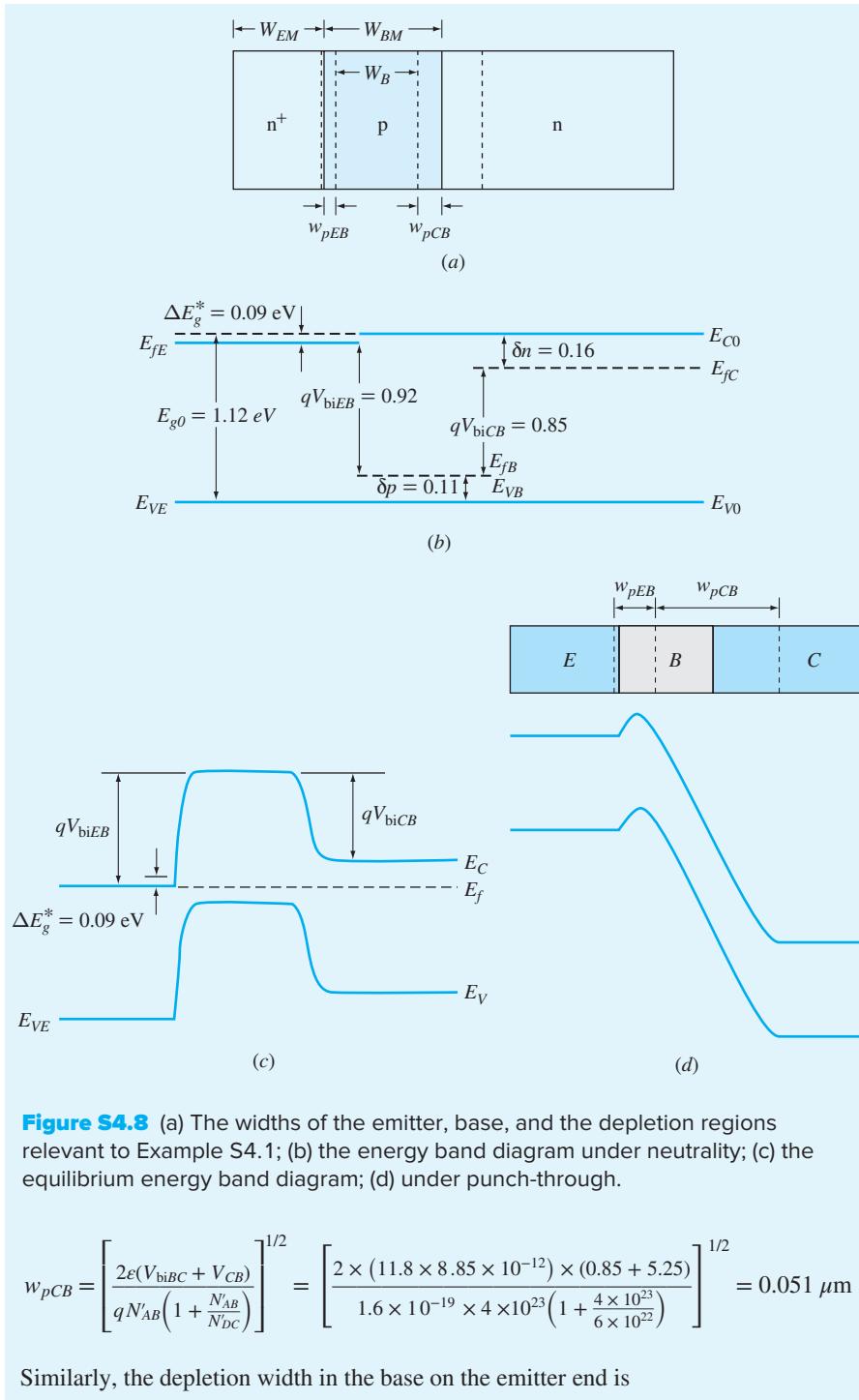
Similarly in the collector, $\delta n = 0.16$ eV and the base-to-collector built-in voltage is

$$V_{biCB} = E_g - \delta n - \delta p = 0.85 \text{ V}$$

Alternatively, from Equation (5.13),

$$V_{biCB} = kT \ln \frac{N'_{DC} N'_{AB}}{n_i^2} = 0.85 \text{ V}$$

The equilibrium energy band diagram is shown in Figure S4.8c; for the specified punch-through voltage of $V_{CE} = 6$ V, the collector-base reverse bias at $V_{BE} = 0.75$ V is $V_{CB} = V_{CE} - V_{BE} = 6 - 0.75 = 5.25$ V. At this bias, the depletion region width on the base side of the C-B junction is



$$w_{pEB} = 0.024 \mu\text{m}$$

At punch-through, (d), the depletion regions meet. To prevent this from happening at a value of V_{CE} less than 6 V, the minimum metallurgical base width is

$$W_{BM} = w_{pEB} + w_{pCE} = 0.024 + 0.051 = 0.075 \mu\text{m}$$

S4.4 AVALANCHE BREAKDOWN

We have considered the case where the collector-base junction voltage is low enough that the carrier multiplication effect is negligible, or $M = 1$ and $\alpha = \gamma\alpha_T$. At sufficiently high reverse voltage, current multiplication occurs and thus $M > 1$. In this case, for avalanche to occur, $M = 1 + 1/\beta$. Note that for $\beta = 100$, for avalanche breakdown $M = 1.01$, or relatively little multiplication is required to cause avalanche breakdown.

Although avalanche is a different effect, the I_C - V_{CE} characteristics for avalanche breakdown look similar to those for punch-through indicated in Figure S4.7c.

S4.5 HIGH INJECTION

Until now, we have treated the BJT in the low-level injection condition. By this, we mean that the electron concentration in the base is everywhere much less than the hole concentration, despite injection of electrons into the p-type base. The energy band diagram of Figure S4.9a illustrates the low-injection case. Under operation, an excess electron concentration Δn is injected into the base. This excess concentration of negative charges tends to make the base more negative. This, in turn, creates a field that attracts excess holes, Δp , from the base contact. To achieve neutrality, Δp must equal Δn . Under low injection $\Delta p \ll N'_A$; however, this makes almost no difference to the relative hole concentration. Thus, the back injection of holes from the base to the emitter is essentially unaffected.

At large forward bias, such that Δn is not small compared with $p_{B0} = N'_A$, as in Figure S4.9b, it is possible for the number of holes, $p = p_{B0} + \Delta p = N'_A + \Delta n$, to be changed significantly. Since at large forward bias the barrier to hole injection is small, many of these excess holes have enough energy to be injected into the emitter, thus increasing both I_{pE} and I_{pB} . The increase in I_{pE} decreases the injection efficiency.

The injection efficiency, γ , is

$$\gamma = \frac{I_{nE}}{I_E} = \frac{I_{nE}}{I_{nE} + I_{pE}} = \frac{1}{1 + \frac{I_{pE}}{I_{nE}}}$$

which approaches $\gamma = \frac{1}{2}$ as $I_{pE} \rightarrow I_{nE}$. This results in an emitter current and thus a collector current that vary as $e^{\frac{qV_{BE}}{2kT}}$.

The increase in I_{pB} means an increase in the total base current I_B , which, since $\beta = I_C/I_B$, causes β to decrease from its low-injection value.

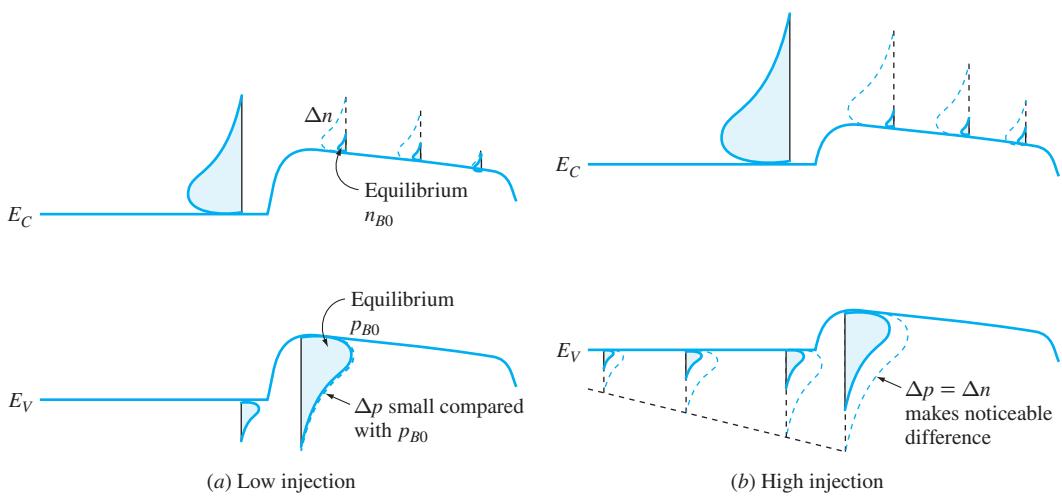


Figure S4.9 To achieve space charge neutrality, electrons injected into the base (Δn) draw an equal number of excess holes (Δp). (a) For low injection, these excess holes have negligible influence on the base-emitter hole current. (b) Under high injection the base-emitter barrier for holes is small enough that many of these excess holes are injected into the emitter.

S4.6 BASE PUSH-OUT (KIRK) EFFECT

There is a second reason β is reduced under high injection. At high currents, the effective base width *increases*, with a concurrent *reduction* in β . This effect, often referred to as the *base push-out effect* or the *Kirk effect* [3] is discussed with the aid of Figure S4.10.

When we considered the low-injection condition we assumed the electron concentration at W_B to be zero, and we neglected the small electron concentration inside the base-collector transition region. This is incorrect because the base electron current density, which is equal to the collector current density J_C , is

$$J_C = J_{nB} = -q n_B(x) v(x) \quad (\text{S4.9})$$

where $v(x)$ is the velocity of the carriers. Since in the high-field C-B junction the electron velocity is limited by its saturation velocity v_{sat} we know

$$v(x) \leq v_{\text{sat}} \quad (\text{S4.10})$$

From Equation (S4.9), if $v(x)$ has a maximum for a given current, then the electron concentration has a minimum, even near and within the CB transition region:

$$n \geq \frac{|J_C|}{|qv_{\text{sat}}|} = n_{\min} \quad (\text{S4.11})$$

Thus n_{\min} is the minimum electron concentration in the depletion region.

For the low-injection condition (small J_C), this value of n_{\min} is small enough to be ignored. For high injection, however, it must be considered. As electrons enter the high-field transition region, they are accelerated toward the collector.

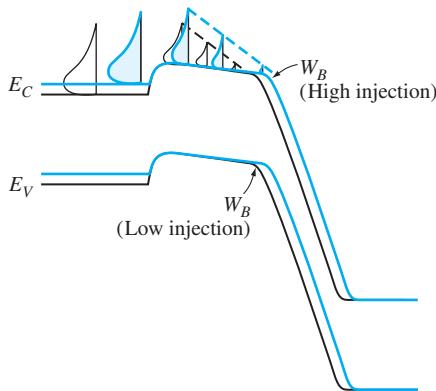


Figure S4.10 At high currents the electron charge density near the collector edge of the base, and for a short distance into the base-collector transition region, must be considered. This results in an increased effective base width and a decreased β .

Their velocity increases with increasing x until they reach their saturation velocity. The effect of this negative charge is to cause the electron potential energy E_C (and also E_V) to increase within the transition region. This effectively increases the effective base width W_B , with a corresponding decrease in β .

For currents large enough such that $n_{\min} > N'_{DC}$, where N'_{DC} is the donor concentration in the n collector, the *apparent* base region extends well into the n collector region and can approach the n⁺ collector region, where N'_{DC} is much larger. With increasing J_C then, the high-field region shifts into the collector, as indicated in Figure S4.10. To avoid appreciable base push-out, it is desirable to design for

$$J_C \leq 0.3|qN'_{DC}v_{\text{sat}}| \quad (\text{S4.12})$$

EXAMPLE S4.2

Estimate the maximum collector current density J_C for an n collector doping of $N'_{DC} = 5 \times 10^{16} \text{ cm}^{-3}$ to avoid excess base push-out.

■ Solution

From Equation (S4.12),

$$\begin{aligned} J_{C\max} &\approx 0.3|qN'_{DC}v_{\text{sat}}| = 0.3(1.6 \times 10^{-19} \text{ C})(5 \times 10^{16} \text{ cm}^{-3})(10^7 \text{ cm/s}) \\ &= 2.4 \times 10^4 \text{ A/cm}^2 \\ &= 0.24 \text{ mA}/\mu\text{m}^2 \end{aligned}$$

From the above, it can be seen that typically J_C is less than 1 mA/ μm^2 .

S4.7 RECOMBINATION IN THE Emitter-Base JUNCTION

Just as for a pn junction diode, some electron-hole recombination occurs in the transition region of the emitter-base junctions of a BJT. The $B-E$ current resulting from this recombination never reaches the collector and thus does not contribute to I_C . It does, however, increase both I_E and I_B from those values found by the earlier models. Since $\alpha = I_C/I_E$ and $\beta = I_C/I_B$, these increases in I_E and I_B reduce the values for α and β . Figure S4.11 shows plots of I_C and I_B as functions of base-emitter voltage. This is referred to as a *Gummel plot*. The collector current I_C is essentially equal to the emitter electron current I_{nE} that is injected into the base and which is proportional to $e^{qV_{BE}/kT}$. Thus on the semilog scale of Figure S4.11, the I_C-V_{BE} plot is a straight line of slope $(q/kT)\log e$.

The base current has two components, injection current and recombination current. Injection current results from holes injected into the emitter and varies

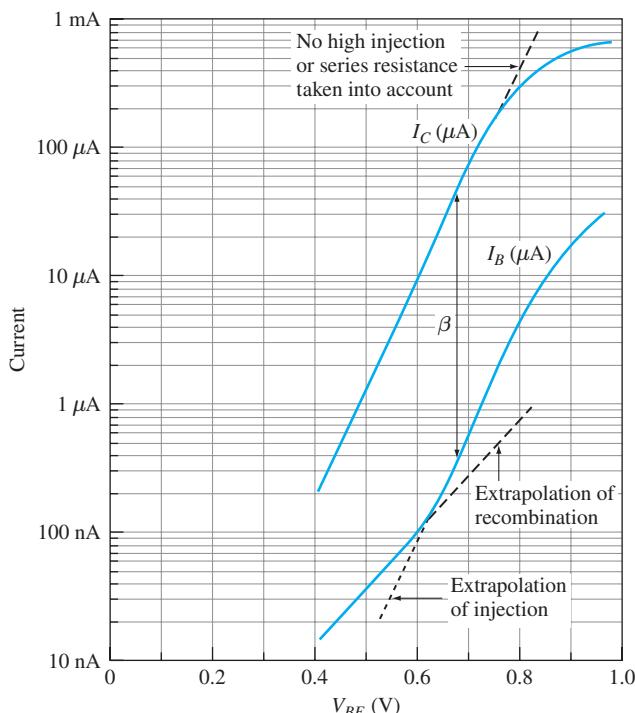


Figure S4.11 Gummel plot for a BJT. Except at high current, the $\log I_C-V_{BE}$ plot is a straight line of slope $(q/kT) \log e$. The I_B-V_{BE} plot is a straight line of slope $(q/kT) \log e$ at low currents, changing to $(q/kT) \log e$ at higher currents. The deviation at the higher currents results from a portion of V_{BE} being dropped across the base series resistance (and for power transistors with small base resistance, high-injection effects).

as $e^{qV_{BE}/kT}$. The recombination current results from the base supplying the holes to support the electron-hole recombination in the emitter-base transition region. The recombination current varies as $e^{qV_{BE}/nkT}$, where n is an empirical parameter and $n \approx 2$. Since the injection current varies more rapidly with V_{BE} than the recombination current, injection predominates at higher base-emitter voltage while the recombination current predominates at lower voltages.

Recall that $\beta = I_C/I_B$. Thus, on this semilog plot, the vertical distance between the I_C and the I_B plots is proportional to β . For analog circuits, BJTs are normally operated in the region of near-constant β .

At high currents, the slope reduces as a result of series resistance, and as discussed in Sections S4.5 and S4.6, β is again reduced because of high-injection effects.

S4.8 OFFSET VOLTAGE IN BJTS

The I_C - V_{CE} characteristics of a BJT in the common-emitter configuration are shown in Figure S4.12a. It appears that $I_C = 0$ at $V_{CE} = 0$ for all values of I_B . In reality, however, the curves do not pass through the origin, as seen in part (b) of the figure. There is a voltage offset where each curve passes through $I_C = 0$. Here we explore this.

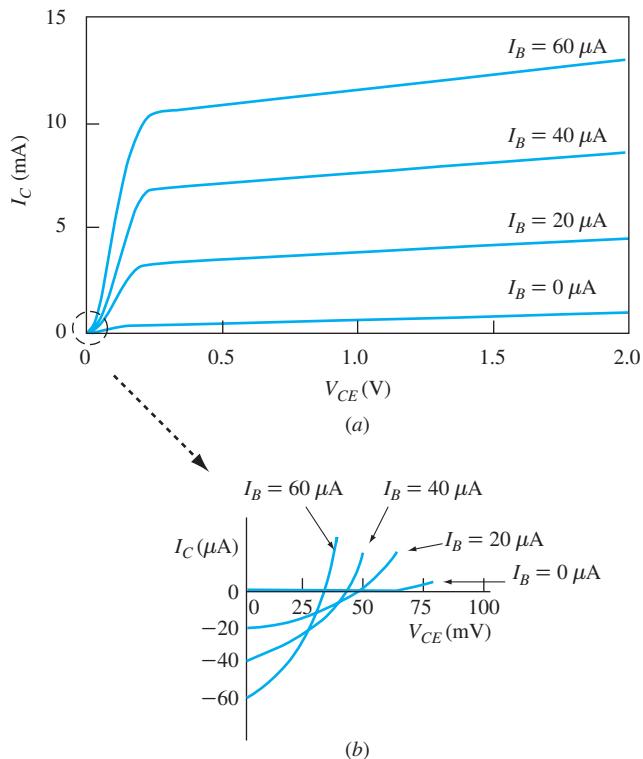


Figure S4.12 I_C - V_{CE} characteristics of an npn BJT (a) and the expanded characteristics near the origin (b).

When $V_{CE} = 0$, both the E-B and B-C junctions are forward-biased, as long as $I_B > 0$. Thus hole current flows from the base into the emitter and also into the collector (negative I_C according to convention). Since the emitter is much more heavily doped than the base and the base is more heavily doped than the collector, $I_{BC} \gg I_{BE}$. Because I_{BE} is very small, it is primarily generation current and thus does not contribute to β .

For a given value of I_B , with increasing V_{CE} , V_{BE} is reduced, and V_{BC} also starts to decrease, reducing I_{BC} (if V_{CE} increases enough, the base-collector junction will become reverse biased). Still holding I_B fixed, because $I_B = I_{BC} + I_{BE}$, when I_{BC} decreases, I_{BE} increases and the base-emitter junction becomes increasingly forward biased. That increases the injection current, thus increasing β , until I_{BE} consists primarily of injection current and β becomes constant. (See Figure S4.11.)

The offset voltage depends on the doping levels in base, emitter, and collector and on the base width, but it is typically about 25 to 50 mV, hardly noticeable in the normal experimental $I_C - V_{CE}$ characteristics.

S4.9 LATERAL BIPOLAR TRANSISTORS

The bipolar junction transistors described so far are vertical, in that the carrier flow from emitter to collector is normal to the surface, Figure S4.13a. More recently, silicon lateral bipolar transistors have been reported [4]. Figure S4.13b and c illustrate the structure of complementary symmetric lateral bipolar transistors on insulator (SOI). In this structure, emitters and collectors are heavily

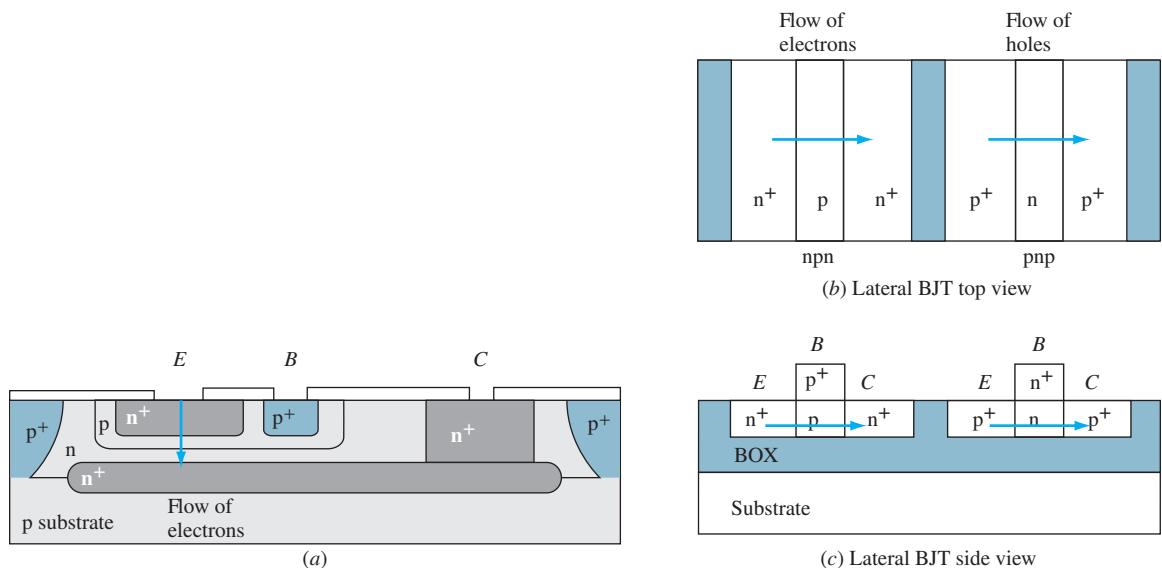


Figure S4.13 Schematic illustrating the structure of complimentary lateral bipolar transistors on oxide (SOI). (a) Vertical transistor for comparison; (b) top view; (c) side view.

doped while the base regions are more lightly doped. Such a device has advantages over the vertical BJT:

1. Because the emitter and collector in the device are equally doped and their widths are equal, a device can operate equally in forward-active and reverse-active modes.
2. Because of the heavy collector doping there is no base push-out (Kirk) effect.
3. Due to the SOI structure, the emitter and collector junction capacitance is small, resulting in higher frequency response. Values of f_T of 350 GHz and f_{max} of 700 GHz have been reported.
4. The fabrication process flow is similar to that for CMOS. This facilitates the manufacture of bipolar transistors and CMOS transistors on the same chip, a process called BiCMOS.
5. Lateral BJTs can be fabricated with much smaller dimensions than vertical BJTs with the same current densities. The smaller size means shorter connecting wires and smaller transistor and wire capacitance, thus higher frequency operation (see Figure S4.13).
6. The **off-to-on** switching time is greatly reduced because the heavy emitter and collector doping with respect to that of the base reduces the minority carrier storage in collector and emitter in the **off** condition.

S4.10 SUMMARY

In this chapter some second-order effects important at higher currents and voltages are discussed. Because of the $I_B R_B$ voltage drop in the base region under the emitter, the base-emitter and collector-emitter currents are “crowded” toward the base contact, thus increasing the high-injection effects. For high collector voltage, a portion of the transition region of the base-collector junction extends into the base, thus reducing the effective base width from its metallurgical width. With increasing collector voltage, the effective base width decreases and thus β increases (Early effect). Under high-injection conditions, such as in power transistors, the electronic charge associated with the high current causes the effective base width to increase. Thus at a given collector voltage, β decreases with increasing current (Kirk effect). Carrier recombination in the emitter-base transition region results in reduced β at low currents.

A symmetric lateral bipolar transistor structure on SOI was briefly described which has several potential advantages over the conventional vertical BJT.

S4.11 REFERENCES

1. J. L. Lary and R. L. Anderson, “Effective base resistance of bipolar transistors,” *IEEE Trans. on Electron Devices*, ED-32, pp. 2503–2505, 1985.

2. J. M. Early, "Effect of space-charge layer widening in junction transistors," *Proc. IRE*, 40, pp. 1401–1406, 1952.
3. C. T. Kirk Jr., "A theory of transistor cutoff frequency (f_T) falloff at high current densities," *IEEE Trans. Electron Devices*, ED-9, pp. 164–174, 1962.
4. Tak H. Ning and Jin Cai, "On the performance and scaling of symmetric lateral bipolar transistors on SOI," *IEEE Journal of the Electron Devices Society*, 1, no. 1, pp. 21–27, 2013.

S4.12 REVIEW QUESTIONS

1. Explain the physical origin of the Early effect.
2. High-injection conditions tend to reduce the current gain in a BJT. Three reasons were discussed in this chapter. Name them and describe the physics of each.
3. What advantages do lateral transistors have over the conventional vertical transistors?
4. What is the base push-out effect, and why is it eliminated in a lateral BJT?
5. Current gain in a BJT is reduced at high currents. Why?

S4.13 PROBLEMS

S4.1 Fill in the missing steps to verify Equation (S4.8).

S4.2 We previously chose an equivalent resistance in the base using the average voltage drop across the base. Here we will use the power dissipated in a resistor, $P = I^2R$. We define a new equivalent resistance based on the power P_B that is dissipated by the base current flowing through a lumped resistance $R_B = P_B/I_B^2$, where P_B is the power dissipated in the base. Find an expression for R_B . Assume I_B to be small enough that Equation (S4.6) is valid.

S4.3 For the transistor whose I_C - V_{CE} curves are shown in Figure PS4.1, find the Early voltage.

S4.4 An npn transistor has $N'_{DE} = 10^{20}$, $N'_{AB} = 5 \times 10^{17}$, and $N'_{DC} = 10^{16}$. For $V_{BE} = 0.75$ V, how small can the base width W_{BM} be to keep the punch-through voltage above 12 V?

S4.5 As indicated in Section S4.6, to avoid excessive base push-out, J_C is less than about $1 \text{ mA}/\mu\text{m}^2$. Consider a BJT with an emitter width (L) of $0.5 \mu\text{m}$ and a base sheet resistance of $10 \text{ k}\Omega/\square$. For $\beta = 100$, find I_B and the lateral voltage drop in the intrinsic base region and discuss the current crowding effect for this device.

S4.6 For the transistor whose I_C - V_{CE} curves are shown in Figure PS4.2, explain why the lines are closer together as I_B increases.

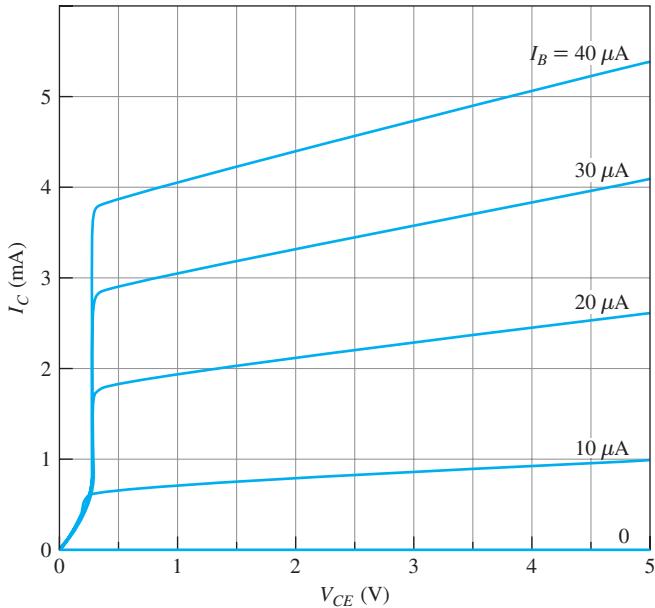


Figure PS4.1

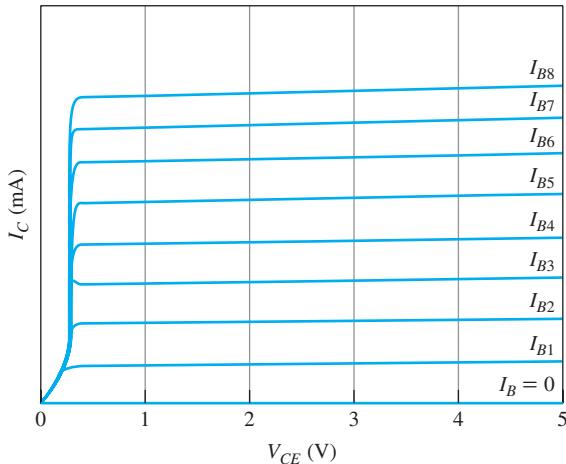


Figure PS4.2

- S4.7** Find β for the device whose Gummel plot is shown in Figure PS4.3.
- S4.8** Why is base width modulation (Early effect) more pronounced in a lateral BJT than in a vertical BJT?
- S4.9** In an npn BJT, for positive base current, the collector current is negative for zero V_{CE} (Figure S4.13b). Explain this.

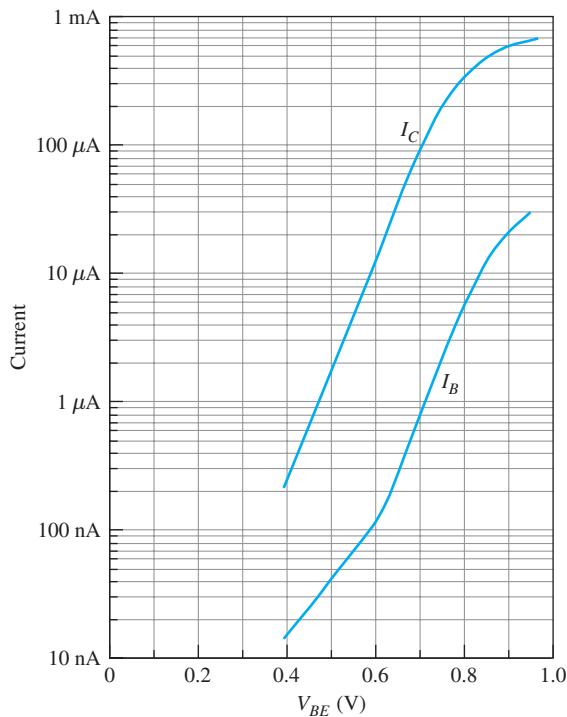


Figure PS4.3

Optoelectronic and Power Semiconductor Devices

One could argue that the two areas in which semiconductors have made the most profound changes in our lives are computing and communication. Increasingly, optical communication is becoming the fastest and in many cases the most cost-effective approach to transferring information from one place to another. The bandwidths that have turned the Internet and the World Wide Web into household tools (and playgrounds) are in large part due to optical communications.

In Chapter 11 we explore the role of semiconductors in photonics, examining some common optical devices such as photodetectors, light-emitting diodes, lasers, and imagers.

At the same time society increasingly depends on electrical and electro-mechanical devices. These devices need power. Cell phones need chargers. Washing machines need electric motors and the means to control them. Electric cars need to get their energy from the grid, and the grid itself needs to be controlled, monitored, and switched. Chapter 12 discusses a variety of semiconductor devices that can switch or control high voltages and high currents. ■

11

CHAPTER

Optoelectronic Devices

11.1 INTRODUCTION AND PREVIEW

To begin our discussion of optoelectronic devices, let us first consider the frequency spectrum, Figure 11.1, from the subaudio region, $f \approx 10$ Hz to the X-ray region, $\lambda \approx 1$ nm. The electromagnetic (EM) frequencies of interest begin with AM radio in the kHz range. For semiconductor optoelectronics, the ultraviolet, visible and near-infrared regions are of primary importance. The photon wavelengths and energies in the visible region are shown on an expanded scale, and the two most important wavelengths for fiber optics are also indicated.

The success of optical communication is not due just to the advances in optical fibers, but also to the concurrent development of the optical sources (laser diodes and light-emitting diodes, LEDs) and photodetectors that are fast, cheap, and reliable. Additionally, LEDs are becoming increasingly important not just for displays but also for lighting. These devices are the subjects of this chapter.

11.2 PHOTODETECTORS

The term *photodetector* usually describes a diode (or transistor) that is used to measure the amount of light energy present. When we say “measure” we imply that the output signal should be a function of the light intensity.¹ For example, a fiber-optic receiver contains a photodetector whose output current is proportional to the detected light intensity. The rest of the receiver contains amplification and decision circuitry.

The case of a generic photodetector is considered first. Once the design trade-offs are understood, some specific photodetector types are considered.

11.2.1 GENERIC PHOTODETECTOR

Figure 11.2 shows a generic photodiode structure. It contains a pn junction and is illuminated from the top. Photons with energy greater than the semiconductor

¹Actually, the term *intensity* is used loosely here. Technically, intensity is in the nonintuitive units of power per solid angle, and is usually used to describe light emitters. For photodetectors, the actual quantity we need is called *irradiance* (measured in W/m^2).

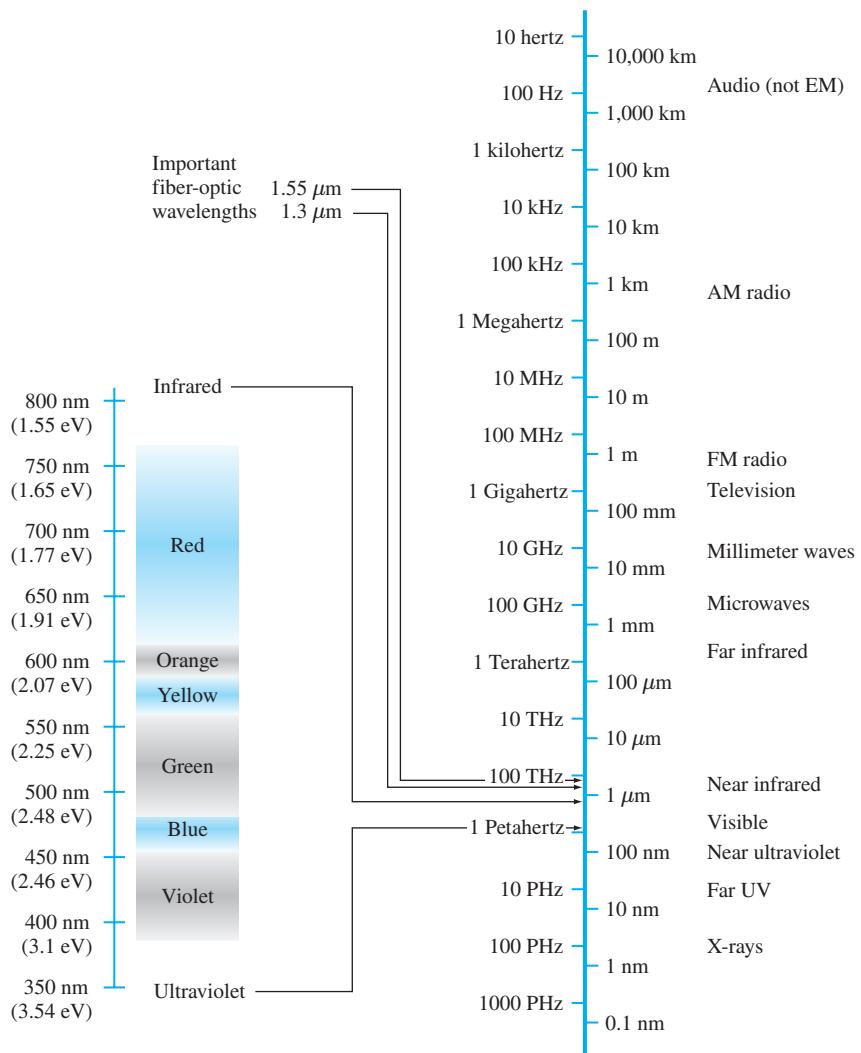


Figure 11.1 The electromagnetic spectrum. (The frequencies of acoustic waves are given for comparison although they are not electromagnetic waves.)

band gap create electron-hole pairs. Electrons and holes are separated by the field in the junction depletion region and flow through top and bottom contacts to an external circuit.

It is important to not block the light with the top contact, so contact is often in the form of a ring. To get to the contact from the n^+ region, the current flows laterally in the top layer; therefore that layer's sheet resistance should be small. A small sheet resistance can result from a thick layer or from heavy doping, or a combination. In practice, however, it is required that little light be absorbed in this layer, so it is made thin (a fraction of a micrometer) and thus it must be heavily doped.

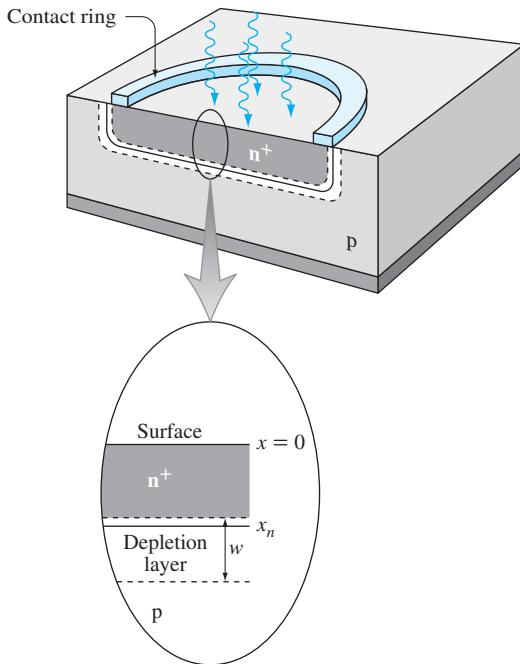


Figure 11.2 A generic photodiode.

As the light penetrates into the semiconductor, it will be absorbed. If the photon flux at a given distance into the material is $F_L(x)$, then the variation of F_L with x within the semiconductor is proportional to $F_L(x)$:

$$\frac{dF_L(x)}{dx} = -\alpha F_L(x) \quad (11.1)$$

where α is the *absorption coefficient*. This can be solved to find

$$F_L(x) = F_L(0)e^{-\alpha x} \quad (11.2)$$

The absorption coefficient α then gives the distance at which the flux is reduced to $1/e$ (37 percent) of its surface value.

When the light is incident on the surface, however, there is a partial reflection. This *Fresnel reflection* occurs when light is incident on an interface between any two materials of different refractive indices. For normal incidence, the reflection coefficient R , or the fraction of incident photons (or fraction of the power) reflected at the interface between material 1 and material 2, is²

$$R = \frac{F_{L\text{reflected}}}{F_{L_i}} = \left(\frac{n_1 - n_2}{n_1 + n_2} \right)^2 \quad (11.3)$$

²Compare this expression, which gives the probability that a given photon is reflected at a step index change between two materials, with the probability that an electron is reflected from a step potential change between two materials. Ultimately the mathematics in deriving these two is the same.

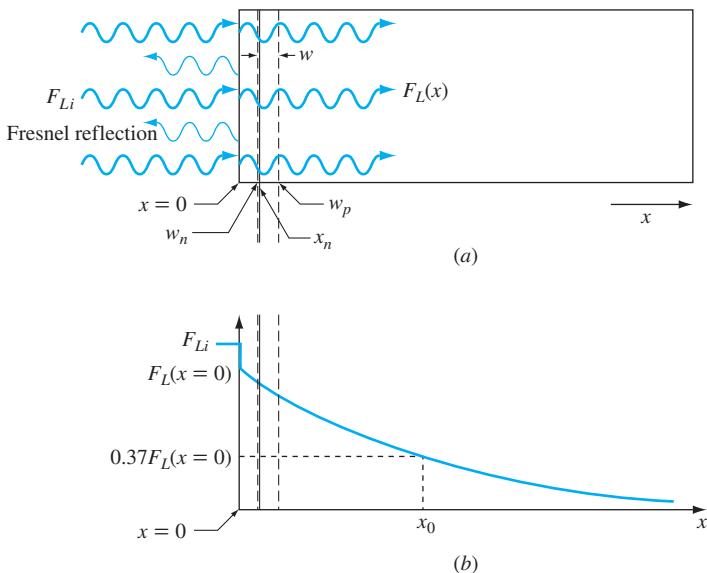


Figure 11.3 Variation of photon flux with distance. (a) A physical diagram showing the sample and the depletion region; (b) a plot of the the flux as a function of distance. There is a loss due to Fresnel reflection at the surface, followed by the decaying exponential loss due to absorption. The photon penetration depth x_0 is defined as the depth at which the photon flux is reduced to e^{-1} of its surface value.

where F_{Li} is the incident photon flux, and the n 's are the refractive indices of the materials. The transmission coefficient T at the surface is

$$T = 1 - R \frac{4n_1 n_2}{(n_1 + n_2)^2} \quad (11.4)$$

Figure 11.3 shows the light incident on the semiconductor and the variation of the optical flux as a function of depth into the material.

EXAMPLE 11.1

Find the variation in light flux with depth in a Si n^+ p junction illuminated with monochromatic light of 1.42 eV (near infrared, see Figure 11.1), such as might be emitted by a GaAs light-emitting diode.

Solution

If the incident flux is F_{Li} , then the flux at the surface of the semiconductor is

$$F_L(0) = (1 - R)F_{Li} \quad (11.5)$$

The refractive index of Si is a function of photon energy, but around the visible region, it is about $n_{Si} = 3.6$. The reflection coefficient between air ($n = 1$) and Si then is, from

Equation (11.3),

$$R = \frac{(n_{\text{air}} - n_{\text{Si}})^2}{(n_{\text{air}} + n_{\text{Si}})^2} = \frac{(1 - 3.6)^2}{(1 + 3.6)^2} = 0.32$$

or about 32 percent of the incident photons are reflected. The Fresnel reflection at the surface of the semiconductor can thus result in a considerable loss of photons. To reduce the surface reflection in practice, thin layers of transparent dielectrics having refractive indices intermediate between those of air and the semiconductor are deposited on the semiconductor surface.

For our uncoated example, the flux that penetrates into the silicon is

$$F_L(0) = (1 - 0.32)F_{Li} = 0.68 F_{Li}$$

Now we can compute the absorption in the silicon. Figure 11.4 shows absorption coefficients as a function of photon energy for several semiconductors. At this photon

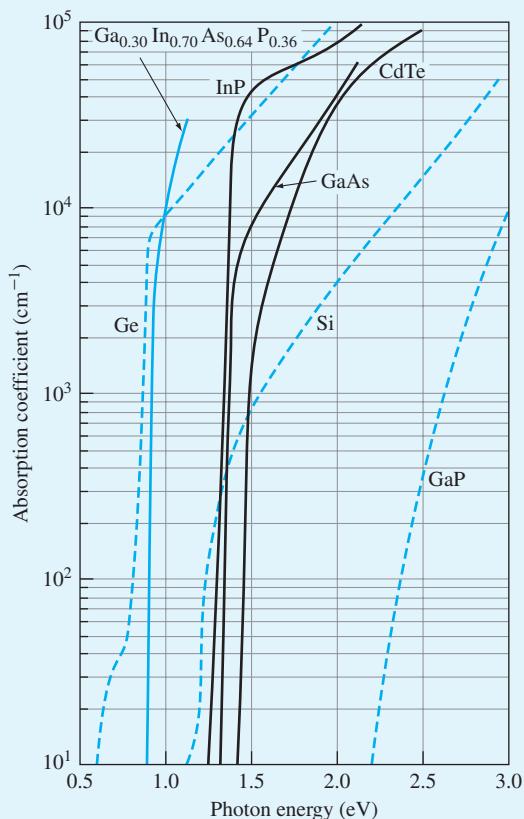


Figure 11.4 Absorption coefficients of some semiconductor materials. The indirect-gap materials are shown with a broken line. Based on data from [1, 2].

energy, the absorption coefficient α in Si is on the order of $4 \times 10^2 \text{ cm}^{-1} \approx 0.04 \mu\text{m}^{-1}$. From Equation (11.2), then, the light is reduced to 37 percent of the surface value at a depth x_0 of

$$\begin{aligned}\frac{F_L(x)}{F_L(0)} &= 0.37 = e^{-\alpha x_0} \\ \ln(0.37) &= -1 = -\alpha x_0 \\ x_0 &= \frac{1}{0.04 \mu\text{m}^{-1}} = 25 \mu\text{m}\end{aligned}$$

Let us assume that each photon absorbed creates an electron-hole pair. If the electron and hole are generated deep within a quasi-neutral region where there is no electric field, they will diffuse in random directions and eventually recombine, producing no net current. To obtain a photocurrent in a pn junction, minority carriers created in the quasi-neutral region must diffuse to the junction. Once at the junction, the electric field will accelerate the carriers across the junction, producing current.

EXAMPLE 11.2

Compare the absorption depth with the minority carrier diffusion length for the n⁺p device of Example 11.1. Can the carriers diffuse to the junction to create a photocurrent? Assume $V_a = 0$, the n⁺ region is 0.3 μm thick, and the p region doping is 10^{17} cm^{-3} .

Solution

The n⁺ layer thickness (0.3 μm) is much less than the photon penetration depth (25 μm). The absorption in this region, and thus its contribution to the photocurrent, can be ignored. For $V_a = 0$, the junction width is 0.11 μm, and this region also contributes negligibly to photocurrent.

In the quasi-neutral p region, the minority carrier (electron) diffusion length at this doping level is 110 μm. This is appreciably greater than the photon penetration depth, thus a large fraction of the optically produced electrons diffuse back to the junction where they are collected and contribute to current.

The rate at which the electron-hole pairs are generated, G_L , is proportional to the photon flux density, $G_L(x) = \alpha F_L(x)$. This is still assuming every photon generates an electron-hole pair. We can find the electron concentration in the steady state by solving the continuity equation for electrons in the p region. In this case we have an optical generation term and a recombination term for the optically produced excess carriers:

$$\frac{1}{q} \frac{dJ_n(x)}{dx} + G_L(x) - \frac{\Delta n_p}{\tau_n} = 0 \quad (11.6)$$

where J_n is the photocurrent and Δn_p is the excess (photogenerated) electron concentration. Since the p region is uniformly doped, there is no field ($\mathcal{E} = 0$) and the total electron current is due to diffusion:

$$J_n = q D_n \frac{dn}{dx} \quad (11.7)$$

Recalling that $n = n_0 + \Delta n$, and that n_0 is a constant, Equation (11.6) becomes

$$D_n \frac{d^2 \Delta n(x)}{dx^2} + \alpha(1 - R)F_{Li}e^{-\alpha x} - \frac{\Delta n(x)}{\tau_n} = 0 \quad (11.8)$$

Solving Equation (11.8) for $\Delta n(x)$, the resultant photocurrent can be obtained by evaluating Equation (11.7) at a convenient location such as $x = 0$. This is treated in more detail in the section on solar cells.

In the preceding discussion, the contributions to the photocurrent from absorption in the transition region and in the n^+ surface were neglected, since for this example these regions are thin compared with the total penetration depth. For higher absorption coefficients, the penetration depth is reduced and absorption in these regions must be considered. Let us consider the absorption in the junction first.

In the depletion region the electric field is high enough that virtually all of the carriers generated there are accelerated out of the region before they can recombine, so all contribute to photocurrent. We can find this by integrating the absorption over the junction width:

$$J_D = q\alpha \int_{x_n}^{x_n+w} F_L(x)dx = qF_{Li}(1 - R)e^{-\alpha x_n} [1 - e^{-\alpha w}] \quad (11.9)$$

where J_D is the photocurrent produced in the depletion region and x_n and w are defined in Figure 11.2.

For high absorption coefficients α , the light is absorbed closer to the surface, and the n^+ surface layer can contribute to photocurrent also. Surface states, however, cause the photogenerated carriers in this thin region to have a high probability of recombining at the surface, and thus the contribution of this region to the photocurrent is low.

The locations at which the carriers are produced also affect the response time. Because of the high field in the depletion region, the response time is much faster for carriers generated there than for the carriers generated in the quasi-neutral regions. The carriers in the quasi-neutral regions must diffuse (a slow process) to the junction to contribute to current. The current contributed by optical generation in the depletion region is sometimes referred to as the *prompt photocurrent*.

Next let us consider the effect of photocurrent on the $I-V_a$ characteristics of the photodiode. The total current is the sum of the dark current and the photocurrent I_L :

$$I = I_{\text{dark}} + I_L \quad (11.10)$$

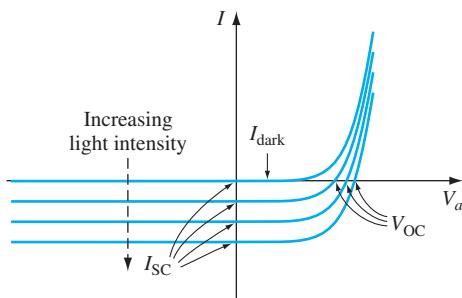


Figure 11.5 The I - V_a characteristics of a solar cell with varying illumination as a parameter. As the intensity increases, the short-circuit current I_{SC} increases linearly, but the open-circuit voltage V_{OC} increases sublinearly.

Neglecting the voltage drop across the series resistance, the junction voltage is $V_j = V_{bi} - V_a$. For the cases in which the surface layer and the depletion region contribute a negligible photocurrent, I_L is independent of junction voltage, meaning that the effect of illumination is to translate the dark I - V_a characteristics in the $-I$ direction by the amount I_L . (The photocurrent is in the same direction as the reverse current.) This is illustrated in Figure 11.5. The short-circuit photocurrent I_{SC} (at $V_a = 0$) and the open-current voltage V_{OC} (at $I = 0$) are also shown. While I_{SC} is directly proportional to light intensity, the open-circuit voltage V_{OC} increases sublinearly (logarithmically) with intensity.

The dark current is the diode current discussed in Chapter 5; it just has a different name when discussing photodiodes. It is the sum of injection current and generation-recombination current, and is given by

$$I_{dark} = I_0 (e^{qV_a/nkT} - 1) \quad (11.11)$$

Setting $I = 0$ in Equation (11.10), with the aid of Equation (11.11), the open-circuit voltage V_{OC} is

$$V_{OC} = \frac{nkT}{q} \ln \left(1 + \left| \frac{I_L}{I_0} \right| \right) \quad (11.12)$$

As we will see in the next section, for solar cells it is advantageous to make V_{OC} as large as possible. From Equation (11.12) it appears that a large n would be desirable to obtain a large V_{OC} . Just the opposite is the case, however, because a large I_0 is associated with a large n , and in actuality V_{OC} decreases with increasing n . Further, the reverse dark current I_0 reduces the signal-to-noise ratio in the detection process and must be minimized.

Two figures of merit for photodetectors are *quantum efficiency* η_Q and *responsivity* R_{ph} . The quantum efficiency is defined as the photoinduced carrier flux density J_L/q passing the junction per incident photon flux density F_{Lb} , or

$$\eta_Q = \frac{J_L/q}{F_{Li}} \quad (11.13)$$

The responsivity is defined as the output current density per watt of incident optical power per unit area. The energy per photon is $h\nu$ and so

$$R_{ph} = \frac{J_L}{h\nu F_{Li}} = \frac{q\eta_Q}{h\nu} \quad (11.14)$$

EXAMPLE 11.3

What is the quantum efficiency and responsivity for the prompt response of an InGaAs pn photodiode whose junction is $0.2 \mu\text{m}$ below the surface, and whose depletion layer is $w = 2 \mu\text{m}$ thick at a reverse bias of 10 V? The incident light has a wavelength of $1.55 \mu\text{m}$. At this wavelength the absorption coefficient of this material is about 10^4 cm^{-1} , the refractive index is 3.4, and no antireflection coating is used.

■ Solution

The Fresnel reflection loss in going from air to the semiconductor is

$$R = \left(\frac{n_{air} - n_{semi}}{n_{air} + n_{semi}} \right)^2 = \left(\frac{1.0 - 3.4}{1.0 + 3.4} \right)^2 = 0.30$$

Of this, $(1 - R) = 0.7$ of the incident power remains. The photon flux density remaining after absorption in the surface layer is

$$F_L(x_n) = (1 - R)F_{Li}e^{-\alpha x_n} = 0.7F_{Li}e^{-(10^4 \text{ cm}^{-1})(2 \times 10^{-5} \text{ cm})} = 0.573 F_{Li}$$

Of this, a fraction,

$$1 - e^{-\alpha w} = 1 - e^{-(10^4 \text{ cm}^{-1})(2 \times 10^{-4} \text{ cm})} = 1 - 0.135 = 0.865$$

is absorbed in the depletion region. Thus the total quantum efficiency for the prompt response is

$$\eta_Q = (1 - R)e^{-\alpha x_n}(1 - e^{-\alpha w}) = 0.573 \times 0.865 = 0.495$$

The corresponding responsivity is

$$R_{ph} = \frac{q}{h\nu}\eta_Q = \frac{q\lambda}{hc}\eta_Q = \frac{(1.6 \times 10^{-19} \text{ C})(1.55 \times 10^{-6} \text{ m})}{(6.62 \times 10^{-34} \text{ J} \cdot \text{s})(3 \times 10^8 \text{ m/s})}(0.495) = 0.62 \text{ A/W}$$

*11.2.2 SOLAR CELLS

Solar cells are photodetectors that are used to generate dc power. [1] As such, they are of reasonably large area, on the order of several square centimeters. To reduce the series resistance due to lateral current flow in the thin surface

layer, digitated metal contacts are used on the illuminated side as indicated in Figure 11.6. Power is generated in the cell and is dissipated in a load; in this case the load resistance is R_L .

The device's $I-V_a$ characteristics and its load line for a given incident light intensity are shown in Figure 11.7. The load line has a slope of $-1/R_L$, and since the applied voltage is zero (the point is for the solar cell to be a power source, not to require one), the load line goes through the origin. Choosing the load to give maximum power for a given illumination, the operating point is (I_m, V_m) as shown in the figure. Note that the power dissipated by a device is $P = IV$, but in

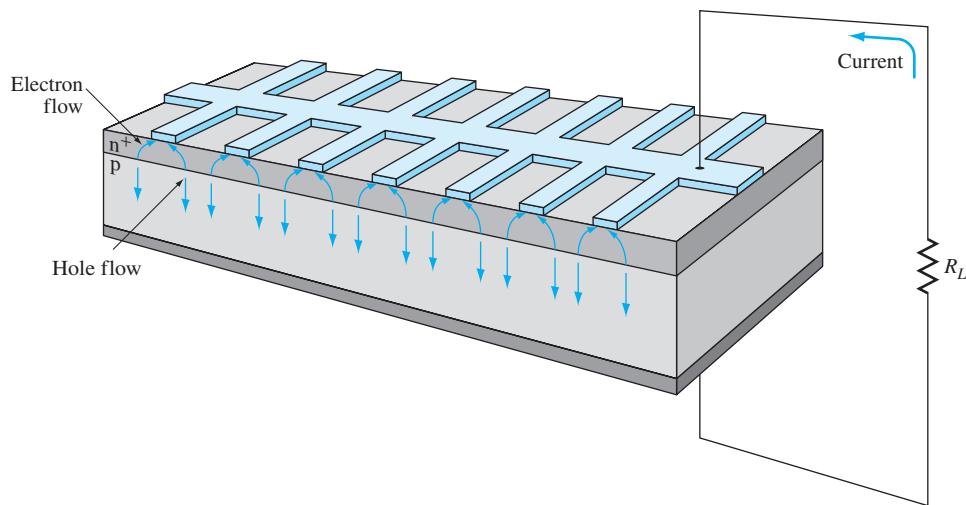


Figure 11.6 A solar cell with digitated contacts. The load resistor is R_L .

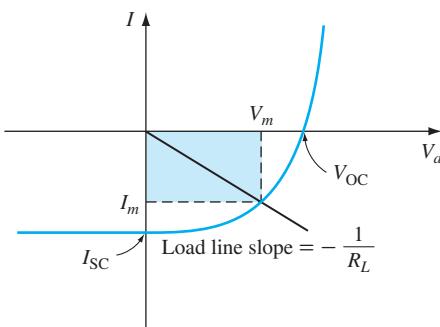


Figure 11.7 The $I-V_a$ characteristic of a solar cell. The maximum power is obtained at $P_m = I_m V_m$.

this case V is positive and I is negative. Therefore the power *dissipated* by the solar cell is negative, meaning it produces power.³

The maximum power output P_m is

$$P_m = I_m V_m \quad (11.15)$$

and is represented by the shaded area of Figure 11.7.

The power conversion efficiency η of a solar cell is defined as the maximum output electrical power divided by the incident optical power P_{Li} :

$$\eta = \frac{P_m}{P_{Li}} \times 100 \quad \text{percent} \quad (11.16)$$

A parameter called the *fill factor*, FF, measures how well the shaded box in Figure 11.7 fills the quadrant IV portion of the I - V characteristic. The fill factor is defined as

$$\text{FF} = \frac{I_m V_m}{I_{SC} V_{OC}} \quad (11.17)$$

A typical value of FF is on the order of 0.7. The efficiency can be expressed in terms of the fill factor by

$$\eta = \text{FF} \frac{I_{SC} V_{OC}}{P_{Li}} \quad (11.18)$$

Note that the power conversion efficiency is η and the quantum efficiency η_Q is given by Equation (11.13).

The power conversion efficiency of a solar cell depends on a number of parameters that affect I_{SC} and V_{OC} . One of these is the spectrum of the output from the sun, since the spectrum affects the absorption. The solar spectrum is shown in Figure 11.8 for AM0 (air mass zero) and AM1 (air mass one). Air mass zero is the radiant energy outside the earth's atmosphere, as seen by orbiting satellites. Passing through the atmosphere alters the spectrum, and AM1 refers to the solar spectrum that has passed through one atmosphere at sea level for the sun directly overhead. The difference in the two curves results from scattering of the incident light and absorption in the earth's atmosphere.

For maximum conversion efficiency, the band gap of the material, the absorption coefficient spectrum, and the minority carrier lifetime are all important considerations. The band gap matters because those incident photons with energy less than the band gap are not absorbed and thus cannot contribute to photocurrent. The absorption spectrum affects the probability that a photon will create an electron-hole pair, and the minority carrier lifetime controls the diffusion

³Recall that all the other I - V characteristics in this book appear in only the first and third quadrants, where the I - V product (dissipated power) is positive.

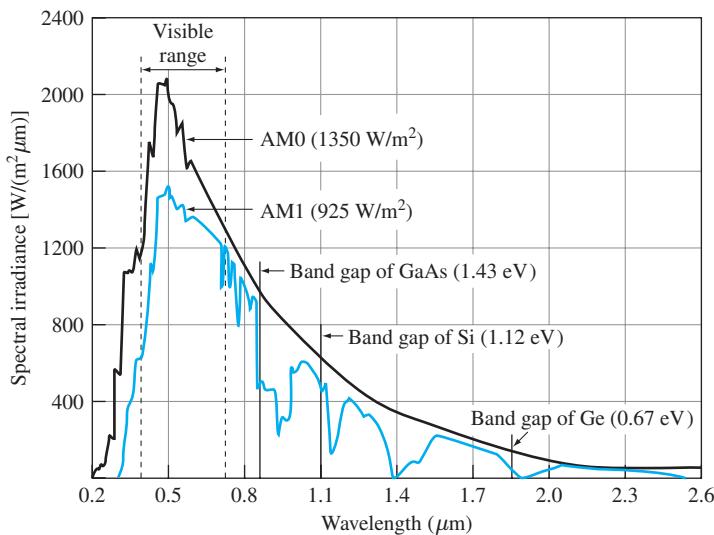


Figure 11.8 The solar spectrum. On earth at sea level, normal incidence, the spectrum is AM1 (one atmosphere). Satellites and other objects outside the atmosphere are exposed to AM0.

length, and thus the probability of collection of the optically generated minority carriers.

In silicon solar cells, about 20 percent of the incident solar power is lost because that much of the solar spectrum power consists of photons with energy below the band gap. Those photons with energy larger than the band gap are absorbed, but one photon creates only one electron-hole pair. The electron relaxes to the bottom of the conduction band, and the hole relaxes to the top of the valence band. Thus, a fraction of the photon energy greater than the band gap is lost as heat (phonons). This causes another 40 percent of the incident energy to be lost in Si devices. Thus for Si cells the maximum theoretical conversion efficiency η_{\max} is on the order of

$$\eta_{\max} \approx 100 - 20 - 40 \approx 40\%$$

Actual conversion efficiencies are approximately half this value.

Consider a Si n⁺p solar cell as shown schematically in Figure 11.9. Let $x = 0$ at the metallurgical junction (we have moved this reference point from the surface, where it was earlier). The width of the p region, W_p' , is made much larger than L_n , the electron diffusion length. We ignore the photocurrent produced in the thin n⁺ region. We wish to calculate the short-circuit current produced by photon absorption in the p region for monochromatic light with absorption coefficient α .

The photon flux density (number of photons entering the p region per unit area per second) is $F_L(1 - R)e^{-\alpha x_n}$. Then at position x within the p region, the photon flux density is $F_L(1 - R)e^{-\alpha x_n}e^{-\alpha x}$ and Equation (11.8) becomes

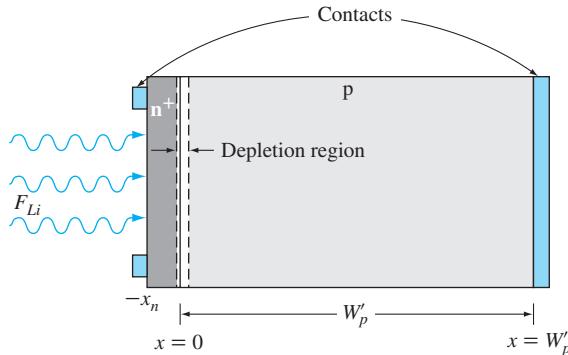


Figure 11.9 A solar cell illuminated from the left.

$$D_n \frac{d^2 \Delta n(x)}{dx^2} + \alpha F_{Li} (1 - R) e^{-\alpha x_n} e^{-\alpha x} - \frac{\Delta n}{\tau_n} = 0 \quad (11.19)$$

The solution to this equation is

$$\Delta n(x) = C_1 e^{x/L_n} + C_2 e^{-x/L_n} - \frac{\alpha \tau_n F_{Li} (1 - R) e^{-\alpha x_n} e^{-\alpha x}}{\alpha^2 L_n^2 - 1} \quad (11.20)$$

where C_1 and C_2 are constants. To find these, we note that since we are calculating short-circuit current, $V_a = 0$. Also, at $x = 0$, the junction, there is a depletion region, so

$$\Delta n(0) = 0$$

If the silicon is thick enough that $W'_p \gg L_n$, we can approximate $W'_p = \infty$ and

$$\Delta n(W'_p) = 0$$

With these boundary conditions, then,

$$C_1 = 0$$

and

$$C_2 = \frac{\alpha \tau_n F_{Li} (1 - R) e^{-\alpha x_n}}{\alpha^2 L_n^2 - 1}$$

Then the excess carrier distribution with depth is

$$\Delta n(x) = \frac{\alpha \tau_n F_{Li} (1 - R) e^{-\alpha x_n}}{(\alpha^2 L_n^2 - 1)} (e^{-x/L_n} - e^{-\alpha x}) \quad (11.21)$$

We can now find the current density crossing the junction from

$$J_n = q D_n \left. \frac{d \Delta n}{dx} \right|_{x=0} \quad (11.22)$$

Taking the derivative of Equation (11.21), we obtain

$$J_n = q D_n \tau_n \alpha F_{Li} (1 - R) e^{-\alpha x_n} \frac{\alpha - \frac{1}{L_n}}{(\alpha^2 L_n^2 - 1)} \quad (11.23)$$

Since $D_n \tau_n = L_n^2$ and $(\alpha^2 L_n^2 - 1) = (\alpha L_n + 1)(\alpha L_n - 1)$, Equation (11.23) becomes

$$J_n = \frac{q \alpha L_n (1 - R) F_{Li} e^{-\alpha x_n}}{(\alpha L_n + 1)} \quad (11.24)$$

EXAMPLE 11.4

Find the quantum efficiency of a Si solar cell for $\lambda = 1 \mu\text{m}$ and near the peak solar energy at $\lambda = 0.5 \mu\text{m}$. Ignore the photocurrent contribution from the n^+ and depletion regions. The cell parameters are:

$$R = 0.2$$

$$x_n = 0.4 \mu\text{m}$$

$$W'_p = 500 \mu\text{m}$$

$$N'_A = N_A = 10^{17} \text{ cm}^{-3}$$

Solution

We begin with Equation (11.13), using J_n for the photoinduced current J_L :

$$\eta_Q = \frac{J_n}{q F_{Li}}$$

Substituting for J_n from Equation (11.24), we obtain

$$\eta_Q = \frac{\alpha L_n (1 - R) e^{-\alpha x_n}}{(\alpha L_n + 1)} \quad (11.25)$$

We must determine α and L_n . We can use Figure 11.4 to find the absorption coefficients, but we need to express our wavelengths in terms of the photon energies. From the expression

$$E_{\text{ph}}(\text{eV}) \lambda(\mu\text{m}) = 1.24$$

Then

$$E_{\text{ph}}(1) = \frac{1.24}{1} = 1.24 \text{ eV} \quad (\text{at } \lambda = 1 \mu\text{m})$$

$$E_{\text{ph}}(0.5) = \frac{1.24}{0.5} = 2.48 \text{ eV} \quad (\text{at } \lambda = 0.5 \mu\text{m})$$

From Figure 11.4, the absorption coefficients in Si at these wavelengths are

$$\alpha(1) = 100 \text{ cm}^{-1} = 10^{-2} \mu\text{m}^{-1}$$

$$\alpha(0.5) = 10^4 \text{ cm}^{-1} = 1 \mu\text{m}^{-1}$$

The electron (minority carrier) diffusion length at this doping level is

$$L_n = 73 \mu\text{m}$$

Since W'_p is appreciably larger than L_n , Equations (11.24) and (11.25) are reasonable approximations.

Then η_Q becomes, from Equation (11.25),

$$\eta_Q(1) = \frac{10^{-2} \times 73 \times (1 - 0.2)e^{-10^{-2} \times 0.4}}{(10^{-2} \times 73) + 1} = \frac{0.73 \times 0.8 \times 0.996}{1.73} = 34\%$$

$$\eta_Q(0.5) = \frac{1 \times 73 \times (1 - 0.2)e^{-1 \times 0.4}}{(1 \times 73) + 1} = \frac{73 \times 0.8 \times 0.67}{74} = 53\%$$

We considered two specific wavelengths in Example 11.4, but to obtain the total quantum efficiency of the cell, one would perform a weighted average of η_Q over the solar spectrum.

The power efficiency η of the cell will be less than the quantum efficiency η_Q . This is because the excess photon energy (greater than that of the band gap) does not contribute to the photocurrent.

To design good solar cells, then, we observe that the quantum efficiency is dependent on the absorption coefficient and the minority carrier diffusion length. The diffusion coefficient should be as large as possible to maximize the collection of photogenerated carriers.

We also recall that the optical penetration depth is equal to the reciprocal of the absorption coefficient. A small value of α results in deep penetration and thus requires large diffusion lengths. A large α , on the other hand, results in absorption in the surface n^+ layer, but as we indicated in the previous section, many of the carriers generated here recombine at the surface; they do not contribute to photocurrent. For the two wavelengths considered in the example, the optical penetration distance is 100 μm (large) for $\lambda = 1 \mu\text{m}$ and 1 μm (short) for $\lambda = 0.5 \mu\text{m}$. A fraction $(1 - e^{-\alpha x_n})$ of the photons that enter the cell do not reach the p-type region. This amounts to 0.4 percent loss at $\lambda = 1 \mu\text{m}$, but a significant 33 percent at $\lambda = 0.5 \mu\text{m}$ in the example.

11.2.3 THE pin (PIN) PHOTODETECTOR

In a solar cell, most of the optically generated current results from minority carrier diffusion to the pn junction. Since diffusion is a relatively slow process, the speed of response to changes in light is limited. While this is of no concern in solar cells where the illumination and thus the output is dc, it severely limits the frequency response of pn photodetectors used in optical communication applications. One method to increase the speed for such an application is to use a pin (often referred to as PIN) detector.

The pin diode has a layer of intrinsic (or lightly doped) material between the n- and p-type layers. This structure is designed to extend the physical length of

the depletion region to increase the collection of the optically produced electron-hole pairs in this region. Figure 11.10a shows the structure, and Figures 11.10b and c show the equilibrium and reverse bias energy band diagrams normal to the plane of the junction respectively.

At equilibrium, it appears as though there are two junctions and two depletion regions. When a reverse bias is applied, however, the energy band diagram looks like Figure 11.10c. We see that there is an electric field throughout the

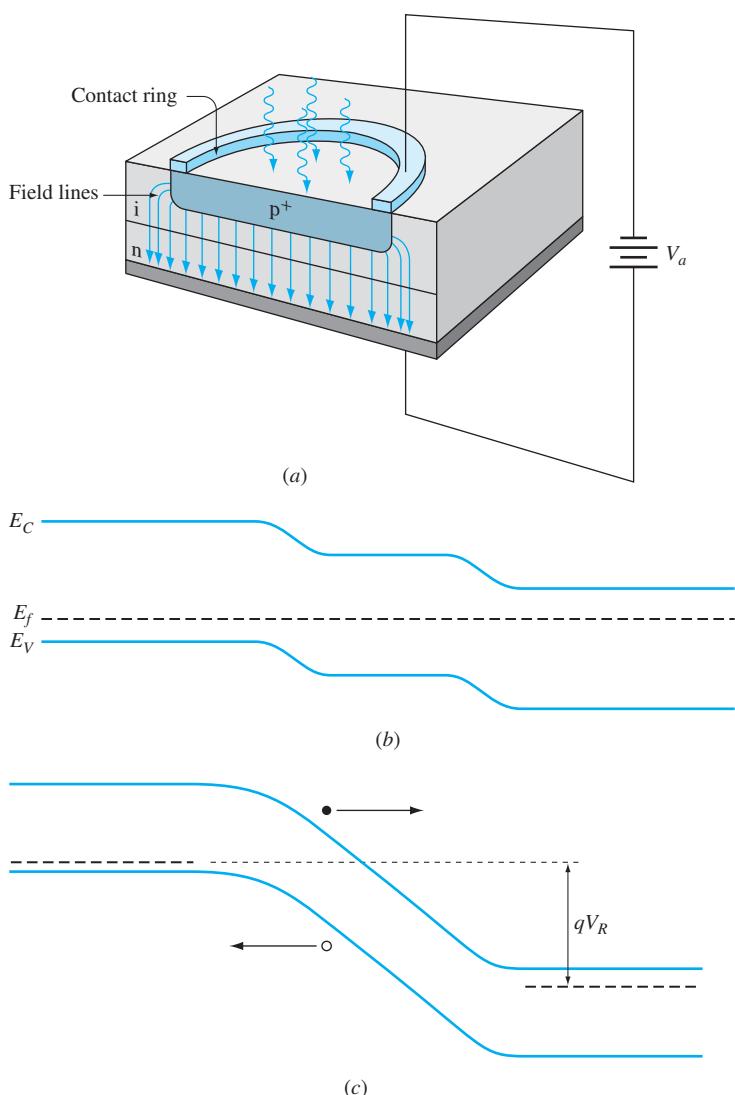


Figure 11.10 The pin diode. (a) The structure; (b) equilibrium energy band diagram; (c) energy band diagram under reverse bias.

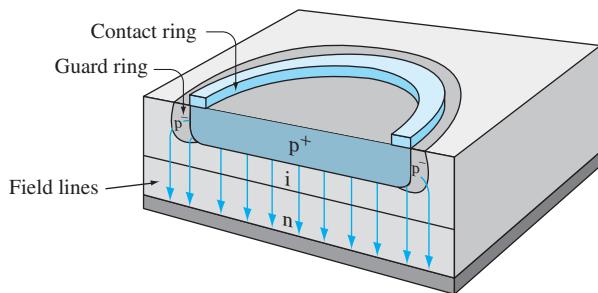


Figure 11.11 A pin structure will break down at the edge of the p^+ region where the electric field lines concentrate. A guard ring prevents premature breakdown.

entire intrinsic region. This is because the intrinsic region is inherently resistive as a result of the small number of carriers available for current transport. This high resistance means a large part of the voltage is dropped across this region. The n- and p-type regions are much more conductive, so field dropped in those regions is negligible.

In a pin photodetector, the surface layer (in this case the p^+ layer) is made thin enough that little optical absorption occurs there. The intrinsic layer is thick enough (greater than the photon penetration depth) that most of the optical absorption occurs in the depletion region. Therefore most of the photogenerated carriers contribute to the prompt response.

Notice, however, that in Figure 11.10a at the corners of the p region, the electric field lines are close together and thus the field is higher here. The pin structure of Figure 11.10a has a problem of breakdown in these corners. To permit a higher reverse voltage and increased speed, the field in these regions is reduced by using a lightly doped p-type *guard ring*, as shown in Figure 11.11. Because some of the depletion region exists in this lightly doped guard ring, the maximum field in this region is reduced.

11.2.4 AVALANCHE PHOTODIODES

Another commonly used type of photodiode in communications is the avalanche photodiode (APD). It uses some of the strategies of the pin diode. The major difference in an APD is that it has internal gain.

These devices are operated with reverse voltages on the order of 200 V, so there is a very high electric field in the depletion region. An electron excited to the conduction band by an incident photon experiences very rapid acceleration, resulting in impact ionization and carrier multiplication. Both electrons and holes are multiplied, producing an amplified response to the photon. Normally the voltage applied is slightly less in magnitude than the avalanche breakdown voltage, where the carrier multiplication factor M is large but the dark current

is not excessive. Gains on the order of 50 are common. The responsivity of an avalanche photodiode, then, is

$$R_{\text{ph}} = M \left(\frac{q \eta Q}{h\nu} \right) \quad (11.26)$$

where M is the multiplication factor.

Avalanche photodiodes are often used in telecommunications, especially in low-light situations, because of their high sensitivity. The avalanche process, however, does take some time, so APDs cannot achieve the high speeds that pins offer. Another trade-off with APDs is that carriers that are generated in the depletion region by thermal rather than by optical processes are also amplified, creating noise. Also, because the impact ionization process has some randomness to it, it creates extra noise.

11.3 LIGHT-EMITTING DIODES

In this section we reverse the optical process and consider emission instead of absorption. Light-emitting diodes (LEDs) are used for a variety of applications from displays to illumination to fiber-optic communication links.

11.3.1 SPONTANEOUS EMISSION IN A FORWARD-BIASED JUNCTION

LEDs operate by spontaneous emission. Electrons in the conduction band and holes in the valence band have finite lifetimes. When they recombine, the excess energy of the electron is released, either as light (a radiative transition), as phonons (nonradiative), or as a combination of the two.

To achieve significant emission, a situation is required where there are many electrons at elevated energy states (i.e., in the conduction band) and holes (for them to recombine with) present in the same physical area. A typical approach is to use a double-heterostructure pn junction, as shown in Figure 11.12a. This junction has a wide-band-gap p side and a wide-band-gap n side, with a narrow-band-gap material in between. The result is a potential well for electrons and another for holes. Under forward bias, excess electrons diffuse across the depletion region from the n side, and holes diffuse across in the other direction. The carriers tend to get caught and confined in the wells, increasing the probability of recombination.

For optical transitions in semiconductors, not only energy but also wave vector (K) must be conserved. Therefore, LEDs should be made of direct-gap materials, as shown in Figure 11.12b. These include gallium arsenide, indium phosphide, and many others. Ternary and quaternary compounds (having three and four components respectively) may also be direct gap. Figure 11.13 shows, for example, the band gap of $\text{GaAs}_x\text{P}_{1-x}$ as a function of the As concentration x . GaAs is a direct-gap material, but GaP is indirect. The $\text{GaAs}_x\text{P}_{1-x}$ alloy makes a transition from direct to indirect gap where $x = 0.55$. GaAsP can be used to make LEDs for wavelengths ranging from about 870 nm (infrared) corresponding to $x = 1$ to about 630 nm (red) for $x = 0.55$.

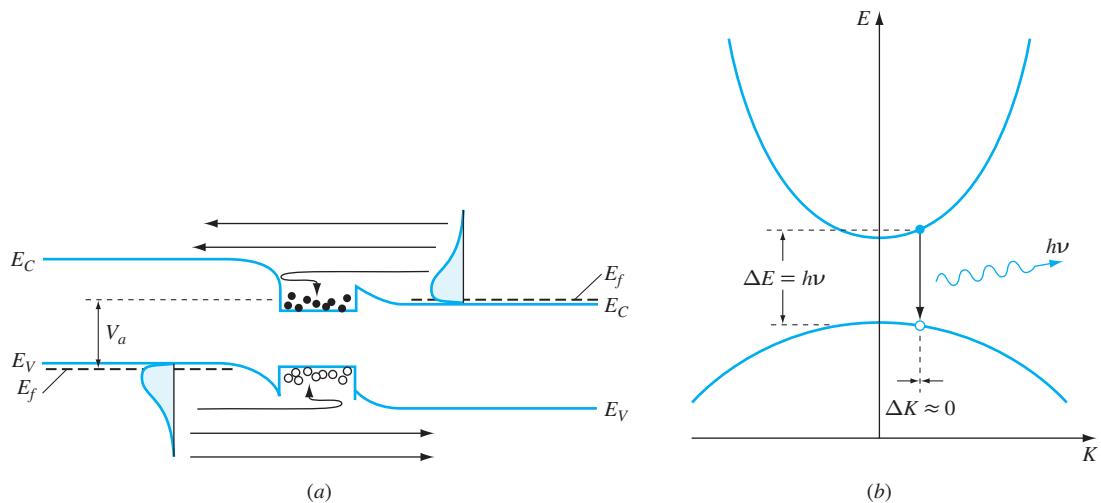


Figure 11.12 (a) A double-heterostructure LED pn diode. The potential wells for electrons and holes capture carriers and increase the probability of recombination. (b) The E - K diagram reminds us that K must also be conserved. Thus LEDs are usually made from direct-gap materials.

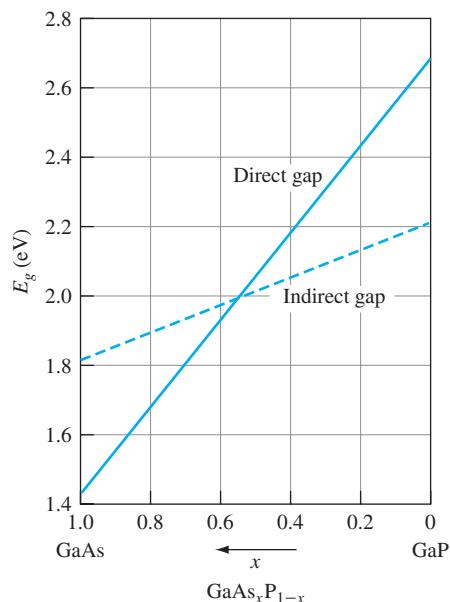


Figure 11.13 The band gap in the GaAsP system. For mole fractions of As less than about 0.55, the material is indirect gap.

Figure 11.14 shows a historical perspective of the development of semiconductor LEDs in the visible range. The first display/indicator LEDs were GaAsP ($\text{GaAs}_x\text{P}_{1-x}$), with efficiency lower than Thomas Edison's first light bulb. In the 1970s and 1980s improvements were made by using GaP doped first with zinc and oxygen, then later with nitrogen. These important LEDs exploit isoelectronic traps (discussed in online module OM11). Further improvements were obtained by going to quaternary compounds and heterojunctions (e.g., AlInGaP with a thin layer of GaAs or GaP). More recently, semiconductor nitrides have been used to make efficient blue and even ultraviolet LEDs. Green LEDs are difficult to make efficient even in the nitrides, leading to the so-called "green gap" in high-power LEDs for lighting, for example.

Another approach to making LEDs has been to use organic polymers (OLEDs). These use electrons and holes in a manner similar to semiconductors, and they have the additional advantage of being flexible, so thin, conformal displays can be envisioned. Currently OLEDs are used in some television sets, cell phones, and other portable devices because they consume relatively little power.

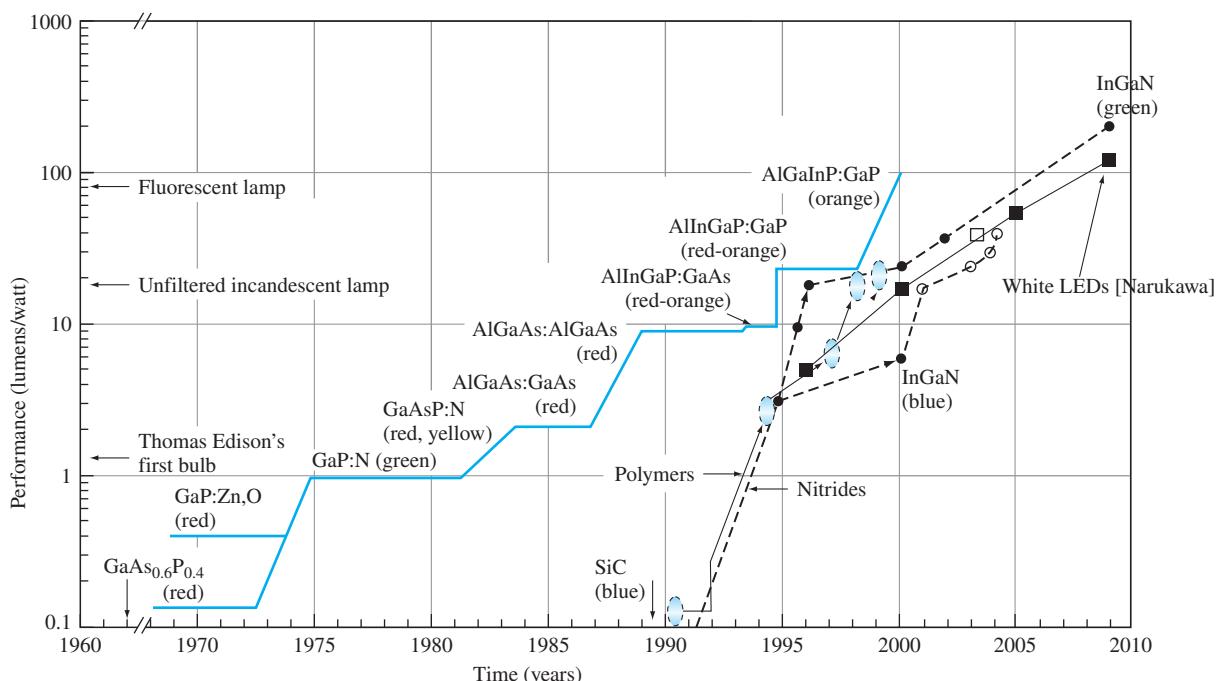


Figure 11.14 A historical view of the development of visible LEDs. Note that the white LEDs use an internal phosphor to produce white light; the LEDs themselves are typically blue, violet, or ultraviolet. (Sources of data: J. R. Sheats, et al., *Science*, 273, no. 5277, pp. 884–888, 1996, © 1996 AAAS, with updates from Yukio Narukawa, et al., "White light emitting diodes with super-high luminous efficacy." *J. Phys. D: Appl. Phys.* 43 (2010), article id. 354002, and data from the University of California Santa Barbara Solid State Lighting Center.)

They suffer, however, from fast degradation, which lowers their brightness over time. Organic semiconductors are beyond the scope of this book.

*11.3.2 BLUE, UTRAVIOLET, AND WHITE LEDs

For many years, LEDs were available only in green, yellow, orange, red, and infrared. The blue LED remained an elusive holy grail, because until blue LEDs existed, LED technology could not be used for color displays, which require red, green, and blue wavelengths to display a full range of colors.

In the late 1990s, however, gallium nitride of respectable quality became available. As a result, LEDs are now commercially available into the ultraviolet region. Incorporating indium into the GaN (InGaN) produces blue LEDs.⁴

All of the LEDs described so far emit only one particular color. White LEDs, which contain many wavelengths, are discussed in Section 11.3.4.

11.3.3 INFRARED LEDs

LEDs are not only used for displays. They are also frequently used as the optical source in fiber-optic systems. Figure 11.15a shows schematically a fiber-optic link coupling a light source (LED or laser) with a photodiode. Figure 11.15b

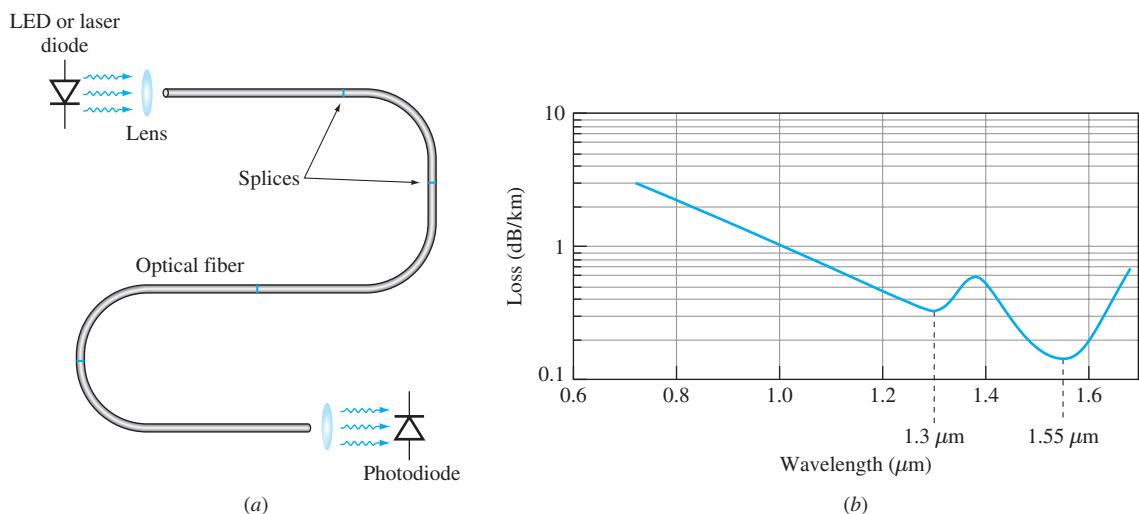


Figure 11.15 (a) A fiber-optic link contains a light source, generally either an LED or a laser diode, a fiber that may contain multiple splices, and a photodetector, usually either a pin diode or an avalanche photodiode. (b) The optical loss of glass used in optical fibers. Loss is least at the favored wavelengths of $1.3\text{ }\mu\text{m}$ and $1.55\text{ }\mu\text{m}$.

⁴Isamu Akasaki, Hiroshi Amano, and Shuji Nakamura received the Nobel Prize in Physics for their work in developing blue (InGaN) LEDs.

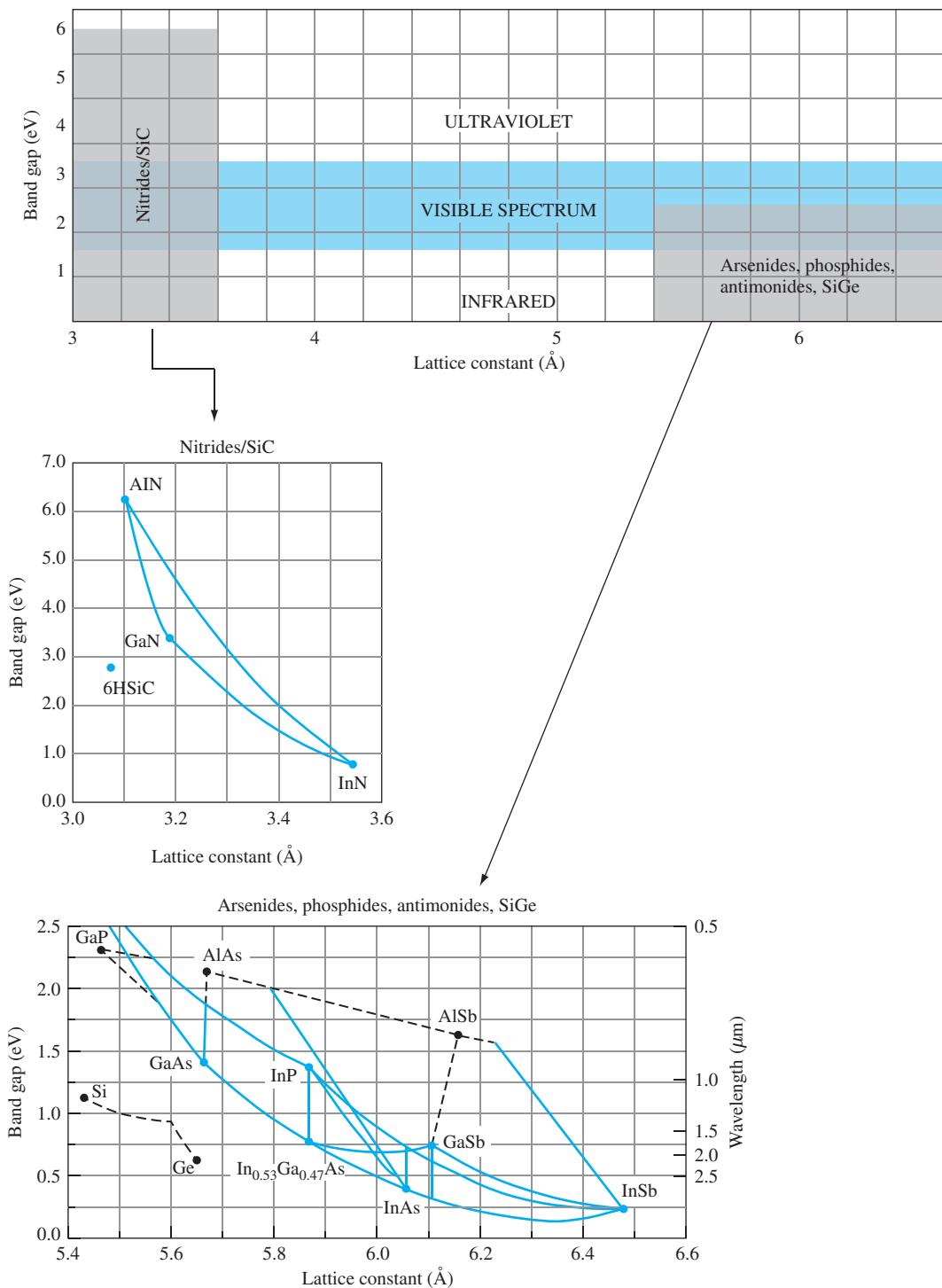


Figure 11.16 The lattice constants of several common semiconductors. The solid lines indicate direct-gap materials; the dashed lines are indirect-gap. (SiC is indirect.)

shows the absorption spectrum of the glass used in optical fibers.⁵ This shows that the best wavelengths to obtain low loss in fibers are $\lambda = 1.3 \mu\text{m}$ and $\lambda = 1.55 \mu\text{m}$, which are in the near infrared.

To determine what material systems would be best for fabricating sources at these wavelengths, we consult Figure 11.16, which shows the band gap and emission wavelengths for various compound semiconductors. It suggests materials such as $\text{In}_x\text{Ga}_{1-x}\text{As}$, $\text{InAs}_x\text{P}_{1-x}$, and some antimonides. The particular materials selected also depend on the lattice constants—the emitting materials must be grown onto a commonly available substrate.

EXAMPLE 11.5

Given that GaAs and InP are commonly available substrates, what material and composition should be used to produce an LED that emits at $1.3 \mu\text{m}$?

■ Solution

We require material whose lattice constant is equal to that of a good substrate, and at the same time has a band gap corresponding to $1.3 \mu\text{m}$:

$$E(\text{eV})\lambda(\mu\text{m}) = 1.24 \quad \text{or} \quad E_g = \frac{1.24}{1.3 \mu\text{m}} = 0.95 \text{ eV}$$

All the materials with the same lattice constant (Figure 11.16) as GaAs have band gaps that are larger than this, so GaAs is not an appropriate substrate.

For InP, however, there is a ternary compound of GaAs-InAs that is lattice matched. This occurs at $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$. It does not, however, emit at $1.3 \mu\text{m}$, but emits at $\lambda = 1.65 \mu\text{m}$. By increasing phosphorus content, though, one can move toward the InP point. The movement is not simply vertical, because creating the quaternary compound InGaAsP means some sort of interpolation between the GaAs-InAs curve and the InAs-InP curve. The final result turns out to be $\text{In}_{0.76}\text{Ga}_{0.24}\text{As}_{0.55}\text{P}_{0.45}$.

Let us now consider some of the characteristics of spontaneous emission. The emission is a random event, and the light reflects that randomness in the following ways: The direction of propagation of the photons is random, the timing of the emission (the phase of the photon) is random, and the polarization of the light is random.

What is not entirely random is the energy of the light. It is controlled by the energy distributions of the electrons and holes. Consider Figure 11.17. The minimum energy a released photon can have in a band-to-band transition is theoretically equal to the band-gap energy. In fact, this particular transition cannot occur, because it requires an electron at the very bottom edge of the conduction band

⁵The peak between 1.3 and $1.55 \mu\text{m}$ is due to water ions in the glass and has been essentially eliminated in modern fibers.

and a hole at the very top energy of the valence band. The density-of-states functions are zero at those two energies. The peak concentration of electron energies is $\frac{1}{2}kT$ above E_C , and for the holes $\frac{1}{2}kT$ below E_V . Thus the most probable emission energy is slightly higher than the band gap. Transitions above and below this peak value are also possible, just less probable. This means that an LED actually emits over a small range of wavelengths, typically about 50 to 100 nm wide. Figure 11.17b shows a typical spectrum of an LED.

Next, let us look at the physical structure of an LED. Figure 11.18 shows a double-heterostructure surface-emitting LED. We have already said that the emission occurs in the narrow-band-gap layer (refer to Figure 11.12), called the *active layer*. Since the direction in which the photon travels is completely random, the emission is uniformly distributed in all directions. Photons that are emitted upward or downward will enter the wider-band-gap material, which cannot absorb the emitted photons (their energy is too small to overcome the band gap). Thus these layers are transparent to the emitted light. Light that is emitted along the active layer will be reabsorbed.

Only light emitted through the surface will be used, but there are other sources of loss. One of these is Fresnel reflection, discussed earlier with respect to photodiodes, and the other is total internal reflection loss. Photons emitted at sufficiently low angle with respect to the surface (Figure 11.18) can be internally reflected according to Snell's law. The critical angle, measured with respect to the semiconductor surface, is

$$\theta_{\text{cr}} = \cos^{-1}\left(\frac{n_{\text{air}}}{n_{\text{semi}}}\right) \quad (11.27)$$

where the n 's are the refractive indices. Only photons striking the surface outside this angle will be transmitted across the surface and into the air.

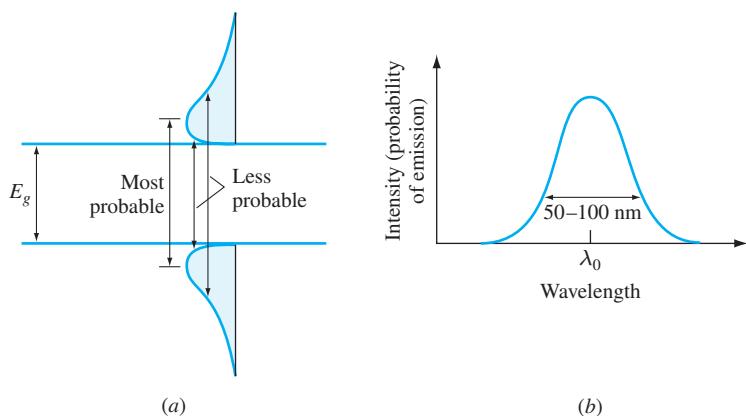


Figure 11.17 (a) The intensity of recombination at a given wavelength depends on the distributions of electrons and holes in energy; (b) the resulting emission spectrum.

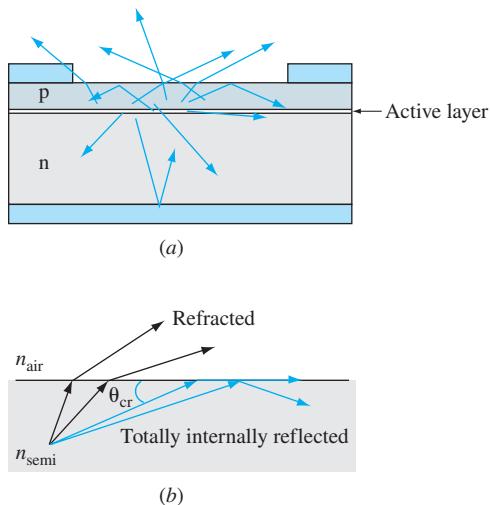


Figure 11.18 A generic surface-emitting LED. Some photons are lost by reabsorption in the bulk, and Fresnel reflection from the surface. (b) Light incident on the boundary at a sufficiently low angle can be totally internally reflected, resulting in additional loss.

Finally, since the light is emitted evenly in all directions, it is essential to bring the light-collection device (a lens or a fiber) as close as possible to the junction to capture as much light as possible. Figure 11.19 shows a *Burrus* LED structure, in which a well is etched in the LED surface and an optical fiber is inserted into the well and epoxied in place.

In general, optical coupling from LEDs to a fiber results in light being lost, but the Burrus structure helps appreciably.

Another solution for coupling the light to an optical fiber is to use an edge-emitting diode. It so happens that the smaller-band-gap materials tend to have higher refractive indices. From Equation (11.27), we see that when light goes from a higher to a lower index, it may be totally internally reflected. Figure 11.20 shows how a thin, high-index layer (the active layer) can also be a waveguide. Those photons that are traveling at shallow enough angles will tend to be reflected at either edge of the active layer.

In this structure, the light leaves the semiconductor from the edge of the chip rather than its surface. This has advantages and disadvantages. The light is emitted from a relatively small region, making it easier to couple the light into a fiber. On the other hand, to get access to the edge, the chip must be accurately sawn or cleaved away from the rest of the wafer, and the tiny devices are difficult to handle.

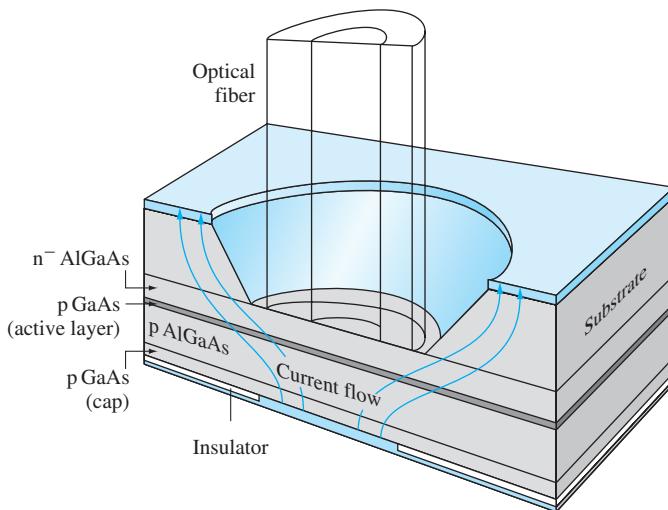


Figure 11.19 A Burrus-type LED. This one uses a double heterostructure to confine the carriers, making recombination more efficient. The etched opening in the LED helps align and couple an optical fiber.

One would expect from Figure 11.20 that most of the light in an edge-emitting LED would be reabsorbed, since it is confined to the emitting layer. That is not, in fact, the case. To understand why we must take a closer look at waveguides.

Consider the optical waveguide of Figure 11.21. It consists of a narrow region of refractive index n_1 , called the *core*, and is surrounded on either side by regions of lower index, n_2 , called the *cladding*. The structure shown could be a waveguide as is found in edge-emitting LEDs and lasers, or it could be an optical fiber. The optics is the same. We will assume we are discussing an LED, however, and thus every photon originates in the high-index material of the core (which is the same as the active layer).

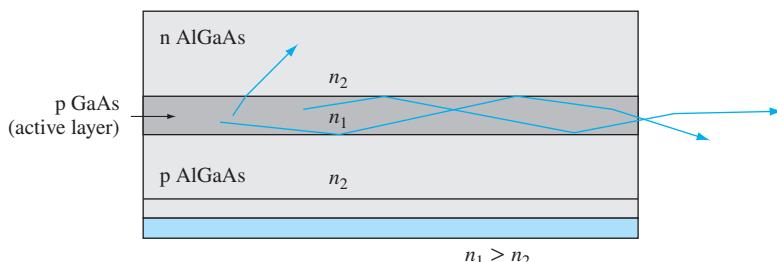


Figure 11.20 In an edge-emitting LED, the higher-index active layer acts as a waveguide for photons traveling at less than the critical angle.

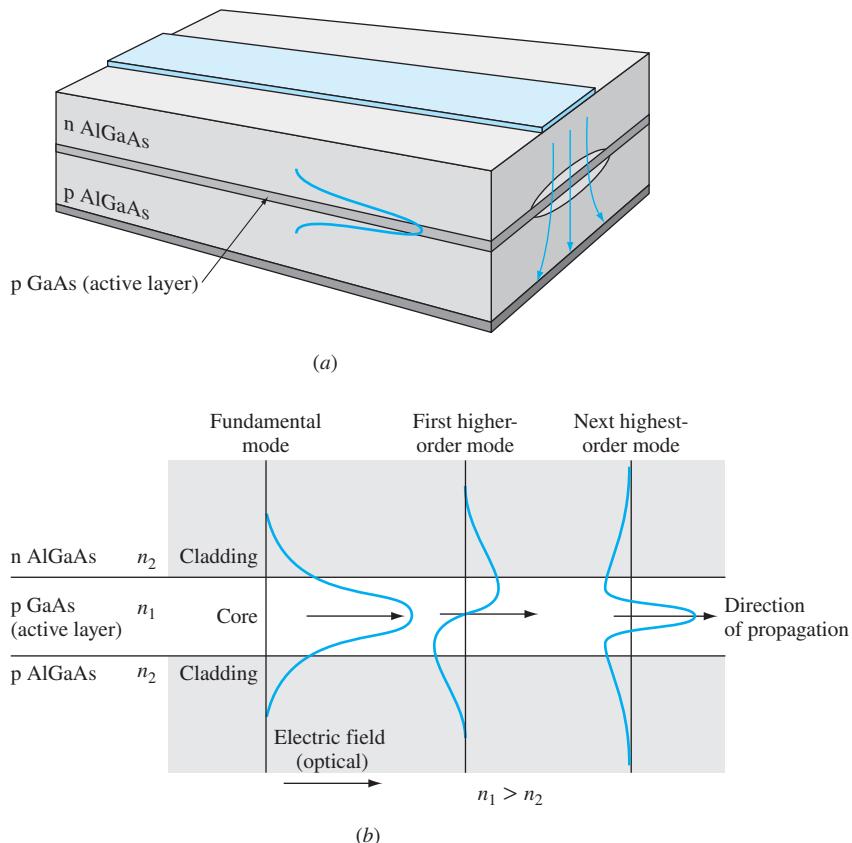


Figure 11.21 The edge-emitting LED's waveguide (a) supports only certain transverse modes, whose field distributions are shown in (b). In practice, only the first mode is allowed. It is not completely confined to the active layer, thus its absorption is reduced.

Although Figure 11.20 showed rays being totally internally reflected, the ray model does not apply to very narrow cores, on the order of $1 \mu\text{m}$ or less. Instead, we have to go to the full vector electromagnetic wave description and solve Maxwell's equations for all regions, matching boundary conditions. Such a derivation is left for other courses, but the result is that the waveguide supports certain modes. The figure shows the electric field distribution for the first three modes, along with their directions of propagation. (In practice, however, the active layer is made thin enough that only a single transverse mode is supported.) Note that even though the mode is centered in the small-band-gap layer, some of the mode's energy is actually carried in the transparent cladding layers. Thus the absorption that the photons in this mode experience is actually some average of the high absorption of the core layer and the low absorption of the cladding layers.

Edge-emitting LEDs are commonly used for fiber-optic applications. The big advantage is that while a surface-emitting LED emits in a wide angle, desirable for lamps and displays, the edge-emitting LED emits into a smaller angle, making coupling to fibers much more efficient. As we shall see in Section 11.4, a very similar structure can be used to produce laser diodes.

11.3.4 WHITE LEDs AND SOLID-STATE LIGHTING

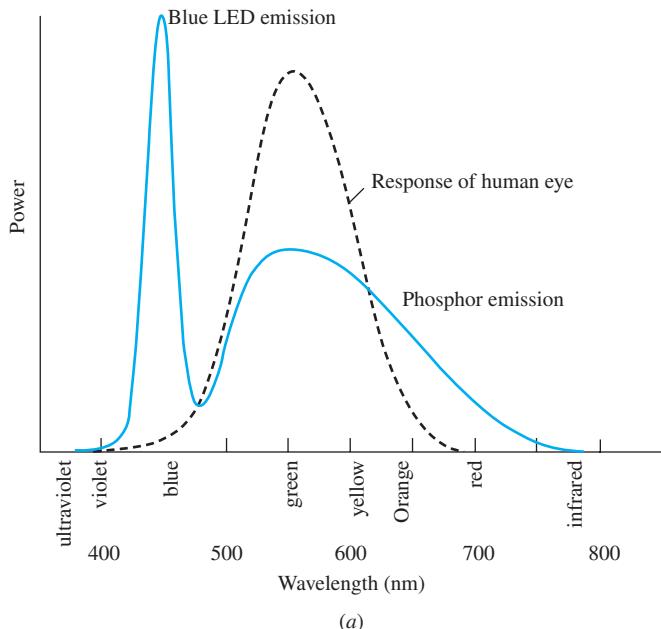
All of the LEDs described so far emit only one particular color. White “LEDs” have been created, however, by using a blue LED and a phosphor. Figure 11.22 shows a typical spectrum of the blue LED and phosphor. Some of the blue light is absorbed by the phosphor and reradiated in the green and yellow. The resulting combination of blue LED light and broader-spectrum light from the phosphor, which may contain wavelengths into the red, appears reasonably white to the human eye. A second, but less commonly used approach is to include three LEDs in a single package, Figure 11.22b. Red, green, and blue (RGB) combine to appear white.

Solid-state (LED) lighting is expected to overtake incandescent and compact fluorescent lighting in the near future. LEDs use much less power for the same amount of light emitted, and they are expected to be far more reliable. Currently about 30 percent of the electricity used in the United States is for lighting, so the energy savings can be considerable.

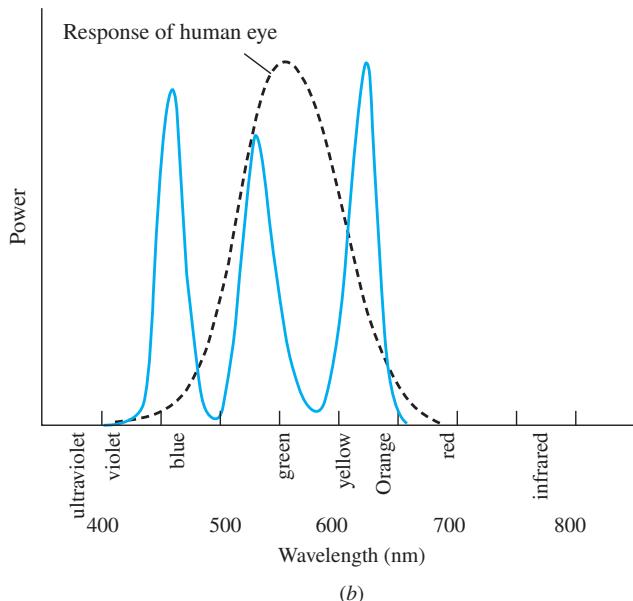
To compare the optical light produced per unit of electrical power consumed, one has to take into account the responsivity of the eye. The human eye is more sensitive to green light than to either red or blue, as seen in Figure 11.22. The perceived “brightness” of a source is measured in lumens, which measures radiant flux in the visible range but corrects it for the response of the human eye.

An incandescent bulb, which generates light by being heated, generates most of its light in the infrared region, but since people can’t see it, it can’t be considered to contribute to the lighting. Therefore, with lighting, the concept of luminous *efficacy* is used. Rather than compare radiant power out to electrical power in, the luminous flux (in lumens) is compared to electrical power in (in watts). Thus a white LED, which produces most of its light in the visible region, has a higher efficacy than an incandescent bulb. The first column of Table 11.1 compares the luminous efficacy of some common light sources.

The luminous efficacy does not tell the whole story, however. Efficacy measures the lumens per watt but does not differentiate between blue, green, and white light. The color is important for many lighting situations, too. For example, a low-pressure sodium lamp has a luminous efficacy of up to 200 lm/W, but the light out is fairly narrow-band yellow. While it is used for street lights where color doesn’t matter much, it is not used for indoor lighting because people and food won’t look right. Early LED lamps were considered too “blue,” but they have recently been made much “warmer.” The ability of a light source to accurately reveal the color of an object, compared to a blackbody radiator such as the sun, is called the “color-rendering index,” or CRI. A perfectly rendering source



(a)



(b)

Figure 11.22 (a) Spectrum of a phosphor-based LED. (b) Three LEDs combine to create white light.

Table 11.1 Some typical values of luminous efficacy in lumens/watt, color-rendering index (CRI), and correlated color temperature (CCT)

	Luminous efficacy (lm/W)	Color-rendering index (CRI)	Correlated color temperature (K)
Incandescent bulb	10	100	2700
Compact fluorescent bulb	60	60	5000
Fluorescent tube T8 (including ballast)	100	62	3000–6500
LED, white	150	80	2700–7000

would have a CRI of 100. White LEDs are around 80, incandescent bulbs are close to 100, and a cool white fluorescent bulb is in the 60s. The low-pressure sodium lamp is near zero. Currently most LED lights still do not make objects look as “natural” as sunlight.

The color temperature of a source is another way to describe the quality of the light. The temperature of a blackbody source determines the spectrum of light emitted from the source. The power radiated per unit area of surface into a unit solid angle per unit wavelength is given by

$$I(\lambda) = \frac{2hc^2}{\lambda^5 \left(e^{\frac{hc}{\lambda kT}} - 1 \right)} \quad (11.28)$$

where λ is the wavelength of the electromagnetic radiation and T is the temperature. Figure 11.23 shows the spectrum for several temperatures. It can be seen that hotter color temperatures correspond to bluer shades of light. This can be confusing, since we think of red being a warmer color than blue (perhaps because fire is red and ice is blue), but that is just a perception. The color temperature of white LEDs can be adjusted to some degree by choice of phosphor. LED color temperatures are commercially available ranging from 2700 K to 7000 K.

The concept of color temperature only really applies to a true blackbody source, and it only corresponds to the actual temperature in the case of a blackbody. For example, the color temperature of an incandescent bulb (a realistic blackbody) corresponds to the actual filament temperature. For a non-blackbody such as an LED, a correlated color temperature (CCT) is specified, which is determined by comparing the color as perceived by people to the temperature of the most similar-appearing blackbody. Thus, the color temperature does not include any direct information about the actual spectrum of the source nor how well the light will reproduce the actual colors of objects. For that, the color-rendering index discussed earlier is needed.

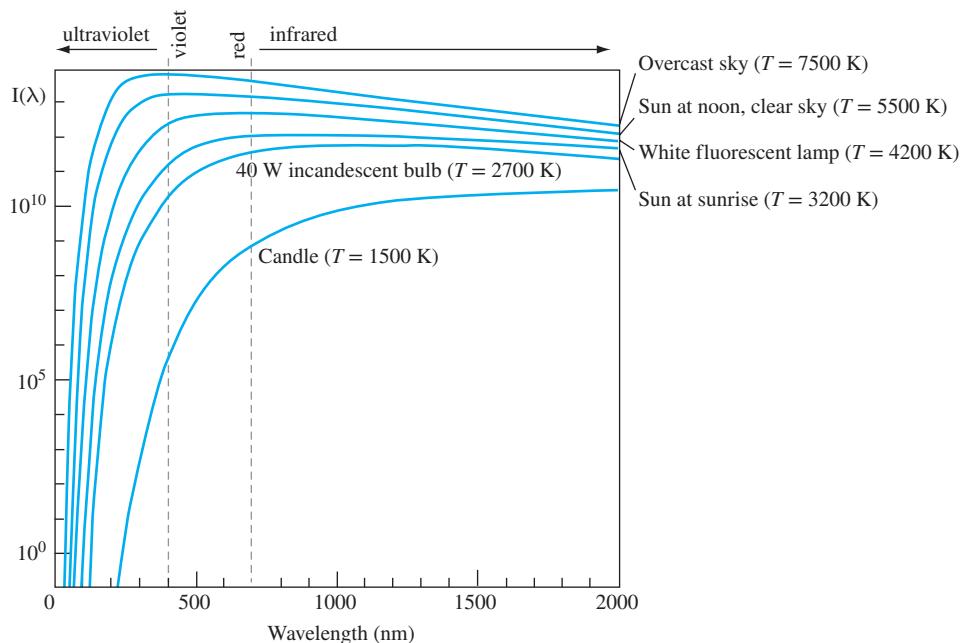


Figure 11.23 Color temperature of various light sources

11.4 LASER DIODES

The key difference between LEDs and laser diodes is that lasers operate by stimulated emission rather than spontaneous emission.⁶ For stimulated emission to occur, population inversion must be achieved. That is, there must be a large enough number of electrons at excited energies (in semiconductors, that means in the conduction band), and a large enough population of empty states at lower energies for the electrons to fall to (holes in the valence band), that the probability of stimulated emission exceeds the probability of absorption by an amount great enough to overcome other losses in the cell. In an optical system, if the number of photons coming out is greater than the number that went in, the system is said to have optical gain.

For lasers to operate, two things are required:

1. Gain
2. Feedback

These two operate together to produce lasing.

⁶The term LASER refers to Light Amplification by Stimulated Emission of Radiation similar to the acronym MASER for Microwave Amplification by Stimulated Emission of Radiation for a device which amplifies an input microwave signal. However, since there is no light input to the semiconductor device, the output is due to oscillation within the device and a more accurate acronym would be LOSER, but the accepted term is LASER.

11.4.1 OPTICAL GAIN

Optical gain exists when the number of photons leaving a system is greater than the number entering (the opposite of optical loss, such as in absorption, where the number of photons is reduced). We have said that there are three optical processes: spontaneous emission, absorption, and stimulated emission. Any of these can happen in a semiconductor, and indeed at any given time all are occurring.

Let us consider spontaneous emission first. Consider two electron levels, E_2 and E_1 as shown in Figure 11.24. If the number of electrons in the upper state is N_2 , then the upper state is being depleted at the rate

$$\frac{dN_2}{dt} = -A_{21}N_2 \quad \text{spontaneous emission} \quad (11.29)$$

where A_{21} is the Einstein rate coefficient for spontaneous emission. Note that the actual rate is negative, meaning that if the only process occurring is spontaneous emission, N_2 will decrease. In fact, the rate constant A_{21} is related to the spontaneous lifetime

$$\tau_{\text{radiative,spont}} = \frac{1}{A_{21}} \quad (11.30)$$

In absorption, the number of electrons N_1 in the lower level decreases and N_2 increases. The rate equation is

$$\frac{dN_2}{dt} = B_{12}N_1g(\nu) \quad \text{absorption} \quad (11.31)$$

where $g(\nu)$ is called the *lineshape* function. It arises because absorption requires an incident photon to excite the process. That photon must have the correct energy to be absorbed—it must have an energy equal to the energy difference between an occupied lower state and an available upper state. The lineshape describes the probability of absorption in a given material as a function of light frequency ν . For example, in Figure 11.24 there are two discrete levels, and thus only one frequency could be absorbed. In this case $g(\nu)$ would be a delta function. By uncertainty, though, no energy level is infinitely narrow, so every lineshape has some width.

In a semiconductor, instead of discrete states there is a valence band and a conduction band, each with its own density of states. The distribution and occupancy of these governs the probability of absorption as a function of photon energy.

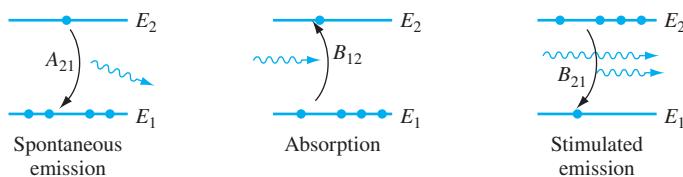


Figure 11.24 Optical processes revisited.

The same probabilities govern spontaneous emission, so the lineshape function is the same as the spectral emission of the spontaneous emission. Thus, the distribution function shown in Figure 11.17b is the lineshape function.

Finally, we consider stimulated emission. We expect that the stimulated emission process will depend on the number of carriers at excited energies that are available for recombination, and it will also depend on the frequencies of the incoming photons. The incoming photons in a semiconductor laser actually originate inside the cavity, from spontaneous emission. Thus the rate equation is:

$$\frac{dN_2}{dt} = -B_{21}N_2g(\nu) \quad \text{stimulated emission} \quad (11.32)$$

Notice that in the Einstein coefficients A_{21} , B_{21} , and B_{12} , the first subscript indicates the initial state and the second indicates the final state.

At thermal equilibrium, electrons are being excited into the conduction band at the same rate as they recombine. Thus

$$\frac{dN_2}{dt} = -\frac{dN_1}{dt} \quad \text{equilibrium} \quad (11.33)$$

Furthermore, the states may have some degeneracy. Let them have g_2 and g_1 states respectively. The g 's are the degeneracies for each state—the number of ways that state can be occupied without violating the Pauli exclusion principle. If electrons can occupy the lower level in g_1 different ways (e.g., because of different quantum numbers such as angular momentum, i.e., g_1 states), and the upper level similarly has a degeneracy of g_2 , then

$$g_2B_{21} = g_1B_{12} \quad (11.34)$$

If a photon with an appropriate wavelength enters the material, stimulated emission is more probable than absorption when

$$\frac{N_2}{g_2} > \frac{N_1}{g_1} \quad (11.35)$$

This situation is called *probability inversion*, although the term *population inversion* is normally used. It is not a normal situation, since all electrons seek their lowest energies. We expect that normally $N_1 > N_2$. To create population inversion, electrons must somehow be artificially induced to be concentrated at high energies, which for semiconductor lasers means excited into the conduction band. At equilibrium in an n-type material there are many electrons, but few holes for them to recombine with. As we saw earlier when we discussed LEDs, excess electrons and holes can be injected across a double heterojunction, placing large numbers of both in the same physical space. The same technique is used in laser diodes. A double heterostructure is employed and the junction is forward biased. The difference is that the excess carrier concentrations must be much higher than are used in LEDs if lasing is to be achieved. In fact, every laser diode is also an LED. When the junction is forward biased by a small amount, a small current flows. Electrons and holes cross the junction where they recombine. Since at low currents, the population is not inverted, the emission is

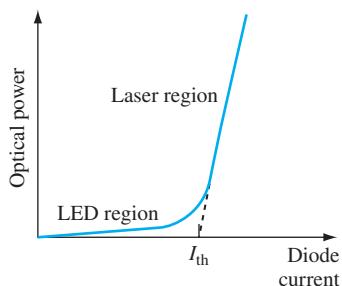


Figure 11.25 The power-current curve of a laser diode. Below threshold, the diode is an LED. Above threshold, the population is inverted and the light output increases rapidly.

primarily spontaneous. As the current increases, the light production increases linearly as shown in Figure 11.25 (LED region).

As the current continues to increase, however, the population can invert. Photons are already present in the junction from the spontaneous emission, and the distribution of their energies matches the lineshape function. Therefore, under population inversion, these photons can stimulate more emission from the excited electrons in the junction. As the inversion increases, the rate of stimulated emission also increases. When the stimulated emission is sufficiently large to overcome not only absorption but other losses in the cavity (such as through the end mirrors discussed in the next section), the light output increases dramatically. There is a distinct current threshold, as reflected in the power-current curve. Below this threshold, the laser behaves as an LED, and above threshold it is a laser.

To maintain a population inversion, current must be continually supplied. Otherwise all the available electrons will be quickly used up. In fact, the lasing process is so fast that above threshold, electrons are demoted back to the valence band by stimulated emission almost as soon as they arrive. Thus the output intensity is limited by the arrival rate of the electrons and holes, or, in other words, is proportional to the current.

11.4.2 FEEDBACK

It would seem, then, that any LED could be made to lase simply by increasing the current. This is not the case, however. Consider the edge-emitting LED of Figures 11.20 and 11.21. Spontaneously emitted photons will have random directions, so many of the emitted photons will not be traveling along the junction. The gain, however, only exists in or near the junction, so those photons will not be amplified. Spontaneously emitted photons that happen to be traveling along the junction plane, however, will remain in the gain region. They will be amplified, as long as the population is inverted to make the probability of stimulated emission greater than the probability of absorption. Still, the chip is short

and the photons don't spend very much time in the gain region before they leave the chip. Furthermore, as we saw before, the optical mode extends into the wide-band-gap layers, so only part of the mode actually overlaps with the gain region.

Optical feedback is used to increase the total optical amplification, by making the photons pass through the gain region multiple times. The optical feedback typically comes from two mirrors, one at each end of the laser. This arrangement is called a *Fabry-Perot cavity*. These mirrors are partially reflecting and partially transmitting, as shown in Figure 11.26. They are often just the cleaved crystal facets of the semiconductor material itself. The Fresnel reflection is significant, since the refractive index of the semiconductors is appreciably higher than that of air. Thus some percentage of the photons striking the mirror will be reflected back into the laser cavity. These photons will be reflected back and forth inside the laser, and on each reflection the light will be further amplified. After a few passes the optical field will be very strong indeed. The intensity will level off when the rate at which carriers are used up by amplification just offsets the rate at which carriers are supplied. To increase the intensity, the rates must be increased by increasing the current.

An interesting thing happens in a mirror cavity, however. The electric fields of the light will interfere on successive round trips. Figure 11.27 shows that certain wavelengths will interfere constructively so that the fields from successive passes add, while others interfere destructively, canceling themselves out. Individual photons emitted spontaneously at the nonresonant wavelengths can still be amplified, as long as $g(\nu)$ is large, but after a few trips through the cavity the interference causes these photons to die out.

On the other hand, photons that originate spontaneously and happen to be at the resonant wavelengths will reinforce themselves after multiple trips through the cavity, and they continue to be amplified on every pass. Thus the optical field is very strong at these wavelengths. These resonant wavelengths, shown in Figure 11.28 for several values of mirror reflectivity, are called the *longitudinal modes*—the “longitudinal” part comes from resonating along the length of the cavity.⁷ Notice from the figure that the sharpness of the resonance is related to the reflectivity of the mirrors.

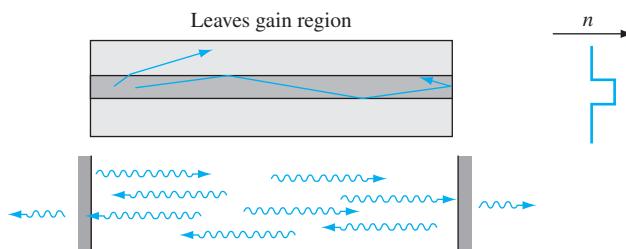


Figure 11.26 The ends of the chip form partially reflective mirrors, which allows the photons to be reflected back and forth and thus be exposed to gain for a longer period of time.

⁷There can be transverse modes too, but laser diodes are usually designed to operate in a single transverse mode, so they are not discussed here.

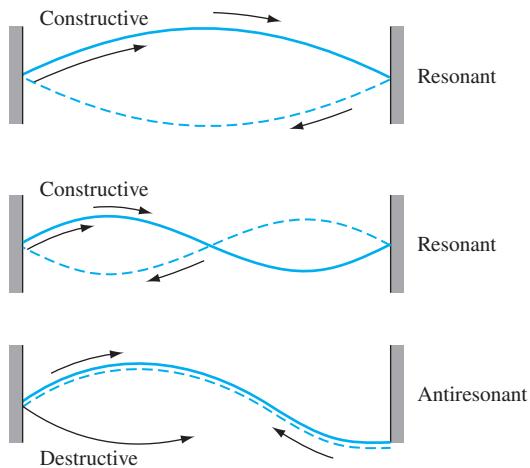


Figure 11.27 Wavelengths that are integer multiples of half the cavity's length can resonate, interfering constructively. Other wavelengths die out eventually.

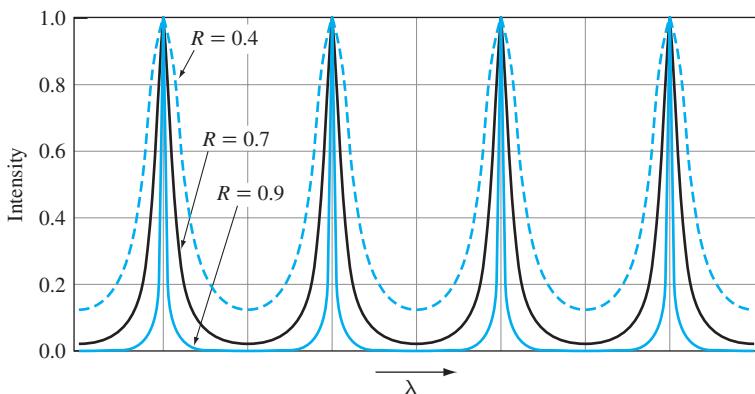


Figure 11.28 The resonances of a Fabry-Perot cavity. The width of the resonances depends on the reflectivity R of the mirrors.

The resonant wavelengths of a Fabry-Perot cavity are given by

$$\lambda = \frac{2nd}{q} \quad (11.36)$$

where d is the length of the cavity, n is the refractive index of the material inside the cavity, and q is an integer.⁸

⁸Note that the symbol q usually signifies electronic charge. To be consistent with much of the laser literature, we used the same symbol here to indicate which longitudinal mode is being considered. It should be clear from context which q is which.

EXAMPLE 11.6

A double heterostructure Fabry-Perot edge-emitting laser in the AlGaAs-GaAs system emits in the neighborhood of 900 nm. The chip is 300 μm long. What is the wavelength difference between two adjacent modes?

■ Solution

We start by finding out in what neighborhood of q we are operating. From Equation (11.35) we see that we will need the refractive index n . In a double heterojunction laser, the lowest band gap and highest index material will be the active layer. In the AlGaAs/GaAs system, GaAs has the lowest band gap (Figure 11.16). It also has the highest refractive index. The refractive index of GaAs is about 4.3 and the addition of Al reduces the index somewhat.⁹ Solving for q we have

$$q = \frac{2nd}{\lambda} = \frac{2(4.3)(300 \times 10^{-6} \text{ m})}{900 \times 10^{-9} \text{ m}} = 2866.7$$

Since q must be an integer, there must be a cavity mode for $q = 2866$ and another for $q = 2867$. Their wavelengths are

$$\lambda_2 = \frac{2nd}{2866} = 900.2 \text{ nm}$$

and

$$\lambda_2 = \frac{2nd}{2867} = 899.9 \text{ nm}$$

The spacing between modes is thus $900.2 - 899.9 = 0.3 \text{ nm}$.

As we saw in Figure 11.28, the spectral width of the modes is influenced by the mirror reflectivity R . For a GaAs laser in which the mirrors are the uncoated cleaved facets, the reflectivity is low ($R = 0.38$), and the resonances are quite broad. Coating the facets can make more highly reflective mirrors, and in this case the resonance peaks can be quite narrow. The result is that the laser beam is more coherent. The use of higher reflectivity mirrors also reduces the lasing threshold current and thus increases the efficiency.

11.4.3 GAIN + FEEDBACK = LASER

In a laser, then, there are two effects at work. If the probability of stimulated emission exceeds the probability of absorption, there is optical gain. In addition, there is optical feedback from the mirrors to allow the photons to pass through the gain region multiple times. Let us see then, how lasing occurs.

To have stimulated emission, we must have some initial photon to start the process off. Photons are always being produced spontaneously, with a probability

⁹Since the light mode will actually extend in the lower index cladding regions as well, the index of refraction that the mode experiences is actually some average of the two indices. We will simplify the problem by assuming the index of GaAs.

determined by the lineshape function (gain curve), shown again in Figure 11.29a. Some of the spontaneously emitted photons will be traveling out of the junction plane, and since they don't remain in the gain region they are lost, eventually reabsorbed or emitted from the surface.

Spontaneously emitted photons that happen to travel along the junction plane can be amplified. Also recall that the double heterostructure acts as a waveguide, so photons traveling at small angles will be reflected back into the active region. All of the spontaneously emitted photons traveling along the junction plane are amplified at first. The stimulated photons are identical to the original photons in wavelength, phase, and direction. Initially, the emission is mostly spontaneous and the output spectrum looks like the gain curve. These photons continue to travel along the junction plane and continue to be amplified. When they reach the end of the waveguide, however, they encounter a partially reflecting mirror. Some percentage of the photons is transmitted, and those photons become part of the laser emission beam; the rest are reflected back along the cavity. These are further amplified, and at the other end there is another partial mirror where some more photons are transmitted. Modes near the peak of the lineshape function, however, have more gain and thus are amplified most.

Now the cavity effect comes into play. After several passes, the fields start to add constructively or destructively. The resonant wavelengths will be amplified and the electromagnetic field associated with those wavelengths will grow rapidly, Figure 11.29b. After a few passes, the interference builds up and the modes start to emerge.

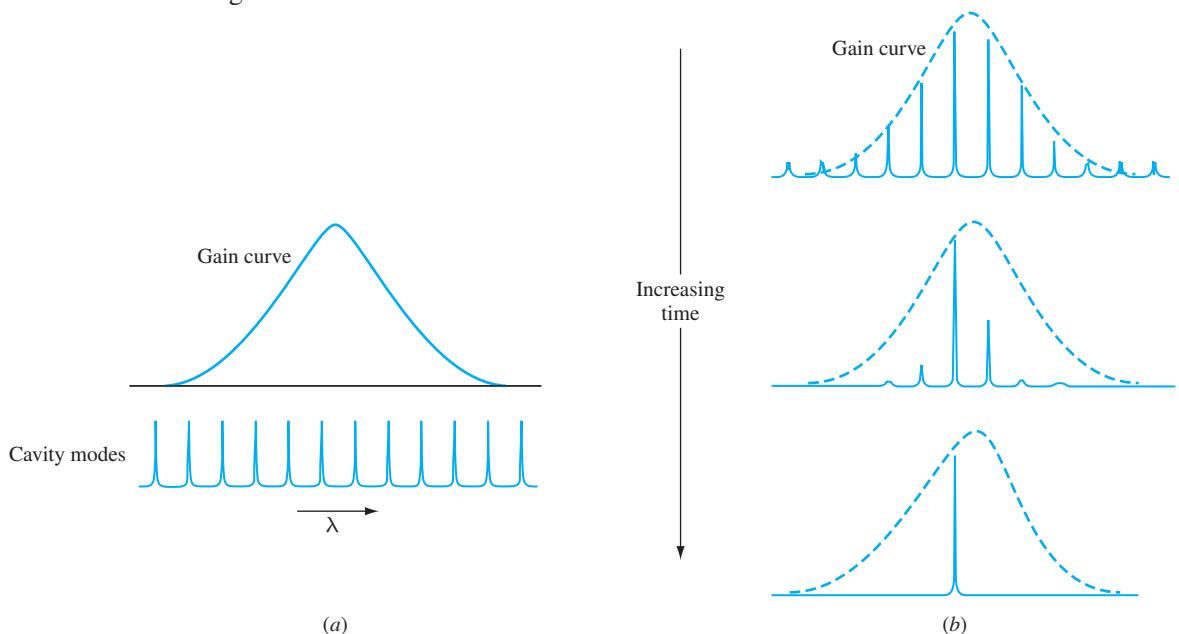


Figure 11.29 Development of lasing. (a) The gain distribution is the same as the spontaneous emission spectrum. (b) Only the photons at the resonance will amplify. The ones near the center of the gain curve will amplify the fastest.

Eventually the center mode will be so large in amplitude that it is stimulating new photons just as fast as electrons become available, Figure 11.29b. This one mode can, in some circumstances, use up all the electrons as fast as they are delivered, since there are many photons at the required energy. This single mode operation is actually preferred, because it means a more coherent beam, and one that can carry a higher bandwidth of data.

Also notice that if the gain in a laser diode is low (low current level), the photons may be lost through the mirrors at a rate faster than they can be amplified during one pass. If that happens, there is no lasing because there is no net gain. In other words, population inversion does not guarantee lasing; the inversion must be great enough to also overcome the cavity loss.

To reduce the loss, the mirrors must be made as reflective as possible. If, for example, the reflectivity is 99 percent, and the laser is emitting 3 mW (enough power to cause eye damage), then the power inside the laser is 99 times larger, or 297 mW.

11.4.4 LASER STRUCTURES

Various structures are used to make laser diodes. For example, the double-heterostructure is often made such that the active layer is thin enough to become a quantum well (the energy states become discrete). Figure 11.30 shows several different double-heterostructure (DH) single-quantum-well (SQW) energy band diagrams and their accompanying refractive index diagrams. The accompanying optical field distributions are also shown. The shaded areas indicate the size of the core; the greater the overlap of the mode with this area, the greater the gain that it sees. Notice from Figure 11.30 that the narrower the well, the less confined the optical mode is. Since only the part of the mode that actually overlaps the active layer sees gain, high confinement is desirable (unlike in the LED, where it contributes to excessive reabsorption).

The number of states in a quantum well, as well as the energies of those states, depends on the width and depth of the well. The energy difference between the allowed states determines the emission wavelength of the laser. The width of the well is determined by the thickness of the active layer, and the depth of the well is controlled by the energy difference in the conduction band (for electrons) and the valence band (for holes). These, in turn, are determined by the choice of materials. This offers opportunities for *bandgap engineering*, in which optical and electrical properties of materials can be tuned by these parameters.

A narrow well is required to produce discrete states, but it has a low probability of capturing electrons or holes, because electrons and holes can easily pass across the top of the well. A narrow well also means poor overlap between the mode and the gain. Figure 11.31 shows two common structures for improving the performance of lasers over a single quantum well design. One is called the *graded-index separate confinement heterostructure* (GRINSCH) laser (a). The active layer is a SQW small-band-gap material as before. In the cladding layers, however, the composition of the alloys is gradually changed. On the energy band diagram, this produces a sort of funnel to direct carriers into the potential

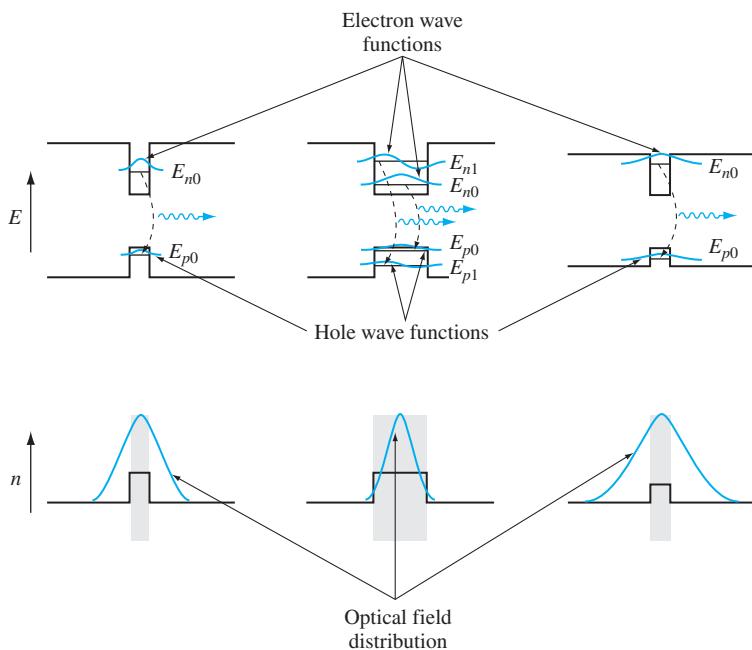


Figure 11.30 Adjusting the depth and width of quantum wells to select the wavelength of emission is one form of band-gap engineering. The shaded areas indicate the width of the well to illustrate the degree of confinement of the mode.

well of the active layer. The result is higher gain for a fixed current level. The GRINSCH laser not only helps confine the carriers but it also affects the distribution of the optical field, improving the overlap between the optical mode and the gain region. Another way to use narrow wells while still improving the optical overlap and carrier confinement is to use a multiple-quantum-well (MQW) structure, Figure 11.31b.

Laser diodes need not be edge emitting, however. Vertical-cavity surface-emitting lasers (VCSELs)¹⁰ use two mirrors, one above the active layer and one below (Figure 11.32). The mirrors here are multiple layers of alternating semiconductors of different refractive indices. These are known as *dielectric mirror stacks* or *distributed Bragg reflectors* (DBRs).

These dielectric mirrors use constructive interference between the Fresnel reflections at each dielectric interface. Stacks of 50 to 100 layers are not uncommon, and high values of reflectivity can be made—approaching 100 percent.¹¹

¹⁰Often pronounced “vick-sells.”

¹¹The constructive interference comes from a careful choice of the thickness of each layer; if different thicknesses are chosen, the Fresnel reflections could interfere destructively, annihilating any reflection and producing an antireflection layer.

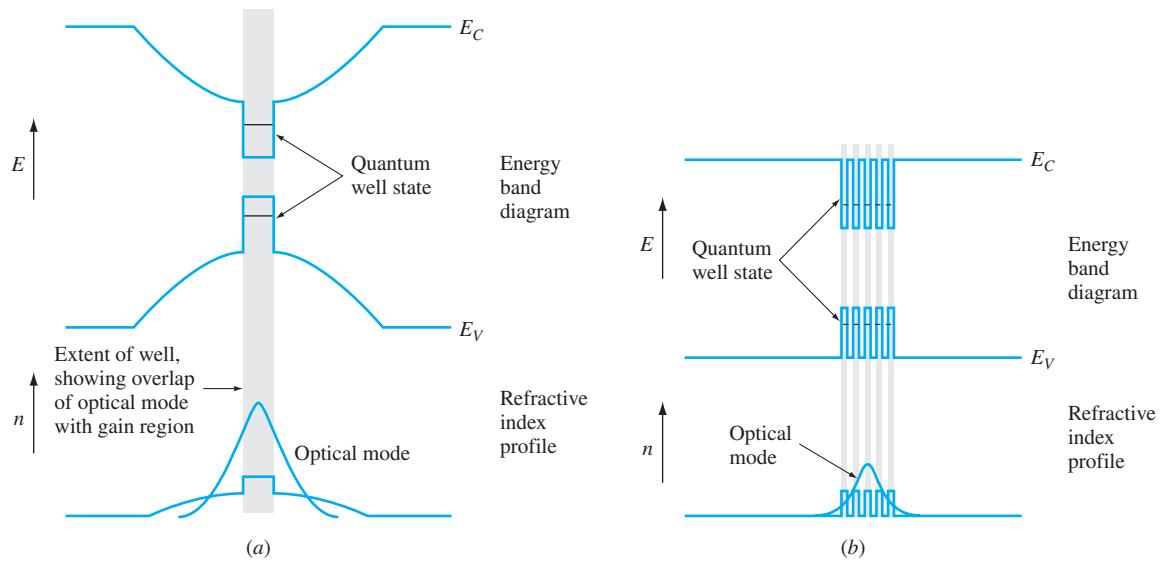


Figure 11.31 (a) A GRINSCH structure helps funnel the carriers into the wells to improve the probability of recombination. (b) A multiple quantum well structure has the advantage of single states, like the SQW, but improves carrier capture.

One of the advantages of VCSELs is that the output beam is easier to couple into optical fibers than the beam from edge-emitting lasers. To see why this is so, consider the edge-emitting Fabry-Perot laser diode in Figure 11.33. The active layer is very thin, on the order of $0.1 \mu\text{m}$. The width of the lasing spot is usually

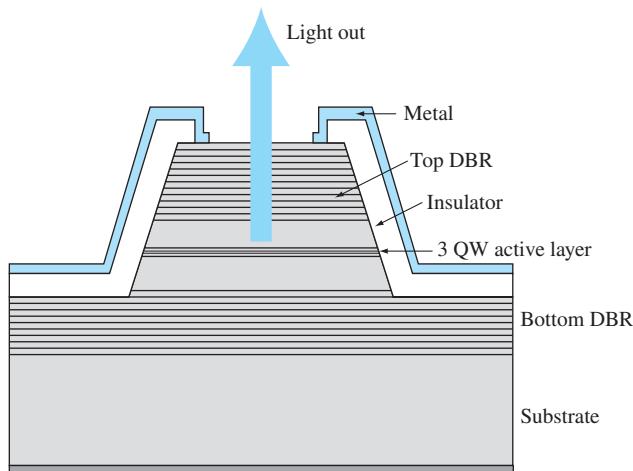


Figure 11.32 A vertical cavity surface-emitting laser.
(Source: Adapted from Ueki et al., *IEEE Photonics Technology Letters*, 11, no. 12, pp. 1539–1541, 1999, © IEEE.)

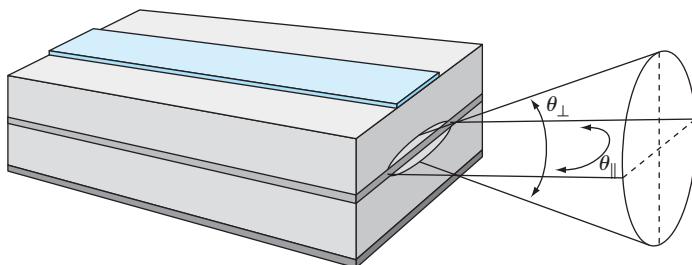


Figure 11.33 The output pattern of an edge-emitting laser is elliptical and widely divergent.

wider than this; a few micrometers is typical. So the lasing spot, if one looks in the near field (right up against the output facet of the chip), appears elliptical.

In the far field, however, the picture is much different. Recall that when light passes through an aperture, it is diffracted. The smaller the aperture, the larger the diffraction angle. The aperture of the beam at the output facet is much smaller in the direction perpendicular to the junction plane, so the angle of spread, θ_{\perp} , of the beam is much wider than the angle in the plane parallel to the junction, θ_{\parallel} . Typical values of θ_{\perp} are 20 to 40°, and values of θ_{\parallel} are in the range of 5° to 20°. Therefore, in the far field, the beam from an edge-emitting laser (or LED) is elliptical, oriented perpendicularly to the near-field ellipse. As might be imagined, coupling an elliptical beam to a circular fiber inevitably leads to losses. Vertical cavity lasers have wide beam angles also, since they are also small, but the beams are circular and can be imaged onto a fiber core efficiently with a lens. Also, with VCSELs the laser does not have to be cleaved from the rest of the chip, making it possible to integrate electronics such as logic and drive circuitry, modulators, and photodetectors on the same substrate, producing optoelectronic integrated circuits (OEICs).

Finally, there are other approaches to providing optical feedback apart from using a Fabry-Perot cavity. One common structure is a distributed feedback (DFB) laser, shown in Figure 11.34. A corrugated layer is obtained below the

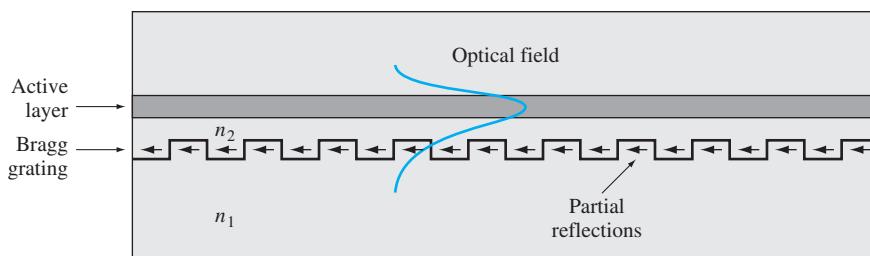


Figure 11.34 The distributed feedback (DFB) laser uses a grating to provide continuous feedback along the laser cavity.

active layer by etching a periodic structure and then filling in the corrugations with a regrowth of a material with a different refractive index. The optical field extends across this periodic variation in refractive index. At each step, there will be a small Fresnel reflection. These repeated reflections accumulate and interfere, producing a laser beam of very narrow spectral width. This type of structure is often used in lasers intended for RF applications.

11.4.5 OTHER SEMICONDUCTOR LASER MATERIALS

Although we have primarily discussed lasers in the AlGaAs and InP systems, many other semiconductors are becoming technologically important for laser diodes. These include III-V compounds such as InGaAsSb, InAsPSb, InGaAsP, and AlGaAs, for the red and near infrared; IV-VI semiconductors including PbSnTe, PbSSe, and the II-VI HgCdTe for the 5- to 17- μm range; and for the blue into ultraviolet, II-VI compounds like ZnCdTe and ZnTeSe as well as III-V nitrides are used. Some semiconductors and their wavelength ranges are shown in Figure 11.35. Difficulties in using some of the materials lie in finding an appropriate lattice-matched substrate on which to grow the layers.

11.5 IMAGE SENSORS (IMAGERS)

Semiconductor-based image sensors (imagers) are widely used in applications such as digital cameras and camcorders. Here we briefly describe two types of such sensors. These are charge-coupled devices (CCDs) and CMOS¹² image sensors using photodiodes for light detection. We first describe a method of charge transfer in a CCD device. Then its use is illustrated in a CCD image sensor along with an alternate CMOS structure for image sensing.

11.5.1 CHARGE-COUPLED DEVICES (CCDs)

A charge-coupled device (CCD) is essentially an analog shift register. It consists of a linear array of MOS capacitors as indicated in Figure 11.36. The basic principle of charge-coupled devices involves the movement of charge from one physical location to another in a controlled manner by the use of sequenced clock pulses. While CCDs are capable of performing numerous electronic functions such as storing data, signal processing, and logic operations, the most important application is in optical image sensing (e.g., in digital cameras). Here we simply indicate one method of charge transfer.

We take the case of a two-phase CCD. In these devices, the potential well under each electrode has two different depths—shallow under a portion of the electrode and less shallow under the other portion. One method to build such a structure is to selectively ion-implant acceptors into the p substrate as indicated in Figure 11.37a. In the region of increased acceptor concentration, the

¹²The name CMOS comes from the CMOS circuitry associated with the processing of the detected signals.

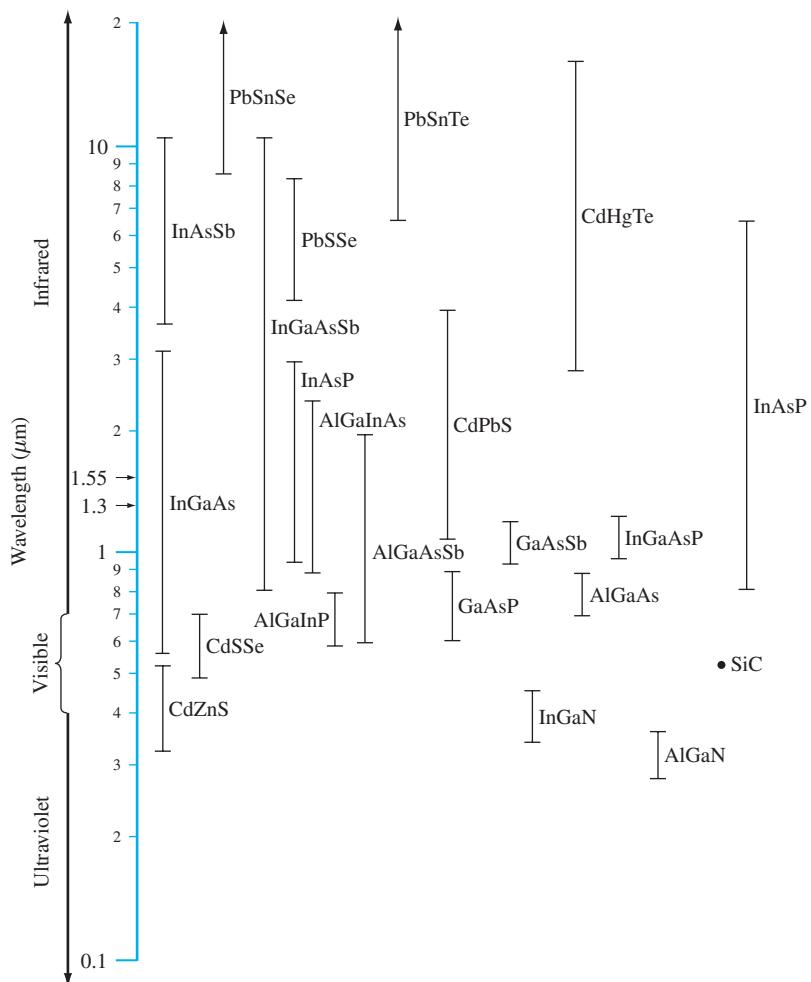


Figure 11.35 Some semiconductor materials and their wavelength ranges. Based on data from [3-5].

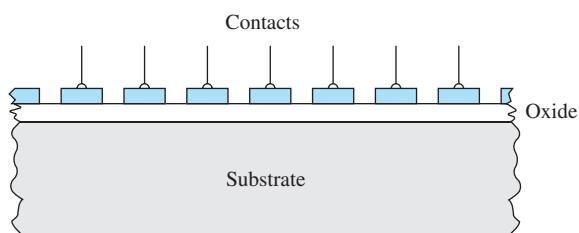


Figure 11.36 A linear array of adjacent MOS capacitors form a CCD shift register.

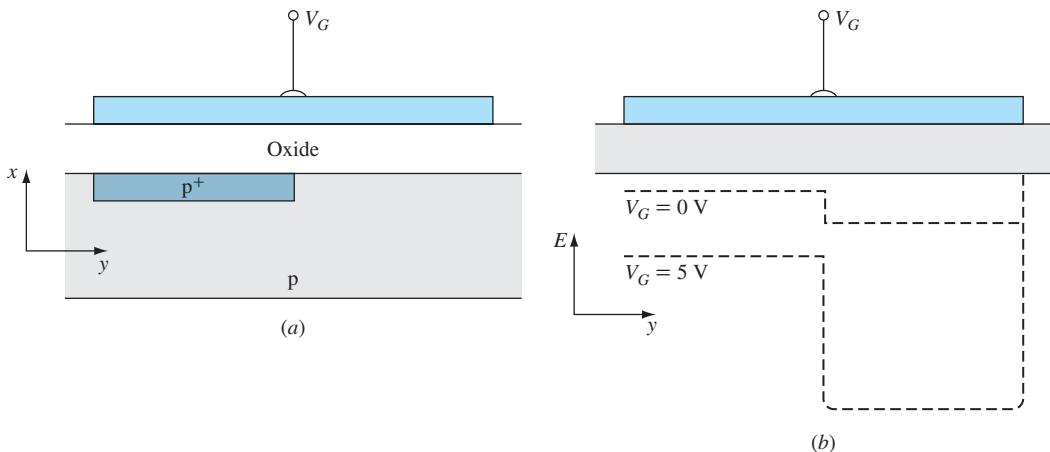


Figure 11.37 (a) Physical diagram of a MOS capacitor for use in a CCD and (b) the corresponding hybrid diagram for empty wells. The required asymmetry in the potential well is achieved by ion implanting acceptors to increase the doping concentration and thus reduce the depth of the potential well in one section of the capacitor.

surface potential is reduced (and ϕ_{ox} increased), as shown in the hybrid diagram of Figure 11.37b for the two cases of $V_G = 0 \text{ V}$ and 5 V .

We illustrate the operation of this two-phase shift register by assuming the gates connected to the timing signal of phase 1 to be at 5 V , and those connected to phase 2 to be at 0 V as shown in Figure 11.38. These two voltages, combined with the stepped doping, create four different well depths in the array.

We also assume that there are various numbers of electrons in the wells under the gates of phase 1, as in Figure 11.38a. For example, these could have been optically generated, resulting from the light of an optical image being focused on the CCD. To transfer electrons one-half step to the right, the voltage on phase 1 is made 0 V and that of phase 2 is 5 V . This increases the well energy under the capacitors of phase 1 and decreases it for those of phase 2. The electrons then transfer to the phase 2 gates as shown in Figure 11.38b. The hybrid diagram after the transfer is completed is indicated in (c). Next, the two voltages are cycled again and the charges transfer another half-step to the right. As the voltages cycle on the two phases, the electron “signals” keep being transferred to the right. These signals are then detected by the use of a MOSFET at the output of the shift register.

11.5.2 LINEAR IMAGE SENSORS

There are a number of variations of image sensor structures. Here we first describe CCD and CMOS linear image sensors and then simple area image sensors such as those used in digital cameras.

Consider the structure of Figure 11.39a, which consists of a number of MOS capacitors with optically transparent gate contacts (labeled “Light sensors”),

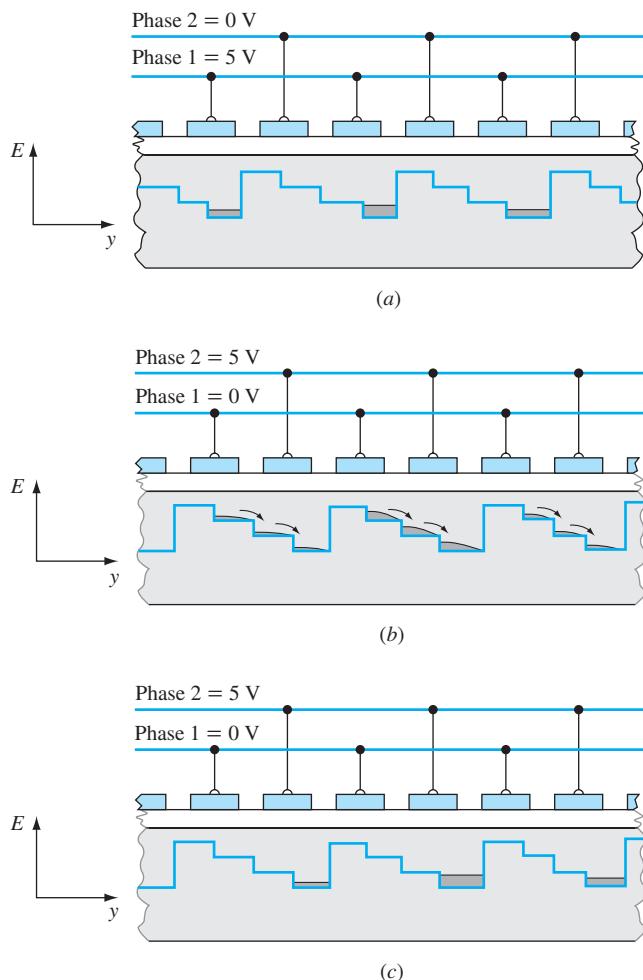


Figure 11.38 Hybrid diagram of a section of a two-phase CCD. In (a) the electrodes connected to phase 1 are at 5 V and those connected to phase 2 are at 0 V. Any mobile charge is confined to the wells under phase 1. In (b), phase 1 is made 0 while phase 2 is at 5 V. Charge flows from the phase 1 capacitors to those of phase 2. In (c) the charge shifting process is completed; the charges now reside in a well one-half step to the right.

a transfer gate, a CCD shift register described above, and a readout device (a charge-to-voltage converter) such as a MOSFET (not shown). Since the output voltage of the charge-to-voltage converter is proportional to the charge magnitude (an analog signal), it is normally followed by an analog-to-digital converter for use in a digital camera. In operation, the light-sensing capacitor is biased

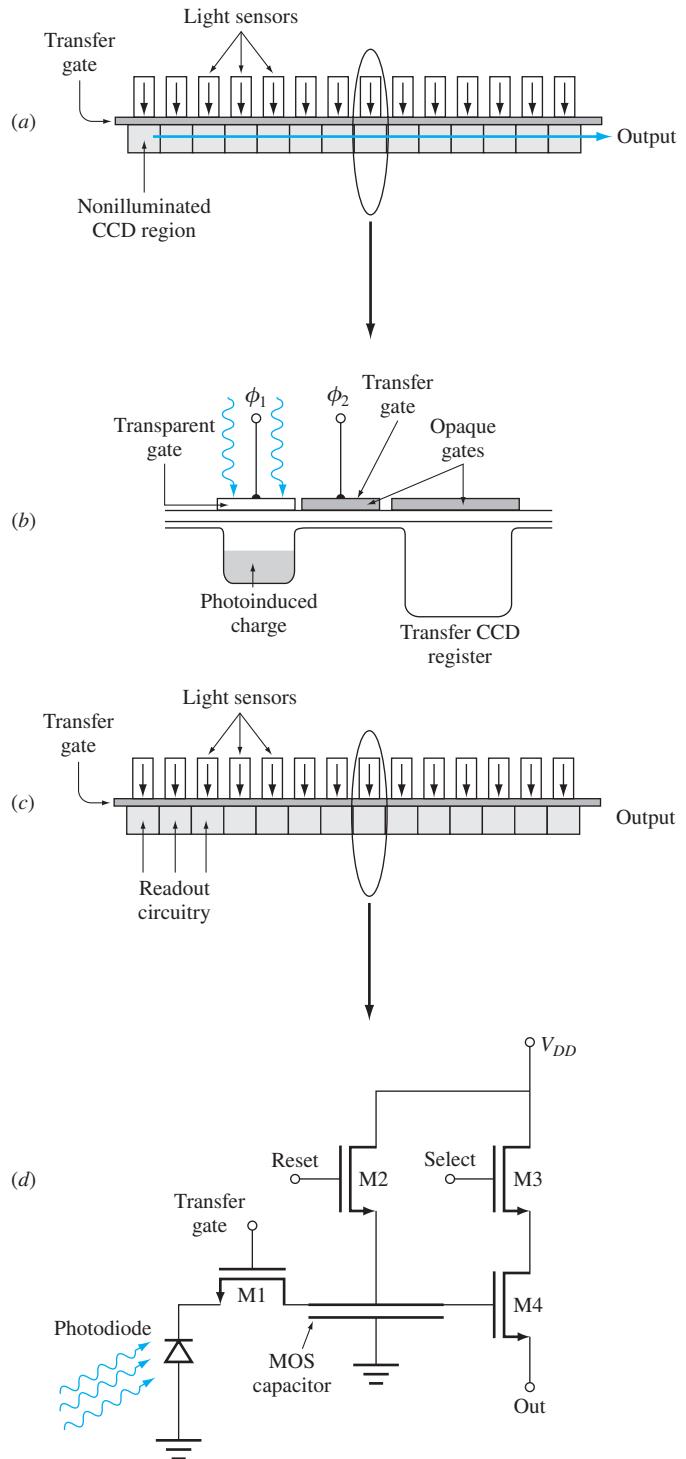


Figure 11.39 (a) Structure of a linear CCD sensor; (b) hybrid diagram of one stage; (c) structure of a linear CMOS sensor; (d) readout circuitry of a single stage.

to produce a potential well. The transfer gate is biased to form a barrier that confines the photogenerated electrons as indicated in Figure 11.39b. Under illumination, the charge in the well increases with time and is proportional to the light intensity. After a given light integration time, the ϕ_1 contact is pulsed low while ϕ_2 is pulsed high. The accumulated charge is then shifted to a well under the transfer gate, from which it is transferred to the readout device. Phases ϕ_1 and ϕ_2 are then returned to their original values and the cycle continues. The time dependence of the output voltage is then a measure of the light intensity as a function of position.

Figure 11.39c illustrates a linear CMOS sensor. Here the charge-to-voltage conversion is accomplished at each pixel by the circuitry illustrated in Figure 11.39d. Associated with each pixel is a photodiode for detecting light, four MOSFETs, and a MOS capacitor.

During the period of light detection, M1, M2, and M3 are **off** and the capacitor is charged to V_{DD} (e.g., 2 V). The voltage on the photodiode (open-circuit voltage) depends on the light intensity and the time of integration. To detect the photodiode voltage and thus its charge, M1 is turned **on**. Electrons then flow from the diode through M1 to the capacitor, thus discharging the capacitor and reducing its voltage to a value dependent on the charge it received. After the charge is transferred (e.g., 1 μ s), M1 is turned **off** and the photodiode again detects the incident light. Meanwhile, the voltage on the capacitor, which represents the light signal, is detected by turning **on** M3 (via a “Select” signal), which puts the drain voltage of M4 at V_{DD} . This activates M4. The output current is then a measure of the capacitor voltage or the detected light. After readout (e.g., 1 μ s), M3 is turned **off** and M2 is turned **on** to recharge (reset) the capacitor to V_{DD} , and then turned **off**. After the prescribed light integration time, M1 is turned **on** and the cycle repeats.

11.5.3 AREA IMAGE SENSORS

A number of the preceding linear imagers can be incorporated into area imagers with the structures of Figure 11.40. In the CCD imager (a), after the light integration time, all the photosensing wells are simultaneously emptied into their respective (vertical) shift registers. The charge in the wells of these shift registers is transferred to the output (horizontal) shift register and then transferred to the output. The output signal voltage waveform is then a function of the two-dimensional position dependence of the light intensity.

A 4×4 pixel CMOS area imager is shown schematically in Figure 11.40b. (Digital cameras have millions of pixels.) As in the linear imager, the charge-to-voltage conversion is accomplished at each pixel.

Both CCD and CMOS imagers are in extensive use. CCD imagers have a greater light sensitivity, but because of the time required to traverse the long charge-coupled registers they are relatively slow compared to CMOS imagers, which avoid such registers. CCD imagers require special fabrication processes, but CMOS imagers can be made in conventional CMOS facilities, making them considerably less expensive. The power consumption of CMOS imagers is also

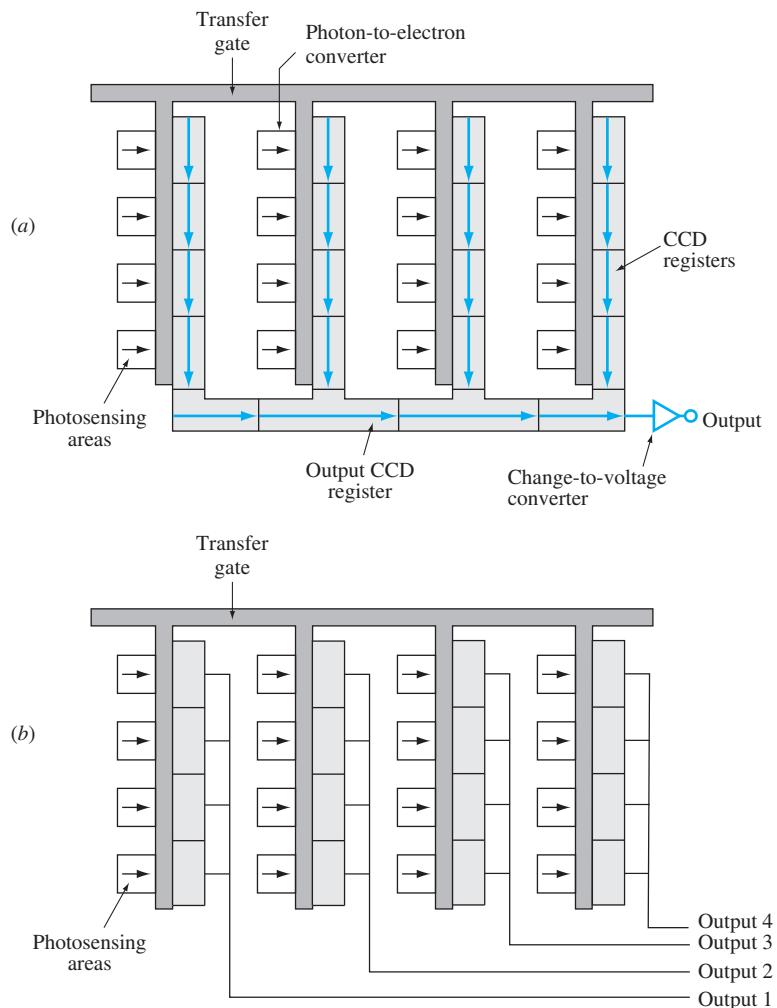


Figure 11.40 Schematic of a 4×4 area CCD imager (a) and of a CMOS imager (b).

about 1 percent of that of CCD imagers. CMOS imagers are therefore normally used in consumer applications such as cell phones and digital cameras. CCD imagers are used where greater sensitivity and higher quality are required, as in cameras for professional photographers.

11.6 SUMMARY

In this chapter we examined some of the optical properties of semiconductor devices and saw how diodes can be used for both light detection and light sources.

For solar cells and photodetectors, only those photons that are absorbed in or within a diffusion length of the junction produce photocurrent. Thus successful diode structures either have their junctions close to the surface where the light is absorbed or the surface layer may be a wide-band-gap material that is transparent to the incident radiation. It is also advantageous to make the depletion region wide to increase photocarrier collection. A very common structure for this purpose is the pin diode, in which the middle intrinsic layer effectively extends the depletion width. In this case most of the photogenerated current contributes to the prompt response.

High-speed photodetectors are operated under reverse bias. This widens the depletion region, increasing absorption, and the electric field in the junction helps speed up the response time. Under reverse bias, the current is proportional to the light input even under varying loads.

Light-emitting diodes and lasers, on the other hand, are operated under forward bias. By injecting electrons and holes across the junction, both types of carriers are made available in the same physical region for more efficient recombination. The use of a double heterostructure to trap carriers increases the efficiency of both lasers and LEDs.

The double heterostructure also helps to confine the light to the gain region in a laser diode. This is because the refractive index of the narrow-band-gap material is different (usually higher) than that of the surrounding wide-band-gap material. Light encountering a boundary from high to low index can be totally internally reflected.

In a laser, two mirrors at either end of the cavity cause the light to reflect back and forth in the gain medium, so the optical field can be amplified many times. Still, only certain wavelengths will add constructively and actually lase. The narrow spectral width of diode lasers arises from the narrow resonances of the Fabry-Perot cavity rather than from the spectral width of the gain.

III-V semiconductors are the most common for semiconductor sources, but advances in materials technology are creating new possibilities over a wide range of optical wavelengths.

Charge-coupled imagers and CMOS-based imagers were briefly discussed. Such devices are used in digital cameras and camcorders.

11.7 REFERENCES

1. J. J. Loferski, “The first forty years: a brief history of the modern photovoltaic age,” *Progress in Photovoltaics: Research and Applications*, Vol. 1, pp. 67–78, 1993.
2. H. Melchior, “Demodulation and photodetection techniques,” in F. T. Arecchi and E. O. Schultz-Dubois, eds., *Laser Handbook*, Vol. 1, North-Holland, Amsterdam, pp. 725–835, 1972.
3. E. Kapon, *Semiconductor Lasers II: Materials and Structures*, Academic Press, New York, p. 73, 1999.

4. B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, John Wiley & Sons, New York, p. 633, 1991.
5. J. Singh, *Semiconductor Devices: Basic Principles*, John Wiley & Sons, New York, p. 460, 2001.

11.8 REVIEW QUESTIONS

1. Why are photodiodes typically reverse biased?
2. Why are solar cells operated in the fourth quadrant?
3. Explain how light energy is converted to electrical current in a photodiode.
4. What is the purpose of the intrinsic region in a pin diode?
5. What is dark current? From what mechanism(s) does it arise?
6. What is the difference between quantum efficiency and responsivity?
7. What is meant by air mass zero?
8. What is the difference between spontaneous emission and stimulated emission? On which does a light-emitting diode operate?
9. Why are direct-gap materials used for lasers?
10. How is optical feedback achieved in lasers?
11. What governs the spectral width of laser diodes?
12. Why is the gain curve the same as the lineshape function?
13. Explain how the double heterostructure improves the efficiency of lasers. (*Hint:* There is an electrical reason and an optical reason.)
14. Explain in your own words the operation of a CCD imager.

11.9 PROBLEMS

- 11.1 Consider an ultra-fast photodetector operating in the neighborhood of 60 GHz. How many cycles of green light are there in a single cycle of 60 GHz? Can a photodetector be used to follow the oscillations of the electromagnetic field associated with this light?
- 11.2 Light with wavelength $\lambda = 620$ nm is incident on a sample of CdTe.
 - a. Where in the spectrum does this radiation lie?
 - b. At what depth is the incident flux (neglecting Fresnel loss) reduced to 10 percent of its value at the surface? 1 percent?
 - c. The wavelength is changed to 830 nm. What color is this? Now how deep does the light penetrate (to the 10 percent level)?
- 11.3. Find the change in conductivity for a film of pure InP of area 1 cm^2 and thickness $25 \mu\text{m}$ when it is illuminated by a light beam with $\lambda = 700$ nm and strength 1 mW/cm^2 . Ignore reflection. Take $\mu_p = 200 \text{ cm}^2/\text{V}\cdot\text{s}$ and $\mu_n = 5400 \text{ cm}^2/\text{V}\cdot\text{s}$.
- 11.4 a. Calculate the Fresnel reflection at normal incidence for light going from air to glass ($n = 1.5$).

- b. Explain why you can see into a store window and see your reflection at the same time, but at night looking out your window from a lighted room you can see only your reflection.

11.5 Show that Equation (11.8) follows from Equation (11.6).

11.6 For the circuit shown in Figure P11.1,

- Plot the $I-V_a$ characteristic for the diode with photocurrent $I_L = 100 \mu\text{A}$, $200 \mu\text{A}$, and $300 \mu\text{A}$. Let $I_{\text{dark}} = I_0 = 10^{-14} \text{ A}$ and the ideality factor $n = 1$.
- On your graph, also plot the load lines for $V_a = +5 \text{ V}$ and $V_a = -5 \text{ V}$.
- Find the current flowing through the circuit for each load line and plot it against I_L . Recalling that the photocurrent I_L is proportional to the intensity of the light, under which bias regime should one operate photodiodes if one wants the output current to be proportional to intensity?

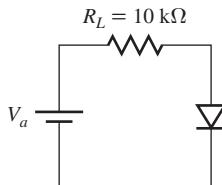


Figure P11.1

11.7 A photodiode is made of $\text{Al}_{0.3}\text{Ga}_{0.7}\text{As}$. The refractive index at $\lambda = 850 \text{ nm}$ is 3.423. If the junction depth is $0.2 \mu\text{m}$, and the junction width is $1.5 \mu\text{m}$, find the quantum efficiency η_Q and the responsivity R_{ph} . Assume the light is incident from air. Let $\alpha = 10^4 \text{ cm}^{-1}$.

11.8 For the solar cell whose $I-V$ characteristics are shown in Figure P11.2, find I_{sc} , V_{oc} , and η . The incident power is 15 mW.

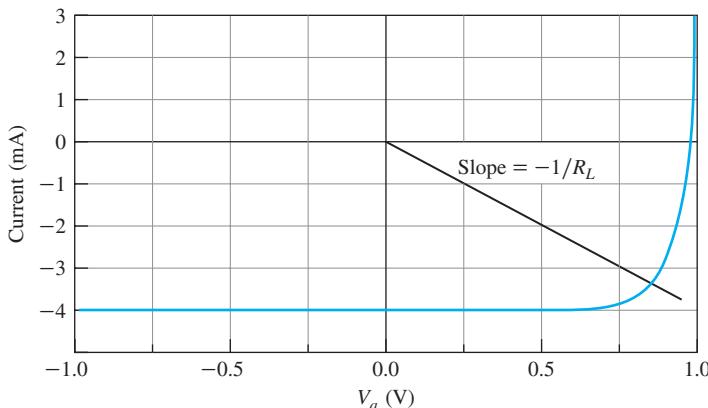


Figure P11.2

- 11.9** Solar cells made from GaAs can be significantly more efficient than Si (35% vs. 20%) because GaAs is a direct-gap material. Yet virtually all solar panels are made from silicon. Why?
- 11.10** If a photon of wavelength at the solar spectrum peak of $\lambda = 0.5 \mu\text{m}$ (green) is absorbed by Si, the electron and hole have excess energy as shown on the energy band diagram of Figure P11.3. If both carriers scatter down (or up) to the band edges, what percentage of the absorbed energy is lost as phonons?

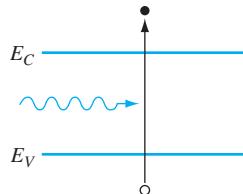
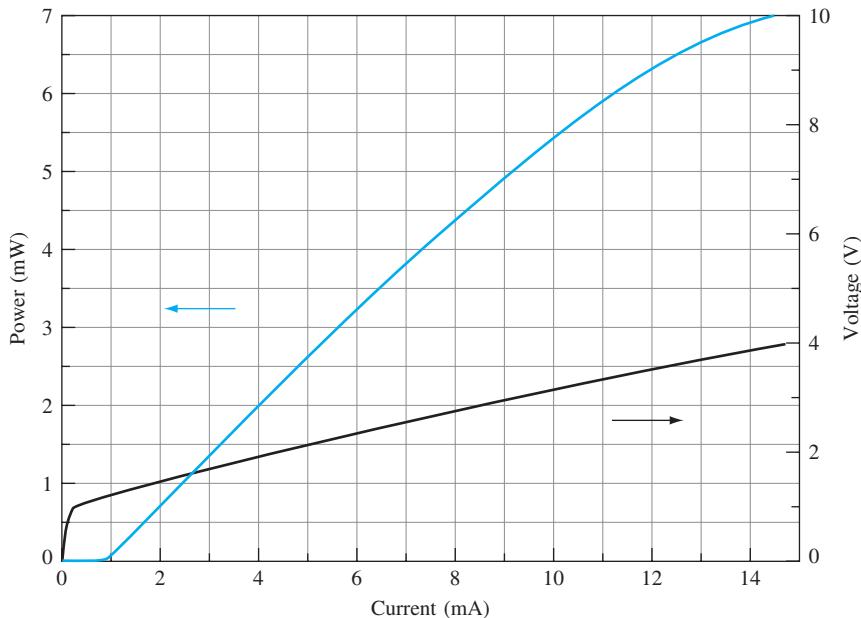


Figure P11.3

- 11.11** If the doping in the p region of an n⁺p photodiode is decreased, one would expect the diffusion length in a solar cell to increase. Verify (or contradict) this by repeating Example 11.4 with $N'_a = 10^{16} \text{ cm}^{-3}$. For a factor of 10 change in doping, what was the change in η_Q ?
- 11.12** If a GaAs photodiode has a junction depth of 0.2 μm , and if light absorbed in the surface layer is considered lost to surface recombination, what is the total fractional loss in photons in the surface layer? Let the photon energy be 1.4 eV, and repeat for $E_{ph} = 1.8 \text{ eV}$. Where on the spectrum are these two energies?
- 11.13**
- a. Explain why the pin diode would break down first at the corners if there were no guard ring.
 - b. Draw the energy band diagram for a pin diode under high reverse bias and indicate the breakdown mechanism.
 - c. Draw the energy band diagram for a p⁺p⁻in junction under the same reverse bias. Explain why this structure will break down at higher voltages.
- 11.14**
- a. What should the concentration x of arsenic be in a GaAs_xP_{1-x} LED designed to emit band to band at $\lambda = 680 \text{ nm}$ (Figure 11.13)?
 - b. If you also take into account the fact that the electrons are concentrated slightly above E_C and the holes are concentrated a little below E_V , how much does that change the band gap you would choose?
- 11.15** Recall that the peak of the electron distribution with energy is about $\frac{1}{2}kT$ above E_C , and the peak of the hole distribution is about $\frac{1}{2}kT$ below E_V . If each distribution is approximated as having an overall width of kT , estimate the spectral width of the emission. Assume the material emits at 1.3 μm . (*Hint:* To find $\Delta\lambda$, use $E = hc/\lambda$, and take the derivative $dE/d\lambda$ to obtain an expression for $\Delta\lambda$ in terms of ΔE .)

- 11.16** Optical fiber manufacturers battled the OH^- ion (resulting from water) for years. These ions, when incorporated into the glass, produce a strong absorption at $1.4 \mu\text{m}$ (see Figure 11.15). They have finally managed to nearly eliminate it. Is there a similar absorption in the earth's atmosphere?
- 11.17** What color is a photon of energy 1.0 eV? What semiconductor materials can be used to produce emission at 1 eV? Of these, are any compatible with readily available substrates (e.g., GaAs, GaN, or InP)?
- 11.18** a. Find the frequency difference between the Fabry-Perot resonances of an edge-emitting laser diode chip in which the effective index that the mode sees is 3.4, the wavelength is 850 nm, and the chip length is $100 \mu\text{m}$. If the gain curve is 50 nm wide, how many Fabry-Perot resonances are there in this range for this diode?
 b. In this chapter we implied that a laser will ultimately resonate in only one mode. This was an oversimplification, and for reasons beyond the scope of this book, it is possible for diode lasers to lase in multiple modes. How short should the cavity in part (a) be to ensure that only one mode lasers? A VCSEL lends itself naturally to this dimension.
- 11.19** A diode begins to lase when the gain in the cavity exceeds the losses. One source of loss is the partially reflective mirrors at either end of the cavity. Some percentage of the light power is lost each time the light strikes one of the mirrors. How would the power-current curve of a laser be changed if coatings were added to the facets to increase the reflectivity?

**Figure P11.4**

- 11.20** Lasers are often characterized with an *L-I-V* plot, or one that plots light, current, and voltage, like the one in Figure P11.4 for a VCSEL. The *L-I* (power, or light, versus current) curve uses the left axis, and the *V-I* curve uses the right axis. For the laser shown in the figure, what is the ratio of the optical power emitted to the electrical power dissipated in the device at an operating current of 10 mA? What happens to the rest of the power?

Power Semiconductor Devices

Power devices constitute a special class of semiconductor devices. These are used in automation systems, electric drives, renewable energy conversion, and other applications. These devices must be able to handle very large voltages (kV) and tens or hundreds of amperes without breaking down, must be thermally conductive to make it possible to remove heat as efficiently as possible, and must be able to operate at elevated temperatures. In this chapter, we will look at some common power semiconductor device structures and see why there is currently much interest in moving to wide band gap materials such as SiC and GaN.

12.1 INTRODUCTION AND PREVIEW

The motor in an electric car may be AC or DC, but it still must run on DC batteries. A solar cell farm produces DC energy, but to distribute that energy, it must be converted to AC. In fact (in the USA) it must be converted to 120 V 60 Hz and synchronized with the grid. A wind turbine produces AC energy, but not necessarily at the right frequency. A washing machine might have a DC motor that has to run on the AC line voltage.

In all these cases, the voltages must be changed from AC to DC or DC to AC, converted from one voltage to another, and even transformed from one frequency to another. The conversion circuits (rectifiers and inverters) require semiconductor devices that can handle very large currents and voltages. Thus the semiconductor materials must be able to withstand high voltages without breaking down, high currents without melting, high temperatures without failing, and in some cases high frequencies as well.

Accurate calculations of the electrical properties of power semiconductor devices are quite involved. To illustrate the physical processes involved, we will be making many approximations. These include neglecting impurity-induced band-gap narrowing effects. Further, for simplicity, compensated doping is ignored, so in n-type material N_D represents $N'_D = N_D - N_A$, and in p-type material N_A

represents $N'_A = N_A - N_D$. The electrical characteristics are also quite temperature sensitive. Here, however, for brevity, we consider only room temperature operation.

12.2 RECTIFYING DIODES

In AC-to-DC converters, diodes may be used to rectify the input voltage, which means they require high reverse breakdown voltages and little reverse leakage current. In the forward direction to handle large currents they need to have low voltage drops and low resistance.

12.2.1 JUNCTION BREAKDOWN

For high breakdown voltages, the breakdown mechanism is by avalanche multiplication. Figure 12.1 shows the energy band diagram under reverse bias for simple pn junctions with two different band gaps. We define a critical field \mathcal{E}_{cr} as the electric field at which carriers are accelerated enough that when they scatter, they are likely to create free carriers in the opposite band. From the figure we can see that at the voltage shown, this can happen for the junction in the narrow-gap material but not for the wide-gap material. The junction in the wide band gap material can sustain much larger voltages without breaking down. Wide band gap materials will also exhibit far lower leakage currents because the reverse current before breakdown is due entirely to thermally generated carriers.

In carrier multiplication, carriers gain enough kinetic energy between collisions to create hole-electron pairs. The carriers thus generated can do likewise, and the multiplication factor can become infinite. This process is called *avalanche*.

Let us examine the multiplication. First, we consider the (artificial) condition of constant field within the transition region. Let P be the probability that either a hole or an electron creates an electron-hole pair while it traverses the junction.¹

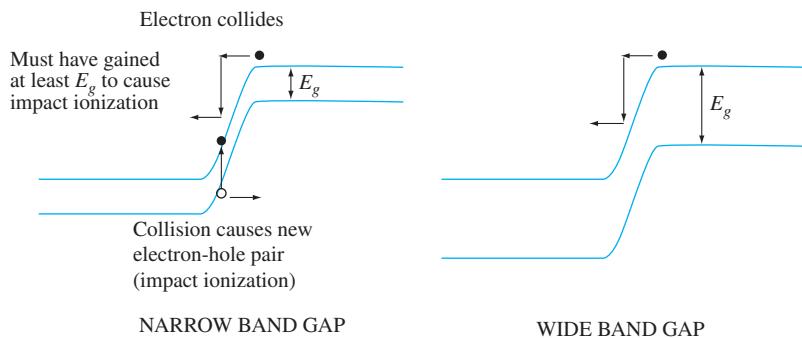


Figure 12.1 For avalanche breakdown to occur, a carrier (electron shown) must be accelerated to a kinetic energy higher than the band gap. This requires a much greater field in the wide band gap material.

¹For simplicity, we assume that the probabilities for holes and electrons are equal.

Let n_{in} be the number of electrons entering the transition region from the p side. Then there will be Pn_{in} ionizing collisions resulting in $n_{in}(1 + P)$ electrons arriving at the n side. But Pn_{in} holes are also generated, which generate $P(Pn_{in}) = P^2n_{in}$ pairs, etc., or the number of total carriers crossing the junction is

$$n_{in}(1 + P + P^2 + P^3 + \dots)$$

This can be expressed as (since $P < 1$)

$$\frac{n_{in}}{1 - P}$$

The carrier multiplication factor, M , is defined as the number of electron-hole pairs that a single carrier can produce:

$$M = \frac{1}{1 - P} \quad (12.1)$$

In an actual pn junction, the field is a function of position in the transition region, and thus so is P . The probability P then becomes $\int_0^{W_D} \alpha dx$, where W_D is the junction depletion region width,² and the impact ionization coefficient α is defined as the number of electron-hole pairs created by a carrier traversing a unit distance in the direction of the electric field in the depletion region. While the impact ionization coefficient is different for electrons and holes, for simplicity, we assumed them to be equal. The carrier multiplication factor M is then

$$M = \frac{1}{1 - \int_0^{W_D} \alpha dx} \quad (12.2)$$

Table 12.1 shows the band gap and critical field for a variety of undoped semiconductors. GaN and SiC are clear winners. These values for critical field are highly dependent on crystal quality, and theoretical values for GaN as high as 5 MV/cm have been predicted. Also shown are the thermal conductivity, electron mobility, and saturation velocity for electrons. SiC has a very high thermal conductivity, so heat

Table 12.1 Some material properties of Si, 4H-SiC, and GaN³

	Band gap E_g (eV) (room temperature)	Critical field \mathcal{E}_{cr} (MV/cm)	Relative permittivity ϵ_r	Thermal conductivity (W/cm·K)	Electron mobility μ_n (cm ² /V·s)	Hole mobility μ_p (cm ² /V·s)	Saturation velocity v_{sat} (10 ⁷ cm/s)
Si	1.124	0.3	11.8	1.3	1330	495	1
4H-SiC	3.26	2.2	10	3.7	1000	115	2.5
GaN	3.437	3.3	9.7	1.3	1000	70	2.5

²Here the term W_D is used for junction depletion width since power diodes designed for high breakdown voltages, for reverse bias the depletion region extends primarily across a lightly doped *drift* layer of width W_D .

³Several values have been reported for hole mobility in GaN. The value used here is near the maximum reported.

can be removed comparatively easily. In practice, however, the heat removal rate and thus the operating temperature in SiC power devices is limited by the packaging.

The breakdown voltage, V_{br} can be calculated as a function of doping level using the impact ionization coefficient of electrons and holes. The impact ionization coefficient is a strong function of band gap and electric field. The power law approximation of the impact ionization coefficient to the electric field is $\alpha \propto \mathcal{E}^n$ where the factor n has been reported to be equal to 7 for Si and 4H-SiC, and equal to 7, 8, and 9 (9.2) for GaN.⁴ For a given p⁺n junction, the impact ionization occurs primarily in the n region near the junction where the field is greatest. Here we use $n = 7$ for the three semiconductors considered. Since the breakdown voltage, V_{br} in a p⁺n junction has the form $V_{\text{br}} \propto N_D^{-\frac{(n-1)}{(n+1)}}$, we use the relation $V_{\text{br}} = C_1 N_D^{-6/8} = C_1 N_D^{-3/4}$ where $C_1 = 5.3 \times 10^{13}$, 3.0×10^{15} , and 6.1×10^{15} respectively for Si, 4H-SiC, and GaN. When using this formula, the n-region doping level N_D is in cm⁻³ and V_{br} is in volts.

Since a power diode is designed to have a specific breakdown voltage, it is convenient to use V_{br} as the independent variable. Thus

$$N_D = \left(\frac{1}{C_1}\right)^{-4/3} V_{\text{br}}^{-4/3} \quad (12.3)$$

Next, we define the critical field, \mathcal{E}_{cr} as the maximum junction field at breakdown. To find an expression for \mathcal{E}_{cr} in terms of the breakdown voltage, we first observe that the applied reverse voltage V_a is

$$V_a = V_j + IR \quad (12.4)$$

where V_j is junction voltage. Then since for reverse voltage, the current I is small, $V_a \approx V_j$, and at breakdown $V_a = V_{\text{br}}$. Thus

$$\mathcal{E}_{\text{cr}} = \sqrt{\frac{2qN_D V_j}{\epsilon_s}} \approx \sqrt{\frac{2qN_D V_{\text{br}}}{\epsilon_s}} \quad (12.5)$$

Similarly

$$W_D = \sqrt{\frac{2\epsilon_s V_j}{qN_D}} \approx \sqrt{\frac{2\epsilon_s V_{\text{br}}}{qN_D}} \quad (12.6)$$

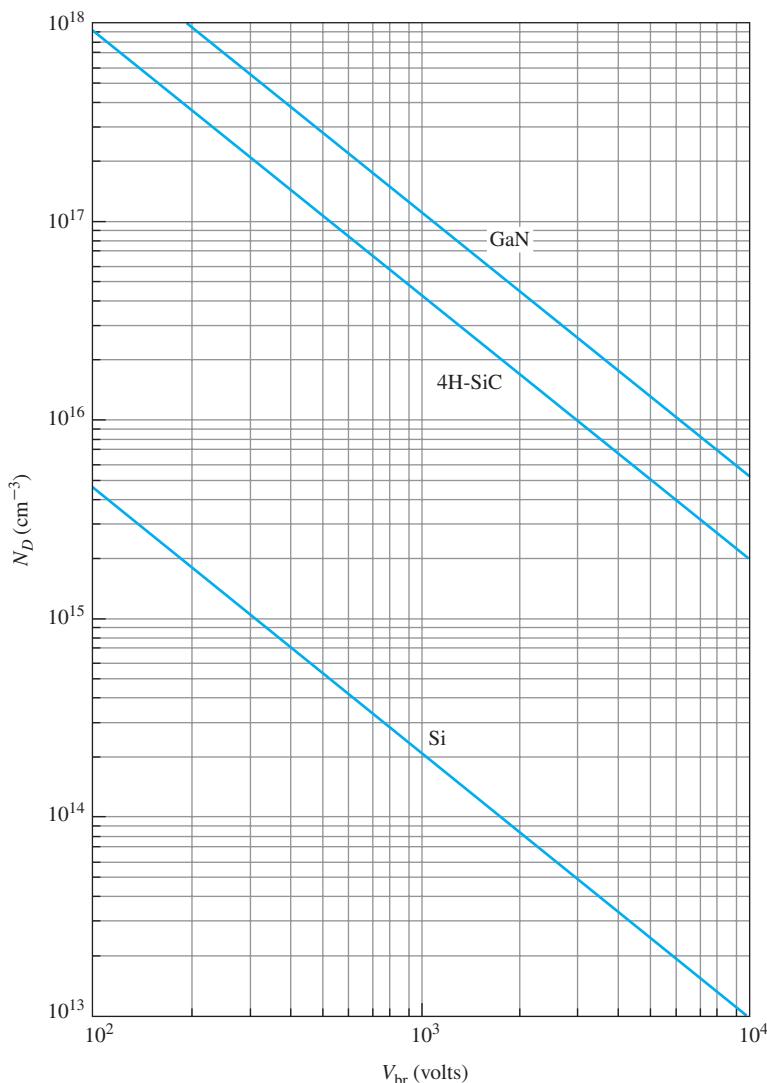
$$V_{\text{br}} \approx \frac{\epsilon_s \mathcal{E}_{\text{cr}}^2}{2qN_D} \approx \frac{qN_D W_D^2}{2\epsilon_s} \quad (12.7)$$

The dependence of N_D , \mathcal{E}_{cr} , and W_D on V_{br} are indicated in Table 12.2 for Si, 4H-SiC, and GaN. In these calculations, the carrier mobility is assumed to be independent of doping level. These results are plotted in Figure 12.2 for breakdown voltages between 100 V and 10 kV. Thus, for a desired breakdown voltage,

⁴The impact ionization coefficient in GaN is not well understood due to the high defect concentration in current material.

Table 12.2 Relationships between V_{br} , N_D , W_D , and \mathcal{E}_{cr}

	Si	4H-SiC	GaN
$N_D(\text{cm}^{-3})$	$2.0 \times 10^{18} V_{br}^{-4/3}$	$4.3 \times 10^{20} V_{br}^{-4/3}$	$1.1 \times 10^{21} V_{br}^{-4/3}$
$W_D (\text{cm})$	$2.6 \times 10^{-6} V_{br}^{7/6}$	$1.1 \times 10^{-7} V_{br}^{7/6}$	$8.8 \times 10^{-8} V_{br}^{7/6}$
$\mathcal{E}_{cr} (\text{V/cm})$	$7.8 \times 10^5 V_{br}^{-1/6}$	$1.3 \times 10^7 V_{br}^{-1/6}$	$2.0 \times 10^7 V_{br}^{-1/6}$

**Figure 12.2 (a)** Donor concentration for an abrupt one-sided p⁺n junction as functions of breakdown voltage at room temperature for Si, 4H-SiC, and GaN.

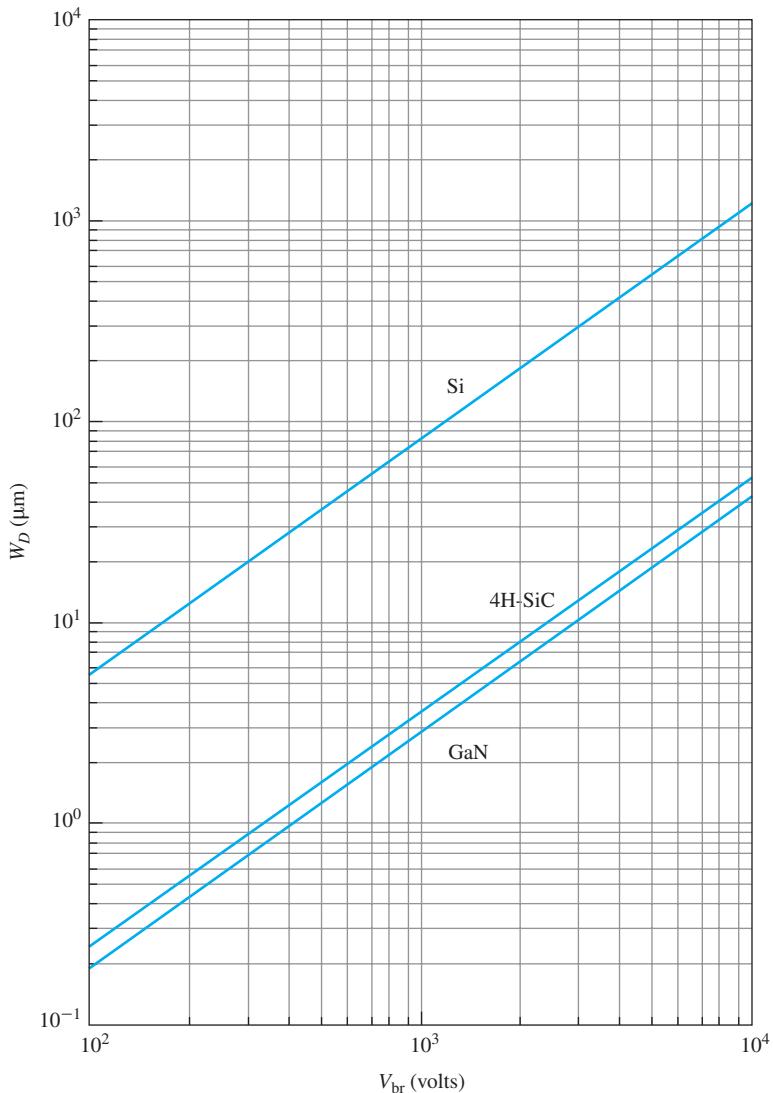


Figure 12.2 (b) Depletion region width for an abrupt one-sided p⁺n junction as functions of breakdown voltage at room temperature for Si, 4H-SiC, and GaN.

the doping level and the width of the depletion region can be determined. The larger the required breakdown voltage, the lighter the doping on the n side and the wider the depletion region.

For power diodes, in addition to having a large reverse breakdown voltage V_{br} , the ability to carry a large forward current is also desired. For both of these reasons, a p-type:intrinsic:n-type (pin) structure or a Schottky diode with an intrinsic

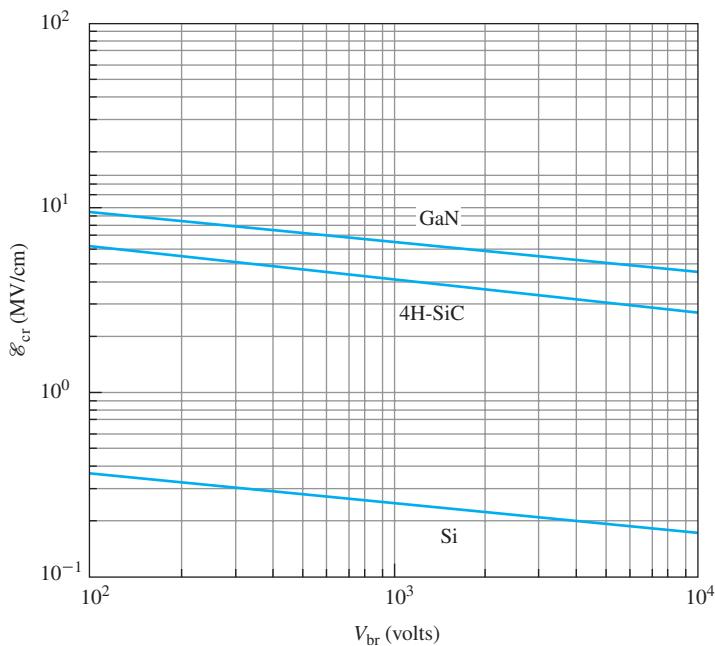


Figure 12.2 (c) Critical field, for an abrupt one-sided p^+ n junction as functions of breakdown voltage at room temperature, for Si, 4H-SiC, and GaN.

layer between the metal and an n-type region is often used for power diodes. In practice, the intrinsic layer is actually lightly doped n type. In a Schottky diode there is a lightly doped n (n^-) region (base) between a heavily doped n (n^+) region and a metal (Figure 12.3a). In a pin junction diode, the lightly doped n^- region is between the p^+ and n^+ regions (Figure 12.3b). The corresponding energy band diagrams at equilibrium are shown in (c) and (d) along with the carrier distribution versus energy functions (sails). At the junction between the n^- region and metal contact (Schottky), and at the p^+n^- junction (pin) there is a depletion region, virtually all in the n^- region.⁵ In the energy band diagrams, for simplicity it is assumed that the Fermi level is at $E_C(n^+)$ and at $E_V(p^+)$.

In both cases, at the n^-n^+ contact, the electron concentration is large and thus of low resistance. The n^-n^+ voltage changes little with forward current for the Schottky diode where current is by electrons. In the pin diode, however, the voltage reduces with increasing current since the current is carried by both electrons and holes, as will be discussed shortly.

⁵Between the n^+ -metal and p^+ -metal regions there are also depletion regions. But because of the heavy doping of the n^+ and p^+ regions, these contacts are nearly short circuits due to electron tunneling, and thus this barrier and the metal contacting the n^+ and p^+ regions are not shown on energy band diagrams.

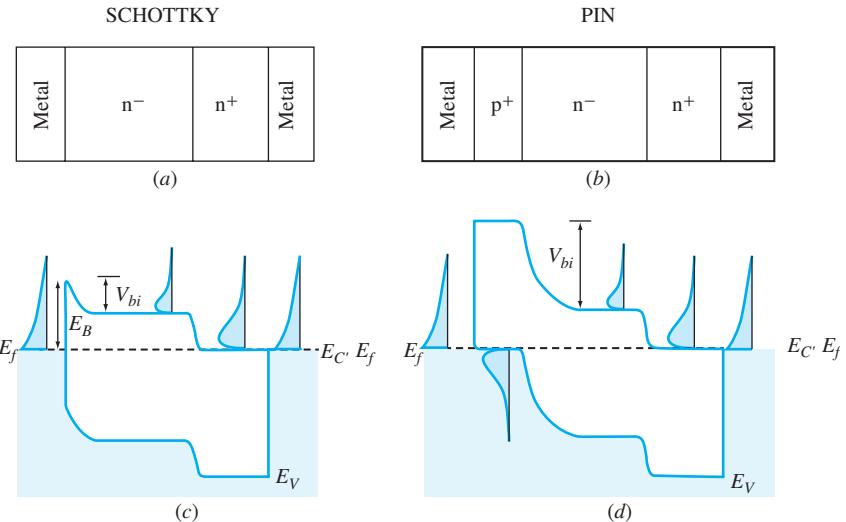


Figure 12.3 Cross section of Schottky (a) and pin (b) diodes and their energy bands at equilibrium, (c) and (d) (not to scale).

Figure 12.4a shows the cross section of a power pin diode and the electric field distribution in the junction for various voltages. In the parallel plate approximation (no edge effects considered), as the reverse bias increases, the peak electric field increases and the field extends further into the depletion region. If the critical field is reached before the depletion reaches the n⁺ region, the diode is said to have a “non-reach-through” design. In non-reach-through diodes, the thickness of the lightly doped layer is chosen so that the depletion region reaches the n⁺ region just as the critical field is reached, or $W_D = W_{cr}$, where W_D is the n⁻ region or drift region width. This region is often referred to as the *base*. The blocking voltage, or reverse breakdown voltage V_{br} , is equal to the area under the electric field curve (recall that $V = -\int \mathcal{E} dx$) when the critical field is reached. Thus for a one-sided p⁺n junction or a Schottky diode,

$$V_{br} = \frac{\mathcal{E}_{cr}(W_D - W_{bi})}{2} \approx \frac{\mathcal{E}_{cr} W_D}{2}$$

Rearranging,

$$W_D = \frac{2V_{br}}{\mathcal{E}_{cr}} = \sqrt{\frac{2\epsilon_s V_{br}}{qN_D}} = \frac{\epsilon_s \mathcal{E}_{cr}}{qN_D} \quad (\text{non-reach-through}) \quad (12.8)$$

Once the applied voltage V_a reaches the avalanche breakdown voltage V_{br} , multiplication occurs, and a large reverse current flows.

Equation 12.8 and Figure 12.2(a) show that the breakdown voltage in the non-reach-through diode depends on the doping level in the lightly doped base

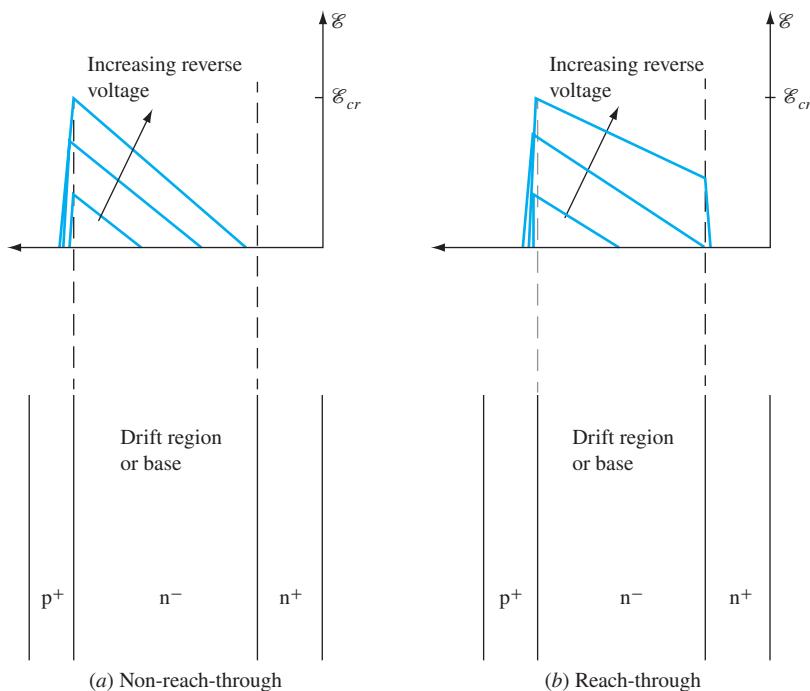


Figure 12.4 (a) Non-reach-through breakdown diode; (b) Reach-through design.

region. For avalanche to occur, the electric field must be greater than the minimum field \mathcal{E}_{\min} required for impact ionization to occur. In fact, the field must be at least this high over a distance of several carrier mean-free paths (ΔW) from the maximum field (\mathcal{E}_{cr}) at the p^+n^- or Schottky junction, so that the carriers have a chance to impact atoms and start the avalanche process. Figure 12.5 shows the fields for two doping levels of the n^- region, N_{D1} (black) and N_{D2} (color), where $N_{D2} > N_{D1}$. Because the field changes faster with position for the device with doping, carriers gain more kinetic energy between collisions. Thus, impact ionization is increased and the width of the region (ΔW) required for avalanche is reduced, or $\Delta W_2 < \Delta W_1$. The fractional reduction of ΔW , however, is less than the fractional reduction of W , and the result is that $\mathcal{E}_{cr2} > \mathcal{E}_{cr1}$, or \mathcal{E}_{cr} increases with increased n^- doping.

The disadvantage of the very thick and lightly doped (n^-) region is a higher forward or “on” resistance. This can create significant heating in power devices as the forward current densities can be large. Thus, a trade-off is sometimes made to make the base (middle, n^-) layer thinner or increase its doping for a smaller on-resistance, but at the expense of reduced breakdown voltage. In this case the depletion region “reaches through” to the highly conductive n layer before the critical field is reached, as shown in Figure 12.4b.

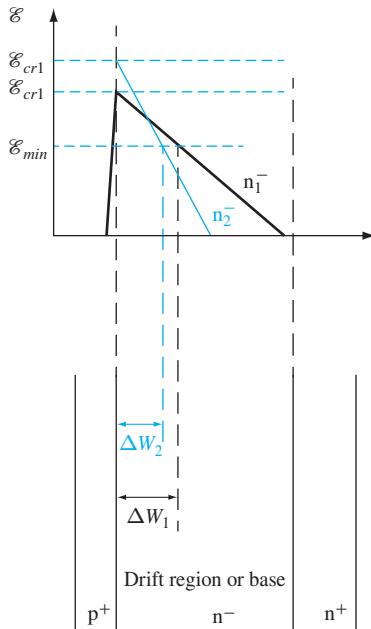


Figure 12.5 Electric field at avalanche (\mathcal{E}_{cr}) as a function of position in the n^- region of a non-reach-through pin diode for two values of N_D , where $N_{D2} > N_{D1}$. For impact ionization to occur over the region ΔW , the critical field increases with increasing doping.

Let W_D be the width of the lightly doped base region, and let W_{cr} be the width corresponding to the critical field. Then if V_{cr} is the breakdown voltage corresponding to \mathcal{E}_{cr} , then

$$\frac{V_{br}}{V_{cr}} = \frac{\int_0^{W_D} \left(\frac{\mathcal{E}_{cr}}{W_{cr}} x + \mathcal{E}_{cr} \right) dx}{\int_0^{W_{cr}} \left(\frac{\mathcal{E}_{cr}}{W_{cr}} x + \mathcal{E}_{cr} \right) dx} = \frac{W_D \mathcal{E}_{cr} - \frac{q N_{D(n^-)} W_D^2}{2 \epsilon_s}}{\frac{1}{2} \mathcal{E}_{cr} W_{cr}} = \left(\frac{W_D}{W_{cr}} \right) \left(2 - \frac{W_D}{W_{cr}} \right) \quad (12.9)$$

(Reach-through design)

We will not consider the reach-through diode further.

The breakdown voltage is calculated using an infinite parallel plane as in Figure 12.4. In reality, the diode areas are finite, and the edges of the diode have to be designed carefully to avoid concentrating the electric field. A guard ring on a pin diode can be used for that purpose. In power diodes, the guard ring is usually

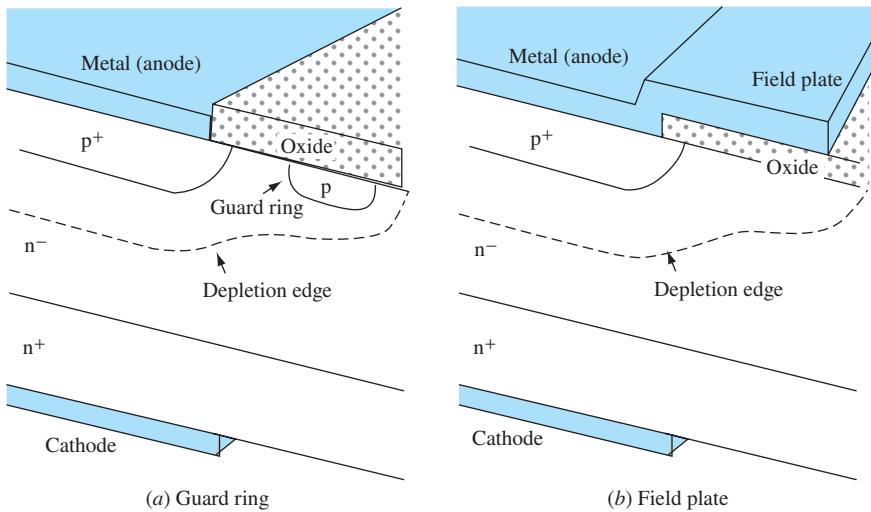


Figure 12.6 To reduce premature breakdown, a guard ring (a) or field plate (b) is used to extend the depletion laterally in planar devices.

a separate p-type ring or rings that are left floating electrically, Figure 12.6a. There is naturally a depletion region around the ring. As the reverse bias on the diode junction increases, the associated depletion region expands, eventually meeting the depletion region around the guard ring, effectively extending it laterally and spreading out the electric field. Another technique is the use of field plates, Figure 12.6b. The field plates can be left floating, or connected to the anode. When connected, as shown in the figure, under forward bias the field plate has negligible effect. Under reverse bias, the negative voltage on the plate repels the electrons in the semiconductor, extending the depletion region.

Power diodes, which need to carry large currents, have large areas, sometimes extending to the edge of the chip or even wafer. At the silicon/air boundary, the electric field lines bow outward, Figure 12.7a. Since the electric field lines are perpendicular to the depletion region boundary, this has the effect of bending the depletion region edge. The electric field lines are closer together near the edge of the wafer, indicating a higher electric field leading to premature breakdown. This can be counteracted by beveling the edge of the wafer, Figure 12.7b. There we see the ionized acceptors in the p⁺ region and the ionized donors in the n⁻ region. To see how this works, imagine that in the beveling process, some of the n⁻ region is removed, and with it some number of positive charges (ionized acceptors). Because the number of negative charges on the p⁺ side of the depletion region has to equal the number of positive charges on the n⁻ side, the depletion region has to be slightly wider near the edge. This bends the field lines apart. With beveling, breakdown voltages near the bulk values can be obtained.

Schottky and Pin Diodes, Reverse Bias Figure 12.8 shows a cross section of a Schottky diode (a). The corresponding energy band diagrams for two values of

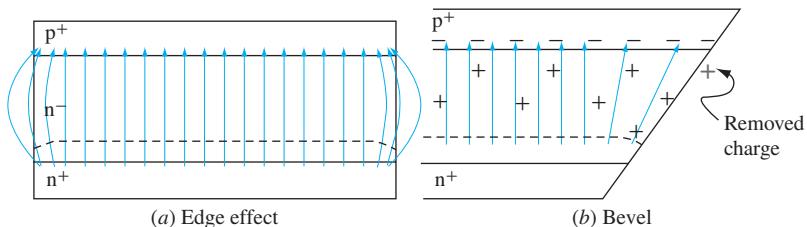


Figure 12.7 (a) When the junction extends all the way to the edge of the semiconductor, the fringing fields tend to concentrate the field lines, increasing the field strength at the edges. (b) A bevel can be used to counteract this effect and allow higher breakdown voltages. The plus and minus signs represent positive and negative charge.

reverse bias are shown in (b) and (c).⁶ In (b) the depletion region extends part way to the n⁺ region. In (c) it extends to the n⁺ region. To minimize the series resistance (important for large currents at forward bias), for a given breakdown voltage, the n⁻ drift region width W_D should be just smaller than that at the critical breakdown, i.e., $W_D < W_{cr}$, and in the calculations it is assumed that $W_D = W_{cr}$.

EXAMPLE 12.1

What is the n⁻ region width W_D for a 4H-SiC Schottky device with a breakdown voltage of 1000 volts?

Solution

From Table 12.2,

$$W_D = 1.1 \times 10^{-7} V_{br}^{7/6} = 1.1 \times 10^{-7} \times (10^3)^{7/6} = 3.4 \times 10^{-4} \text{ cm} = 3.4 \mu\text{m} \quad (12.10)$$

This agrees with Figure 12.2b.

This analysis is also applicable to a pin diode. The difference in V_{bi} makes a negligible difference in the result.

Forward Bias Although the breakdown analyses for Schottky and pin diodes under reverse bias are similar because of low currents, the case for forward bias current is considerably different due to the fact that in the Schottky diode the current is carried by majority carriers while in the pin diode the majority and minority carriers contribute to current.

12.2.2 SPECIFIC ON-RESISTANCE

We saw in the previous section that the breakdown voltage was a function of N_D , which determined the base width, and was independent of device area. The forward current, however, does depend on device area. Thus, rather than discussing

⁶For simplicity, the image-induced barrier lowering effect is neglected.

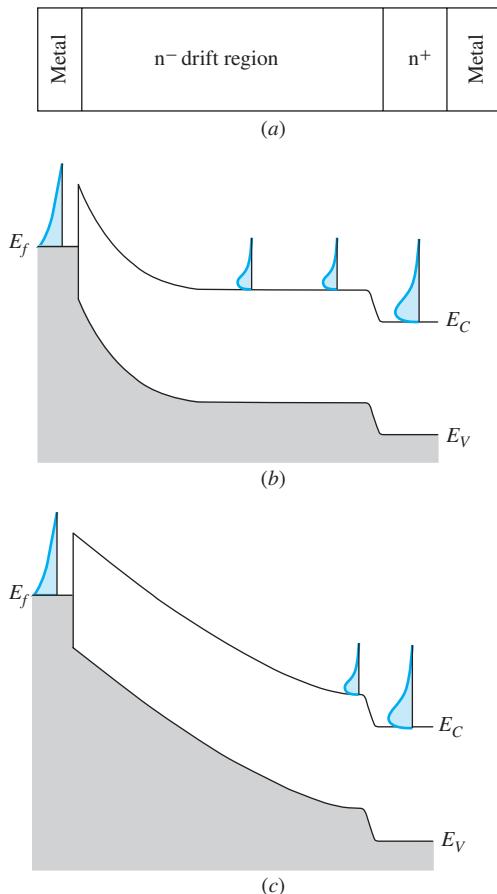


Figure 12.8 Cross section of a Schottky diode
(a) and energy band diagrams for two values of reverse bias.

the $I-V_a$ characteristics, for comparing devices it is convenient to consider the $J - V_a$ characteristics, which are independent of area.

Let V_D be the voltage across the drift region. Since $V_D = IR_D = JAR_D = JR_{\text{on,sp}}$, where R_D is the resistance of the drift region, the term AR_D is defined as the specific on-resistance ($R_{\text{on,sp}}$). The specific on-resistance has units of ohms × area or $\Omega \cdot \text{cm}^2$.

Recall that the resistivity of an n-type semiconductor is

$$\rho = \frac{1}{q\mu_n n} \approx \frac{1}{q\mu_n N_D} \quad (12.11)$$

To find the specific on-resistance, we multiply the resistivity by the length of the region of interest. In power diodes, the large breakdown voltage requires

that the n⁻ drift region width (W_D) be large. For forward bias the depletion region width is a small fraction of W_D . Thus, it is generally approximated that the drift region is equal to W_D . Multiplying the resistivity by the thickness of the drift region W_D gives

$$R_{\text{on,sp}} = \frac{W_D}{q\mu_n N_D} \quad (12.12)$$

But, since $W_D = \frac{2V_{\text{br}}}{\mathcal{E}_{\text{cr}}}$ and $N_D = \frac{\epsilon_s \mathcal{E}_{\text{cr}}^2}{2qV_{\text{br}}}$, we can write

$$R_{\text{on,sp}} = \frac{4V_{\text{br}}^2}{\epsilon_s \mu_n \mathcal{E}_{\text{cr}}^3} \quad (12.13)$$

For a given breakdown voltage, the specific on-resistance varies inversely as the cube of the critical field. Thus wide band-gap semiconductors having a large critical field have a much smaller specific on-resistance, both helpful properties for power devices. Figure 12.9 shows the specific on-resistance as a function of breakdown voltage for Si, 4H-SiC, and GaN Schottky rectifiers. It is seen that 4H-SiC has a specific on-resistance about 2500 times smaller than Si, and GaN nearly half of that of SiC.

We have assumed that the drift region is n type, which is generally the case, in part because of electrons' higher mobility.⁷ This specific on-resistance is just one contributing factor to the overall forward resistance of the diode, but it is a useful figure of merit for comparing devices.

This analysis applies to both Schottky and pin diodes for low currents. At higher current levels, it continues to apply to Schottky diodes. In a pin diode, however, under high injection such as is easily encountered in power devices, the number of minority carriers injected can be substantial, orders of magnitude higher than the doping concentration, thus reducing $R_{\text{on,sp}}$ considerably (conductivity modulation). In the next sections we will compare the forward bias characteristics of Schottky and pin diodes.

J-V_a Characteristics: Schottky Diodes Here we discuss the forward bias J-V_a characteristics of Schottky diodes. The characteristics for pin diodes are somewhat different and are discussed later.

The energy band diagram for the Schottky diode is shown in Figure 12.10 at equilibrium (a), at a small forward bias (b), and at large forward bias (c). The junction voltage V_j and built-in voltage V_{bi} at equilibrium are indicated.

At small applied voltage such that the barrier voltage V_j is still larger than $3kT/q$, the carrier concentration in the base region, at energies above the barrier, varies exponentially with energy, and

⁷In 4H-SiC $\mu_n = 1000 \text{ cm}^2/\text{Vs}$ and $\mu_p = 115 \text{ cm}^2/\text{Vs}$

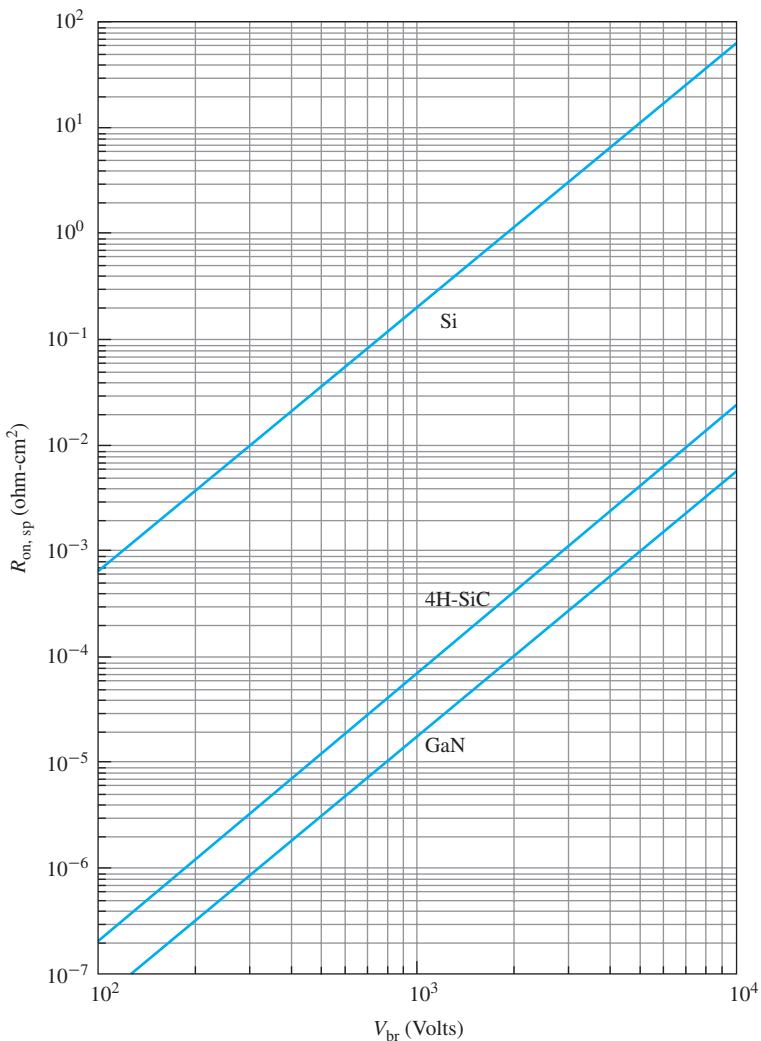


Figure 12.9 Room temperature specific on-resistance as a function of breakdown voltage for Si, 4H-SiC and GaN Schottky barrier diodes.

$$J = J_0 \left(e^{\frac{q(V_{bi} - V_j)}{kT}} - 1 \right) = J_0 \left(e^{\frac{q(V_a - JR_{on,sp})}{kT}} - 1 \right) \quad (12.14)$$

where the voltage across the junction $V_j = V_a - JR_{on,sp}$, and V_{bi} is the built-in voltage.

For larger applied voltage, the junction voltage is reduced to $V_j < 3kT/q$, and we see that the carrier variation with energy above the barrier is less than exponential, so now J increases less rapidly with V_j . This effect, along with the $JR_{on,sp}$ voltage drop, causes the current to increase less rapidly with V_a . In the limit of $V_j = 0$, $V_a = V_{bi} + JR_{on,sp}$, or

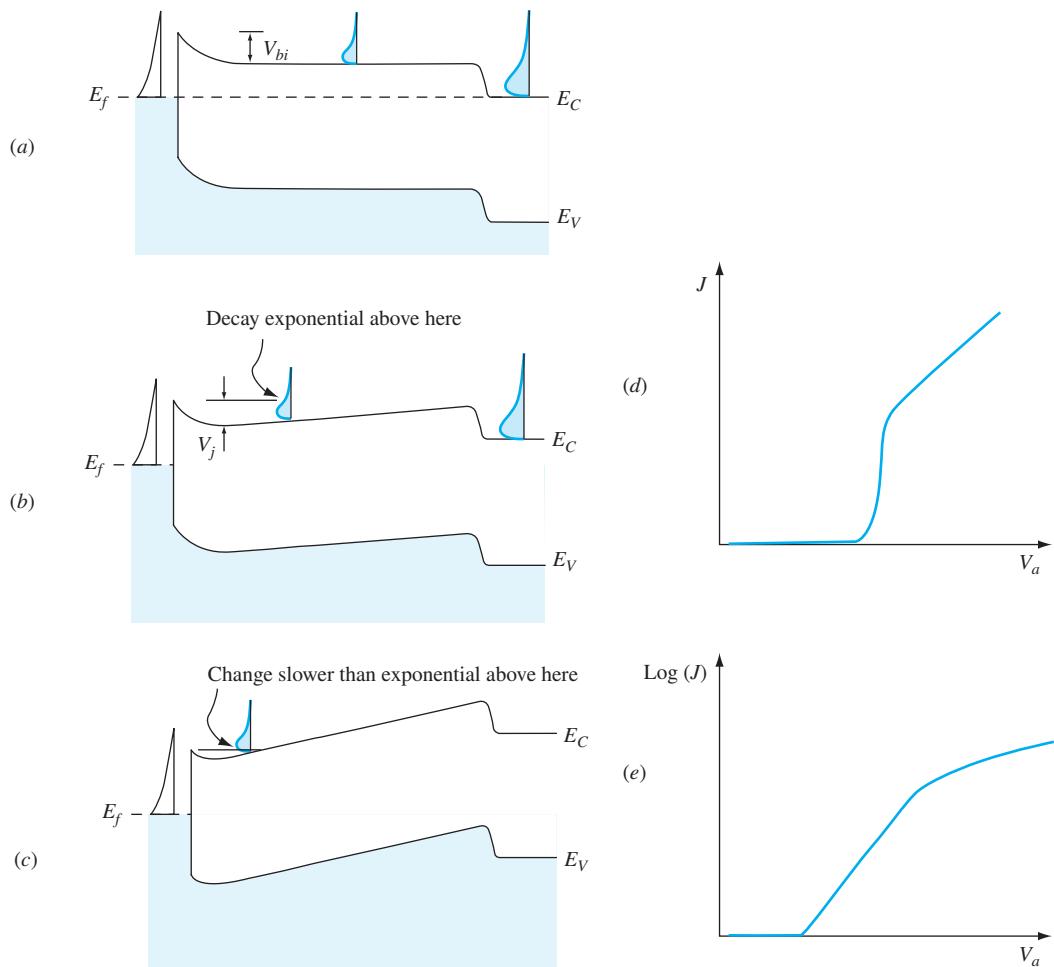


Figure 12.10 Energy band diagram of a Schottky diode at equilibrium (a), at small forward bias (b), and at large forward bias (c). The J - V_a curves are shown on a linear scale (d) and a log scale (e).

$$J = (V_a - V_{bi})/R_{on,sp} \quad (12.15)$$

For large V_a , such that $V_a \sim V_{bi}$,

$$J = V_a/R_{on,sp} \quad (12.16)$$

The forward bias J - V_a characteristics of a Schottky diode are shown in Figure 12.10 on a linear scale (d) and on a semi-log scale in (e). Because of the large range of currents in power devices, it is convenient to plot $\log J$ vs. V_a .⁸

⁸This is especially useful for discussion of the merged pin–Schottky diodes discussed in Section 12.2.4.

J-V_a Characteristics: Bipolar Pin Diodes Next we examine the forward characteristics for the bipolar diode case. Energy band diagrams for a pin diode are shown in Figure 12.11 at equilibrium (a), at an intermediate bias (b), and at high bias (c). From (a) it is seen that the barriers for electrons and holes in the n⁻p⁺ junction are equal. But since the hole concentration in the p⁺ region is much larger than the electron concentration in the n⁻ region, for a small forward bias, current is primarily by holes injected into the n⁻ base. This positive charge tends to lower the n⁻ bands from their equilibrium energy, which reduces the n⁺n⁻ barrier. This results in electrons being injected into the n⁻ conduction band. Because of the $JR_{on,sp}$ drop in the drift region, however, the n⁻ bands are tilted slightly, which accelerates electrons left and holes right in the diagram of Figure 12.11b.

Under very high injection such as is easily encountered in power devices (c), the number of minority carriers injected into the n⁻ region can be substantial, orders of magnitude higher than the doping concentration. Under high injection, both n and p in the drift region are large, increasing the conductivity. This effect is called *conductivity modulation*.

The steady-state carrier concentration, ($n+p$) as a function of position is shown in Figure 12.12. In the steady state, the carrier recombination is supplied by carrier injection from the end contacts. Because the electron concentration in the p⁺ region is much smaller than in the n⁻ region and the hole concentration in the n⁺ region is much less than in the n⁻ region, recombination in the p⁺ and n⁺ regions contributes little to the current, and the approximation is made that the current results entirely from recombination in the drift region. This is a good approximation for $W_D > 3L$, where L is the carrier diffusion length.

From Equation 12.12, $R_{on,sp} = \frac{W_D}{q\mu N_D} = \frac{W_D}{q\mu_n n}$ where n is the electron concentration in the drift region under low injection. The specific resistance in the drift region of a pin junction in the high injection region is given by

$$R_{on,sp} = \frac{W_D}{q(\mu_n + \mu_p)n_a} \quad (12.17)$$

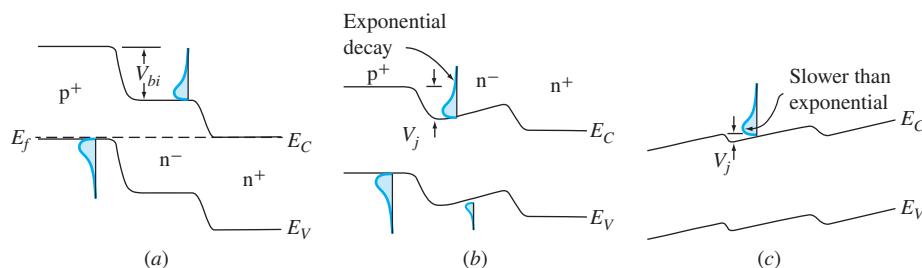


Figure 12.11 Energy band diagrams for a pin diode at equilibrium (a), at moderate forward bias (b), and at high forward bias (c).

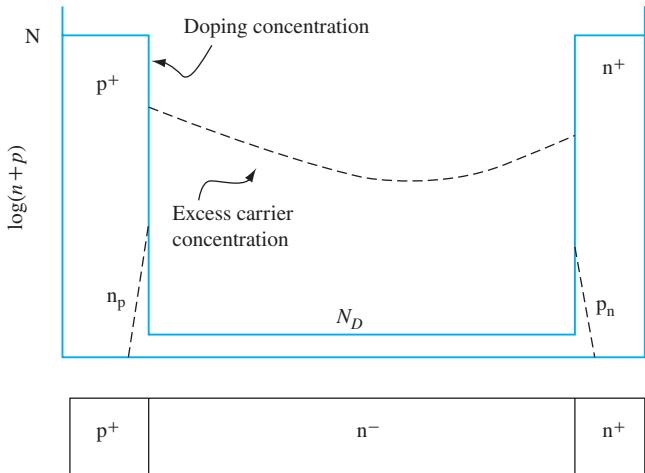


Figure 12.12 Steady-state carrier concentration in a pin diode under forward bias.

where n_a is the carrier density averaged across the drift region (we assume $n \approx p$ as well).

The recombination rate in the drift region of the pin diode is

$$r \approx n_a / \tau_{HL} \quad (12.18)$$

where τ_{HL} is the carrier lifetime under high injection, and we have used a lower-case r for the recombination rate to avoid confusing it with R for resistance. In steady state, the recombination is supplied by the injection of carriers from the ends, so the total current density J is

$$J = \int_0^{W_D} q r dx = \frac{q n_a W_D}{\tau_{HL}} \quad (12.19)$$

where we have used Equation (12.18).

The voltage across the drift region is $V_D = IR_D = JR_{on,sp}$. Multiplying the current density J by the specific resistance $R_{on,sp}$, we obtain

$$V_D = J \cdot R_{on,sp} = \left(\frac{q n_a W_D}{\tau_{HL}} \right) \left(\frac{W_D}{q(\mu_n + \mu_p)n_a} \right) = \frac{W_D^2}{\tau_{HL}(\mu_n + \mu_p)} \quad (12.20)$$

This indicates that at high currents, with conductivity modulation the voltage V_D dropped across the drift region is independent of the current density,

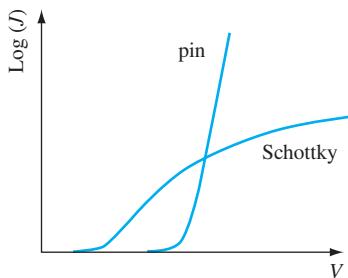


Figure 12.13 Comparison of pin and Schottky diode $J-V_a$ characteristics.

unlike the Schottky diode.⁹ Thus, while the Schottky diode has a lower turn-on voltage, at high current levels the voltage drop across the pin diode is lower, as shown in Figure 12.13. The curves include the voltage across the drift region and the forward voltage of the junctions.

EXAMPLE 12.2

Find V_D , the voltage across the drift region, for a 4H-SiC power pin diode having a breakdown voltage of 1.5 kV and carrier lifetime of 5 ns.

■ **Solution**

$$\text{From Eq. 12.20, } V_D = \frac{W_D^2}{\tau_{HL}(\mu_n + \mu_p)}$$

From Table 12.2, $W_D = 1.1 \times 10^{-7} V_{br}^{7/6}$, and

$$V_D = \frac{(5.6 \times 10^{-4})^2 \text{ cm}^2}{5 \times 10^{-9} \text{ s} \cdot (1000 + 115) \text{ cm}^2/\text{V} \cdot \text{s}} = 5.6 \times 10^{-2} \text{ V} = 56 \text{ mV}$$

In the Schottky diode, there is no conductivity modulation because there is no injection of minority carriers. As the voltage increases, the current increases, but not because the carrier density increases—rather because the carriers in the drift region are moving faster. Because a thick drift region needed for high-voltage blocking is necessarily more resistive than a thinner one, Schottky diodes (in silicon) typically are designed for blocking voltages under a hundred volts. The resistive drift region also accounts for the deviation from exponential of the $I-V$ characteristics at higher currents for the Schottky diode.

⁹The voltage across the drift region does increase slightly with increasing current due to two effects not considered in this analysis: First, at very high current densities the carrier lifetime decreases due to increased recombination, and second, the carrier mobilities decrease due to carrier-carrier scattering.

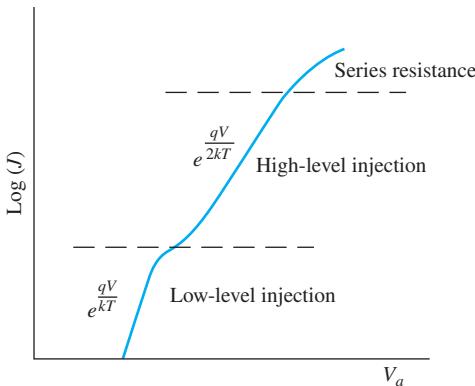


Figure 12.14 J-V_a characteristics indicating the onset of conductivity modulation and the effect of the device ohmic resistance.

The forward J-V_a characteristics of a pin diode are shown in Figure 12.14. At low currents such that the injected hole density is much less than the electron concentration ($N_D(n^-)$) in the drift region, the current density varies as $e^{\frac{q(V_a - JR_{on,sp})}{kT}}$ where $R_{on,sp} = \frac{W_D}{q\mu_n N_D}$. The conductivity modulation is negligible. As the current increases such that the injected hole concentration exceeds $N_D(n^-)$, the diode enters the high-injection region and the current density varies as $e^{\frac{q(V_a - JR_{on,sp})}{2kT}}$, where because of conductivity modulation, $R_{on,sp} = \frac{W_D}{q(\mu_n + \mu_p)n_a}$ and the voltage across the drift region is $V_D = JR_{on,sp} = \frac{W_D}{\tau_{HL}(\mu_n + \mu_p)}$. Note that here V_D is independent of current.

At very high bias, the rate of current increase decreases with applied voltage as indicated in Figure 12.14. This is explained with the aid of Figure 12.11. At large V_a , the voltages across the junctions approach zero and the carrier density-energy functions are less than exponential at the junctions. Therefore the rate of carrier injection across the junctions decreases with changing voltage. At this large bias the barriers to electrons and to holes are small such that electrons and holes from the drift region are injected into the p⁺ and n⁺ regions respectively. This reduces the carrier concentration, n_a , in the drift region, which increases the resistance of the drift region. Therefore the voltage drop across the drift region increases as the current increases. In the limit of zero V_j , the current $I = V_a/R$ or $J = V_a/R_{on,sp}$, where R is the sum of the resistances in the p⁺, n⁻, and n⁺ regions.

12.2.3 TRANSIENT LOSSES

Power diodes control the direction of current flow in circuits used in various applications. Part of the time they operate in the reverse (**off** or blocking) state and part of the time in the forward (**on** or conducting) state. The transition of the diode from

off to **on** results in a transient overshoot of device voltage before settling down to its steady-state value. This is referred to as *forward recovery*. In switching from **on** to **off**, the charge stored in the drift region's excess carrier concentration must be extracted before reaching the **off** state, producing a large transient current. This process is referred to as *reverse recovery*. In power electronics it is common to have an inductive load (e.g., an electric motor). Since the inductor voltage is $V = L(dI/dt)$, this can have a significant influence on turn-on and turn-off times.

Turn-On Transient (Forward Recovery) During the time a pin diode is being switched from reverse (**off**) to forward (**on**) bias, the voltage across the diode can be much larger than the steady-state (**on**) voltage. Figure 12.15 shows the current waveform and the resulting voltage across a pin diode. The diode starts in the reverse bias state. There is a wide depletion region containing a large number of fixed charges (ionized donors). When the applied voltage turns positive, current begins to flow. Because of the finite rate for the diffusion of minority carriers, the n^- drift region requires some time to become populated. Thus, in this time interval only a portion of the fixed charges (donors) are neutralized. If the current increases rapidly, a portion of the base has a high resistance. This *IR* drop can generate a significant transient forward voltage, and it can be compounded by voltages generated in the inductances $V = L(dI/dt)$ if the switching rate dI/dt is high.

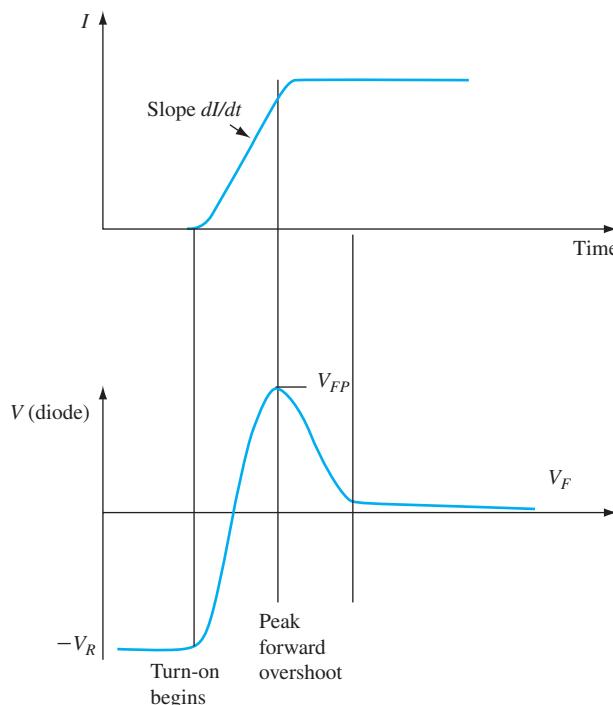


Figure 12.15 Turn-on transient in a power pin diode.

As the injected carrier concentration builds up in the drift region, the resistance decreases due to conductivity modulation, and eventually the diode settles down to its steady-state value V_F . There will thus be a trade-off between reducing switching time (i.e., the rate of change of current with time, dI/dt) and reducing the forward peak voltage V_{FP} . Note that this forward voltage overshoot, which can be tens or even hundreds of volts in power diodes, is normally negligible in signal diodes, since those typically have small depletion regions and thus little on-resistance, and they carry nowhere near as much current. In power pin diodes, the forward voltage overshoot multiplied by the large forward current results in power loss every time the diode is switched **on**. We will see in the next section that the power losses associated with turn-off are actually much greater.

The turn-on overshoot for a Schottky barrier diode is much less than for a pin diode. In practice, discrete Schottky diodes are typically doped more heavily in the drift region than a comparable bipolar diode, to keep the forward resistance down.

Turn-Off Transient and Reverse Recovery An even more serious problem can occur when a power rectifier is switched from conducting to blocking. When a diode's voltage is switching positive to negative, there is a pulse of negative current as excess minority carriers are removed from the depletion region. For signal diodes, the diode can be considered to be in series with a resistor, but in power circuits we have to take inductance into account. As in turn-on, when the applied voltage is changed instantaneously, a current flows through the inductance at a constant rate of change dI/dt .

Suppose the applied voltage is switched negative at $t = t_0$, Figure 12.16. First, the charge stored in the excess carrier concentration in the drift region has to be removed. Until the excess carriers are gone, the junctions are still forward-biased. Some carriers diffuse back into the n^+ and p^+ regions, and others recombine in the drift region. The current direction is positive as shown in the figure. When the excess carriers are removed, the current at that instant is zero. Now the depletion regions form. Electrons are removed toward the n^+ region and holes toward the p^+ region, creating a current in the reverse direction during time t_1 . Note that the rate of current change is still dI/dt , and the slope is negative. This creates a negative induced voltage in the circuit inductance, which reduces the magnitude of the voltage across the diode. When the junction voltage is equal to the applied reverse voltage, the depletion region is fully established. The reverse current has reached its maximum, $L(dI/dt) = 0$, and the diode voltage is equal to the applied voltage. The rate of change of the current is zero, but the current itself is not zero. As it begins to decay to its equilibrium value, the sign of $L(dI/dt)$ changes, creating a voltage in the circuit that *increases* the magnitude of the voltage across the diode, creating a voltage overshoot.

A parameter called “snappiness,” sometimes called a “soft” factor, is defined to describe how quickly the reverse characteristic “snaps” back to V_R after the overshoot. It is given by

$$S = \frac{t_2}{t_1} \quad (12.21)$$

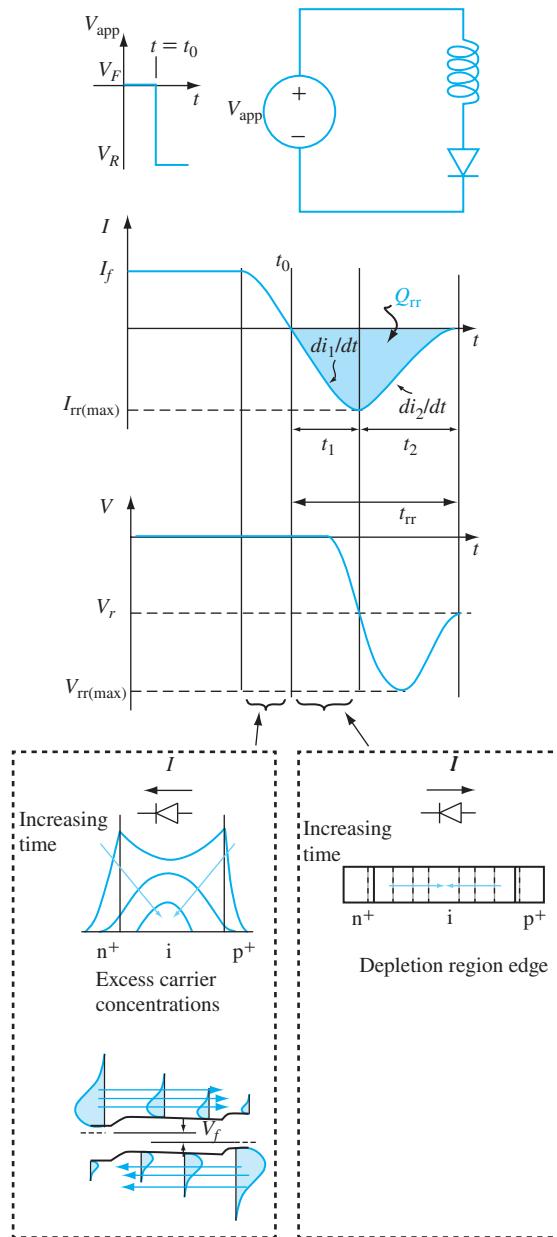


Figure 12.16 Reverse recovery waveform in a bipolar power diode.

Recalling that $Q = \int Idt$, the amount of charge Q_{rr} that must be moved during the reverse recovery is the area under the current curve, as shown in Figure 12.16. If we define $t_{rr} = t_1 + t_2$, the amount of charge can be estimated as

$$Q_{rr} \approx \frac{1}{2} I_{rr(\max)} t_{rr} \quad (12.22)$$

For abrupt junction non-reach-through diodes, the reverse recovery time is approximately [3]

$$t_{rr} \approx \frac{2 V_{br}}{\mathcal{E}_{cr}} \sqrt{\frac{2qI_F}{kT(\mu_n + \mu_p) \frac{di_1}{dt}}} \quad (12.23)$$

$$I_{rr(\max)} \approx \frac{2 V_{br}}{\mathcal{E}_{cr}} \sqrt{\frac{2qI_F (di_1/dt)}{kT(\mu_n + \mu_p)}} \quad (12.24)$$

if recombination in the drift region is neglected. Since the peak current $I_{rr(\max)}$ is the slope di_1/dt times the time t_1 , combining (12.22) and (12.23) we have

$$Q_{rr} = \frac{di_1}{dt} \frac{t_{rr}^2}{2(S+1)} \quad (12.25)$$

In the bipolar pin diode, the switching time is comparatively slow because the excess carriers in the drift region have to be injected (by diffusion) in the case of turn-on, and have to be extracted (again by diffusion) or recombine during turn-off. Often the semiconductor is doped with recombination centers (e.g., Au or Cu) to shorten the carrier lifetime and improve switching speed. A trade-off is that the recombination centers also act as generation centers and can increase leakage currents.

Let us compare the turn-off transient for the pin diode to that of a Schottky power diode. For Schottky diodes, there is only one type of excess carrier in the drift region (majority carriers), and they are supplied by thermionic emission. Figure 12.17 shows the transient waveforms (a) and the energy band diagram under forward bias (b). When the applied voltage switches, the current starts toward zero but cannot switch instantaneously, again because of circuit inductance. As the current decreases, fewer electrons are flowing over the barrier (c). Eventually the drift region is empty of excess carriers and fully depleted. As a result, there is little or no reverse recovery current.

To summarize pin versus Schottky power rectifiers, Schottky diodes typically have lower turn-on voltage and faster switching speeds, but they can also have a higher specific on-resistance. As the blocking voltage increases, the on-resistance of the Schottky diode becomes excessive and for very high blocking voltages, pin diodes are less resistive. Schottky diodes generally have greater power losses during forward conduction and higher reverse leakage currents, but pin diodes have greater power losses during transients.

Silicon Schottky diodes are commercially available with breakdown voltages in the neighborhood of 1500 V but are more commonly found in the <100 V region, while 4H-SiC and GaN pin and Schottky rectifiers can have blocking voltages exceeding 10 kV.

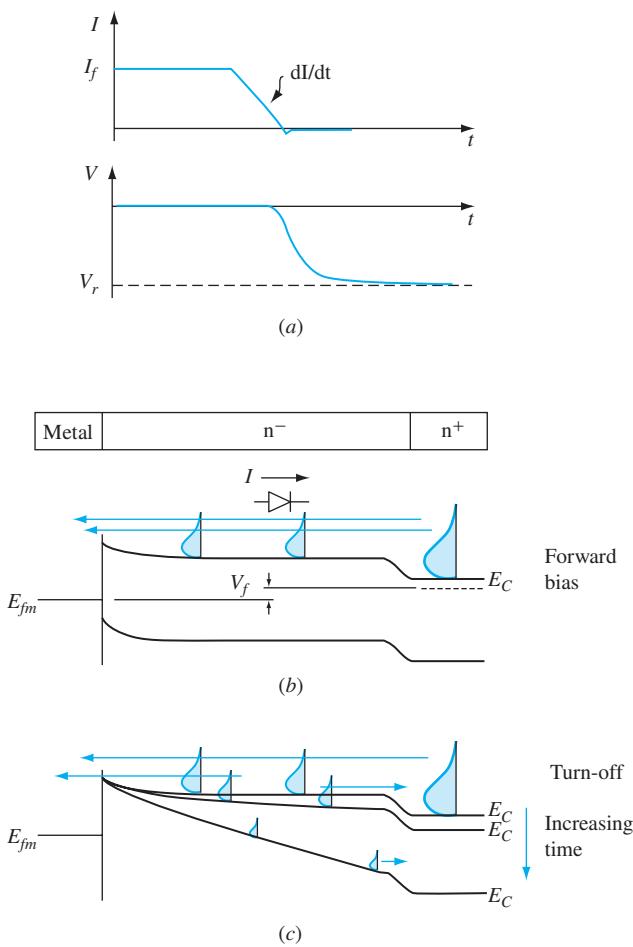


Figure 12.17 Turn-off waveforms in a Schottky barrier diode. There is little or no reverse recovery (a). Under forward bias (b), electrons are thermionically emitted over the barrier. When the applied voltage goes negative, the circuit inductance produces a dI/dt and the finite rate of change means the current cannot change instantaneously. The current is still positive but reduces to zero (c). As the current decreases, the number of carriers emitted over the barrier decreases. Excess electrons in the drift region are also swept out.

12.2.4 MERGED PIN-SCHOTTKY (MPS) DIODES

A technique for obtaining the “best of both worlds” for the forward bias characteristics is the merged pin–Schottky (MPS) diode in Figure 12.18a. Here there are Schottky and pin structures periodically arrayed throughout the device. Figure 12.18b shows the calculated $J-V_a$ relation of a Si non-reach-through MPS

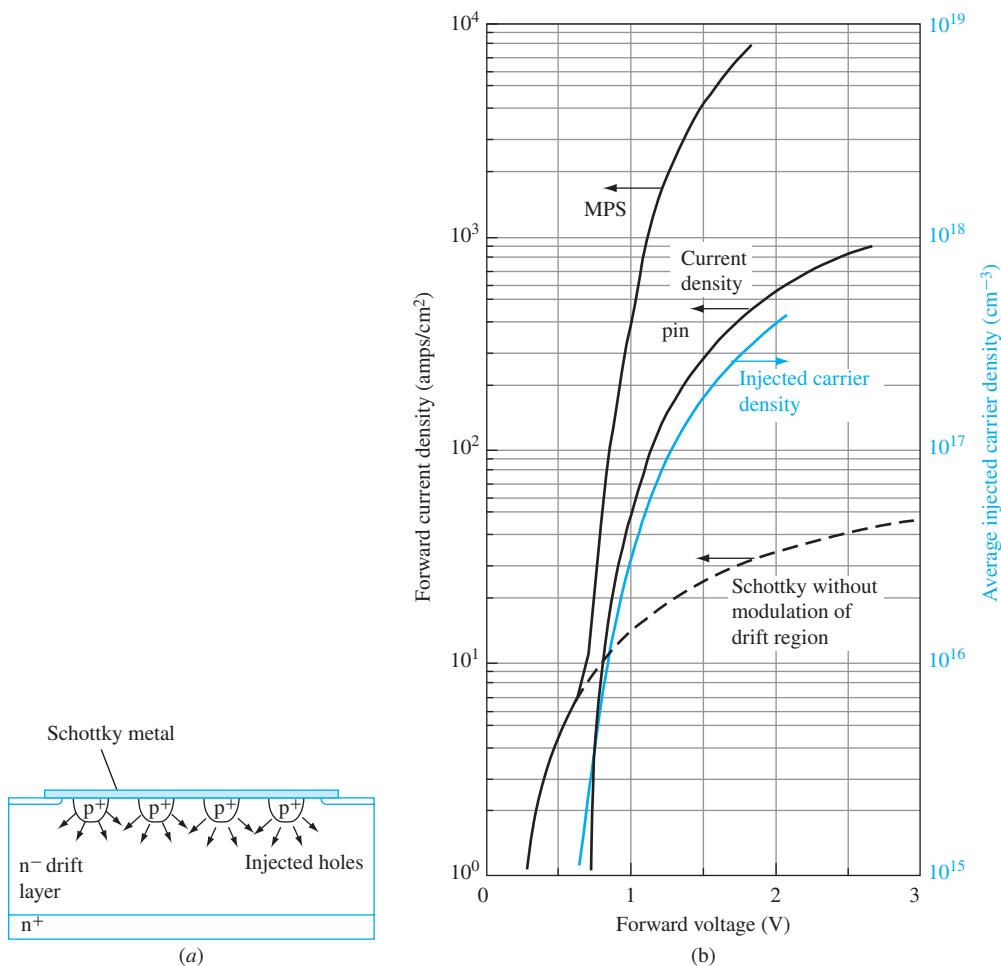


Figure 12.18 (a) A merged pin-Schottky power diode. The injected holes from the pin reduce the resistance for both pin and Schottky devices. (b) Calculated $J-V_a$ characteristics of a Si MPS rectifier. The characteristics of the Schottky, pin, and merged PIN device are indicated. (Adapted from [4]).

rectifier having a reverse breakdown voltage of 400 V.¹⁰ For small forward bias, the pin diode is not turned **on**, but the lower built-in voltage of the Schottky barriers allows current to flow. As the forward voltage increases, the p^+n^- junctions turn **on**, the injected carrier density increases, which reduces the series resistance (and $R_{on,sp}$) of the base regions due to conductivity modulation. Because of the proximity of the pin and Schottky structures, holes injected into the base from

¹⁰The current density in Figure 12.18(b) is that of the entire MPS. Since the Schottky and pin devices have equal areas, the area of each device is halved.

the pin structure reduce the n⁻ base resistance (and thus $R_{on,sp}$) of both pin and Schottky structures, and thus the voltage across the base is independent of current. Since the built-in voltage of the Schottky is less than that of the pin device, and both will have equal $R_{on,sp}$ after the pin turns on, the Schottky will always have greater current than the pin, typically about an order of magnitude greater. It can be seen that for high forward currents, the applied voltage is considerably higher than the built-in voltage of either device, so most of the voltage is dropped across the series resistance. The same current could be obtained by simply using a resistor. A resistor, however, would not provide the required high reverse blocking voltage. Because the pin current is much lower than the total current, for a given forward current, the stored charge and the switching speed are an order of magnitude smaller than those of a pin device.

12.3 THYRISTORS (npnp SWITCHING DEVICES)

Next we discuss multiple-junction bipolar devices. While power bipolar transistors (BJTs) with three layers (npn or pnp) and two interacting junctions exist, they are not widely used since the introduction of better devices (see IGBTs later). Here, however, we consider devices with three interacting pn junctions. The general name for this class of device is *thyristor*. First we discuss the four-layer diode switch, which, like a diode, has two leads (anode and cathode) and conducts and blocks current depending on the signal applied. The silicon-controlled rectifier (SCR) has the same four-layer structure, but a third contact is added (the gate) to allow the device to be switched **on** by a control signal applied to the gate. Usually when people refer to “thyristors” they are talking about the SCR. The SCR can be turned on by the external control signal but cannot be turned off except by the signal being switched. A variation on the SCR, called the gate turn-off thyristor (GTO), is modified such that it can be turned on *and* off by the gate signal. A semiconductor-controlled switch (SCS) can also be turned on and off; it has two control contacts.

12.3.1 THE FOUR-LAYER DIODE SWITCH

Figure 12.19a shows the structure of the simplest thyristor, an npnp diode. Here the intermediate p and n floating layers are relatively thin—appreciably less than a minority carrier diffusion length. We call the top terminal (the p layer) the anode (*A*), and the bottom n layer is the cathode (*K*). The I_A - V_{AK} characteristics of such a diode are indicated in Figure 12.19b. We use I_A for the anode current while V_{AK} is the anode voltage with respect to the cathode. Note this characteristic is distinctly different from that of a two-layer diode. The *I-V* characteristics look similar in the reverse region, but in the forward region there is a forward breakdown, whose character is different from that of a usual junction breakdown. The forward-bias breakdown voltage V_{BF} and the reverse-bias breakdown voltage V_{BR} are indicated.

Using this I_A - V_{AK} characteristic, we first discuss the circuit behavior of this switch, and we then describe the physics of its operation.

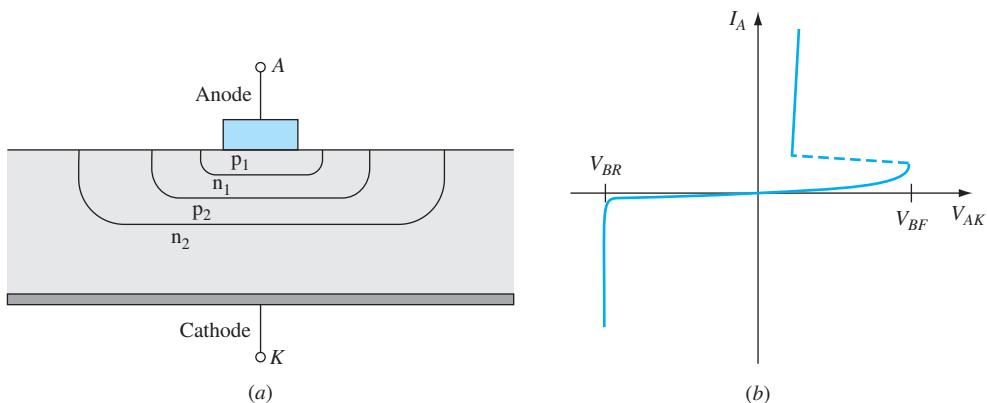


Figure 12.19 (a) The cross section of an npnp four-layer device; (b) the I_A - V_{AK} characteristic.

Figure 12.20a indicates an npnp diode circuit with an adjustable supply voltage V_S and a load resistance R_L . The I_A - V_{AK} characteristic and its load lines are shown in Figure 12.20b for V_S increasing from zero. For $V_S = 0$, $I_A = 0$. As V_S increases, the current is given by the intersection of the I_A - V_{AK} characteristic and the instantaneous load line. Note that for V_{S2} and V_{S3} there are three intersections; however, since V_S is increasing from zero, the operating point is that indicated on the lower branch, since there is no way to change to another current value.

For V_{S4} , however, the load line intersects the I_A - V_{AK} characteristic only at one point—at high current and low voltage. Thus for $V_S = V_{S4}$ the diode switches from a high-voltage, low-current (**off**) state to a low-voltage, high-current (**on**) state. For $V_S > V_{S4}$, the device operates on the upper branch.

Once the device is in the **on** state, as V_S is reduced, the operating point follows the path shown in Figure 12.20c. It will remain on until I_A becomes smaller than the *holding current* I_H , and will then switch to the **off** state. If V_S is then increased again, the characteristics will follow the lower branch until V_S exceeds V_{S4} as indicated in Figure 12.20b and the device switches again from **off** to **on**.

Note that this switching results from a negative differential resistance (negative slope) in the I_A - V_{AK} characteristics. There is no such negative resistance for V_{AK} negative and so, for this polarity of V_S , no switching occurs.

Now let us explain (qualitatively) the npnp device characteristic with the use of energy band diagrams. Figure 12.21a shows the energy band diagram for V_S positive while Figure 12.21c is for negative V_S . We will come back to part (b) of the figure.

For positive V_S there are two forward-biased pn junctions and one reverse-biased junction. Because the n₁ and p₂ layers are thin, holes injected into the narrow n₁ region from the p₁ region are collected in the narrow p₂ region. Since p₂ is electrically floating, it becomes positive from the collected holes. Positive potential for electrons means lower energy, so the p₂ section moves down on the energy band diagram. This reduces the p₂n₂ barrier, allowing more electrons to

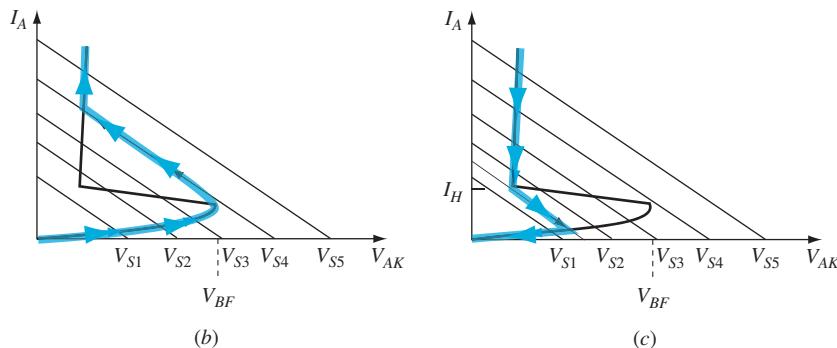
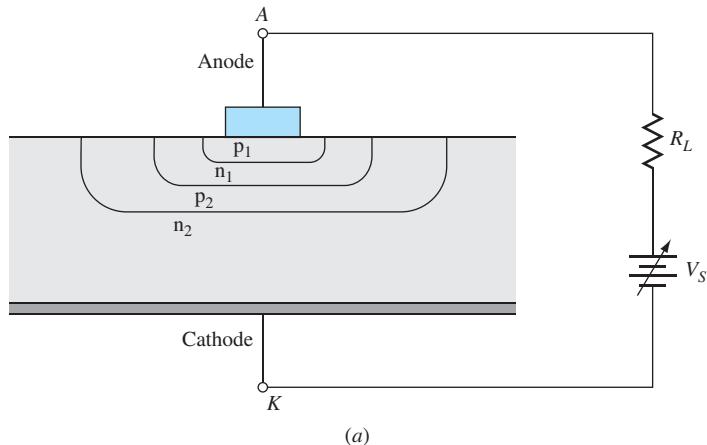


Figure 12.20 The npnp four-layer diode with R_L load line in operation. (a) A varying voltage is applied from anode to cathode. (b) As V_S increases from zero, the voltage across the device increases up to V_{BF} . Then the device switches to a low-voltage, high-current state. (c) When the V_s is decreased, the operating point follows the path shown, remaining in a low-voltage, high-current state until the holding current I_H is reached.

be injected from n_2 into p_2 . These electrons are then collected by n_1 , causing it to become negative, effectively moving it upward on the diagram. That reduces the p_1n_1 barrier, which causes more holes to be injected from p_1 into n_1 , and collected by p_2 . When the voltage V_S is low, this regenerative effect continues until the injection current in the forward-biased junctions is just equal to the recombination current (device **off**), Figure 12.21a. Thus, the overall current is limited to some small value. When V_S increases above some amount, the feedback causes the current to increase very rapidly, and the current is limited by the $I_A R_L$ drop in the circuit. This is the situation indicated in part (b). Here all junctions are forward biased.

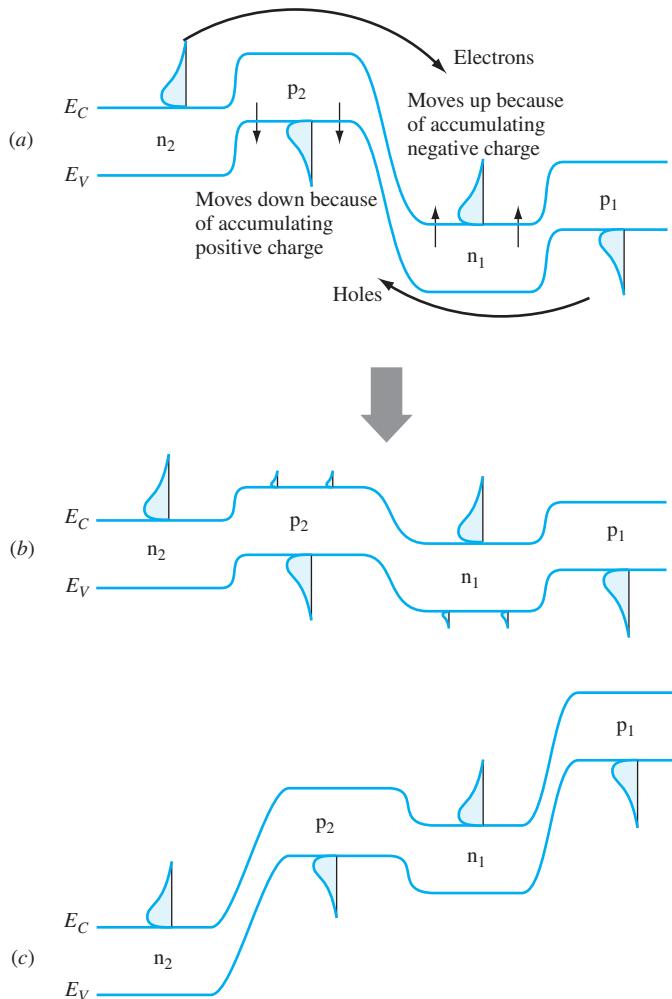


Figure 12.21 The energy band diagram of the npnp device.
 (a) Under positive V_S , two of the junctions are forward biased but the center one is reverse biased, so little current flows through the device. (b) Injected electrons accumulate in the n₁ region and injected holes accumulate in the p₂ region, causing the reverse bias across the junction between those two areas to decrease, producing the low-voltage, high-current state shown. (c) Under reverse bias, two of the junctions are reversed biased.

For V_S negative, Figure 12.21c, there are two reverse-biased junctions and only one that is forward biased. Carriers trapped in n₁ and p₂ do reduce the n₁p₂ barrier, but that has no effect on current injected from cathode or anode.

12.3.2 TWO-TRANSISTOR MODEL OF AN NPNP SWITCH

For V_S positive, the energy band diagram of Figure 12.21a, repeated in Figure 12.22a, appears to be that of a pnp transistor in series with an npn transistor.

The figure indicates the emitters, bases, and collectors of these two transistors. The circuit schematic in Figure 12.22b shows the two transistors explicitly.

In general for a transistor,

$$\begin{aligned} I_C &= \alpha I_E + I_{C0} \\ I_B &= (1 - \alpha)I_E - I_{C0} \end{aligned} \quad (12.26)$$

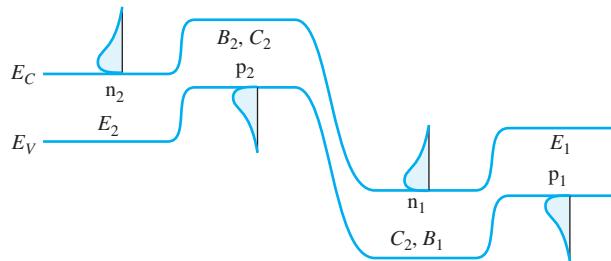
where I_{C0} is collector-base current with $I_E = 0$. Rearranging each of these, we have

$$\begin{aligned} I_{B1} &= (1 - \alpha_1)I_{E1} - I_{C01} \\ I_{C2} &= \alpha_2 I_{E2} + I_{C02} \end{aligned} \quad (12.27)$$

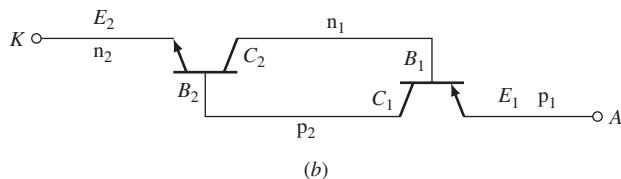
But $I_{B1} = I_{C2}$. Also the emitter current I_{E1} is the same as the anode current I_A , and the other emitter current is the same as the cathode current. Furthermore, the anode current has to equal the cathode current, so we have $I_{E1} = I_{E2} = I_A = I_K$, from which we obtain the relation

$$I_A = \frac{I_{C01} + I_{C02}}{1 - (\alpha_1 + \alpha_2)} \quad (12.28)$$

The current gains α_1 and α_2 , however, are functions of I_A . In the forward active region of a regular transistor, they are close to unity, but here the transistors never get to that mode of operation. For small I_A , recombination current



(a)



(b)

Figure 12.22 The npnp device can be thought of as an npn transistor connected to a pnp transistor. (a) The energy band diagram for small forward bias; (b) the equivalent circuit.

in the forward-biased junction transition regions predominates, and the α 's are small. This produces a small current. When I_A is large enough that $(\alpha_1 + \alpha_2) = 1$, I_A attempts to become infinite but is limited by the $I_A R_L$ drop.

12.3.3 SILICON-CONTROLLED RECTIFIERS (SCRs)

While the four-layer diode discussed is useful for describing the switching mechanism, the device itself is of limited use because the applied voltage must be changed to initiate the **off-to-on** transition. In other words, the signal to be switched must itself be changed, instead of being switched by another voltage as in the transistor. A much more common type of thyristor is a variation of this npnp device, the silicon-controlled rectifier¹¹ or SCR. In fact, when people talk about “thyristors” they usually mean the SCR, even though the SCR is a special case.

The structure of an SCR is indicated schematically in Figure 12.23a. It is similar to that of a four-layer diode except that an additional contact is made to one of the interior regions—in this example, to p_2 . This region is called the gate because a current pulse applied to this region is used to initiate the **off-to-on** transition. Figure 12.23b shows the SCR device circuit symbol.

The energy band diagram for the **off** state is indicated in Figure 12.24a. For this case $(\alpha_1 + \alpha_2) < 1$. If the gate (p_2) is made positive by applying a positive current pulse, the p_2 section will move down on the energy band diagram, and the $n_2 p_2$ junction barrier will be reduced, allowing more electrons to be injected into p_2 . These extra electrons are collected by n_1 , causing that (floating) region to become negative, which in turn reduces the $p_1 n_1$ barrier and increases the hole current injected from p_1 to n_1 . This increase in current results in an increase in α_1 and α_2 . Thus, for I_G large enough to cause $(\alpha_1 + \alpha_2) = 1$, [Equation (12.28)], the

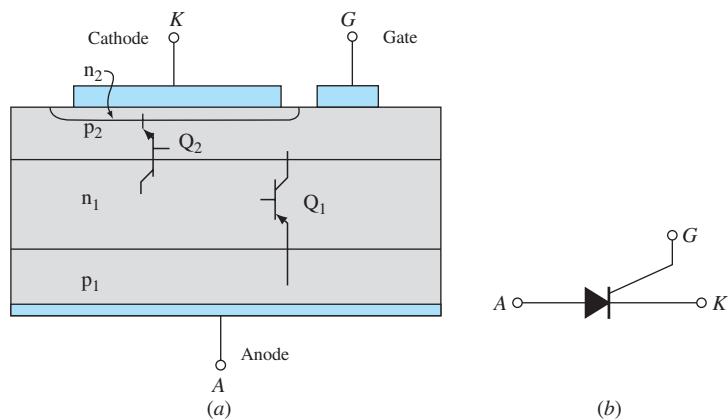


Figure 12.23 (a) The silicon-controlled rectifier is an npnp structure with a gate connection at one of the internal layers; (b) the circuit symbol.

¹¹Sometimes referred to as a *semiconductor controlled rectifier*.

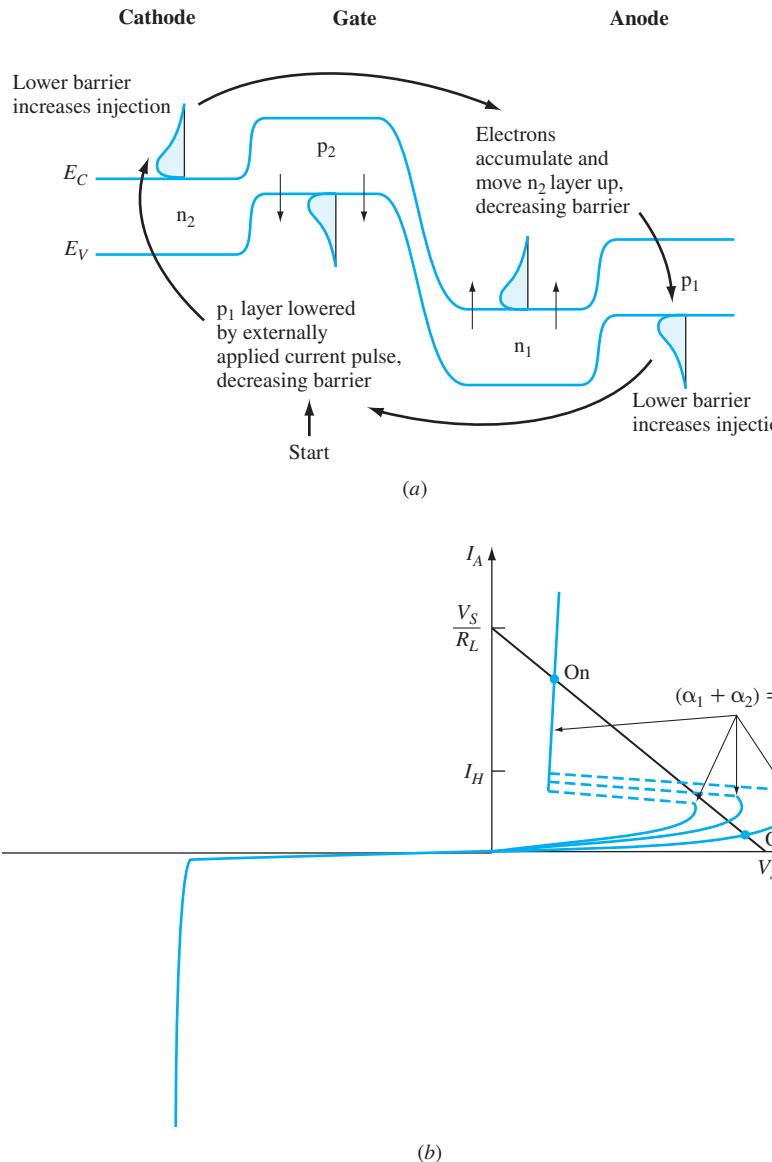


Figure 12.24 (a) The SCR energy band diagram and (b) operating characteristics. In this case, switching is activated by the application of a gate current I_G , which alters the I_A-V_{AK} characteristic so the device can switch.

device will switch from **off** to **on**. This is illustrated in Figure 12.24b for three values of gate current. For $I_G = 0$ the characteristics are those of a pn_n diode. Increasing I_G decreases the device voltage V_{AK} at which $(\alpha_1 + \alpha_2) = 1$, thus changing the switching point.

Modifying Equation 12.28 for the four-layer diode to include the gate current of the SCR results in

$$I_A = \frac{\alpha_2 I_G + I_{C01} + I_{C02}}{1 - (\alpha_1 + \alpha_2)} \quad (12.29)$$

where we have again recognized that I_A is the emitter current of Q1, and that I_G adds to the collector current of Q1, resulting in the additional $\alpha_2 I_G$ term in the numerator.

Note that once the device is **on**, it will remain **on** until V_S is reduced to the point where I_A reaches I_H , the holding current, or the minimum value of I_A such that $(\alpha_1 + \alpha_2) = 1$.

Let us look at the blocking characteristics briefly. When V_{AK} is positive and the device is not turned **on**, the n_1p_2 junction is reverse biased. The lightly doped drift region n^- is depleted. Under negative V_{AK} , it is the p_2n_1 junction that is reverse biased, and again the drift region is providing the blocking. Typically forward and reverse blocking voltages are comparable in SCRs. An SCR can be used to control the current (and thus the power) in a load by the timing of the gate pulse. Consider the circuit of Figure 12.25. The ac power line feeds a load whose average current is controlled by the SCR. The time delay circuit controls the timing of the input cycle, in which a positive pulse is applied to the gate (at times t_1, t_3, t_5, \dots). These pulses trigger the SCR to turn on, and it conducts for the remainder of the positive cycle until the current reduces to the holding current, I_H (t_2, t_4, \dots). The SCR remains off during the negative of the input cycle and is turned on again by the next gate pulse.

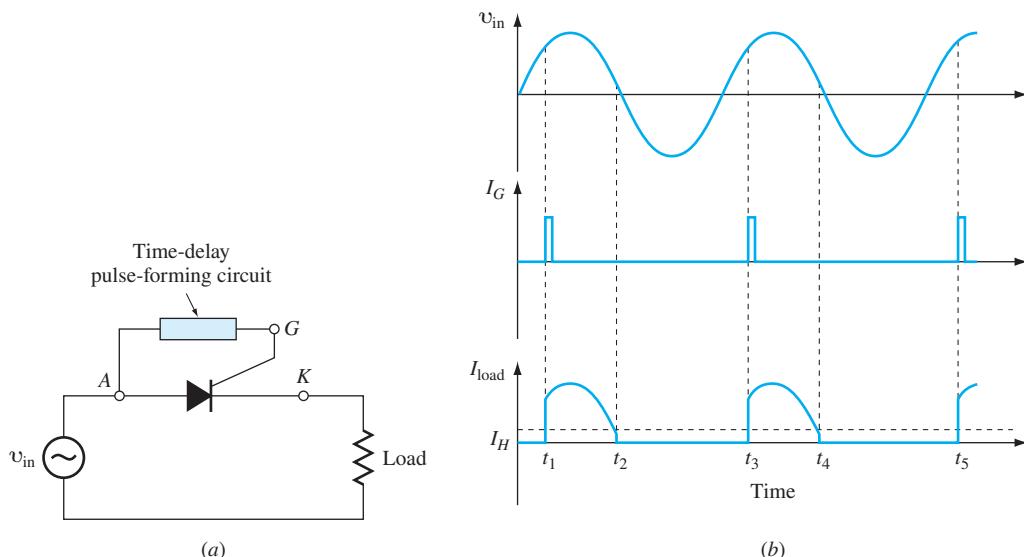


Figure 12.25 (a) Use of an SCR to control the power delivered to a load. (b) Input voltage, gate, and load current waveforms.

12.3.4 TRIAC

In an SCR the load current flows for less than half of the input cycle. A more useful thyristor is the TRIAC, whose structure is indicated schematically in Figure 12.26a. It consists of two SCRs connected in antiparallel, permitting current to flow during both halves of the input ac cycle. Figure 12.26b shows the input voltage and the output load current for a TRIAC for the gate current pulses indicated. In (a), I_1 and I_2 indicate the current flow from terminal T_2 to T_1 via the $p_2n_2p_1n_1$ SCR and I_2 from T_1 to T_2 via the $p_1n_2p_2n_4$ SCR.

Considering terminal T_1 to be at ground or the reference potential, the TRIAC can be turned **on** (conducting) for four different conditions: for positive current, from T_2 to T_1 , and for negative current (T_1 to T_2). For positive T_2 the TRIAC can be turned **on** if the gate current pulse is either positive or negative.

Here we consider the case for T_2 voltage and the gate trigger current to be positive. The equivalent circuit for this case is shown in Figure 12.27. When the positive gate current pulse is applied to the base (p_1) of the $n_1p_1n_2$ transistor, this transistor turns **on**, which causes current to flow out of the base of the $p_2n_2p_1$ transistor, turning it on. The net result is that there is a $pn\!pn$ SCR, and current flows from T_2 to T_1 .

For T_2 negative, when the gate current is pulsed negative, the n_3p_1 junction is forward biased, injecting electrons into the p_1 layer. Some of these electrons reach the n_2 region, lowering its potential. This n_2 layer acts as the base of the $p_2n_2p_1$ transistor. The p_2 region also acts as a base for the $n_4p_2n_2$ transistor. In effect then, there is an $np\!np$ SCR between T_2 and T_1 .

TRIACs are quite useful for ac power control. Common examples are dimmer switches to control the ambient light level and variable-speed motors.

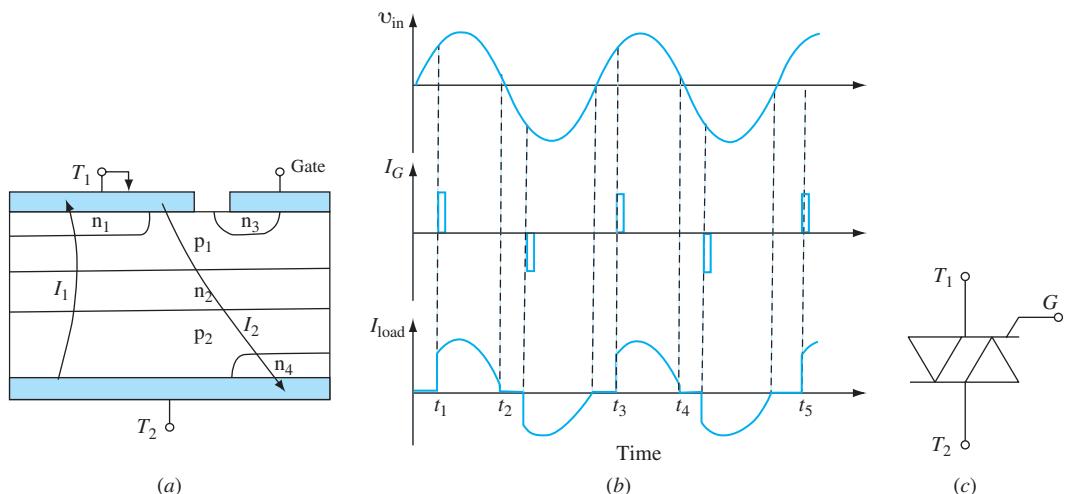


Figure 12.26 (a) Simplified structure of a TRIAC, (b) input voltage, gate, and load current waveforms, and (c) circuit symbol for a TRIAC

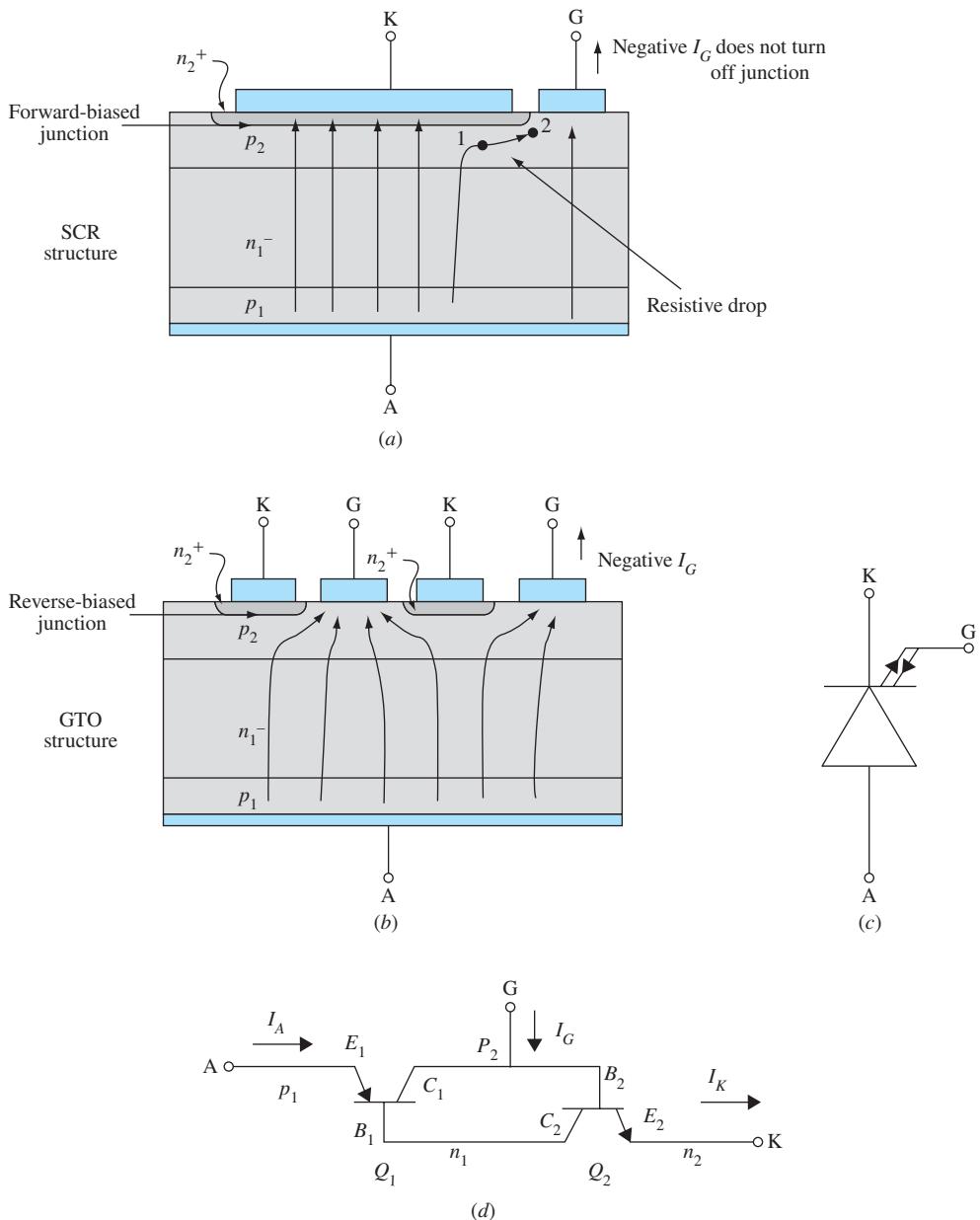


Figure 12.27 (a) A negative current at the gate does not turn off the SCR because the resistive drop laterally in the p_2 layer causes the potential at point (1) to be higher than at point (2), preventing the n_2p_2 junction from becoming back-biased over most of the cathode (K) region. (b) In a gate turn-off thyristor, the cathode contacts are made narrow, and the gate contacts interleaved with them, allowing the n_2p_2 junction to become reverse-biased and shutting off the anode-cathode current flow. (c) the circuit symbol for the GTO. (d) the transistor model.

12.3.5 GATE TURN-OFF THYRISTORS (GTOs)

The SCR structure can be modified to allow the gate control to turn the device **off** as well as **on**. Such a device is called a gate turn-off thyristor, or GTO. The SCR and GTO are both turned **on** by applying a positive current to the gate. Let us examine what happens when a *negative* gate current is applied.

In this case the gate current is *increasing* the barrier between the gate and the cathode (the p₂n₂ junction), which one might expect to have the effect of turning Q₂ **off** and stopping the current flow. Figure 12.27a shows the conventional SCR structure. We notice that the gate area in an SCR is very small compared to the anode and cathode areas. Current flowing out of the gate is being supplied by the anode, but most of that current would have to flow sideways in the p₂ layer to reach the gate terminal. Because the p₂ layer is somewhat resistive, there is a voltage drop between the center of the cathode and the gate. The potential at point 1 is higher than at point 2, meaning that although the p₂n₂ junction near the gate will be reverse-biased, which is the desired effect, the potential in the p₂ region under most of the cathode is still high enough to forward-bias the p₂n₂ junction, and the thyristor continues to conduct.

The solution is to modify the contact structure so that the cathode and gate electrodes are narrow and interdigitated, Figure 12.27b. Now a negative gate current will be able to flow through the gate contact to reverse-bias the gate-cathode junction more completely (the voltage applied to the gate is negative), and the device can be turned **off**. The circuit symbol for the GTO suggests ability to control the device with gate currents in either direction.

If we think of the GTO as a three-terminal device, and of the anode as the emitter, the cathode as the collector, and the gate as the base, we can define a sort of β as

$$\beta = \frac{I_K}{I_G} \quad (12.30)$$

Here, we are interested in the turn-off condition, where I_G is negative, so

$$\beta_{\text{turnoff}} = \frac{I_K}{-I_G} \quad (12.31)$$

From Equation (12.27), neglecting the leakage current I_{C02}

$$I_{E2} = I_K = \frac{I_{C2}}{\alpha_2} \quad (12.32)$$

Combining this with Equations (12.29) and (12.31), again neglecting the leakage currents, we obtain

$$\beta_{\text{turnoff}} = \frac{\alpha_2}{(\alpha_1 + \alpha_2) - 1} \quad (12.33)$$

If we don't want the turn-off current to be large, we want to maximize β_{turnoff} . That means we want α_2 to be large (as close to 1 as possible), but then to keep the

denominator small, we want α_1 to be small. This means we want to make transistor Q_1 to be a poor one. It turns out that we are doing this already. Recall that

$$\alpha = \gamma \alpha_T M$$

and the base transport factor is

$$\alpha_T \approx 1 - \frac{W_B^2}{2L_{nB}^2}$$

We already have a thick base region (the n_1^- layer) in thyristors because we need a high blocking voltage and thus a wide drift region. To make Q_1 into a poor transistor, we would also want L_{nB} to be small, implying increased doping. For high blocking voltage, however, we must dope the drift region lightly, so this works against us. The result is a higher voltage drop in the conducting state than a normal SCR, but we have gained the ability to turn the device both **on** and **off** with an external control signal.

12.4 THE POWER MOSFET

So far we have discussed power diodes and thyristors. Diodes are turned **on** and **off** by the signal itself. SCRs can be turned **on** with the application of a separate control signal, and GTOs can be turned both **on** and **off** with a separate control signal, which makes their operation—qualitatively, anyway—somewhat transistor-like. In this section and the next, we will introduce actual power transistor structures—first power MOSFETs, followed by the insulated gate bipolar transistor (IGBT) in Section 12.5.

MOSFET devices used for digital or analog circuitry only see low voltage. Power MOSFETs require high blocking voltage. For reverse blocking voltages greater than about 10 V, this is accomplished by a lightly doped n^- drain extension or drift region. There are several structures for power MOSFETs. Figure 12.28a shows the cross section of a vertical double diffused MOSFET (VDMOSFET or simply VDMOS).¹² The source (n^+) is on the top surface, and the drain is on the bottom. The source n^+ well is embedded in a larger p-well. The channel forms under the gate where the p-region meets the surface. The current path is shown. MOS power transistors are normally **off**, and a gate voltage is applied to turn them **on** i.e., they are enhancement MOSFETs. The channel length L is indicated. We recall that the gain is proportional to W/L , so a large channel width is obtained by arranging the source regions in geometric patterns and having many of them in parallel. This provides the ability to carry large currents.

In the blocking mode the gate voltage is zero,¹³ so that the channel is **off**. The drain voltage is positive, so the pn^- junction is reverse-biased. Because the

¹²Originally, the n^+ source and the adjacent p-region were diffused into an n^- substrate. Currently, these diffusions are often replaced by ion implantation.

¹³Here the source is assumed to be the reference, or at zero potential.

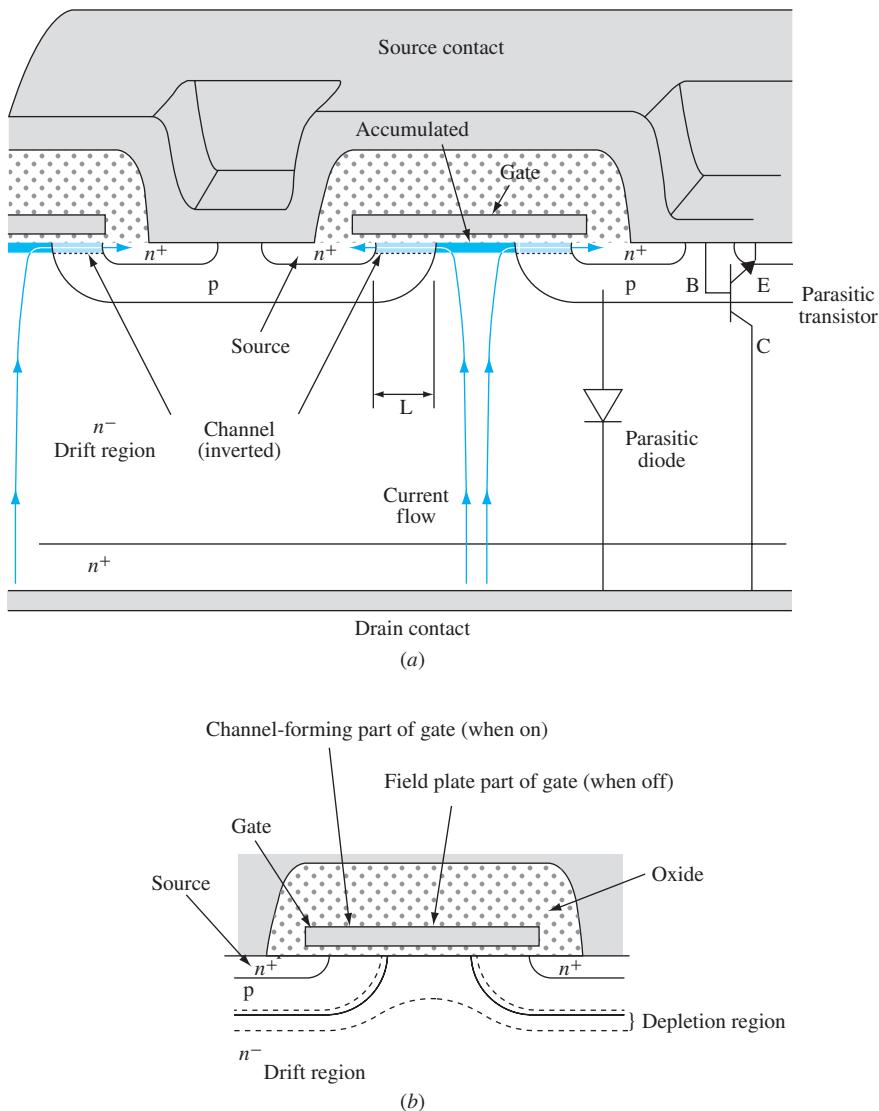


Figure 12.28 The power VDMOSFET. (a) Current flow when transistor is on; (b) close-up of gate electrode showing the region that acts as a field plate.

n⁻-drift region has a lower doping level than that of the p-region, most of the junction depletion region is in the drift region. The thickness of the drift region is determined by the n⁻ doping level and the specified blocking voltage.

To turn the device **on**, with a positive voltage on the drain, the gate voltage is made positive, creating a conducting channel. Electrons then flow from n⁺ source through the channel and then through the n⁻ drift region to the drain.

Because the device is unipolar (only electrons contribute to current), there is no conductivity modulation and the drift resistance is large, similar to that of a Schottky barrier device.

Notice that there is a parasitic npn transistor between the drift region and the source. The p-region is shorted to the n⁺ source in the middle by the metal contact, which shorts out the base-emitter junction and prevents the parasitic transistor from ever turning **on**, because for the parasitic transistor $V_{BE} = 0$. There remains, in that case, a parasitic diode, whose forward voltage limits the reverse voltage the MOSFET can support. The MOSFET's forward breakdown is set by the drift region thickness and doping as it was for the Schottky rectifier. Premature breakdown at the pn⁻ junction is avoided by extending the gate electrode over the region where the n⁻ material reaches the surface (Figure 12.28b). In this way the gate electrode acts as a field plate, depleting the n⁻ material near the surface. This connects the depletion regions of two adjacent pn junctions, avoiding a tight curvature that could increase the electric field locally.

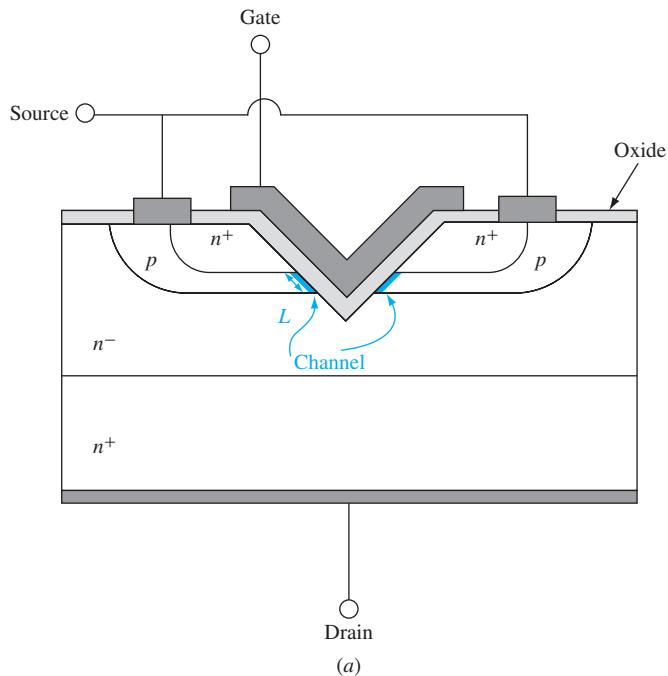
For power MOSFETs with breakdown voltages less than about 10 V, the high-resistance n⁻ drift layer can be short enough that the resistance in the drift region is still small, and the channel resistance predominates. The shorter the channel, the lower the $R_{on,sp}$. Two variations of the VDMOS described above but with even shorter channels are shown in Figure 12.29. These are referred to as VMOS (a) and UMOS (b), (sometimes called a trench MOSFET). The channel is now vertical, and is shortened because the thickness of the p region (channel length) can be made much less than the horizontal dimensions along the surface as in the VDMOS of Figure 12.28 (a).

In VDMOSFETs the on-resistance is the sum of the channel resistance and the drift region resistance. For blocking voltages greater than about 30 V, however, the resistance of the drift region predominates, and the results for a Schottky diode can be used to approximate N_D , W_D , and $R_{on,sp}$ for power MOSFETs.

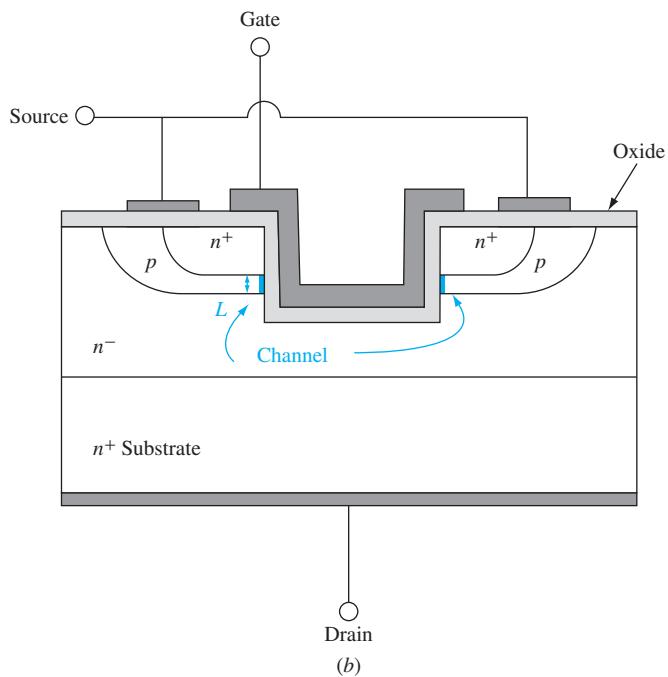
For devices with higher breakdown voltages, the resistance of the drift region predominates, and the drift region must be quite thick.¹⁴ This, however, increases the specific on-resistance, thus limiting the **on** current. Figure 12.30 shows the cross section of a *superjunction* MOSFET, a structure that greatly reduces $R_{on,sp}$. In the superjunction MOSFET, deep p⁺ trenches are fabricated in n-type material. For devices with high breakdown voltages, this structure allows the devices to carry high on-current.

A MOSFET is operated with drain positive with respect to the source. For the superjunction MOSFET of Figure 12.30 with $V_G = 0$, no channel exists. Since the p⁺ trench is connected to the source, the p⁺n junctions are reverse biased, with the depletion region being primarily in the n drift region. The n region is thin enough that it is fully depleted, resulting in high resistivity in this blocking state.

¹⁴In Si VDMOS, for a breakdown voltage of 600 V, the resistance of the drift region is on the order of 30 times that of the channel.



(a)



(b)

Figure 12.29 Cross section of a power VMOS (a) and a UMOS device (b).

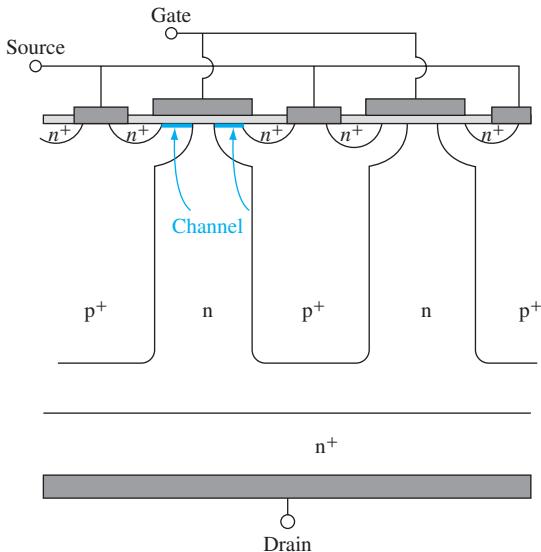


Figure 12.30 Cross section of a superjunction vertical power MOSFET.

For $V_G > V_T$, a channel forms, connecting the n⁺ source to the n drift region, reducing the depletion region in the drift region and increasing its conductance, reducing the on-resistance.

Thus superjunction MOSFETs can use thinner, more highly doped drift regions. The increased doping and reduced thickness of the drift region results in a reduced $R_{on,sp}$, giving an increased **on** current capability and a high blocking voltage.

The vertical MOSFETs discussed above are normally discrete devices, although for increased current capabilities, often many devices are connected in parallel on a single chip.

Often, however, we want to use a power device on an integrated circuit chip. In this case a planar technology is required. Figure 12.31 shows the cross section of a lateral double diffused MOSFET or LDMOS. The operation is similar to that of the VDMOS except that the current flows laterally from source to drain through the n⁻ drift region.

12.5 THE INSULATED-GATE BIPOLAR TRANSISTOR

The power MOSFET just discussed has some advantages over a bipolar power transistor. It can be turned **on** and **off** with a gate voltage signal (easier to implement than a base current signal such as is needed with bipolar junction transistors). There is no current flow into the gate, creating very high input impedance.

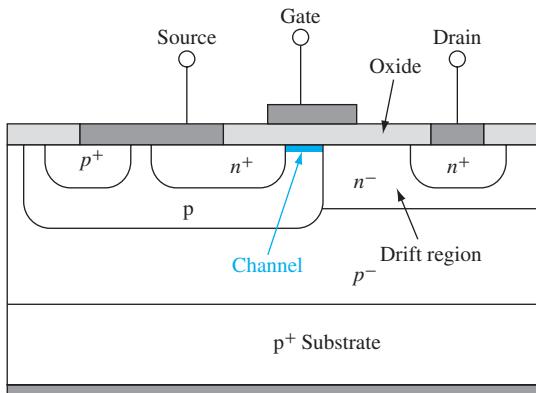


Figure 12.31 Cross section of an LDMOS device.

And, because it is a unipolar device, it can be turned **on** and **off** more quickly than a bipolar device. The on-resistance is high, however. The drift region has to be both thick and lightly doped to support high voltages,¹⁵ as we discussed when we looked at the Schottky rectifier diode, and that creates a high on-resistance. When large currents flow through that resistance, heat is created, so the devices must be made large to dissipate the heat.

The insulated-gate bipolar transistor (IGBT) is a clever device that combines the advantages of a MOSFET with the benefits of a bipolar junction transistor (BJT). Figure 12.32 shows the structure of a generic IGBT.¹⁶ It is basically the same as the MOSFET but with the addition of a p⁺ layer on the bottom in place of the n⁺ layer. Let us walk through its operation.

Off state: When the drain (anode) voltage is positive with respect to the source (cathode), and the gate voltage is below threshold, $V_G < V_T$ Fig. 12.32, there is no channel. No current flows, and the device is in region 1 on the I -V characteristics of Figure 12.33. Note that the p layer near the top is shorted to the n⁺ wells, so as long as there is no conducting channel from the n⁺ to the n⁻ region, the device looks like a pnp BJT in the **off** state. The n⁻ region (transistor base) is floating, and no base current flows. The naming of the emitter and collector terminals is reversed from the actual emitter and collector (from a physics perspective) of the pnp transistor, as we shall see later. The naming was done for historic reasons, so to avoid confusion we will label the junctions J₁, J₂, and J₃. The pnp BJT can block voltages in either direction (this structure is called “symmetric,” and the electric field does not punch through the drift region). Breakdown occurs at the reverse-biased junction, junction J₂ for positive drain voltage and junction J₁ for negative drain voltage.

¹⁵Except for superjunction power MOSFETs.

¹⁶There are several structures used in these devices.

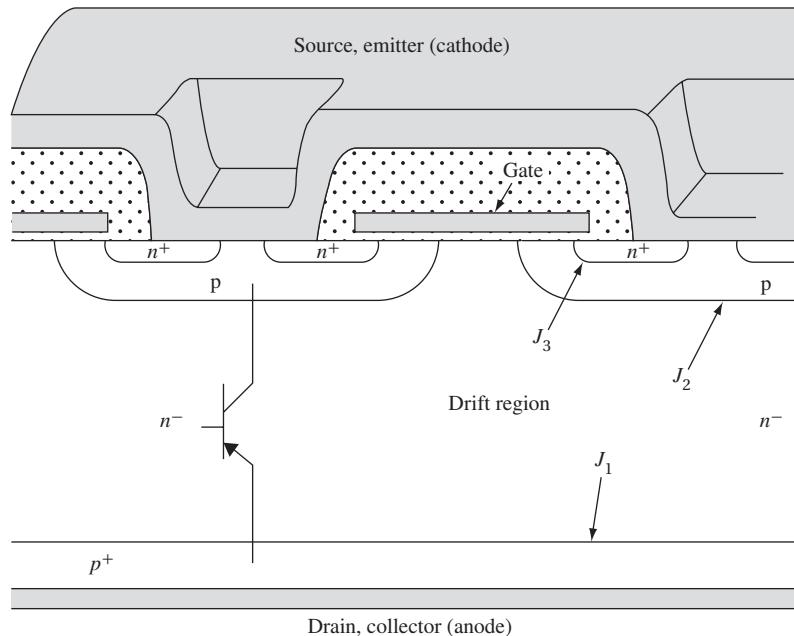


Figure 12.32 Cross-section of the structure of a generic IGBT.

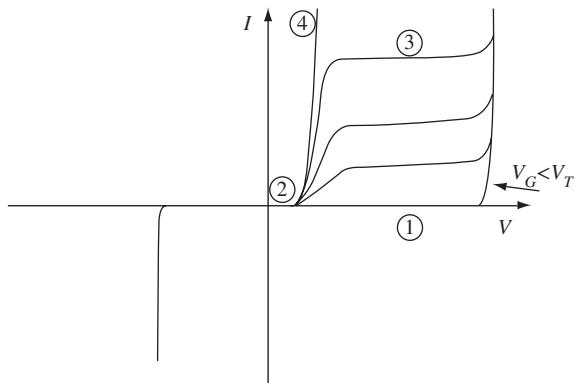


Figure 12.33 Anode to cathode I - V characteristics of the IGBT for different bias conditions, (See text.).

Diode: When $V_G > V_T$ is applied to the gate, Figure 12.34, the channel is enhanced, creating a conducting path between the n⁻ drift region and the n⁺ source region, effectively making the structure look like a pn junction (J_1). Let us consider low positive drain-to-source voltage first. Holes are injected into the n⁻ drift region by junction J_1 , and electrons are being supplied by the FET current. In this regime, the device looks essentially like a diode in series with a FET. There is a forward voltage drop across the diode, and the current increases exponentially with voltage

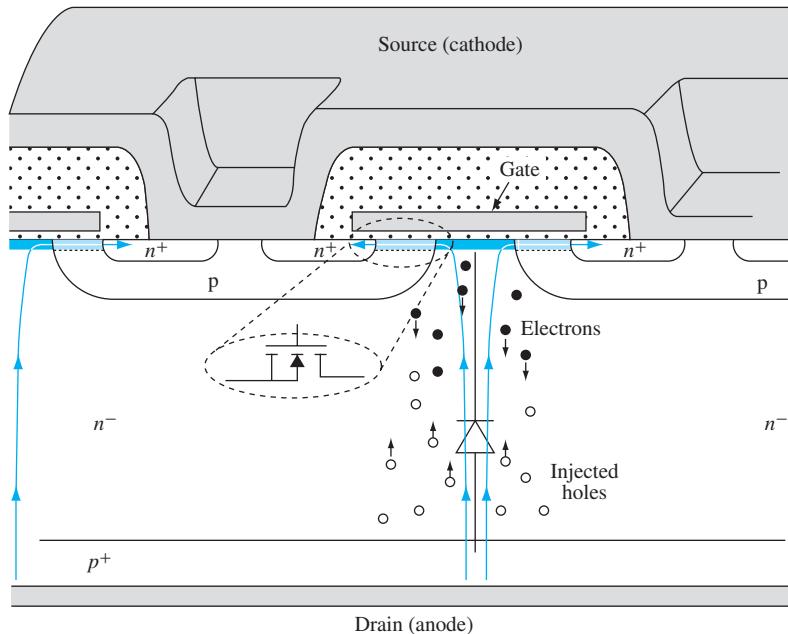


Figure 12.34 Illustration of current flow for an IGBT with $V_G > V_T$.

(region 2 on the I - V curve of Figure 12.33). As the drain-source voltage increases and the current increases, the channel saturates, which limits the current that flows through this path. This causes the I - V characteristics to level off (region 3).

On State: The current flowing out of the n^- drift region through the MOSFET can also be thought of as flowing out of the base of the pnp transistor. The BJT amplifies that current, causing a large current to flow through the collector and emitter of the BJT (Figure 12.35). This large current creates excess carrier densities that reduce the resistance (conductivity modulation) in the drift region. Figure 12.36 shows the equivalent circuit. The p^+ layer is called the “collector” of the IGBT (also called “drain” or “anode”). Because this layer is more highly doped than the n^- drift region or the p-wells, physically the p^+ layer acts as the emitter of the pnp transistor.

Let us label the emitter, base, and collector of the pnp transistor with lowercase letters. The collector current I_c of that transistor is β times the base current, which is also the channel current I_{CH} . The collector current I_C of the IGBT is then

$$I_C = I_c + I_{CH} = (\beta + 1)I_{CH} \quad (12.34)$$

In the **on** state, the IGBT normally operates in region 3 (Figure 12.33), where the channel current (and thus the anode current) is in saturation. The holes injected from the p^+ region into the n^- region flow into the p region and out the source.

Latchup: The lateral flow of holes in the p region develops a lateral voltage drop in the ohmic resistance of this p layer that tends to forward-bias a portion

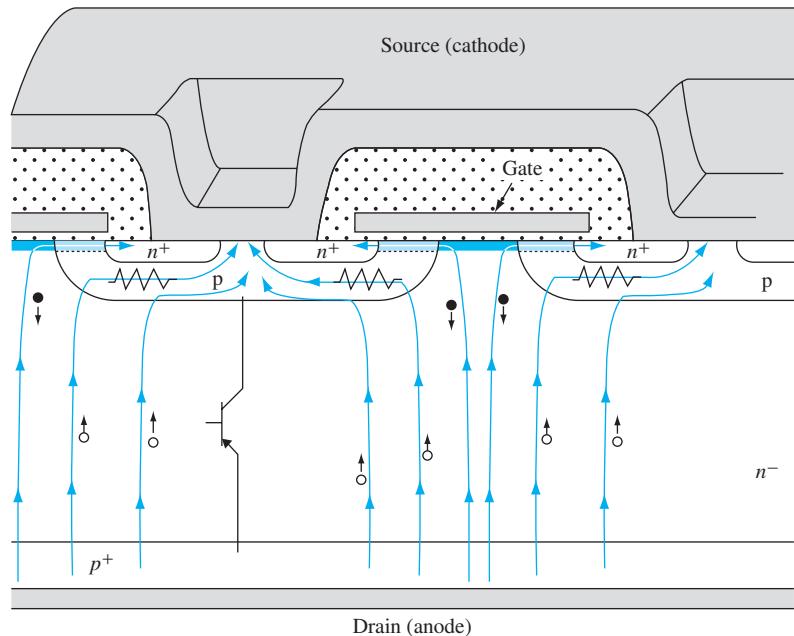


Figure 12.35 Illustration of current flow for $V_G \gg V_T$ and larger V_{DS} . Hole current flows laterally through the p region, turning on the n⁺pn⁻ transistor. The result is a p⁺n⁻pn⁺ thyristor.

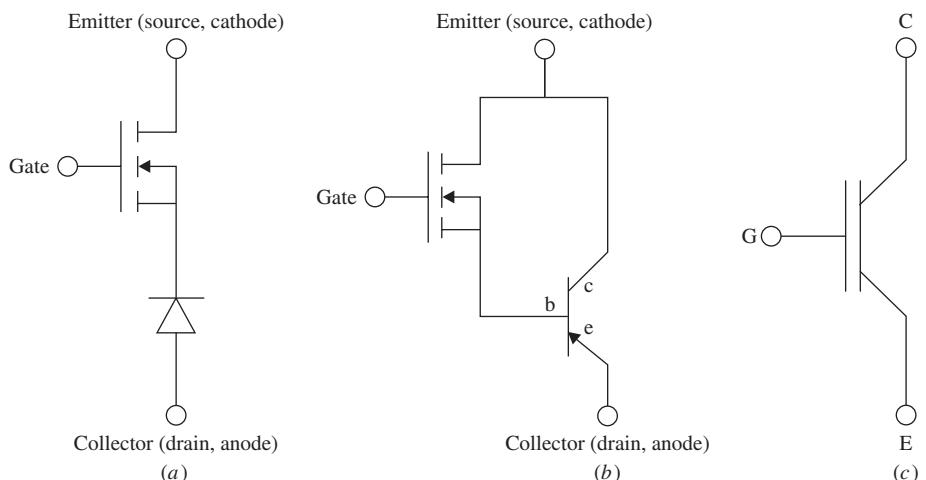


Figure 12.36 IGBT Equivalent circuit for low current (a), high current (b) and (e) and circuit symbol (c).

of this n^+p junction, causing electrons to be injected from the n^+ region into the p region. If this lateral current is large enough, the resultant n^+p voltage turns **on** the n^+pn^- transistor. In this case, both n^+pn^- and p^+n^-p transistors are **on**, resulting in a $p^+n^-pn^+$ thyristor with the characteristics of region 4 in Figure 12.33 (latch-up). Once in latch-up, the gate has no control of the anode current, and the only way to turn **off** the IGBT is to decrease the anode-cathode voltage below the holding voltage, as in a conventional thyristor.

Forward and reverse breakdown and reverse blocking: In the forward direction, if V_{DS} is large enough, the device can break down, limiting the voltage operation range. In the reverse direction, the device is blocking until the reverse breakdown voltage is reached.

In the blocking mode, the voltage across the IGBT is primarily that of the drift region, and thus the results for a Schottky diode can be used to approximate N_D and W_D for a given breakdown voltage.

12.6 POWER MOSFET VERSUS IGBT

Power MOSFETs and IGBTs currently dominate the power semiconductor market in applications such as motor drives, uninterruptible power supplies, and solar inverters. Power MOSFETs and IGBTs have very similar structures except that the substrate (drain) is n^+ in the MOSFET while in the IGBT the substrate (drain, anode) is p^+ (compare Figures 12.28 and 12.32). Both devices are **off** (nonconducting) for zero gate voltage and **on** (conducting) for positive gate voltage. Figure 12.37 indicates some of the boundaries where it is reasonably clear whether the power silicon-based MOSFET or IGBT are preferred.

For silicon-based devices, power MOSFETs are preferred in:

- High-frequency applications ($f > 200$ kHz)
- Low-voltage applications ($V_{br} < 250$ V)
- Low-power output ($P < 500$ W)

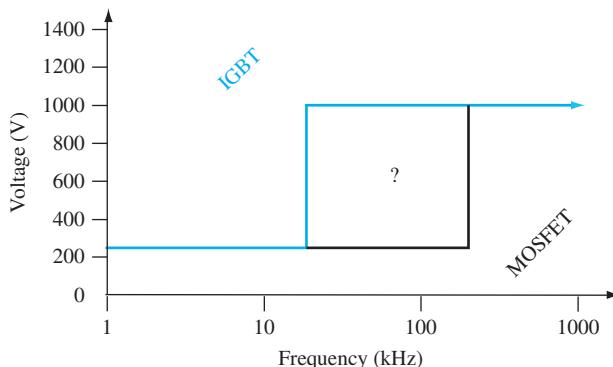


Figure 12.37 Indication of approximate voltage-frequency boundaries for Si-based MOSFETs and IGBTs (neglecting output power).

IGBTs are preferred in:

- Low-frequency applications ($f < 25$ kHz)
- High-voltage applications ($V_{br} > 1000$ V)
- High-power output ($P > 5$ kw)

MOSFETs can operate at higher switching frequencies because conduction in the drift region is by majority carriers (electrons). In an IGBT the presence of minority carriers in the drift region results in a long **on-to-off** switching time and reduced operating frequency.

The specific on-resistance, $R_{on,sp}$, of a MOSFET increases approximately linearly with breakdown voltage, V_{br} . This limits the on-current (heating) at the higher V_{br} . In an IGBT, for a given V_{br} , $R_{on,sp}$ decreases with increasing current (conductivity modulation) and approaches zero at high currents. Thus for high V_{br} and high on-current, device width of MOSFETs must be appreciably larger than in IGBTs. For MOSFETs of practical size, this limits their power handling capabilities.

12.7 SUMMARY

In this chapter, we discussed the physics of semiconductor devices that are specifically designed for high-voltage, high-current operation, which requires significant differences in structure than the usual low-power devices used in analog or digital logic circuits. In particular we saw that to block high voltages, power devices generally require a thick, lightly doped region known as the drift region. This also implies, generally, a vertical structure.

We also saw that wide-band-gap materials such as GaN and 4H-SiC have very high critical breakdown fields, making them well suited to power applications. SiC MOSFETs are commercially available now, with GaN devices just starting to come on the market at the time of this writing. Silicon is still the primary material in use because Si fabrication processes and material growth are very mature.

For rectifiers, the specific on-resistance, the blocking voltage, and transient losses all have to be taken into account when choosing pin versus Schottky junctions. The advantages of both pin and Schottky diodes can be taken advantage of with the merged pin–Schottky (MPS) rectifier, where the conductivity modulation induced by the pin is shared by the Schottky.

We also discussed a variety of power-switching devices. On the bipolar side (current carried by both electrons and holes), we discussed four-layer devices, including thyristors (silicon-controlled rectifiers). The four-layer diode switch, the simplest of the npnp switching devices, has no separate switching control, but instead is turned **on** by the signal itself (the one being switched). A silicon-controlled rectified has a third contact, allowing a separate control signal to turn it **on**, but it is turned **off** by the signal V_S being switched. The gate turn-off transistor (GTO) allows one to turn the device both **on** and **off** via external control. Figure 12.38 shows that thyristors in general can support the highest voltages (currently up to 12,000 V) and appreciable current. They are limited in speed, however. GTOs can support the most current, but with voltages only up to about 4 kV.

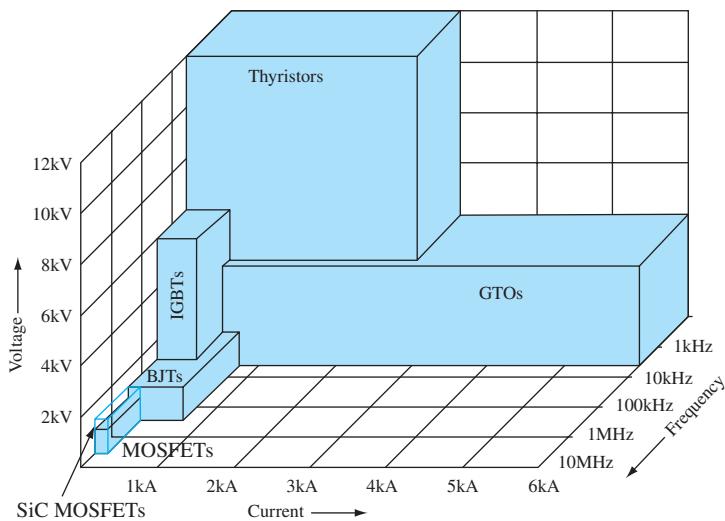


Figure 12.38 The operating parameter of some common power devices. Note that the boxes represent the limits of current, voltage, and frequency that can be achieved with a device, but this does not mean all three extremes can be achieved in any single device.

We next discussed power MOSFETs, which have the advantage (over bipolar transistors) of being controlled by a voltage rather than a current signal. They can be switched faster than any other power device (including BJTs) but cannot support the same high currents or voltages. We saw that an insulated gate bipolar transistor (IGBT), a hybrid of MOSFET and BJT, combines the benefits of both. While not as fast as MOSFETs, they are faster than thyristors and can support higher voltages.

In bipolar devices, turn-off switching speed is limited by the time required to remove minority carriers from the drift regions. In unipolar (MOSFET) devices, although the switching speed is greater than that of bipolar devices, the current density J is limited by the large on-resistance. The on-resistance can be greatly reduced, however, by the use of superjunction VDMOSFETs.

12.8 REFERENCES

1. B. J. Baliga, “Breakdown voltage,” Chapter 3 in *Fundamentals of Power Semiconductor Devices*, Springer, New York, 2008.
2. Z. Z. Bandic, E. C. Piquette, P. M. Bridger, and T. C. McGill, “Design of GaN/AlGaN high power devices,” *Electrochemical Society Proceedings*, 98-12, 1998.
3. N. Mohan, T. M. Undeland, and W. P. Robbins, *Power Electronics: Converters, Applications, and Design*, 3rd ed., Wiley, New York, 2003.
4. B. J. Baliga, “Analysis of a high-voltage merged p-i-n/Schottky (MPS) rectifier,” *IEEE Electron Device Letters*, vol. EDL-8, pp. 407-409, September 1987.

12.9 REVIEW QUESTIONS

- What is the purpose of the lightly doped (drift) n^- region in a power diode?
- Why is the drift region n^- instead of p^- or intrinsic?
- Why is \mathcal{E}_{cr} , the critical breakdown field, larger in GaN than in Si?
- In a power diode, why is the specific on-resistance quoted rather than the actual resistance of the n^- drift region?
- Why is the turn-off transient time larger in a pin diode than in a Schottky diode?
- Why is the turn-off transient time in a merged pin-Schottky diode less than in a pin diode?
- Explain how a four-layer diode works.
- Show how the operating point of a four-layer diode switch moves along the I - V characteristics as the voltage is first increased, then decreased.
- What is meant by latchup?
- Explain why in a pin diode, at high currents the voltage across the drift region is independent of current.

12.10 PROBLEMS

For simplicity, unless otherwise stated, assume that the blocking voltage is equal to the breakdown voltage.

- 12.1** From Figure 12.2a, the variation of N_D versus V_{br} is a straight line on a $\log V_{br} - \log N_D$ plot with the same slope (α) for Si, 4H-SiC, and GaN p^+n diodes. Thus the relation between V_{br} and N_D can be expressed as

$$\log V_{br} = \alpha \log N_D$$

or

$$\alpha = \frac{\Delta \log V_{br}}{\Delta \log N_D}.$$

Also, $\log \mathcal{E}_{cr} = \beta \log V_{br}$ and $\log W_D = \gamma \log V_{br}$

Find β and γ in terms of α .

- 12.2** Find the specific on-resistance ($R_{on,sp}$) for a Si Schottky diode having a blocking voltage (V_{br}) of 100 V.
- 12.3** Determine the specific on-resistance for a 4H-SiC Schottky diode designed to block 100, 600, and 900 V.
- 12.4** Calculate the on-state current density for a Si pin diode when the doping level in the drift region is 10^{14} cm^{-3} and the average injected

carrier density is 10^{15} cm^{-3} . The drift region width is 100 μm and the high-level carrier lifetime is 1 μs .

- 12.5 Calculate the reverse breakdown voltage for the pin diode of Problem 4.
- 12.6 Find the on-state voltage for the pin diode of Problem 4 at an on-state current density of 100 A/cm².
- 12.7 Find the reverse breakdown for a 4H-SiC pin diode having a drift region doping of 10^{16} cm^{-3} .
- 12.8 What is the width of the drift region for the diode of Problem 7?
- 12.9 For a GaN pin diode designed to have a breakdown voltage of 5 kV, find the doping level and width of the drift region.
- 12.10 Find the on-state specific resistances for a Si power MOSFET with breakdown voltages of 20, 50, 100, and 500 V.
- 12.11 Find the drift region doping concentration and thickness for an n-channel Si power MOSFET with a breakdown voltage of 300 V.
- 12.12 Find the on-state specific resistance for a GaN power MOSFET with breakdown voltages of 200, 500, 1000, and 5000 V.
- 12.13 In connection with the npnp device of Figure 12.21, it was stated that at low V_S , the recombination in the forward-biased junction transition regions is significant.
 - a. Explain why the recombination current is significant.
 - b. When recombination cannot keep up with the carrier injection, what happens to the energy band diagram? Explain why the device switches to a high-current state.
- 12.14 We indicated that in an SCR device, applying a positive voltage pulse V_G to the gate (p2 region) could turn the device to the **on** state. Explain how applying a negative current pulse to the n1 region would have the same effect.
- 12.15 Can the SCR be turned off by applying a negative current pulse to the p2 region?
- 12.16 For a Si VDMOSFET designed for a blocking voltage of 250 V, find its $R_{\text{on,sp}}$. Assume that the on-resistance is primarily in the drift region, W_D .
- 12.17 Repeat Problem 16 for a 4H-SiC VDMOSFET with a blocking voltage of 4 kV.
- 12.18 A Si IGBT has a breakdown voltage of 300 V and a high-level carrier lifetime of 1 μs . Find the drift region thickness and the voltage across the drift region at high on-currents.
- 12.19 Repeat Problem 18 for a 4H-SiC IGBT having a breakdown voltage of 4 kV and a high-level carrier lifetime of 5 ns.

A APPENDIX

Constants

Table A.1 Derived units

Energy	Joule: $\frac{\text{kg} \cdot \text{m}^2}{\text{s}^2}$
Force	Newton: $\frac{\text{kg} \cdot \text{m}}{\text{s}^2}$
Power	Watt: $\frac{\text{kg} \cdot \text{m}^2}{\text{s}^3} = \frac{\text{J}}{\text{s}}$
Capacitance	Farad: $\frac{\text{C}^2 \cdot \text{s}^2}{\text{kg} \cdot \text{m}^2}$
Current	Ampere: $\frac{\text{C}}{\text{s}}$
Current density	Ampere/area: $\frac{\text{C}}{\text{s} \cdot \text{m}^2}$
Electric potential	Volt: $\frac{\text{kg} \cdot \text{m}^2}{\text{s}^2 \cdot \text{C}}$
Inductance	Henry: $\frac{\text{kg} \cdot \text{m}^2}{\text{C}^2}$
Resistance	Ohm: $\frac{\text{kg} \cdot \text{m}^2}{\text{s} \cdot \text{C}^2}$
Magnetic induction	Tesla: $\frac{\text{kg}}{\text{s} \cdot \text{C}}$

Table A.2 Some physical constants

Speed of light in vacuum	c	$2.9979 \times 10^8 \text{ m/s}$
Permittivity of vacuum	ϵ_0	$8.8542 \times 10^{-12} \text{ F/m}$ $8.8542 \times 10^{-14} \text{ F/cm}$
Boltzmann's constant	k	$1.38 \times 10^{-23} \text{ J/K}$ $8.62 \times 10^{-5} \text{ eV/K}$
Planck's constant	h	$6.63 \times 10^{-34} \text{ J} \cdot \text{s}$ $4.14 \times 10^{-15} \text{ eV} \cdot \text{s}$
h-bar	\hbar	$1.05 \times 10^{-34} \text{ J} \cdot \text{s}$ $6.59 \times 10^{-16} \text{ eV} \cdot \text{s}$
kT , room temperature ($T = 300 \text{ K}$)	kT kT/q	0.0259 eV 0.0259 V
Free electron mass (at rest)	m_0	$9.11 \times 10^{-31} \text{ kg}$
Electronic charge	q	$1.60 \times 10^{-19} \text{ C}$

Table A.3 Conversion of units

1 eV	1.6×10^{-19} J
1 cm	10^{-2} m 10^8 Å 10^4 μ m
1 Å	10^{-10} m 10^{-8} cm 0.1 nm
0°C	273.18 K
1 tesla	1 Wb/m ² 10^4 gauss

Table A.4 Some semiconductors and their band gaps

Semiconductor	Band structure	Band gap, eV
Si	Indirect	1.12
GaAs	Direct	1.43
Ge	Indirect	0.67
InP	Direct	1.35
AlAs	Indirect	2.16
AlP	Indirect	2.45
AlSb	Indirect	1.6
4H-SiC	Indirect	3.26
GaN	Direct	3.44
GaP	Indirect	2.34
GaSb	Direct	0.81
InAs	Direct	0.36
InSb	Direct	0.18
CdS	Direct	2.42
CdTe	Direct	1.56
CdSe	Direct	1.70
ZnO	Direct	3.35
ZnS	Direct	3.68
ZnTe	Direct	2.25
ZnSe	Direct	2.7
PbS	Indirect	0.41
PbTe	Indirect	0.31

Table A.5 Constants of some semiconductors

	Silicon	GaAs	Ge	InP	GaN	4H-SiC
E_g	1.1242 eV	1.43 eV	0.67 eV	1.35 eV	3.437	3.26
N_C	$2.89 \times 10^{19} \text{ cm}^{-3}$	$4.4 \times 10^{17} \text{ cm}^{-3}$	$1.05 \times 10^{19} \text{ cm}^{-3}$	$5.4 \times 10^{17} \text{ cm}^{-3}$	2.3×10^{18}	1.23×10^{19}
N_V	$3.10 \times 10^{19} \text{ cm}^{-3}$	$8.4 \times 10^{18} \text{ cm}^{-3}$	$4.0 \times 10^{18} \text{ cm}^{-3}$	$6.9 \times 10^{18} \text{ cm}^{-3}$	1.8×10^{19}	4.58×10^{18}
n_i	$1.08 \times 10^{10} \text{ cm}^{-3}$	$2.2 \times 10^6 \text{ cm}^{-3}$	$1.64 \times 10^{13} \text{ cm}^{-3}$	$9.3 \times 10^6 \text{ cm}^{-3}$	1.9×10^{-10}	5×10^{-9}
χ	4.05 eV	4.07	4.0	4.35	4.1	3.1
ϵ_r	11.8 (SiO ₂ :3.9)	13.2	16.0	12.4	8.9	9.7
v_{sat} (intrinsic)	$1 \times 10^7 \text{ cm/s}$ (electrons and holes)	$6 \times 10^6 \text{ cm/s}$ (electrons and holes)	$6 \times 10^6 \text{ cm/s}$ (electrons and holes)		2.5×10^7	2.5×10^7
E_{pho}	0.063 eV	0.034 eV	0.034 eV		91.2 meV	104.2 meV

Table A.6 Effective masses for electrons in units of m_0 , the rest mass of the free electron

	$m_{(K=0)}^*$	m_{\parallel}^*	m_{\perp}^*	m_{ce}^*	m_{dse}^*
Si		0.92	0.197	0.26	1.09
GaAs	0.067			0.067	0.067
Ge		1.64	0.082	0.12	0.56
InP	0.077			0.077	0.077
GaN	0.20			0.20	0.20
4H-SiC		0.29	0.42	0.32	0.77

Table A.7 Effective masses for holes in the valence bands of several semiconductors

	m_{lh}^*	m_{hh}^*	m_{sh}^*	Δ (eV)	m_{ch}^*	m_{dsh}^*
Si	0.16	0.48	0.24	0.044	0.36	1.150
GaAs	0.082	0.45	0.15	0.34	0.34	0.48
Ge	0.044	0.28	0.08	0.29	0.21	0.292
InP	0.08	0.4	0.15	0.11	0.3	0.42
GaN	0.3	1.4	0.6	0.02		
4H-SiC						1.0

Table A.8 Periodic table of the elements

B A P P E N D I X

List of Symbols

a	acceleration, lattice constant
A	area
A_{21}	Einstein coefficient for spontaneous optical emission
A_E	area of emitter junction
B	magnetic field
B_{12}	Einstein coefficient for absorption
B_{21}	Einstein coefficient for stimulated emission
BF	forward current gain (SPICE)
BR	inverse mode current gain (SPICE)
BV	reverse breakdown voltage (SPICE)
c	speed of light
C	capacitance
C'_B	substrate (bulk) capacitance per unit area
C_{GD}	gate-to-drain capacitance
C_{GS}	gate-to-source capacitance
C_{in}	input capacitance
C_j	junction capacitance
C_{jBC}	collector-base junction capacitance
C_{jBE}	emitter-base junction capacitance
C_{JC}	zero-bias base-collector junction capacitance (SPICE)
C_{jD}	drain junction capacitance
C_{JE}	zero-bias base-to-emitter junction capacitance (SPICE)
C_{jS}	source junction capacitance
C_L	load capacitance
C'_{ox}	oxide capacitance per unit area

C_{OD}	drain overlap capacitance
C_{OS}	gate to source overlap capacitance
C_{out}	output capacitance
C_{sc}	stored-charge capacitance
C_{scBC}	collector-base stored-charge capacitance
C_{scBE}	emitter-base stored-charge capacitance
C_w	stray wiring capacitance
C_μ	capacitance between collector and base in hybrid-pi model
C_π	capacitance between base and emitter in hybrid-pi model
d	length of optical cavity
D_n	diffusion coefficient for electrons
D_{nB}	diffusion coefficient for electrons in base (npn)
D_p	diffusion coefficient for holes
D_{pC}	diffusion coefficient for holes in collector (npn)
D_{pE}	diffusion coefficient for holes in emitter (npn)
E	energy
E_a	activation energy
E_0	reference energy; ground state energy
E_A	acceptor energy; acoustic phonon energy
E_A^*	effective acceptor energy
E_B	energy barrier height
E_C	energy at the bottom of the conduction band
E_{Cn}	conduction band edge in n-type material
E_{Cp}	conduction band edge in p-type material
E_{C0}	conduction band edge of non-degenerately doped material
ΔE_C	change in conduction band edge due to degenerate doping
E_D	donor energy
E_D^*	effective donor energy
E_i	intrinsic Fermi level
E_f	Fermi level
E_{fm}	Fermi level in the metal (or polySi)
E_{fn}	quasi-Fermi level for electrons
E_{fp}	quasi-Fermi level for holes
E_{fs}	Fermi level in the semiconductor
E_g	band gap
E_{gn}	band gap in n-type material
E_{gp}	band gap in p-type material

E_{g0}	band gap of nondegenerately doped material
ΔE_g	change in band gap due to degenerate doping
E_g^*	apparent band gap due to degenerate doping
ΔE_g^*	impurity-induced apparent band-gap narrowing
ΔE_{gBE}^*	apparent band-gap narrowing for base-emitter junction
E_K	kinetic energy
E_n	n th energy level
E_P	potential energy
E_{pho}	optical phonon energy
E_{phonon}	phonon energy
E_T	trap energy level
E_V	energy at the top of the valence band
$E_{V\text{bulk}}$	valence band edge in bulk
E_{Vn}	valence band edge in n-type material
E_{Vp}	valence band edge in p-type material
E_{vac}	vacuum energy level
\mathcal{E}	electric field; true electric field
\mathcal{E}_L	longitudinal electric field
\mathcal{E}_{Lc}	critical longitudinal electric field
\mathcal{E}_e^*	effective electric field for electrons
\mathcal{E}_h^*	effective electric field for holes
\mathcal{E}_{\max}	maximum electric field
\mathcal{E}_T	transverse electric field
$\mathcal{E}_{T\text{eff}}$	effective transverse electric field
f	frequency
f_{co}	cutoff frequency
$f(E)$	probability of occupancy of a state at energy level E by an electron
$f_p(E)$	probability of occupancy of a state at energy level E by a hole
f_T	unity current gain frequency
F	force
F_e	force on electrons
FF	fill factor
F_h	force on holes
F_{Hn}	Lorentz force on electrons
F_{Hp}	Lorentz force on holes
F_L	photon flux
F_{Li}	incident photon flux

F_n	electron flux
FO	fan-out
F_p	hole flux
g	degeneracy
$g(v)$	lineshape function
g_d	output conductance
g_{dsat}	saturation output conductance
g_i	degeneracy of i th state
g_m	transconductance
g_{msat}	saturation transconductance
G	generation rate
G_L	optical generation rate
G_n	electron generation rate
$G_{n(\text{th})}$	thermal electron generation rate
$G_{n(\text{op})}$	optical electron generation rate
G_{op}	optical generation rate
G_p	hole generation rate
G_P	small-signal conductance of the transition region
G_{th}	thermal generation rate
GTO	gate turn-off thyristor
h	Planck's constant; emitter width
\hbar	h-bar ($h/2\pi$)
HOT	higher-order term
\hat{i}	unit vector in the x -direction
i_b	small-signal base current
i_c	small-signal collector current
i_d	small-signal drain current
i_e	small-signal emitter current
i_g	small-signal gate current
I	current
I_A	anode current
I_B	base current
I_C	collector current
I_{C0}	collector-base junction leakage current
I_{CT}	collector current (Ebers-Moll)
I_D	drain current
I_{dark}	dark current

I_{Dn}	drain current in NFET
I_{Dp}	drain current in PFET
I_{Dsat}	saturation current
I_{Dsatn}	saturation current in NFET
I_{Dsatp}	saturation current in PFET
I_E	emitter current
I_{EE}	current flowing from emitter to ground (or emitter supply)
I_F	forward current
I_{F0}	forward saturation current (Ebers-Moll)
I_G	gate current
I_K	cathode current
I_{KF}	forward knee current
I_{KR}	inverse knee current
I_L	photocurrent exclusive of dark current
I_m	operating current for a solar cell
I_n	electron current
I_{nC}	electron current that crosses the base from the emitter and reaches the collector (npn)
I_{nE}	electrons injected from emitter to base (npn)
I_0	dark current, leakage current
I_p	hole current
I_{pC}	collector hole current
I'_{pC}	hole current due to electron-hole generation in B-C junction
I''_{pC}	hole current extracted from collector into the base (npn)
I_{pE}	hole current injected from base to emitter (npn)
I_{rec}	recombination current (in base)
I_R	reverse current
I_{R0}	reverse saturation current (Ebers-Moll)
I_{SC}	short-circuit current
j	$\sqrt{-1}$
\hat{j}	unit vector in the y direction
J	current density
J_C	collector current density
$J_{C\max}$	maximum collector current density to avoid excess base push-out
J_D	photocurrent generated in the depletion region
J_{diff}	diffusion current density
J_{drift}	drift current density

J_E	emitter current density
J_{E0}	emitter saturation current density
J_G	generation current density
J_{GR}	generation-recombination current density
J_{GR0}	generation-recombination leakage current density
J_L	photoinduced current density
J_n	electron current density
J_{nB}	electron diffusion current density in base (npn)
J_{ndiff}	electron diffusion current density
J_{ndrift}	electron drift current density
J_{np}	current density due to electrons in p-type material
J_0	diffusion dark current, density; diffusion leakage current density
J_{pn}	current density due to holes in n-type material
J_p	hole current density
J_{pB}	hole current density in base
J_{pC}	hole diffusion current density in collector (npn)
J_{pE}	hole diffusion current density in emitter (npn)
J_{pdiff}	hole diffusion current density
J_{pdrift}	hole drift current density
J_S	a leakage current that is a function of both J_0 and J_{GR}
\hat{k}	unit vector in the z direction
k	Boltzmann's constant
K	wave vector for electrons; propagation constant
K_A	wave vector for acoustic phonon
K_e	wave vector for an electron
K_i	wave vector for incident particle
K_{phn}	wave vector for a phonon
K_{pht}	wave vector for a photon
K_t	wave vector for transmitted particle
K_x	x component of K
K_y	y component of K
K_z	z component of K
\bar{l}	mean free path
L	length; channel length
ΔL	channel length shortening
L_{eff}	effective channel length

L_m	metallurgical channel length
L_{\min}	minimum channel length such that short-channel effects need not be accounted for
L_n	diffusion length for electrons, gate length for NMOS
L_{nB}	diffusion length for electron in base (npn)
L_p	diffusion length for holes; gate length for PMOS
L_{pE}	diffusion length for hole in emitter (npn)
m	mass
m_0	mass of a free electron
m^*	effective mass
m_x^*	effective mass for an electron traveling in the x direction
m_y^*	effective mass for an electron traveling in the y direction
m_z^*	effective mass for an electron traveling in the z direction
m_{ce}^*	conductivity effective mass for electrons
m_{ch}^*	conductivity effective mass for holes
m_{dse}^*	density of states effective mass for electrons
m_{dsh}^*	density of states effective mass for holes
m_e^*	effective mass for electrons
m_h^*	effective mass for holes
m_{lh}^*	effective mass for holes in the light-hole band
m_{hh}^*	effective mass for holes in the heavy-hole band
m_{sh}^*	effective mass for holes in the split-off band
$m_{ }^*$	longitudinal effective mass for electrons
m_{\perp}^*	transverse effective mass for electrons
M	mass of a particle; multiplication factor; collection efficiency
MTTF	mean time to failure
n	electron concentration; quantum number in Bohr model; refractive index; diode quality factor
n_B	electron concentration in base
n_{BC}	base-to-collector recombination current emission coefficient
n_{BE}	base-to-emitter recombination current emission coefficient
n_{B0}	equilibrium electron concentration in base (npn)
$n_{B(\text{norm})}$	normalized electron concentration in base
n^+	heavily doped n-type material
Δn	excess electron concentration
Δn_B	excess electron concentration in base (npn)
Δn_p	excess electron concentration in p-type material
$n(E)$	distribution of electrons with energy

n_{E0}	equilibrium concentration of electrons in emitter
n_i	intrinsic electron concentration
n_{\min}	minimum electron concentration in depletion region
n_{n0}	equilibrium concentration of electrons in n-material
n_p	electron concentration in p-type material
n_{p0}	equilibrium concentration of electrons in p-material
n_0	equilibrium electron concentration
n_s	electron concentration at the surface
N_A	concentration of acceptors
N'_A	net concentration of acceptors ($N_A - N_D$)
N'_{AB}	net concentration of acceptors in base (npn)
N_C	effective density of states of electrons in the conduction band
N_D	concentration of donors
N'_D	net concentration of donors ($N_D - N_A$)
N'_{DC}	net concentration of donors in collector (npn)
N'_{DE}	net concentration of donors in emitter (npn)
N_{ii}	number of implanted impurity atoms per unit area
N_T	concentration of traps
N_V	effective density of states for holes in the valence band
O_{op}	quantum mechanical operator
p	hole concentration; classical momentum
p_{ac}	ac power
p_B	hole concentration in base
p_{B0}	equilibrium hole concentration in base
p_{dc}	dc power
p_E	hole concentration in emitter
p_{E0}	equilibrium hole concentration in emitter (npn)
P_m	maximum power that can be produced by a solar cell at a given illumination condition
p_n	hole concentration in n-type material
p_0	equilibrium concentration of holes
p_{n0}	equilibrium concentration of holes in n-type material
p_{p0}	equilibrium concentration of holes in p-type material
p^+	heavily doped p-type material
p_{C0}	equilibrium hole concentration in collector (npn)
Δp	excess hole concentration
Δp_C	excess hole concentration in collector (npn)
Δp_E	excess hole concentration in emitter (npn)

Δp_n	excess hole concentration in n-type material
$p(E)$	distribution of holes with energy
p_0	equilibrium hole concentration
$P(x, t)$	probability density
P	probability that a carrier generates an electron-hole pair in avalanche breakdown; power
P_{dynamic}	dynamic power dissipation
q	absolute value of the charge of an electron ($q = 1.6 \times 10^{-19}$ C); longitudinal mode number in laser
Q	charge
Q_B	depletion or bulk charge; stored charge in the base
Q_s	stored charge
Q_{ch}	channel charge (per unit area) (mobile channel charges that carry current)
Q_{chlf}	channel charge at low field
Q_{chmin}	minimum channel charge
Q_f	fixed oxide charge
Q_{ii}	implanted charge per unit area
Q_{it}	interface trapped charge
Q_m	mobile ion charge
Q_{ot}	oxide trapped charge
Q_{sR}	reclaimable stored charge
Q_{ss}	surface charge density
Q_V	charge density (charge per unit volume)
Q_{VT}	tunneling charge volume density
r	radius; position; distance
r_b	series base resistance
r_c	collector resistance
r_e	emitter resistance
r_o	output resistance
r_n	radius of n th energy level orbit (Bohr model)
r_μ	feedthrough resistance (between base and collector)
r_π	differential input resistance (BJT)
\vec{r}	position vector
R	resistance; recombination rate; reflection coefficient
R_{\square}	sheet resistance in the base
R_B	effective resistance under the emitter; base resistor
R'_B	resistance from base contact to emitter edge

R_{BD}	drain-to-substrate resistance
R_{BS}	source-to-substrate resistance
R_C	collector resistor
R_D	drain resistance
R_E	emitter resistor
R_H	Hall coefficient
R_{Hn}	Hall coefficient for electrons
R_{Hp}	Hall coefficient for holes
R_L	load resistance
R_{\max}	maximum recombination rate
R_n	electron recombination rate
$R_{\text{on,sp}}$	specific on-resistance
R_p	hole recombination rate; parallel resistance; small-signal resistance of the transition region
R_{ph}	responsivity
R_S	series resistance; source resistance
S	gate voltage swing, snappiness
$S(E)$	density of states
$S_a(E)$	number of states at a particular energy E_a
S_{\min}	minimum voltage swing
$S_V(E)$	density of states for holes
t	time; thickness
t_{BC}	base-collector transit time
t_d	propagation delay time
t_{dr}	rise time
t_{df}	fall time
t_{lh}	low-to-high transition time
\bar{t}_n	mean free time between collisions for electrons
t_{nii}	scattering time for electrons due to ionized impurity scattering
t_{nl}	scattering time for electrons due to lattice scattering
t_{ox}	oxide thickness
\bar{t}_p	mean free time between collisions for holes
t_s	storage time
t_T	base transit time
t_{TF}	base transit time in forward mode
t_{TR}	base transit time in reverse mode
T	temperature, transmission coefficient; period, tunneling probability

$T(t)$	time-dependent part of wavefunction
U_K	Bloch function
v	velocity
v_a	applied ac voltage
v_{be}	small-signal base-to-emitter voltage
v'_{be}	small-signal voltage drop across r_π
v_{ce}	small-signal base-to-collector voltage
v_d	small-signal drain voltage
v_{dn}	electron drift velocity
v_{dp}	hole drift velocity
v_g	group velocity; small-signal gate voltage
v_{\max}	maximum velocity
v_n	velocity of electron in n th energy level
v_p	phase velocity
v_{sat}	saturation velocity
v_{satn}	saturation velocity in n-type material
v_{satp}	saturation velocity in p-type material
V	voltage
V_a	applied voltage
V_A	Early voltage
V_{AF}	Early voltage forward mode
V_{AK}	anode-to-cathode voltage
V_{AR}	Early voltage reverse mode
V_B	substrate-to-source voltage, base voltage
V_{BE}	base-to-emitter voltage
V_{BF}	forward breakdown voltage
V_{bi}	built-in voltage
$V_{\text{bi}BC}$	built-in voltage of collector-base junction
$V_{\text{bi}EB}$	built-in voltage of emitter-base junction
V_{BR}	reverse breakdown voltage
V_{ch}	channel voltage
V_E	emitter voltage
V_C	collector voltage
V_{CC}	collector supply voltage (in bipolar circuits)
V_{CE}	collector-to-emitter voltage
V_{CB}	collector-to-base voltage
V_D	drain voltage

V_{DD}	drain supply voltage in FET circuits
V_{Dn}	drain voltage in NFET
V_{Dp}	drain voltage in PFET
V_{DS}	drain-to-source voltage
$V_{D\text{sat}}$	drain saturation voltage
V_F	forward voltage
V_{FB}	flat band voltage
V_{FP}	forward peak voltage
V_G	gate voltage
V_{GS}	gate-to-source voltage
V_H	Hall voltage
V_{Hn}	Hall voltage for electrons
V_{Hp}	Hall voltage for holes
V_{in}	input voltage
V_j	junction voltage
V_j^n	junction voltage appearing across n side
V_j^p	junction voltage appearing across p side
V_{jCB}	junction voltage of collector-base junction
V_{jBE}	junction voltage of base-emitter junction
V_m	operating voltage for a solar cell
V_{OC}	open-circuit voltage
V_{out}	output voltage
V_ρ	voltage in resistivity measurement
V_R	reverse voltage
V_S	source voltage
V_{SS}	source supply voltage in a MOS circuit
V_{sub}	substrate voltage
V_T	threshold voltage
V_{Tn}	threshold voltage for NFET
V_{Tp}	threshold voltage for PFET
w	depletion region width
w_B	depletion width in bulk or substrate
w_{BC}	width of base-collector depletion region
w_D	drain depletion width
w_n	width of depletion region on n side
w_p	width of depletion region on p side

w_{pEB}	width of p side of depletion region of base-emitter junction (npn)
w_{pCB}	width of p side of depletion region of collector-base junction
w_s	source depletion width
w_T	width of depletion region at source at and above threshold
W	width; gate width
W_B	width of quasi-neutral region in base
W_{BM}	width of base region between metallurgical junctions
W_E	width of quasi-neutral region in emitter
W_{EM}	width of the emitter region between metallurgical junctions
W_n	gate width of NFET
W_{nC}	width of quasi-neutral region in n collector (npn)
W_{nCM}	width of n-collector region between metallurgical junctions
W_p	gate width of PFET
W_B	base width
W_T	tunneling width
x	position
x_B^-	base edge of base-collector depletion region
x_B^+	collector edge of base-collector depletion region
x_n	location of edge of depletion region in n-type material
x_0	location of the metallurgical junction
x_p	location of edge of depletion region in p-type material
y	position, position along channel
y_{be}	admittance from base to emitter
z	position
Z_{in}	input impedance
α	common base current gain; absorption coefficient
α_F	common base current gain for forward operation
α_R	common base current gain for reverse operation
α_T	base transport factor
β	probability that an electron will recombine with a hole in a given time; common emitter current gain in a BJT
β_{DC}	low-frequency current gain
β_F	common emitter current gain for forward operation
β_R	common emitter current gain for reverse operation
γ	ionization energy; injection efficiency
γ_n	ionization energy in n-type material
γ_p	ionization energy in p-type material

δ	fraction of reclaimable charge
δ_n	energy difference between conduction band edge and Fermi level ($E_C - E_f$)
δ_p	energy difference between Fermi level and valence band edge ($E_f - E_V$)
ϵ	permittivity
ϵ_0	permittivity of free space
ϵ_{ox}	permittivity of oxide
ϵ_r	relative permittivity
ϵ_{SiO_2}	permittivity of silicon dioxide
η	doping parameter (W_B/λ); power conversion efficiency
η_Q	quantum efficiency
θ	mobility modulation factor
θ_{cr}	critical angle
θ_{\perp}	divergence angle perpendicular to the junction plane
θ_{\parallel}	divergence angle parallel to the junction plane
λ	wavelength; doping characteristic length
μ	mobility
μ_H	Hall mobility
μ_{lf}	low field mobility
μ_n	electron mobility
μ_{nii}	mobility of electrons due to ionized impurity scattering
μ_{nl}	mobility of electrons due to lattice scattering
μ_0	zero-field mobility under effect of transverse field
μ_p	hole mobility
ν	frequency
ρ	resistivity
σ	conductivity
σ_0	conductivity at equilibrium, in the dark
σ_n	conductivity due to electrons
σ_p	conductivity due to holes
σ_{pc}	conductivity due to photocarriers
τ_0	lifetime (generic)
τ_D	dielectric relaxation time
τ_n	electron lifetime
τ_{nB}	electron lifetime in base (npn)
τ_T	minority carrier transit time
τ_p	hole lifetime
τ_{pC}	hole lifetime in n-type collector (npn)

τ_{pE}	hole lifetime in n-type emitter (npn)
$\tau_{\text{radiative,spont}}$	radiative lifetime for spontaneous emission
ϕ_{ox}	voltage drop across oxide
$\phi_{\text{ox}}^{\text{th}}$	oxide voltage at threshold
ϕ_f	potential difference between Fermi level and intrinsic level (in volts)
ϕ_s	surface potential (in volts)
Φ	work function
Φ_M	work function in metal
Φ_n	work function in n-type material
Φ_p	work function in p-type material
Φ_s	work function in semiconductor
χ	electron affinity
χ_n	electron affinity in n-type material
χ_p	electron affinity in p-type material
χ_s	electron affinity in the semiconductor
ψ	time-independent wave function
ψ_n	time-independent wave function for n th state
ψ_r	time-independent wave function for reflected particle
ψ_t	time-independent wave function for transmitted particle
Ψ	wave function
ω	angular frequency
ω_{phn}	phonon angular frequency

C APPENDIX

Fabrication

C.1 INTRODUCTION

Here we briefly examine some of the fabrication techniques used in the semiconductor industry. Integrated circuits, for example, may combine billions of transistors on a single chip. A process of photolithography is used to define the structures, and various techniques are used to dope the different regions n type or p type. Furthermore, metal or other conductors must be laid down to provide the interconnections, as well as insulating layers to avoid short circuits. The circuits are mass produced on wafers and then cut up into individual chips after being fabricated. The chips are mounted and connections to the outside world are made. Finally the packaging process is completed and the devices are shipped. The emphasis here is on silicon. The entire process starts with the production of ultrapure silicon.

C.2 SUBSTRATE PREPARATION

For semiconductor device fabrication, the industry requires ultrapure and defect-free silicon crystals. Although the semiconductor is intentionally doped with alien elements, very precise control is needed over the concentration of those elements, because other elements may influence the electrical characteristics of the final device. Impurities such as gold and copper can create trap states in the forbidden gap, which can collect and freeze carriers in the lattice.

Similarly, crystalline defects can also create problems. Dangling bonds and other defect states can trap carriers or provide mechanisms by which they can recombine. When carriers recombine, current is lost.

Thus, there is an art as well as a science to fabricating electronic-grade, defect-free, pure silicon. Here some fabrication processes involved in producing high-quality silicon for integrated circuits are discussed.

C.2.1 The Raw Material

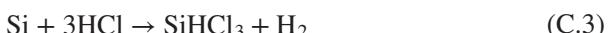
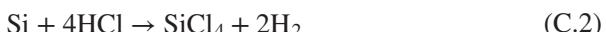
One of the reasons integrated circuit (IC) technology is inexpensive is that silicon is highly abundant. The process of producing silicon substrates actually starts with SiO_2 (silica), which is the primary ingredient of sand.

The silica is heated to about 1800°C in the presence of carbon, to get rid of the oxygen. The carbon reacts with the oxygen to create carbon monoxide, leaving silicon behind:



The resultant Si is on the order of 95 percent pure. Although it is suitable for metallurgical applications, this silicon contains too many impurities to be usable for electronics.

Next, the silicon is reacted with hydrochloric acid. The silicon reacts with the chlorine to produce either silicon tetrachloride (SiCl_4) or trichlorosilane (SiHCl_3), both of which are liquids.



Other impurities, particularly iron, also react with the chlorine, and the resulting compounds can be distilled out. After several distillation steps at different temperatures, the silicon-chlorine compound is pure enough (on the order of 99.9999 percent pure) to proceed to the next step.

The ultrapure SiCl_4 or SiHCl_3 is then reacted with hydrogen to form Si, for example via



Although silicon naturally forms a diamond crystal, at this stage it is polycrystalline. The next step, then, is to grow a single, defect-free crystal from the high-purity polycrystalline silicon. The process of obtaining a single crystal of Si incorporates more impurities, however, particularly oxygen and carbon.

C.2.2 Crystal Growth

To obtain a large single crystal of silicon, one starts with a single crystal seed. It is carefully oriented so that the surface that will be growing is a particular crystal plane, typically (111) or (100). Then the resulting large crystal also has the desired orientation.

One common method for growing a single crystal ingot of silicon is the *Czochralski method*. Here the seed crystal is dipped into a crucible of molten silicon, Figure C.1. The molten silicon may be doped n type or p type in the melt to produce a doped silicon sample. As the seed crystal is slowly pulled out, the silicon in the melt near the seed cools off and crystallizes onto the seed, extending the crystal downward and outward. In this process, the seed crystal is effectively extended, the new material expanding the edges of the seed crystal. The diameter

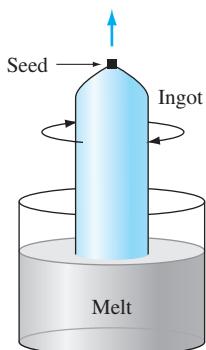


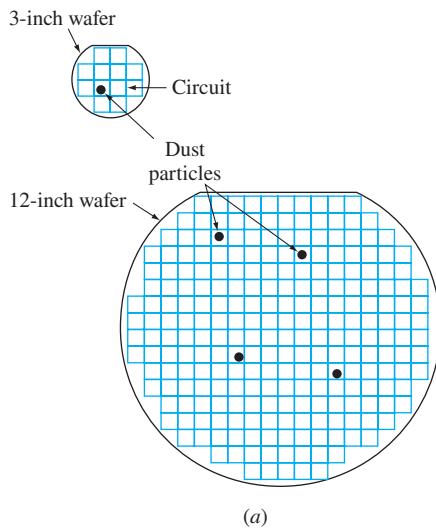
Figure C.1 In the Czochralski method, a seed crystal is pulled slowly from a melt of pure silicon.

of the ingot increases, depending on the rate at which the seed is pulled. When the growing crystal has the desired diameter, the rate of pulling is adjusted so that the diameter remains constant.

The process has to be very carefully controlled, however. If the pull rate slows down, the crystal (boule) gets wider; if it speeds up the boule gets narrower. The crystal is also rotated for better uniformity. Furthermore, the crystal has to be grown slowly (on the order of 50 mm per hour) to avoid defects, and the temperature has to be carefully controlled. This process is difficult but has developed over the years. In fact, one way to date photographs of silicon fab (fabrication) lines is by the wafer size. In the very early days, the boules could only be grown reliably in 1- or 2-inch diameters. Wafers 3 inches in diameter appeared around the 1970s. At the end of the twentieth century, 200-mm (8-inch) wafers were common and 300-mm (12-inch) wafers were in production.

Larger wafers mean greater economy of scale. Consider a very large scale integrated (VLSI) circuit, such as a microprocessor chip with a few million transistors on it. The larger the wafer, Figure C.2a, the more circuits can fit on it, and thus the more economical the process. Figure C.2b shows a 200-mm wafer holding 130 complete circuits (16-Mbyte DRAMs).

Consider what happens if a piece of dust gets on the wafer during the process. A piece of dust can ruin many transistors at a time, but if even one transistor is destroyed, the whole circuit may have to be discarded. In the small wafer, if one of the 12 circuits is rejected, the yield of good devices is about 92 percent. In the large wafer case, for the same level of contaminants per unit area, more



(a)



(b)

Figure C.2 (a) The larger the wafer, the more circuits it can contain, and thus the more cost-effective the process. (b) A 200-mm wafer of 16 Mbyte dynamic random-access memory (DRAM) chips. b. © John Madere/Fuse/Getty Images

circuits may be lost, but the percentage yield is much greater. For the example in the figure, the large wafer has four bad circuits but still produces a better than 98 percent yield.

Furthermore, the ovens must be heated up the same number of times and the processes executed the same number of times to produce a 3-inch wafer as a 12-inch wafer, so the economies of scale are enormous. The downside is that the entire fabrication line has to be designed for a particular wafer size. The tubes of the ovens, the boats (wafer holders), and everything must be matched to the size of the wafers. Therefore, it is expensive to upgrade a given line.

Another method sometimes used to produce single-crystal Si of the appropriate diameter is the *float-zone process*. This method is used to obtain material of greater purity than can be obtained by the Czochralski technique. The starting material is a bar of cast polycrystalline Si. This bar is held in a vertical position as indicated in Figure C.3. Here, an RF coil or other heating device is passed along the ingot, melting the material locally. As the heated zone moves, one end of the melted region is starting to melt and the other end is recrystallizing. As the material solidifies, however, some of the contaminants may preferentially stay in

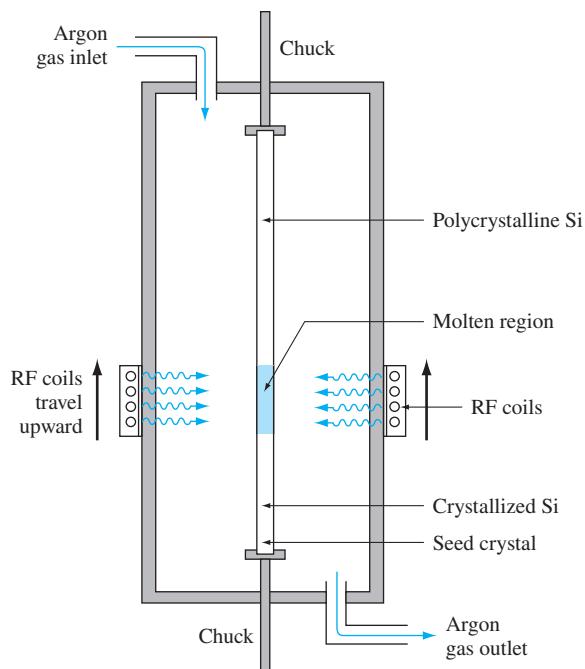


Figure C.3 Float-zone refining. The melted region moves along the boule. As the material on the trailing end of the melted region cools, the impurities preferentially stay in the melt and travel along in the liquid zone to the end, where they are cut off. The process is generally repeated several times.

the liquid rather than the solid phase. This is controlled by the *segregation coefficients* of the impurities. The segregation coefficient is the ratio of a particular substance that remains in the liquid form of another substance compared with the fraction that freezes out.

As the liquid region moves along, the impurities are carried along with it. The process is repeated several times to purify the ingot as much as possible. The impurities accumulate at one end, which can be cut off and discarded.

This float-zone process is more expensive than the Czochralski process, and, because of the difficulties associated with maintaining a large-diameter molten zone, the resultant crystal diameter is less than for the Czochralski technique. Once the boule is ready, it is sawn into wafers, and then polished.

C.2.3 Defects

The silicon crystal must be of very high purity, but it must also be free of crystal defects. Figure C.4 shows three common crystal defects that can occur; there are many others.

The interstitial defect occurs when an extra atom is inserted into the crystal but doesn't necessarily bind to the neighboring atoms. It does, however, distort the lattice, disrupting the periodicity locally and altering the energy band structure.

The vacancy defect occurs when a lattice site is empty. This can distort the lattice and also create dangling bonds. The dangling bonds, in turn, can attract electrons or holes, causing trap states or recombination sites. These traps can collect carriers and reduce the conductivity of the sample.

The edge defect is one type of line defect. Here an extra plane of atoms exists in the lattice and it ends abruptly. This can also cause dangling bonds and distorts the lattice as well.

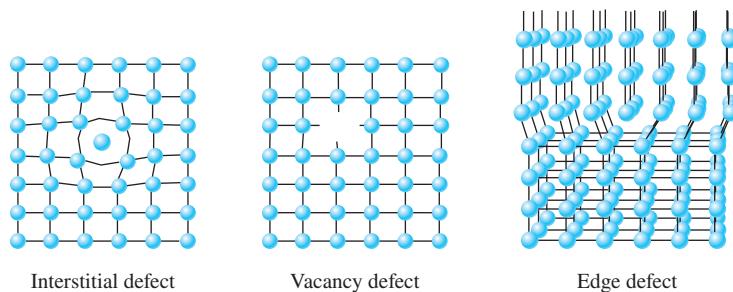


Figure C.4 Some simple crystal defects. The interstitial and vacancy defects are known as *point defects*, and the edge defect is one type of line defect.

C.2.4 Epitaxy

Often, we wish to grow a uniform layer of one type of semiconductor onto a substrate. For example, one can grow a layer of silicon with one doping level onto a substrate of another. We require that the new layer form a continuous crystal with the substrate, so that there are no defects or interface states that can trap carriers. For this, an epitaxial (epi) growth process is used. In epitaxy, a thin layer of crystal is grown, rather than simply deposited onto the substrate wafer, and the substrate wafer acts as a seed crystal.

Lattice Matching In semiconductors, we often wish to create heterojunctions, meaning that one type of crystal is grown atop another. For example, one may wish to construct a heterostructure in which a layer of InP is sandwiched between two layers of InGaP. One cannot purchase InGaP substrates, so one must start with an available substrate material that is lattice matched.

Lattice matching means that the lattice constants of the two materials must be as nearly equal as possible. Consider what would happen if one attempted to grow the material in Figure C.5 on the substrate shown. The upper material has a slightly larger lattice constant. As the atoms land on the surface and try to fit into the crystal structure, they will bond with the atoms in the layer below. Since the upper material's atoms are normally spaced farther apart, defects will occur occasionally unless the lattice constants of the two materials are very well matched.

In addition, even where defects do not occur, there is strain on the lattice near the junction. The epi layer may have to stretch slightly to match the substrate, or it may compress. This strain changes the lattice constant of the epi layer locally, and that in turn will affect the band gap. As the epitaxial layer grows thicker, the atoms will eventually assume their normal spacing, so the layer is strained only near the interface.

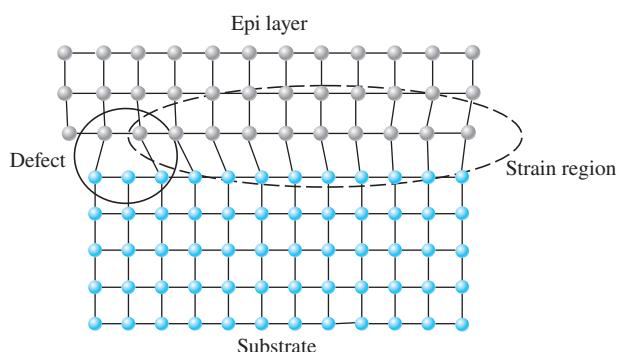


Figure C.5 In heteroepitaxy, the two crystals must be well lattice matched or defects will occur. When the lattice constants are different, one layer may become strained, which changes that material's periodicity and thus the band-gap properties.

If, on the other hand, the epitaxial layer is thin, and another layer of the substrate material is grown on top, then the outer layers can keep the epitaxial layer distorted. Strained-layer epitaxy is being exploited increasingly as engineers get more sophisticated in growth techniques, leading to many band-gap engineering opportunities.

To get back to our example, if one wanted to grow an InGaP-InP-InGaP double heterostructure, one would start with InP, which is a readily available substrate. One would then grow the first layer of InGaP, followed by growth of the InP layer, followed by the InGaP cap layer.

Figure C.6 shows the lattice constants of some common semiconductor crystals. The lines connecting two binary compounds indicate the ternary compounds. For example, tracing the line that connects GaAs to InAs, we can find the lattice constant and band gap for any $In_xGa_{1-x}As$ ternary compound. The x indicates the ratio of indium to gallium in this case. The compound $In_{0.53}Ga_{0.47}As$ has exactly the same lattice constant as InP, making this a popular combination for heteroepitaxy.

Vapor-Phase Epitaxy (VPE) Vapor-phase epitaxy (VPE), or chemical vapor deposition (CVD), is a technique commonly used to grow silicon layers on silicon. The silicon wafers are placed in a chamber, Figure C.7, and exposed to an atmosphere containing, for example, $SiCl_4$. The wafers are heated to about $1200^{\circ}C$ so that the overall reaction



can take place. The silicon produced by this reaction can adhere onto the surface of the crystal and the crystal grows.

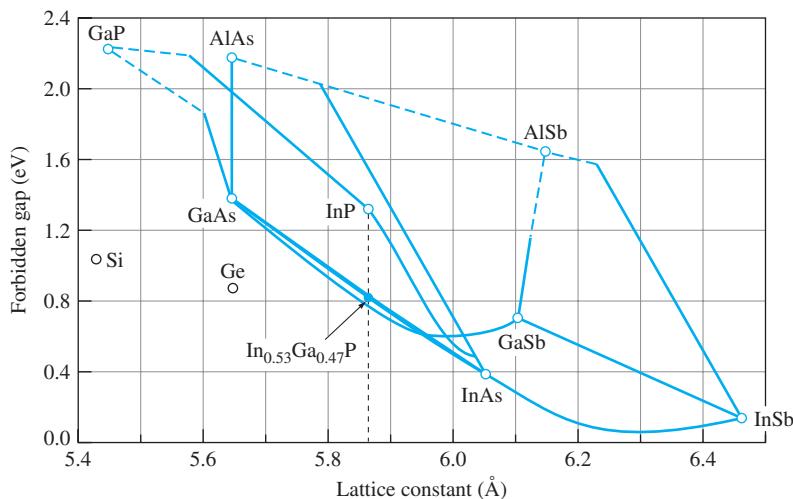


Figure C.6 Energy band gap versus lattice constant for several common compound semiconductors. We can see that the ternary compound $In_{0.53}Ga_{0.47}As$ is lattice matched to InP.

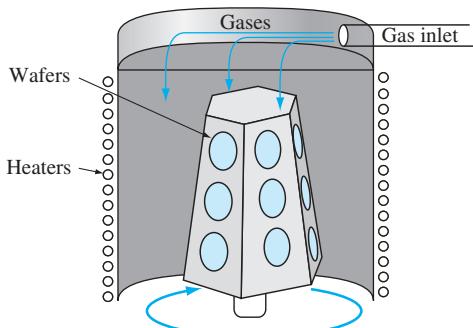


Figure C.7 In vapor-phase epitaxy, the materials are introduced in gaseous form.

VPE is also used to grow III-V compounds. For example, a GaAs substrate can be exposed to an atmosphere of arsine (AsH_3), phosphine (PH_3), and gallium chloride (GaCl) gases. The gallium from the GaCl attaches to the existing crystal in the lattice sites where the column III element normally goes. In the alternative sites, either of the two column V elements can attach. The ratio of As to P in this GaAsP film is controlled by the ratio of the arsine to phosphine gasses.

Metal-Organic Vapor-Phase Epitaxy (MOCVD) Vapor-phase epitaxy does not work ideally for all systems, however. Particularly, compounds containing aluminum are difficult to grow in accurately controlled compositions because the aluminum does not diffuse well on the surface (to find its correct place in the lattice) and because of its high activity, among other reasons. A variation of the epitaxial growth technique that helps with this is called metal-organic chemical-vapor deposition (MOCVD). Here the Ga and Al metals are introduced in organic compounds such as trimethyl gallium, $\text{Ga}(\text{CH}_3)_3$, and trimethyl aluminum, $\text{Al}(\text{CH}_3)_3$. The arsenic and phosphorus are introduced in arsine and phosphine gases as before. MOCVD is capable of growing monolayers (layers one atom thick), which makes possible abrupt changes in composition and highly precise control.

Molecular Beam Epitaxy (MBE) A highly versatile technique for growing epitaxial layers is known as *molecular beam epitaxy* (MBE). It might be considered as sort of a solid-phase epitaxy. The individual elements (and dopants) are heated in their separate crucibles in the MBE machine under high vacuum. As the atoms evaporate, they travel to the substrates and are deposited. The gates to the individual crucibles can be opened and closed to vary the composition of the layers as indicated in Figure C.8.

MBE can also put down monolayers for extremely precise control of material growth.

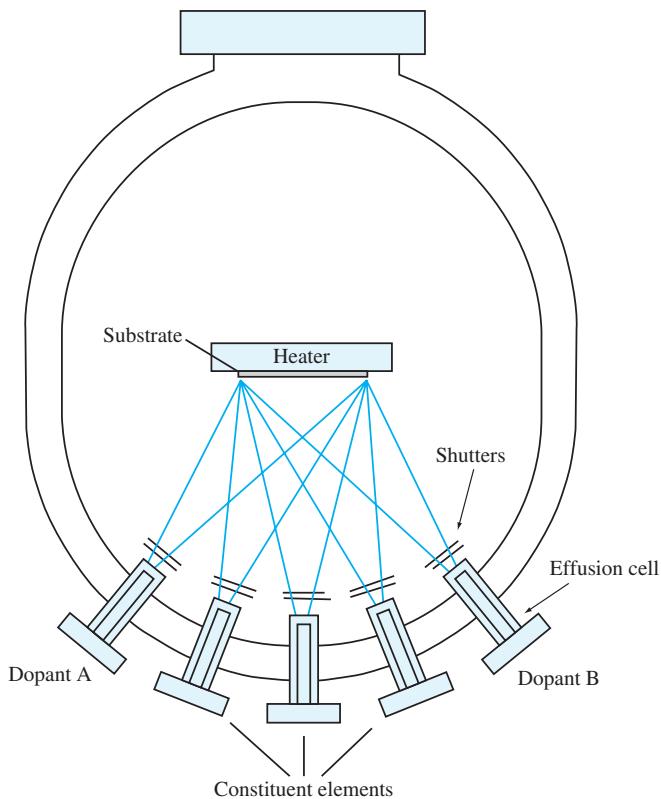


Figure C.8 Schematic diagram of a molecular beam epitaxy (MBE). (Source: Modified from S. K. Ghandi, *VLSI Fabrication Principles: Silicon and Gallium Arsenide*, 2nd ed., p. 294, John Wiley and Sons, 1994.)

C.3 DOPING

A substrate may be n type or p type (or intrinsic), but integrated circuit engineers need to be able to make certain regions be of the opposite type, to create diodes and transistors. There are two major techniques for doping an existing semiconductor crystal: diffusion and ion implantation.

C.3.1 Diffusion

In diffusion, a substrate at an elevated temperature is exposed to an atmosphere containing the desired dopant. For example, to diffuse an n-type layer into a p-type substrate, the wafer is placed in a diffusion furnace containing a gas of an n-type dopant such as phosphorus. The P atoms are in higher concentration in the

atmosphere than in the wafer, so they will diffuse into the surface of the wafer. It requires very high temperatures (800 to 1100°C) for the phosphorus atoms to have enough kinetic energy to work their way into the substrate. The longer the exposure and the higher the temperature, the deeper the diffusion.

In diffusion, the distribution of dopants is not uniform, but is instead more concentrated toward the surface. Figure C.9 shows a plot of the concentration of dopants after diffusion. Note that the material is p type until the phosphorus concentration exceeds the background boron concentration in the p-type wafer. Near the surface where $N_D > N_A$, the material is n type.

A second diffusion, this time with acceptors, could also be performed to make the surface p type, producing a pnp structure. The second diffusion would have to be shallower but with an even higher concentration, to compensate for the donors deposited earlier. Furthermore, when the wafer is heated for the second diffusion, the donors from the first diffusion will diffuse even further into the semiconductor.

Diffusion is not currently used extensively as a primary method for doping semiconductors, because many other processing steps are done at high temperatures, and every time the temperature is raised, whatever dopants exist in the semiconductor will diffuse, always toward regions of lower concentration.

C.3.2 Ion Implantation

A more common way to dope semiconductors is by ion implantation. In this approach, ions of the required dopants are accelerated toward the substrate. The ions have high kinetic energies, from the keV range up to the MeV range. They arrive with such force at the crystal that they are implanted into the surface. The penetration depth may be on the order of a micrometer—about 1800 lattice constants in silicon. This results in significant damage to the crystal; bonds are broken and atoms are dislodged. Thus, ion implantation is followed by an annealing

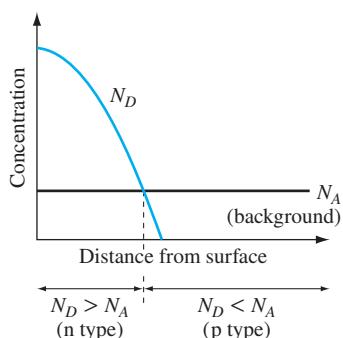


Figure C.9 The dopant profiles after a phosphorus diffusion into a p-type substrate.

step. The crystal is heated so that the atoms can move somewhat easily, breaking bonds and shifting positions. Through this process they tend to fall back and regroup into a natural crystal formation.

Figure C.10 shows a bipolar junction transistor and its dopant profile. This device is fabricated using BiCMOS technology (an integrated circuit that combines

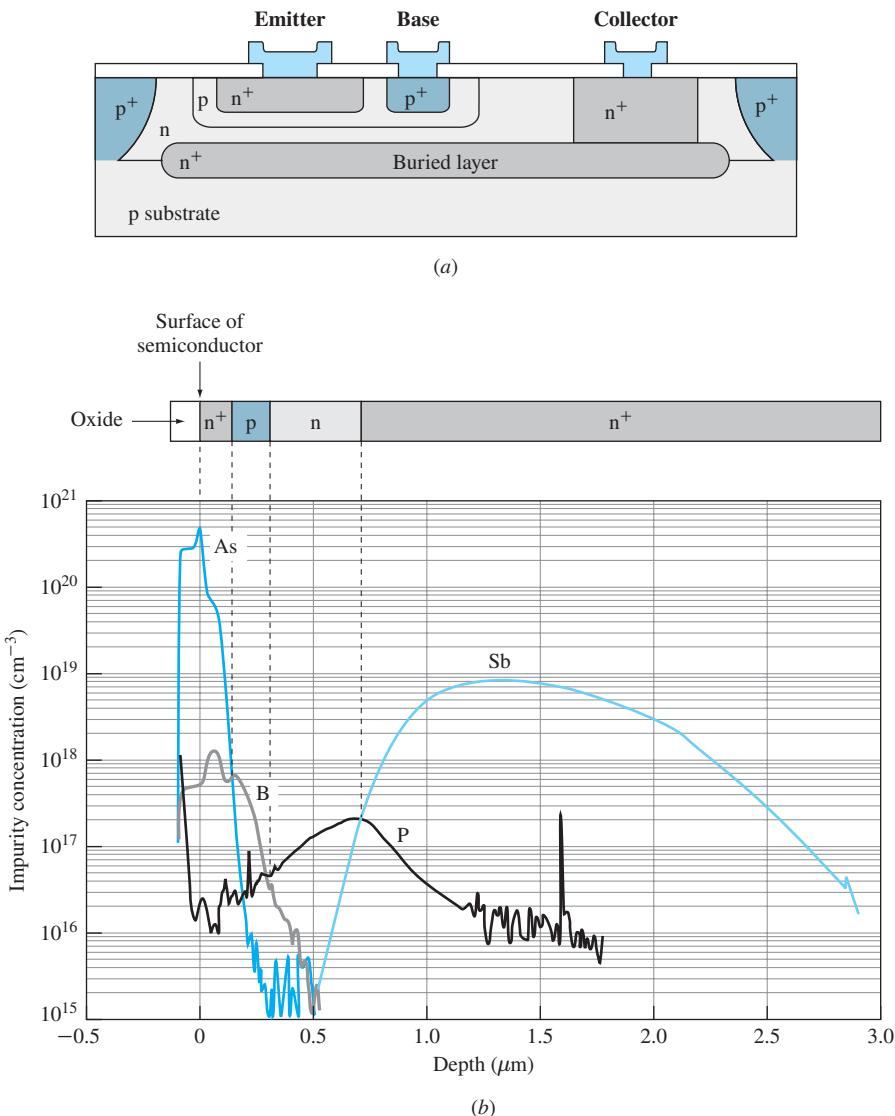


Figure C.10 (a) A cross section of an npn bipolar transistor; (b) a SIMS profile showing the dopant concentration (Source: Data from IBM). The SIMS was done before metallization, and there is an oxide on the surface at this point.

bipolar transistors with complementary metal-oxide-semiconductor field-effect transistors). The plot is measured by a technique known as secondary ion mass spectroscopy (SIMS). It produces a profile of the dopant concentration as a function of depth from the surface. The dopants are normally added by ion implantation.

The deep layer is a heavily doped n-type (n^+) region “buried layer” that is highly conductive. It is doped with antimony, because Sb diffuses slowly in silicon, about an order of magnitude more slowly than the boron and phosphorus used later. Thus, the Sb atoms will not move much during the subsequent processing. The n layer is the *collector* of the transistor, doped with phosphorus. The p-type base (boron) is then implanted. Finally, the surface layer, the *emitter*, must be doped even more heavily so that the donor concentration again exceeds the acceptor concentration. Here arsenic is often used as the emitter dopant because it makes a more uniform emitter-base junction than phosphorus. There is an oxide layer at the surface that also contains dopants, and these also appear on the SIMS plot.

C.4 LITHOGRAPHY

To make transistors and diodes we need to perform the doping only in isolated areas rather than across the entire wafer. In addition, we need to create contacts and connections on the devices. This section indicates how lithography is used to make tiny and precise structures on semiconductor wafers. In lithography, masks with precisely defined features are used to block off parts of the wafer from the various processes.

Let us see how to make a simple diode (Figure C.11). The p-type wafer is covered with a layer of oxide followed by a layer of *photoresist*, Figure C.11a. Photoresist is a photosensitive organic compound, often spun onto the wafer. When it is exposed to light, it undergoes a chemical reaction. For positive photoresist, a developer will remove the photoresist that was exposed to the light. For negative photoresist, the developer removes the unexposed parts. A mask is used, and the photoresist is exposed to light only where the mask is transparent. The photoresist is then developed, leaving holes in the resist.

Next, the wafer is put in the ion implanter. Donor ions embed themselves in the oxide (which is removed later) and the substrate. Where they strike the semiconductor, it becomes n type. In this case an n^+ -type well is created in the p-type substrate. Note that the concentration of donors implanted must exceed the background acceptor concentration to make the material n type where implanted.

Next, a new layer of oxide is grown. This layer is patterned with another mask to allow the ion implantation that creates a p^+ layer for the ohmic p contact.

Another layer of oxide is grown, and another layer of photoresist spun onto the wafer. Now a third mask is used, to form the two holes through which the contact will be made. When the next layer of metal is put down, it will be isolated from the p-type substrate by the oxide left behind.

Finally, the metal is patterned to make electrical contact to the p region and the n region. The contacts are made through the holes in the oxide.

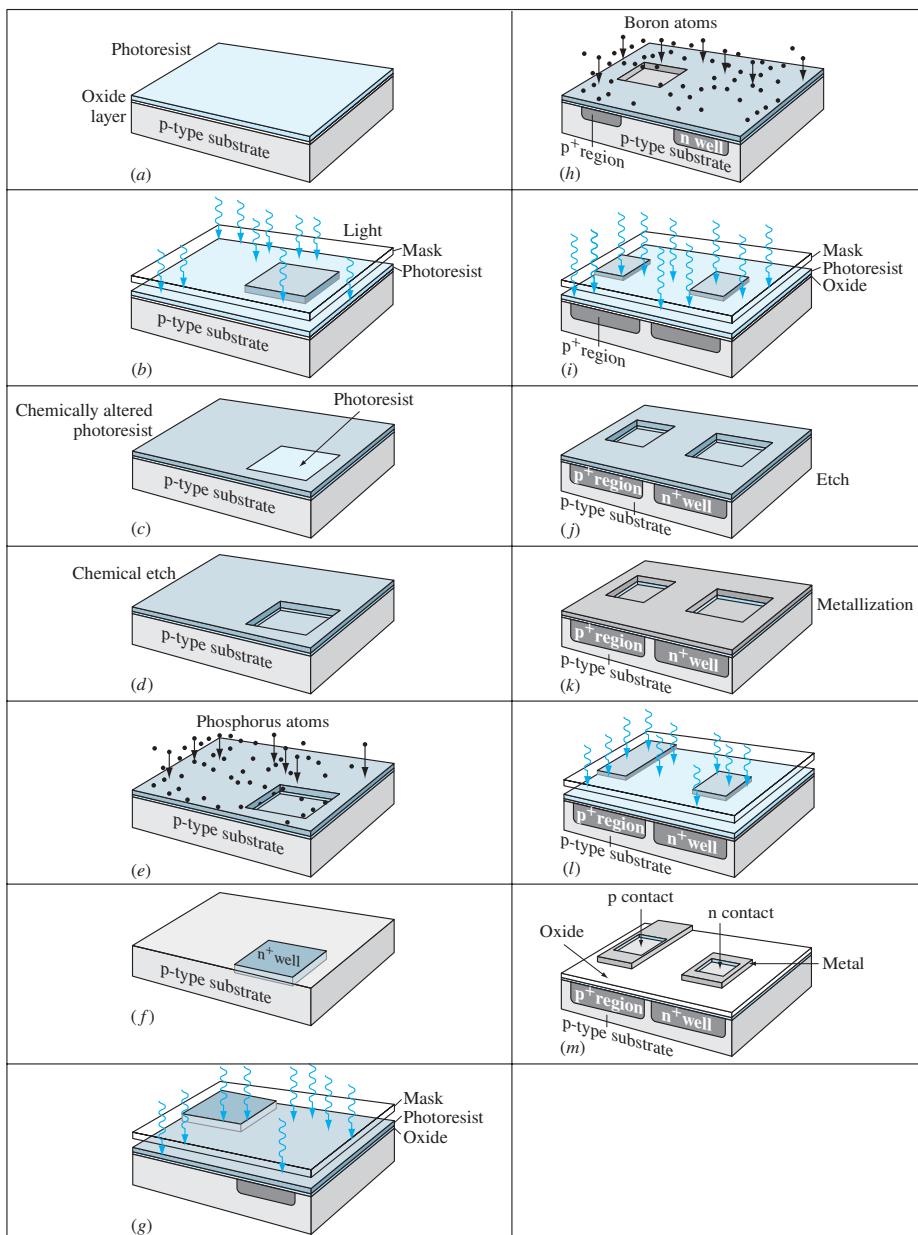


Figure C.11 Photolithography is used to make a diode. (a) A layer of silicon dioxide is grown and then covered with photoresist. (b) This is exposed to light through a mask. (c) The light changes the chemistry of the photoresist where it's exposed. (d) The exposed photoresist and the layers under it are etched away. (e) Phosphorus atoms are ion implanted, leaving the n⁺ structure shown in (f). In (g), a new layer of oxide is deposited, and a mask is used to open a hole in the oxide. (h) A second implantation is done to create a p⁺ contact layer. (i) A third mask is used to pattern the oxide and (j) etched to make contact holes. (k) A layer of metal is deposited over the entire wafer. (l) The metal is patterned using another mask. (m) The final device has two contacts, one to the n region and one to the p substrate.

C.5 CONDUCTORS AND INSULATORS

In addition to the semiconductor materials, insulators and connectors are required to interconnect the individual components and to isolate the components and their electrodes from one another. The conductors are typically metal, usually aluminum, copper, or degenerately doped silicon layers. The degenerately doped layers may be part of the semiconductor crystal or they may be polycrystalline silicon.

C.5.1 Metallization

Metal is often used for interconnections on integrated circuits. Aluminum was originally the principal metal used, because it is inexpensive and highly conductive. It has largely been replaced by copper or copper alloys.

In complex circuits, several layers of metallization are needed to provide all the connections, Figure C.12. These layers are separated by silicon dioxide, which is insulating, and metal “vias” (connections from one layer to the next) are formed by making holes in the oxide and filling with metal such as tungsten.

More recently, copper is replacing the aluminum in the top interconnection layer. Aluminum suffers from electromigration, in which the electrons traveling through the conductor have high enough kinetic energy to dislodge the aluminum atoms. Over time, this process moves enough Al atoms to create voids (open circuits) or bridges (short circuits). The copper atoms are much heavier and thus not as susceptible to electromigration.

Gold is also used in metallization, especially in GaAs. Although Au can be used in silicon, it must be combined with layers of other metals. Gold diffuses easily in silicon and creates deep trap states that can ruin the electrical properties of the semiconductor.

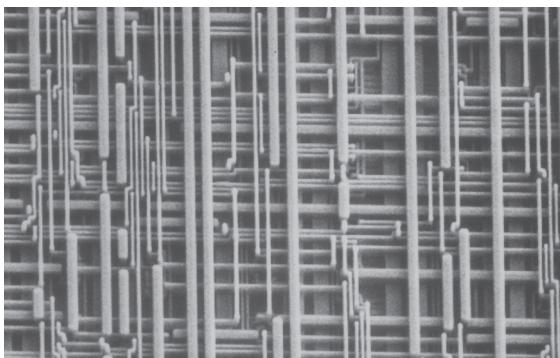


Figure C.12 A scanning electron micrograph of the metal connections of an integrated circuit. All material except the metal has been removed to show the connections. Courtesy of Avago Technologies

C.5.2 Poly Si

Degenerately doped silicon is very conductive and can be used for interconnections as well. For example, heavily doped layers may be used to transport carriers laterally between different structures on a chip. The buried n layer in the bipolar junction transistor of Figure C.10 conducts electrons from the collector region under the emitter laterally to the collector contact.

Polycrystalline silicon is often used on the surface as a conductor. It can be deposited by a pyrolysis process using silane, the reaction for which is



Although Si is deposited, it will form not a single-crystal epitaxial layer but rather a polycrystalline layer, because of the conditions prevailing when the deposition is done. While normally not very conductive, poly Si can be doped to high conductivities, appropriate for such VLSI structures as gate electrodes for MOS technology. It is also used as an interface between metal and the silicon substrate in order to ensure a low-resistance electrical contact.

C.5.3 Oxidation

Silicon dioxide is an electrical insulator and is used to electrically isolate various structures on a chip. One of the reasons silicon is the most widely used semiconductor is the ease with which a “native” oxide can be grown on it—silicon dioxide forms whenever the silicon is exposed to oxygen.

Silicon dioxide can be used to protect areas from diffusion and even ion implantation. The SiO_2 is patterned onto the wafer by photolithography, and the diffusion or implant is done through the holes in the oxide. If the oxide is thick enough, and the diffusivities of the dopant atoms are small enough in the oxide, the dopants will not reach the protected areas of the silicon. Fortunately, the diffusivities in SiO_2 of boron, phosphorus, and antimony, to name a few, are much smaller than their diffusivities in silicon.

Oxidation can be carried out through a wet or dry process. In the dry process, the wafer is exposed to dry O_2 , producing the reaction



Alternatively, in a wet oxidation process, the oxygen is introduced by water vapor, reacting via



The rate of the reaction is controlled by the temperature. Typical temperatures are in the range of 900 to 1200°C. The wet process is considerably faster, while the dry process yields a better-quality interface between the silicon and the oxide. Therefore a common technique is to grow a thin layer of oxide by using the dry process and following it with a thicker wet-oxidation process layer.

When the oxidation process starts, the first oxygen atoms react with the silicon surface. As the oxidation continues, however, the growth of additional SiO_2 occurs at the Si- SiO_2 interface, since that is where more silicon atoms are available

for reaction. Thus, as the oxide grows, the silicon is consumed, as shown in Figure C.13. The surface of the wafer moves up (material is being added) but the surface of the silicon moves downward. The rates of these movements are nearly equal.

Oxide layers can also be deposited, rather than grown. Deposition techniques such as CVD or sputtering, however, do not produce as high-quality an oxide-silicon interface as thermal oxidation does.

Silicon dioxide is used to insulate the various metal layers from each other, as seen in cross section in Figure C.14. Here, there is a surface conductor layer of polysilicon, to which some metal is also connected (Metal 1). Metal 2 consists of several lines going into the page. Notice that some of the Metal 2 lines overlap the edges of the features in the layers below. To avoid the metal being deposited over a drop-off (like that seen in the surface oxide), oxide is grown over the lower layer and then planarized by chemical-mechanical polishing (CMP). The ability to planarize layers has been an important development in the production of complex ICs.

A final comment about oxidation is that the reaction occurs quickly and easily—meaning that as soon as any silicon wafer is exposed to atmosphere, an oxide layer immediately starts to grow, even at room temperature. Thus, care must be taken when transferring wafers from one process to another, or else the oxide must be etched off before the next step is started.

C.5.4 Silicon Nitride

Another insulating material commonly used in a silicon fabrication process is silicon nitride, Si_3N_4 . It is even denser than silicon dioxide, and thus even more impenetrable, which is helpful against rapid diffusers such as water and sodium. This is why Si_3N_4 is often used as a final passivation layer when the chip is complete, to protect it from the environment. Silicon nitride is also used during fabrication as a blocking layer against diffusions and implants.

Silicon nitride can be deposited by using dichlorosilane and ammonia in a liquid-phase chemical-vapor deposition (LPCVD) process:

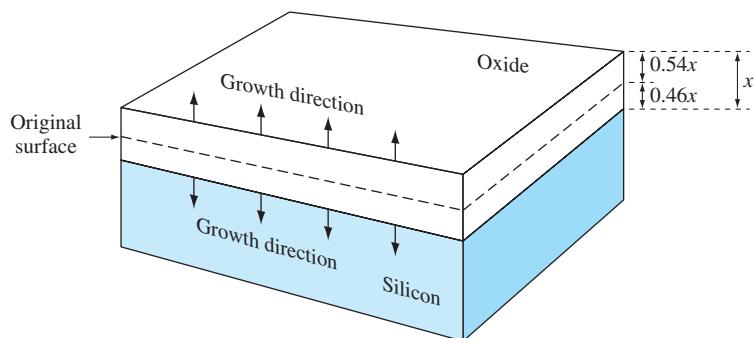
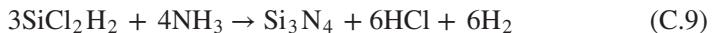


Figure C.13 As an oxide layer grows, the oxide layer expands upward and also downward into the silicon.

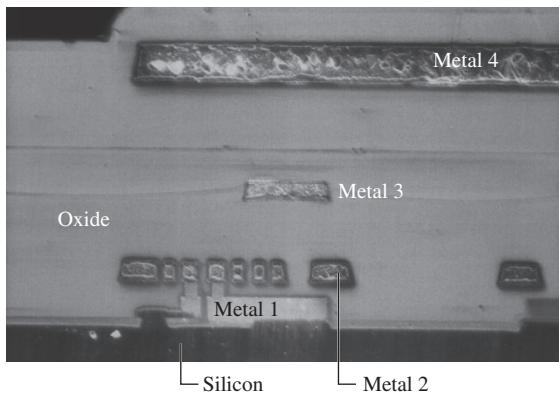


Figure C.14 A cross section of a chip with a polysilicon conductor layer at the surface, followed by four layers of metal, separated by layers of oxide. The oxide layers are planarized by chemical-mechanical polishing (CMP). Courtesy of Avago Technologies

This process is conducted at high temperatures, between 700 and 800°C. When the nitride is to be used as a final passivation and protection layer, however, a high-temperature process can't be used—remember that whenever the temperature is raised, all of the dopants will diffuse. Even more important, the aluminum used in the metal contacts melts at 660°C.

For passivation, then, a low-temperature process is needed, and plasma-enhanced CVD (PECVD) can be carried out with silane and ammonia in an argon plasma. In PECVD the RF energy is transferred to the reactants via the Ar ions and assists the reaction:



In the PECVD process, the silicon nitride film is not stoichiometric, and it may have varying amounts of hydrogen.

C.6 SILICON OXYNITRIDE (SiO_xN_y OR SiON)

Nitrogen is often incorporated into SiO_2 for use as a gate dielectric in MOSFETs. Nitrogen replaces oxygen in the ratio 2N for 3O. A major advantage of SiON over SiO_2 is that it suppresses dopant atoms (e.g., boron) from diffusing from the Si channel into the gate dielectric. It also has a higher dielectric constant than SiO_2 , thus permitting a thicker gate dielectric for a given gate capacitance, reducing gate-to-channel electron tunneling. This increased electron tunneling is, however, partially offset by a reduced band gap as indicated approximately by Equation (C.11) (To show the tunneling dependence explicitly on the bandgap, Equation (C.11) assumes that the electron tunnels at mid-gap energy.)

$$P \approx e^{-\frac{2}{\hbar}\sqrt{\frac{m^*E_g}{2}}t} \quad (\text{C.11})$$

For a given tunneling probability P , a reduced dielectric bandgap E_g requires a larger dielectric thickness t .

Figure C.15 shows the dielectric constant and bandgap of SiON as a function of nitrogen content. In practice an equivalent SiO_2 value of about 50 percent is used with a dielectric constant of ≈ 5.5 .

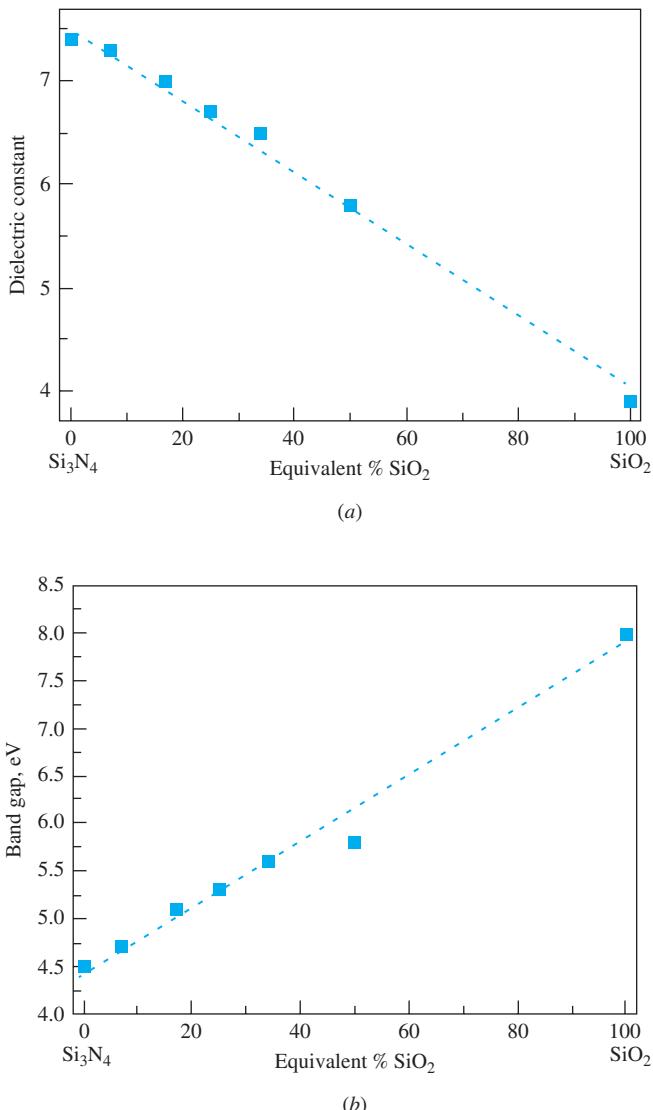


Figure C.15 Dielectric constant (a) and band gap (b) of SiON as a function of nitrogen content. After M. L. Green et. al., “Ultrathin (<4 nm) SiO_2 and Si-O-N gate dielectric layers for silicon microelectronics: Understanding the processing, structure and physical electrical limits,” *Journal of Applied Physics*, 90, no. 5, pp. 2057–2121, 2001.

C.7 CLEAN ROOMS

In all of the processing steps, the need for cleanliness is absolute. The integrated circuits are all made in “clean rooms,” which are specially designed to reduce contaminants. Huge filters clean the air continuously. The air pressure in the clean room is higher than that outside, so when the door is opened, air will flow out and not in. Entry to a clean room is through an anteroom, and the door to the outside hall is closed before the door to the clean room is opened.

The greatest source of contamination is potentially from the workers. Fallen eyelashes, dead skin cells, and contaminants too small to see must be prevented from contacting the wafers. Thus, all clean room personnel wear “bunny suits” (Figure C.16), which cover their clothes, skin, hair, hands, shoes, etc. In Figure C.16, the holes in the floor that help circulate clean air can be seen.

C.8 PACKAGING

Once the chip is complete, it must be packaged so it can be connected to the outside world. The integrated circuit is tested in wafer form, with a series of electrical probes that step and repeat across the wafer. Each circuit is tested, and those that fail are marked with ink.

The wafer is separated into individual chips. It can be sawn or diced. In the dicing process, the wafer is scribed by a diamond tip. The wafer is then placed on a soft backing and gently rolled to snap apart the dice, cleaving the silicon crystal. The defective (inked) devices are discarded.

Next, the functioning parts are attached to a header. Usually, the back of the chip serves as one of the contacts, so the mounting technique must provide electrical connection to the header. Typical examples are conductive epoxy, solder, or a eutectic bond. In the eutectic bond, gold is deposited onto the back of the wafer. The chip is pressed against the header and heated, and ultrasonic energy is applied. This forms a silicon-gold alloy and bonds the chip to the header.



Figure C.16 A production clean room. All personnel must wear special clothing to avoid contaminating the wafers. Courtesy of Atmel Corporation

C.8.1 Wire Bonding

Each chip usually has a series of bonding pads around the outside perimeter. These are large-area contacts (large by device standards, perhaps 100 to 250 μm square), Figure C.17a). Very thin wires (thin by outside world standards, about 15 to 75 μm in diameter¹) are bonded to these pads. These wires are used to make connections to the leads of the package.

The two main types of wire bonds are ball bonds and wedge bonds. In a ball bonder, the wire is brought to the bonding pad inside a capillary tube. The tube is

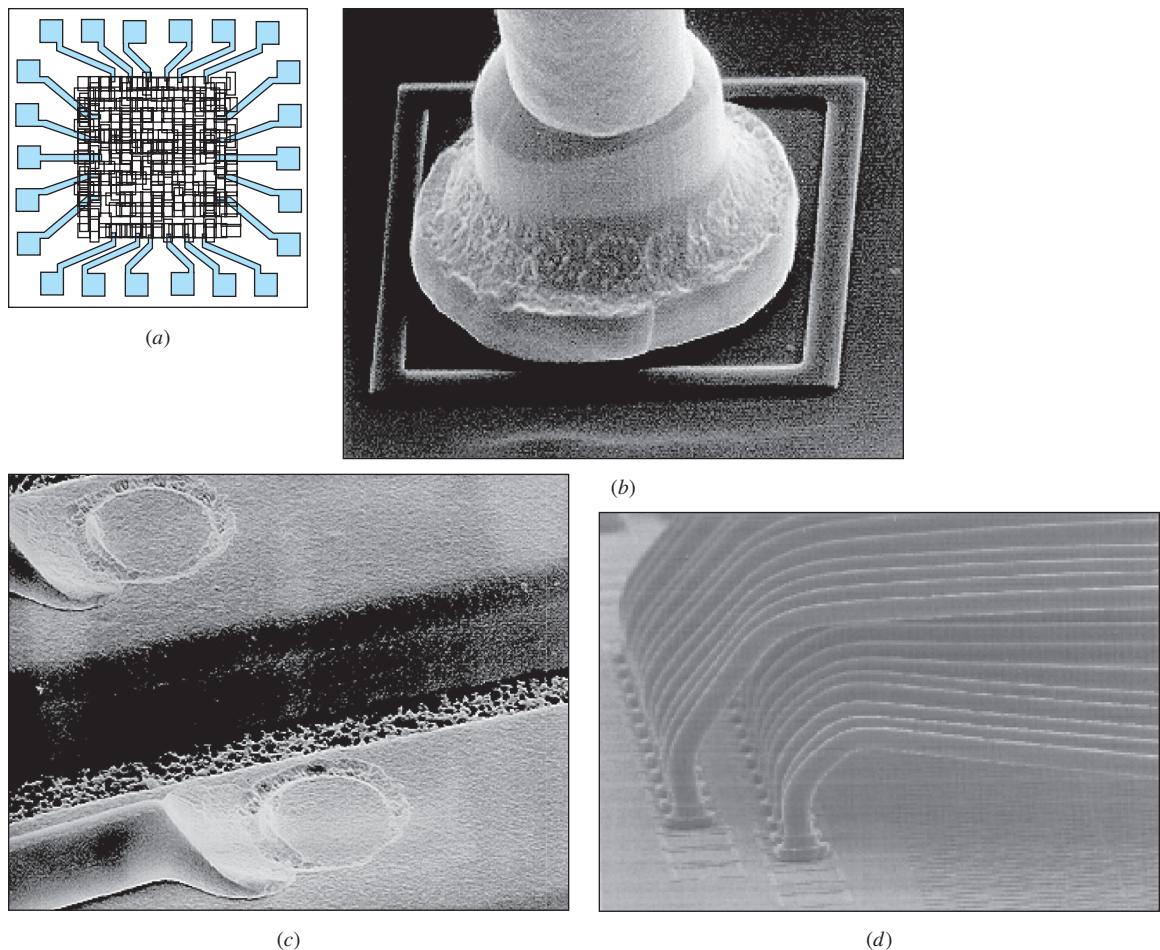


Figure C.17 (a) A chip has bonding pads around its perimeter to which wire bonds can be made; (b) a ball bond; (c) a wedge bond; (d) an array of ball bonds. b-d. Photos used with the permission of the Intersil Corporation.

¹Human hair is about 60 to 100 μm in diameter.

heated, and the end of the wire is melted with an electric spark. The melted wire is pressed onto the pad, forming a ball and making good electrical contact with the pad, Figure C.17b. A wedge bond (Figure C.17c) is formed by using pressure combined with ultrasonic vibration (in lieu of melting). The ultrasound causes the metal of the wire to flow, and the pressure causes it to bond with the metallized contact pad. The vibration from the ultrasound also lets the wire *scrub* its way through any residual surface oxide.

Ball bonding machines can make many bonds in parallel, Figure C.17d.

C.8.2 Lead Frame

The other ends of the wire bonds go to the individual leads of the package. There are many types of packages, from TO-style cans used for individual transistors, to dual in-line packages (DIPs), to ball grid array surface-mount packages for ultrahigh-density circuit boards.

Figure C.18a shows a TO can, typically used for discrete transistors. These come in various sizes and have different numbers of leads. Sometimes there is a window or lens in the can if the package is to be used for an LED, a laser, or a photodetector. For transistors and such, the can is solid metal.

The die is mounted on the header and wire bonded to one or more leads. The ground contact is often made through the bottom of the chip, and the corresponding lead is in contact with the entire can. The other leads are electrically insulated from the surrounding can.

The familiar dual in-line package of Figure C.18b contains a lead frame with various numbers of leads. The die is mounted on a support and wire bonded to the leads as necessary. Then the package is sealed in epoxy (the common black

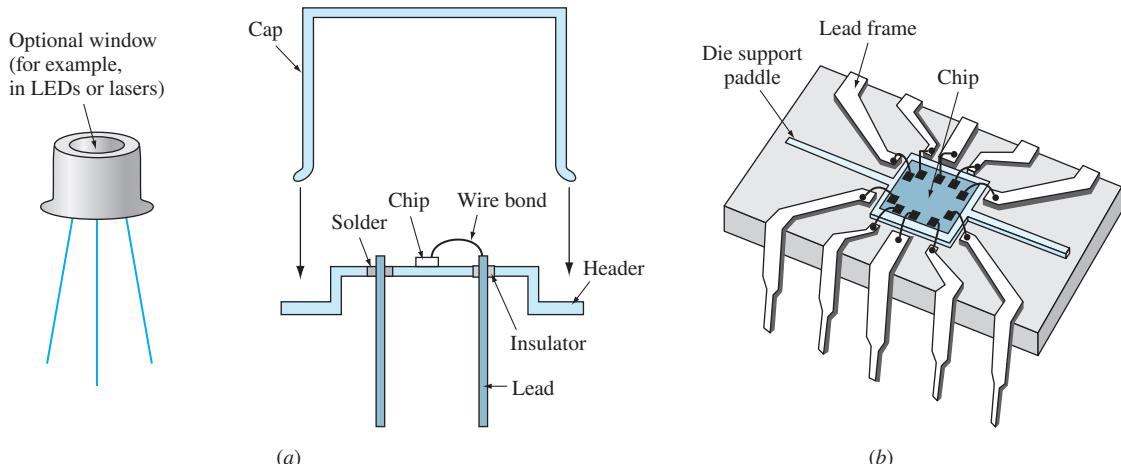


Figure C.18 (a) A TO can used for discrete devices (the window version might be used for LEDs or photodetectors); (b) the lead frame of a dual in-line package (DIP).

package) or, for environmentally demanding situations, the package may be hermetically sealed under a metal cap.

C.8.3 Surface-Mount Packages

The DIP and TO can are just two examples of through-hole-mount packages, those having leads that go through holes on a circuit board. The other class of packages is the surface-mount type, the leads of which are soldered or otherwise attached to contacts on the surface of the chip. Examples are the leadless chip carrier (LCC), the thin small-outline package (TSOP), and the ball grid array (BGA), Figure C.19. The leadless chip carrier has contact pads around

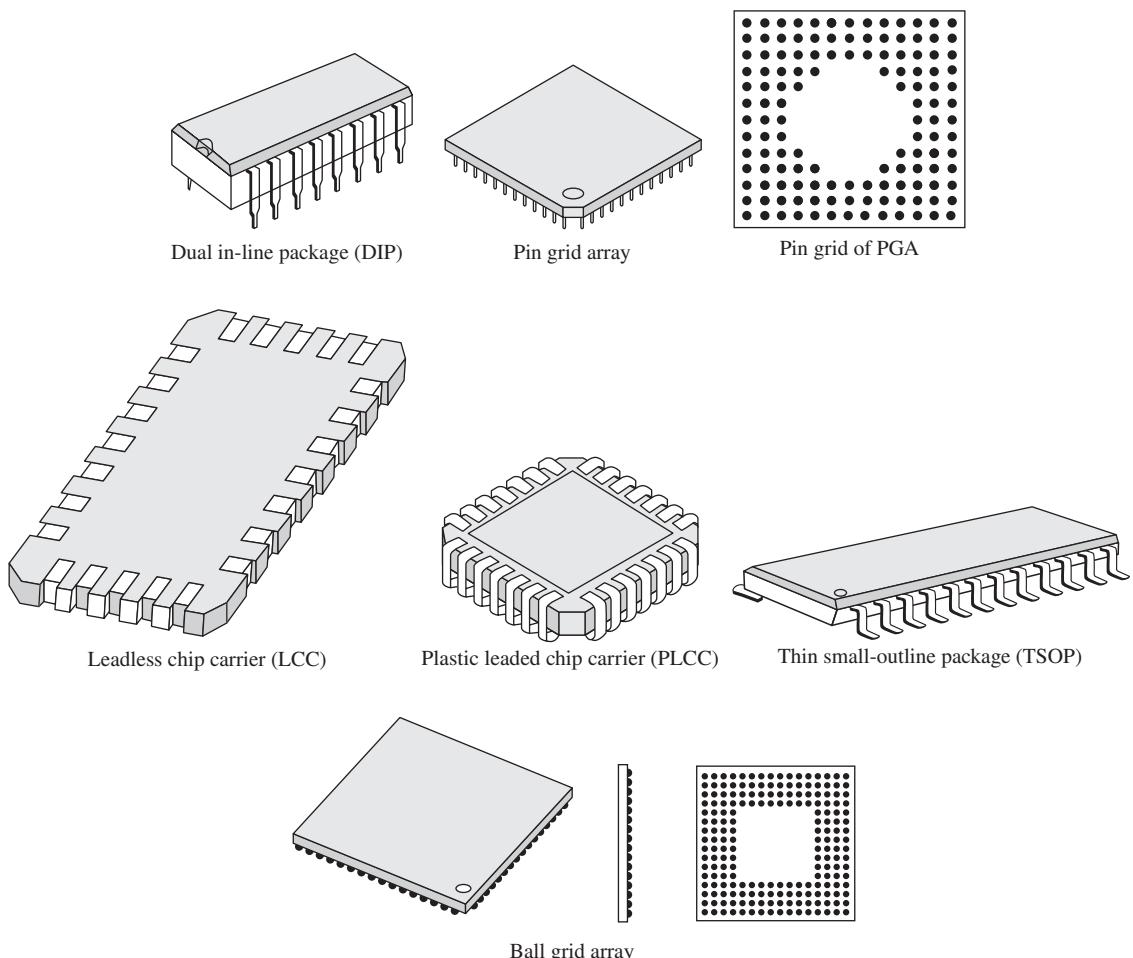


Figure C.19 Various types of integrated circuit packages. The DIP and PGA are through-hole-mount packages; the others mount onto the surface of the circuit board.

the edge of the device that directly contact the circuit board. A variation is the lead chip carrier, which has actual leads that bend underneath the package (J leads, shaped like the letter J) and make contact that way. The thin outline small package (TSOP) has leads that bend outward (gull wings). The ball grid array makes for dense packaging.

C.9 SUMMARY

This has been a very brief outline of the growth, fabrication, and packaging processes used in modern semiconductor manufacturing. We saw that conductors on ICs can be metal or degenerately doped silicon, and that silicon dioxide makes for a very useful and convenient insulator. The packaging continues to evolve, and we showed a few of the many package styles that are available.

Semiconductor fabrication involves not just electrical engineers, but physicists, crystallographers, chemists, materials scientists, mechanical and packaging engineers, and many others.

A P P E N D I X D

Some Useful Integrals

$$y_n = \int_0^\infty x^n e^{-ax^2} dx \quad \text{for } a > 0$$

$$y_n = \int_0^\infty x^n e^{-ax} dx = \frac{\Gamma(n+1)}{a^{n+1}} \quad \text{for } n > -1, a > 0$$

n even

$$y_0 = \frac{1}{2} \sqrt{\frac{\pi}{a}}$$

$$y_2 = \frac{1}{4} \sqrt{\frac{\pi}{a^3}}$$

$$y_4 = \frac{3}{8} \sqrt{\frac{\pi}{a^5}}$$

n odd

$$y_1 = \frac{1}{2a}$$

$$y_3 = \frac{1}{2a^3}$$

$$y_5 = \frac{1}{a^3}$$

$$y_0 = \frac{1}{a}$$

$$y_1 = \frac{1}{a^2}$$

$$y_2 = \frac{2}{a^3}$$

$$y_{-1/2} = \sqrt{\frac{\pi}{a}}$$

$$y_{1/2} = \frac{1}{2} \sqrt{\frac{\pi}{a^3}}$$

$$y_{3/2} = \frac{3}{4} \sqrt{\frac{\pi}{a^5}}$$

$$\Gamma_{(n+1)} = n! \quad \text{if } n \text{ is an integer}$$

$$\Gamma_{1/2} = \sqrt{\pi}$$

$$\Gamma_{(n+1)} = n\Gamma_n \quad \text{if } n > 0$$

$$\int x^m e^{ax} dx = \frac{x^m e^{ax}}{a} - \frac{m}{a} \int x^{m-1} e^{ax} dx$$

E APPENDIX

Useful Equations

GENERAL PHYSICS

$$\mathcal{E} = -\frac{dV}{dx}$$

$$C = \frac{\epsilon A}{t}$$

$$\Psi(x, t) = U_K(x) e^{j[Kx - (E/\hbar)t]}$$

$$F = -\nabla E_P = -\frac{dE_P}{dr}$$

SEMICONDUCTOR MATERIALS

$$\gamma = E_{\text{vac}} - E_V \quad \text{non degenerate}$$

$$\chi = E_{\text{vac}} - E_C$$

$$E_g = E_C - E_V$$

$$n_0 = N_C e^{-(E_C - E_f)/kT} \quad \text{nondegenerate}$$

$$p_0 = N_V e^{-(E_f - E_V)/kT} \quad \text{nondegenerate}$$

$$n_i^2 = N_C N_V e^{-E_g/kT} \quad \text{nondegenerate}$$

$$n_0 p_0 = n_i^2 \quad \text{nondegenerate}$$

$$E_g = kT \ln \frac{N_C N_V}{n_i^2} \quad \text{nondegenerate}$$

$$\delta_n = E_C - E_f = kT \ln \frac{N_C}{N'_D}$$

$$\delta_p = E_f - E_V = kT \ln \frac{N_V}{N'_A}$$

$$J_n = q\mu_n n \mathcal{E} + qD_n \frac{dn}{dx} = q\mu_n \left(n\mathcal{E} + \frac{kT}{q} \frac{dn}{dx} \right)$$

$$J_p = q\mu_p p \mathcal{E} - qD_p \frac{dp}{dx} = q\mu_p \left(p\mathcal{E} - \frac{kT}{q} \frac{dp}{dx} \right)$$

$$\mathcal{E} = \frac{1}{q} \frac{dE_{\text{vac}}}{dx}$$

$$n_0 p_0 = N_C N_V e^{\frac{-(E_{g0} - \Delta E_g^*)}{kT}} = n_i^2 e^{\frac{\Delta E_g^*}{kT}}$$

JUNCTIONS

$$qV_{\text{bi}} = |\Phi_p - \Phi_n|$$

$$V_{\text{bi}} = \frac{kT}{q} \ln \frac{N'_D N'_A}{n_i^2}$$

$$qV_{\text{bi}} = E_g - (\delta_n + \delta_p)$$

$$V_j^n = \frac{qN'_D}{2\epsilon} w_n^2$$

$$V_j^p = \frac{qN'_A}{2\epsilon} w_p^2$$

$$w_n = (x_0 - x_n) = \left(\frac{2\epsilon V_j^n}{qN'_D} \right)^{1/2} \left[\frac{2\epsilon V_j}{qN'_D \left(1 + \frac{N'_D}{N'_A} \right)} \right]^{1/2}$$

$$w_p = (x_p - x_0) = \left(\frac{2\epsilon V_j^p}{qN'_A} \right)^{1/2} \left[\frac{2\epsilon V_j}{qN'_A \left(1 + \frac{N'_A}{N'_D} \right)} \right]^{1/2}$$

$$w = w_n + w_p = \left[\frac{2\epsilon V_j (N'_A + N'_D)}{qN'_A N'_D} \right]^{1/2} \quad \text{pn junction}$$

$$w = \left(\frac{2\epsilon V_j}{qN'_A} \right)^{1/2} \quad \text{n}^+ \text{p junction}$$

$$w = \left(\frac{2\epsilon V_j}{qN'_D} \right)^{1/2} \quad \text{p}^+ \text{n junction}$$

$$\begin{aligned}
V_j &= \frac{qN'_DN'_Aw^2}{2\epsilon(N'_D + N'_A)} && \text{pn junction} \\
n_{p0} &= N'_De^{-qV_{bi}/kT} && \text{nondegenerate} \\
p_{n0} &= N'_Ae^{-qV_{bi}/kT} \\
n_p(x_p) &= n_{p0}e^{qV_a/kT} \\
p_n(x_n) &= p_{n0}e^{qV_a/kT} \\
\Delta n_p(x_p) &= n_{p0}(e^{qV_a/kT} - 1) \\
\Delta p_n(x_n) &= p_{n0}(e^{qV_a/kT} - 1) \\
J &= q\left(\frac{D_nn_{p0}}{L_n} + \frac{D_pp_{n0}}{L_p}\right)(e^{qV_a/kT} - 1) = J_0(e^{qV_a/kT} - 1) && \text{pn junction} \\
J_0 &= q\left(\frac{D_nn_{p0}}{w_B} + \frac{D_pp_{n0}}{L_p}\right) && \text{short p-side diode} \\
J_0 &= q\left(\frac{D_nn_{p0}}{w_{B(p)}} + \frac{D_pp_{n0}}{w_{B(n)}}\right) && \text{both sides short} \\
I &= I_0(e^{qV_a/kT} - 1) \\
R - G &= \frac{np - n_i^2}{\tau_0(n + n_i + p + n_i)} \\
C_j &= A \left[\frac{q\epsilon N'_DN'_A}{2(N'_D + N'_A)(V_{bi} - V_a)} \right]^{1/2} && \text{pn junction} \\
C_j &= A \left[\frac{q\epsilon N'}{2(V_{bi} - V_a)} \right]^{1/2} && \text{one-sided junction} \\
C_j &= \frac{A \left(\frac{\epsilon_n \epsilon_p}{w_n w_p} \right)}{\left(\frac{\epsilon_n}{w_n} + \frac{\epsilon_p}{w_p} \right)} && \text{heterojunction} \\
C_{sc} &= \frac{dQ_{sr}}{dV_a} = \delta \frac{dQ_s}{dV_a} = \delta I \tau_n q/kT \\
t_T &= \frac{(W_B)^2}{2D_n} \\
\mathcal{E}(x) &= \frac{qa}{2\epsilon} \left[x^2 - \left(\frac{w}{2} \right)^2 \right] & - \left(\frac{w}{2} \right) \leq x \leq \frac{w}{2} & \text{linearly graded junction}
\end{aligned}$$

$$w = \left(\frac{12\epsilon V_j}{qa} \right)^{1/3} \quad \text{linearly graded junction}$$

$$V_{bi} = \frac{2kT}{q} \ln \left[\frac{a}{2n_i} \left(\frac{12\epsilon V_{bi}}{qa} \right)^{1/3} \right] \quad \text{linearly graded junction}$$

FIELD-EFFECT TRANSISTORS

$$I_D = W Q_{ch}(y) \mu(y) \mathcal{E}_L(y)$$

$$I_D = 0 \quad V_{GS} < V_T$$

$$C'_{ox} = \frac{\epsilon_{ox}}{t_{ox}}$$

$$I_D = \frac{WC'_{ox}\mu}{L} \left[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2} \right] \quad V_D \leq (V_G - V_T) \quad \text{simple model}$$

$$V_{DSsat} = (V_{GS} - V_T) \quad \text{simple model}$$

$$I_{Dsat} = \frac{WC'_{ox}\mu}{L} \left[\left(V_{GS} - V_T - \frac{V_{DSsat}}{2} \right) V_{DSsat} \right] \quad V_{GS} > V_T, \quad V_{DS} > V_{DSsat} \quad \text{simple model}$$

$$I_{Dsat} = \frac{WC'_{ox}\mu}{2L} (V_{GS} - V_T)^2 = \frac{WC'_{ox}\mu}{2L} V_{DSsat}^2 \quad V_{GS} > V_T, \quad V_{DS} > V_{DSsat}$$

$$|v| = \frac{\mu_{lf} |\mathcal{E}_L|}{1 + \frac{\mu_{lf} |\mathcal{E}_L|}{v_{sat}}} \quad \mu = \frac{\mu_{lf}}{1 + \frac{\mu_{lf} |\mathcal{E}_L|}{v_{sat}}}$$

$$I_D = \frac{WC'_{ox}\mu_{lf}}{\left(L + \frac{\mu_{lf} V_{DS}}{v_{sat}} \right)} \left[(V_{GS} - V_T) V_{DS} - \frac{V_{DS}^2}{2} \right] \quad V_{DS} \leq V_{DSsat} \quad \text{considering velocity saturation}$$

$$I_{Dsat} = \frac{WC'_{ox}\mu_{lf}}{\left(L + \frac{\mu_{lf} V_{DSsat}}{v_{sat}} \right)} \left[(V_{GS} - V_T) V_{DSsat} - \frac{V_{DSsat}^2}{2} \right]$$

$$V_{DS} \geq V_{DSsat} \quad \text{considering velocity saturation}$$

$$V_{DSsat} = \frac{v_{sat}}{\mu_{lf}} L \left[\left(1 + \frac{2\mu_{lf}(V_{GS} - V_T)}{v_{sat} L} \right)^{1/2} - 1 \right] \quad \text{considering velocity saturation}$$

$$\mu_{lf} = \frac{\mu_0}{1 + \theta(V_G - V_T)}$$

$$\begin{aligned}
w_T &= \left[\frac{2 \epsilon_s}{q N'_A} \phi_s(0) \right]^{1/2} \\
V_T &= \frac{\Phi_{MS}}{q} + 2 \phi_f - \frac{Q_f}{C'_{ox}} - \frac{Q_i(2 \phi_f)}{C'_{ox}} - \frac{Q_B(2 \phi_f)}{C'_{ox}} \\
\Delta V_T &= -\frac{Q_{ii}}{C'_{ox}} = +\frac{q N_{ii}}{C'_{ox}} \\
I_D &= I_0 e^{q(V_{GS} - V_T)/nkT} \quad \text{subthreshold} \\
S &= \frac{2.3kTn}{q} \\
P &= C_L V_{DD}^2 f \\
P &= I_{\text{Leakage}} V_{DD} \\
t_d &= \frac{1}{2} \left(\frac{C_L V_{DD}}{2 I_{D\text{satn}}} + \frac{C_L V_{DD}}{2 I_{D\text{satp}}} \right) \\
g_m &\equiv \frac{i_d}{v_g} = \frac{\partial I_D}{\partial V_{GS}} \\
g_d &\equiv \frac{i_d}{v_d} = \frac{\partial I_D}{\partial V_{DS}} \\
g_m &= \frac{W C'_{ox} \mu_{lf} V_{DS}}{L \left(1 + \frac{\mu_{lf} V_{DS}}{L v_{\text{sat}}} \right)} \\
g_{msat} &= W v_{\text{sat}} C'_{ox} \left\{ 1 - \left[1 + \frac{2 \mu_{lf} (V_{GS} - V_T)}{v_{\text{sat}} L} \right]^{-1/2} \right\} \\
g_d &= \frac{W C'_{ox} \mu_{lf}}{L} \left[\frac{(V_{GS} - V_T - V_{DS}) - \frac{\mu_{lf} V_{DS}^2}{2 L v_{\text{sat}}}}{\left(1 + \frac{\mu_{lf} V_{DS}}{L v_{\text{sat}}} \right)^2} \right] \quad V_{DS} \leq V_{DS\text{sat}} \\
g_{dsat} &= 0 \quad V_D \geq V_{D\text{sat}} \\
f_T &= \frac{v_{\text{sat}}}{2 \pi L} \left\{ 1 - \left[1 + \left(\frac{2 \mu_{lf} (V_{GS} - V_T)}{v_{\text{sat}} L} \right)^{-1/2} \right] \right\} \quad \text{saturation} \\
I_D &= \frac{W C'_{ox} \mu_{lf} \left(V_{GS} - V_T - \frac{V_{DS}}{2} \right) (V_{DS} - 2 I_D R_S)}{L \left[1 + \frac{\mu_{lf} (V_{DS} - 2 I_D R_S)}{L v_{\text{sat}}} \right]} \quad V_{DS} \leq V_{DS\text{sat}} \\
&\quad \text{considering series resistance}
\end{aligned}$$

$$\frac{I_{D\text{sat}}(77 \text{ K})}{I_{D\text{sat}}(300 \text{ K})} = \frac{\mu_{\text{lf}}(77 \text{ K})}{\mu_{\text{lf}}(300 \text{ K})} \frac{\left(1 + \frac{\mu_{\text{lf}} V_{DS\text{sat}}}{L v_{\text{sat}}} \right)_{77 \text{ K}}}{\left(1 + \frac{\mu_{\text{lf}} V_{DS\text{sat}}}{L v_{\text{sat}}} \right)_{300 \text{ K}}}$$

$$I_D = \frac{qW\mu_{\text{lf}}N'_D a}{L} \left\{ V_{DS} - \frac{2}{3}(V_{bi} - V_T) \left[\left(\frac{V_{DS} + V_{bi} - V_{GS}}{V_{bi} - V_T} \right)^{3/2} - \left(\frac{V_{bi} - V_{GS}}{V_{bi} - V_T} \right)^{3/2} \right] \right\}$$

bulk channel FETs, below saturation, simple model

$$I_{D\text{sat}} = \frac{qW\mu_{\text{lf}}N'_D a}{L} \left\{ V_{GS} - V_T - \frac{2}{3}(V_{bi} - V_T) \left[1 - \left(\frac{V_{bi} - V_{GS}}{V_{bi} - V_T} \right)^{3/2} \right] \right\}$$

bulk channel FETs, simple model

BIPOLAR JUNCTION TRANSISTORS

$$V_{CE} = V_{CB} + V_{BE}$$

$$I_E = I_C + I_B$$

$$I_E = I_{nE} + I_{pE}$$

$$I_C = I_{nC} + I_{pC}$$

$$I_{nC} = I_{nE} - I_{\text{rec}}$$

$$I_B = I_{pE} + I_{\text{rec}} - I_{pC}$$

$$\alpha = \frac{I_{\text{out}}}{I_{\text{in}}} = \frac{I_C}{I_E}$$

$$\beta = \frac{I_{\text{out}}}{I_{\text{in}}} = \frac{I_C}{I_B}$$

$$\beta = \frac{\alpha}{1 - \alpha}$$

$$\alpha = \frac{I_{nE}}{I_E} \frac{I_{nC}}{I_{nE}} \frac{I_C}{I_{nC}} = \gamma \alpha_T M$$

$$\gamma = \frac{I_{nE}}{I_E}$$

$$\alpha_T = \frac{I_{nC}}{I_{nE}} = 1 - \frac{I_{\text{rec}}}{I_{nE}}$$

$$M = \frac{I_C}{I_{nC}} = 1 + \frac{I_{pC}}{I_{nC}}$$

$\gamma \approx \frac{1}{1 + \frac{p_E(0^-)}{n_B(0^+)} \frac{D_{pE}}{D_{nB}} \frac{W_B}{W_E}}$	prototype (uniformly doped) npn
$\gamma \approx \frac{1}{1 + \frac{n_E(0^-)}{p_B(0^+)} \frac{D_{nE}}{D_{pB}} \frac{W_B}{W_E}}$	prototype (uniformly doped) pnp
$\alpha_T \approx 1 - \frac{W_B^2}{2L_n^2}$	prototype (uniformly doped) npn
$\alpha_T \approx 1 - \frac{W_B^2}{2L_p^2}$	prototype (uniformly doped) pnp
$\gamma \approx \frac{1}{1 + \frac{N'_{AB}}{N'_{DE}} e^{\Delta E_g^*/kT} \times \frac{D_{pE}}{D_{nB}} \times \frac{W_B}{W_E}}$	prototype npn with degenerately doped emitter
$\gamma \approx \frac{1}{1 + \frac{N'_{DB}}{N_{AE}} e^{\Delta E_g^*/kT} \times \frac{D_{nE}}{D_{pB}} \times \frac{W_B}{W_E}}$	prototype pnp with degenerately doped emitter
$\beta \approx \frac{N'_{DE}}{N'_{AB}} \frac{D_{nB}}{D_{pE}} \times \frac{W_E}{W_B}$	nondegenerate emitter (npn)
$\beta \approx \frac{N_{DE}}{N'_{AB}} \frac{D_{nB}}{D_{pE}} \times \frac{W_E}{W_B} e^{-\Delta E_g^*/kT}$	degenerate emitter (npn)
$\Delta E_g^* = E_g^* (\text{base}) - E_g^* (\text{emitter})$	
$\eta = \frac{W_B}{\lambda}$	graded-doping parameter
$\eta = \ln \frac{N'_A(0^+)}{N'_A(W_B)}$	exponential grading of base doping
$J_{nB} = \frac{q D_n n_B(0^+)}{W_B} \left(\frac{\eta}{1 - e^{-\eta}} \right)$	graded base
$\frac{\beta(\eta)}{\beta(0)} \approx \left(\frac{\eta}{1 - e^{-\eta}} \right)$	graded base
$I_F = I_{F0} \left(e^{qV_{BE}/kT} - 1 \right)$	Ebers-Moll Model
$I_R = I_{R0} \left(e^{qV_{BC}/kT} - 1 \right)$	
$I_E = I_F - \alpha_R I_R$	
$I_C = \alpha_F I_F - I_R$	

$$I_E = I_{CT} + \frac{I_F}{\beta_F + 1}$$

$$I_C = I_{CT} + \frac{I_R}{\beta_R + 1}$$

$$I_B = \frac{I_F}{\beta_F + 1} + \frac{I_R}{\beta_R + 1}$$

$$I_{CT} = \frac{\beta_F I_F}{\beta_F + 1} - \frac{\beta_R I_R}{\beta_R + 1}$$

$$R_B = \frac{R_{\square} L}{4h} \quad \text{base resistance}$$

$$J_C \leq 0.3 |q N_{DC} v_{sat}| \quad \text{design rule for base push-out (Kirk) effect}$$

$$C\mu = C_{jBC} \quad \text{Hybrid-pi model}$$

$$C_\pi = C_{jBE} + C_{scBE}$$

$$r_\pi \approx \frac{kT}{qI_B} \approx \frac{\beta_{DC} kT}{qI_C}$$

$$r_\pi = \frac{\frac{kT}{q^2 A_E n_i^2} e^{-qV_{BE}/kT}}{\frac{D_{pE}}{W_E N_{DE}}} \quad \text{npn}$$

$$r_\mu = \frac{1}{\left. \frac{\partial I_C}{\partial V_{CB}} \right|_{V_{BE}}}$$

$$r_0 = \frac{1}{\left. \frac{\partial I_C}{\partial V_{CE}} \right|_{V_{CB}}} = \frac{V_A}{I_C}$$

$$g_m = \left. \frac{\partial I_C}{\partial V_{BE}} \right|_{V_{CE}} = \frac{i_c}{v_{be}} = \frac{qI_C}{kT} = \frac{\beta_{DC}}{r_\pi}$$

$$Q_B = -\frac{qA_E n_B(0^+) W_B}{2}$$

$$dQ_{Br} = \delta dQ_B$$

$$C_{scBE} = \frac{\delta W_B^2 \beta_{DC}}{2 D_{nB} r_\pi}$$

$$f_{co} = \frac{1}{\sqrt{2\pi r_\pi(C_\pi + C_\mu)}}$$

$$\beta(f) = \frac{g_m v_{be}}{i_b} = \frac{\beta_{DC}}{\sqrt{1 + \left(\frac{f}{f_{co}}\right)^2}}$$

$$t_{tB} = \frac{Q_B}{I_{nB}} = \frac{Q_B}{I_C} \quad t_{tB} = \frac{W_B^2}{2D_n}$$

OPTOELECTRONIC DEVICES

$$F_L(x) = F_L(0) e^{-\alpha x}$$

$$R = \left(\frac{n_1 - n_2}{n_1 + n_2} \right)^2$$

$$V_{oc} = \frac{nkT}{q} \ln \left(1 + \left| \frac{I_L}{I_0} \right| \right)$$

$$\eta_Q = \frac{J_L/q}{F_{Li}}$$

$$R_{ph} = \frac{J_L/q}{hvF_{Li}} = \frac{q\eta_Q}{hv}$$

$$FF = \frac{I_m V_m}{I_{sc} V_{oc}}$$

$$\eta = FF \frac{I_{sc} V_{oc}}{P_{Li}}$$

$$R_{ph} = M \left(\frac{q\eta_Q}{hv} \right)$$

$$\frac{dN^2}{dt} = -A_{21} N_2 \quad \text{spontaneous emission}$$

$$\tau_{\text{radiative, spont}} = \frac{1}{A_{21}}$$

$$\frac{dN_2}{dt} = B_{12} N_1 \rho(v)$$

$$\frac{dN_2}{dt} = -B_{21} N_2 \rho(v) \quad \text{stimulated emission}$$

$$\lambda = \frac{2nd}{q}$$

POWER SEMICONDUCTOR DEVICES

$$\xi_{cr} = \sqrt{\frac{2qN_D V_j}{\epsilon_s}} \approx \sqrt{\frac{2qN_D V_{br}}{\epsilon_s}}$$

$$W_D = \sqrt{\frac{2\epsilon_s V_j}{qN_D}} \approx \sqrt{\frac{2\epsilon_s V_{br}}{qN_D}}$$

$$V_{br} \approx \frac{\epsilon_s \xi_{cr}^2}{2qN_D} \approx \frac{qN_D W_D^2}{2\epsilon_s}$$

$$R_{on,sp} = \frac{W_D}{q\mu_n N_D}$$

$$R_{on,sp} = \frac{4V_{br}^2}{\epsilon_s \mu_n \mathcal{C}_{cr}^3}$$

$$R_{on,sp} = \frac{W_D}{q(\mu_n + \mu_p)n_a} \text{ pin device}$$

$$\begin{aligned} I_{B1} &= (1 - \alpha_1)I_{E1} - I_{C01} \\ I_{C2} &= \alpha_2 I_{E2} + I_{C02} \end{aligned} \quad \left. \right\} \text{ npnp two-transistor model}$$

$$I_A = \frac{I_{C01} + I_{C02}}{1 - (\alpha_1 + \alpha_2)} \quad \text{ npnp}$$

A

absolute zero, 18
 absorption, 675
 coefficients, 646, 648
 first look at, 35
 optical, 646
 in a semiconductor, 136–139
 spectrum, 654
 ac operation of a BJT, 543
 ac quantities, representing, 593
 acceptor energy, 67
 acceptors, 66–68
 accumulation region, 313
 activation energy, 526
 active layer, 667, 681
 active mode energy band diagram, 554, 555
 affinity, electron, 230
 air mass zero (AM0), 654, 655
 AlGaAs, passivation layer, 468
 Al:n-Si:
 diode, 326
 metal semiconductor, 324
 Schottky barrier diode, 325
 aluminum on integrated circuits, 782
 AM1 (air mass one), 654, 655
 amorphous materials, 20
 amphoteric impurities, 68
 amplifiers, transistors used as, 357
 analog circuits:
 BJT operation in, 580
 MOSFET analog equivalent
 circuits, 506–511
 transistor operation in, 362, 363
 Anderson model. *See* electron affinity
 model (EAM)
 annealing, during ion implantation, 778
 APDs (avalanche photodiodes), 660–661
 apparent band-gap narrowing, 98–100, 559
 area image sensors, 691–692
 asymmetrical dual gate fully depleted
 silicon-on-insulator MOSFET, 458–460

atom, models of, 4–5
 attractive (centripetal) force between
 particles, 5
 avalanche, 273
 breakdown, 275, 276, 632, 700
 current, 272
 avalanche photodiodes (APDs), 660–661
 azimuthal quantum number, 13

B

B-C junction, 303
 ball bonding, 788
 ball grid array (BGA), 790
 ballistic transport, 423–426
 Balmer series, 35
 band, impurity, 124, 125
 band bending, 169
 due to surface states, 321
 at threshold, 385
 band-gap engineering, 39, 205, 681
 band-gap narrowing:
 apparent, 98–100, 559
 effect of, 562–563
 effect on barrier height, 565
 FET gate, 502
 impurity-induced, 96–98
 band gaps, 17, 20
 of semiconductors, 666, 751
 of solar cell materials, 654
 table of, 17
 band-to-band generation, 134, 135
 band-to-band recombination, 134, 135
 barrier:
 of finite width, 195–198
 of infinite width, 200
 semiconductor-to-metal, 327
 base-collector (B-C) junction, 303
 base-collector transit time, 607
 base push-out effect, 633–634
 base resistance:
 BJT, 623
 intrinsic, 623
 reduction due to current crowding, 626
 base-transit time, 291, 606–607
 base transport efficiency, 551, 557–563
 base width modulation, 627–632
 bathtub curve, 531
 battery, nonuniformly doped
 semiconductor, 168
 BGA (ball grid array), 790
 bias regimes, BJT, 541
 BiCMOS (bipolar-CMOS) technology, 563, 619, 779
 bilateral transmission gates, 530
 BiMOS technology, 618–619
 bipolar devices, 539, 623
 BJTs (bipolar junction transistors), 302, 357, 539
 circuit symbols for npn and pnp, 540
 compared to FETs, 539
 comparison to MOSFETs, 616–618
 doping gradients in, 563–571
 Ebers-Moll common emitter ac
 model for, 590–592
 electrical characteristics of, 543
 energy band diagrams for, 545, 546
 modes of operation, 541
 recombination in the emitter-base
 junction, 635–637
 stored-charge capacitance in, 598–603
 switching on and off, 611–612
 time-dependent analysis of, 590–621
 transconductance of, 617
 Bloch modulation factor, 29
 Bloch theorem, 28
 Bloch wave, 29, 56
 blue LEDs, 664
 body-centered cubic structure, 39
 body effect coefficient, 504
 Bohr, Niels, 4, 5
 Bohr model, 64
 of an atom, 4
 of the hydrogen atom, 5–11
 Bohr radius, 7, 8, 64
 Boltzmann approximations, 76–77, 80

Boltzmann probability function, 76–77
 Boltzmann's constant, 750
 bond diagram, 60
 bonding, wire, 788–789
 bonding pads, 788
 bonds, dangling, 320, 496, 497
 boron, doping silicon with, 66, 68
 boule, 771. *See also* crystals
 bound states:
 of holes, 67–68
 probability of occupancy, 77
 breakdown:
 avalanche, 275, 276, 632, 700
 junction, 700
 reverse in a pn junction, 269
 breakdown voltage,
 275–276
 bridges (short circuits), 782
 Brillouin zone, 30, 56
 broken down, diode reverse current,
 274
 broken-gap heterojunction, 310
 built-in electric field, 168
 built-in voltage, 168, 232
 measurement of, 343
 MOSFET, 369
 bulk charge, 496, 497–498
 bulk value, 393
 bunny suits, 787
 Burrus LED, 668

C

C-V curves, measuring built-in voltage, 343
C-V measurements, 520
 camera, CMOS. *See* imager, CMOS
 capacitance:
 depletion layer, 279–281
 diffusion, 601
 effect on high-speed behavior of BJTs, 598–603
 heterojunction, 332
 junction, 279–281, 342
 MOS, 515–519
 in nonideal junctions, 332
 of a parallel plate capacitor, 391
 stored-charge, 278, 281–285, 601,
 602–603
 capacitance-voltage characteristics,
 343, 370, 372
 capacitor well, 521–523. *See also*
 potential energy well
 capacitors, MOS, 514–520

carrier concentrations:
 at equilibrium, 90–91
 at high temperatures, 91–95
 at low temperatures, 95
 carrier drift velocity, 121
 carrier freezeout, 95
 carrier lifetime, 62
 carrier mobilities, 117–121
 dependence on transverse and longitudinal fields, 404
 for scattering mechanisms, 124
 carrier multiplication, 239, 273
 carrier scattering, 121–123, 211–213
 carrier velocity, 411
 carriers, 62
 cavity effect, 681
 CCDs (charge-coupled devices),
 686–688
 centrifugal force, 6, 7
 channel:
 in an FET, 358
 enhanced, 379
 channel charge:
 density, 390–392
 dependence on longitudinal field,
 493–495
 channel current, equations for, 365
 channel length:
 modulation, 400–404
 channel low-field mobility. *See* low-field mobility
 channel mobility in the long-channel simple model, 393–396
 channel modulation effect, 433
 channel quantum effects, 504–506
 channel resistance, 420–423
 channel voltage, 396
 characteristic equation, 221
 characteristic tunneling distance, 201
 charge carriers in semiconductors, 1–2
 charge-coupled devices (CCDs),
 686–688
 charges:
 bulk, 497–498
 channel, 390–391
 depletion region, 496
 effect on threshold voltage,
 498–499
 fixed oxide, 496
 interface trapped, 496–498
 mobile electron in FET channel, 497
 mobile ion, 496
 oxide fixed, 496
 oxide trapped, 496, 497
 polarization charge density, 473
 refreshing stored, 523
 chemical-mechanical polishing (CMP), 785
 chemical vapor deposition (CVD),
 775, 785
 chips, packaging of, 787–791
 circuit analysis programs, 353
 circuit model, diode, 277
 circuits:
 devices, matching, 445–449
 inverter circuits, 449–453
 inverters, 444–446
 switching, 454
 transistors in, 362–363
 cladding of a waveguide, 669
 clean rooms, 787
 CMOS (complementary MOS), 444
 CMOS, linear sensor, 691
 CMP (chemical-mechanical polishing), 785
 collection efficiency:
 in a BJT, 551
 graded base transistor, 565
 in a prototype npn BJT, 554–555
 collection multiplication factor, 551
 collector resistance, 594, 595
 collisions with channel walls, 405, 406
 combination rate, 145
 common base:
 circuit configuration, 541
 current gain, 550, 580
 I-V characteristics, 548–550
 common emitter:
 circuit configuration, 541
 current gain, 550
 of the Ebers-Moll model
 representation, 581
 I-V characteristics, 548–550
 communication fiber optic link, 38
 compensated materials, 122, 164
 compensated semiconductors, 126
 compensation, 85, 164
 complementary MOS. *See* CMOS
 composition, nonuniform, 170–173
 compound semiconductors, 16
 concentration:
 electron, 73
 equilibrium, 60, 61, 145
 intrinsic, 81
 conductance, small signal, 277
 conduction band:
 in crystalline Si, 15
 E-K structure of, 57–58

- edge, 16, 396, 398
 electrons in, 18
 Fermi level crossing into, 98
 structures, 57–58
- conduction current in the valence band, 70
- conductivity, 116
 in the dark, 145
 of doped Si, 120
 effective mass, 58, 59, 65
 electron, 117
 high-field effects on, 126–130
 of a semiconductor, 360
 total, 117
- conductivity modulation, 715
 IGBT, 743–744
- conductors, 782
- conjugate variables, 206
- conservation of wave vector, 137
- constant effective mass, 52
- constant field scaling, 523–525
- constant voltage scaling, 523
- constants, 750
- continuity equations, 141–144, 255
 for electrons, 142
 for holes, 142
 to obtain steady-state (dc) currents, 552
- conversion efficiency, solar cell, 655
- copper in the top interconnection layer, 782
- Coulomb attraction, 7
- Coulomb force, 5
- Coulomb’s equation, 352
- covalent bonding, 11–13
 in crystalline solids, 14–21
- critical field, breakdown, 710
- crystalline solids, covalent bonding in, 14–21
- crystallographic directions, 40–41
- crystallographic planes, 40–41
- crystallography, 39
- crystals, 39
 defects, 773
 growth, 770–773
 momentum, 32, 53, 137
 one-dimensional, 49–55
 three-dimensional, 55–57
- cubic structure for crystals, 39
- current, 114, 363
 diffusion, 130–133
 drift, 113–117
- current components, 547
- current crowding, 623
- current density, 115–116
 current flow in a pn homojunction, 236–241
- current gain:
 for common base operation, 550
 for the common emitter configuration, 550
- current saturation, 396–400
- current-saturation region. *See*
 saturation region
- current transport factor. *See* transport efficiency, BJT
- current-voltage characteristics:
 NFET, 386–389
 power rectifiers, 715
- cutoff frequency, 511–513
 effect of base field on, 570
 magnitude, 511
- cutoff frequency, BJT, 604
- cutoff mode, BJT, 542
- CVD (chemical vapor deposition). *See*
 vapor-phase epitaxy
- Czochralski method for crystal growth, 770–772
- D**
- dangling bonds, 320, 496, 497
- dark conductivity, 145
- dark current, 651
- DBRs (distributed Bragg reflectors), 683
- dc model of BJT operation, 543
- dc quantities, representing, 593
- de Broglie relation, 26, 50
- Debye length, 516
- defects, 773
- degeneracy of states, 676
- degenerate semiconductors, 79, 95–100
 tunneling in, 351
- degenerately doped n type, 233
- density of states, 72–74, 80
- density-of-states effective mass, 58, 59
 for electrons, 73
- density-of-states functions:
 for electrons in bands, 72–74
 for holes, 72
 for MOSFETs, 503
- depletion approximation, 250, 265, 307
- depletion capacitance, 280
- depletion devices, 382
- depletion-mode MESFET, 475
- depletion region, 232, 313
- in an ideal MOS capacitor, 518
- in an NMOS transistor, 427
- MOSFET, 369
- photocurrent produced in, 650
 in a Schottky barrier, 349
- depletion region charges, 496. *See also* bulk charge
- depletion-type NFET, 381
- depletion-type PFET, 381
- derived units, 751
- deuterium in passivation, 497
- device degradation, 526–530
- DFB laser, 685
- diamond structure, 14, 40
- DIBL (drain-induced barrier lowering) effect, 457
- dicing process, 789
- dielectric constant for SiO_2 , 391
- dielectric mirror stacks, 683
- dielectric mirrors, 683
- dielectric relaxation time, 239, 338
 majority carriers, 338–340
 minority carriers, 341–342
- differential input resistance, 596
- differential junction capacitance, 280
- diffusion, 113, 130
 of dopants, 777–778
 electron, 168, 169
- diffusion capacitance, 601
- diffusion coefficient, 131
- diffusion current, 2, 130–133, 240, 255–257
 electron, 131
 forward bias, 258–262
 holes, 132
 minority carrier, 239, 260
 reverse bias, 262–264
 step junction, 261
- diffusion length, minority carrier, 149–152
- digital circuits:
 BJT operation in, 580
 transistor operation in, 362, 363
- diode quality factor, 270
- diodes, 223–224
 current flow in, 236–241
 effects of temperature on, 291–292
 four-layer, 725–728
 operation of, 228
 tunnel, 241–245
 turning on and off, 287–288
- DIP (dual-in-line package), 789, 790
- dipoles, tunneling-induced, 317, 325–326

direct gap materials, 137, 138, 661, 662
 direct gap semiconductors, 140
 directions, crystallographic, 40–41
 distillation of silicon, 770
 distributed Bragg reflectors (DBRs), 683
 distributed feedback in a laser, 685
 distribution of carriers, 77
 donor energy, 63, 64–65
 donor states, 63
 donors, 62–66
 dopant atoms, 62
 dopants, diffusion of, 777–778
 doped n-type material, 83
 doped Si, 120
 doping, 62, 777–782
 nonuniform, 164–166
 doping concentration:
 measurement of using *C-V* curves, 344
 mobility varying with, 118–119
 doping gradients, 563–571
 doping profile:
 of a graded-base transistor, 567
 for ion implantation, 227
 measurement of, 345
 double-gate silicon-on-insulator, 457–463
 double heterojunction bipolar transistors (DHBTs), 577–579, 615–616
 double heterostructure, 661, 662
 double-heterostructure (DH) single-quantum-well (SQW) energy band diagrams, 682
 double poly Si self-aligned BJT, 608–611
 drain in an FET, 357, 358
 drain-induced barrier lowering (DIBL) effect, 457
 drain saturation voltage, 392
 drain-source voltage, MOSFET, 378
 drain voltage, dependence of threshold voltage on, 428–429
 DRAM (dynamic random-access memory), 521–523
 drift, 113, 168–169
 drift current, 2, 113–117, 365
 drift velocity, 123
 dry oxidation process, 783
 dual-in-line package (DIP), 789, 790
 dynamic power dissipation, of a CMOS circuit, 451

dynamic random-access memory (DRAM), 521–523

E

E-B junction, 303
E-K diagram, 31
E-K relation:
 expanding in a power series, 49–50
 plotting for three-dimensional crystal, 56
 E versus K diagram, 30
 EAM (electron affinity model), 310
 early effect, 627–632
 Early voltage, 403, 597
 Ebers-Moll model:
 BJT, dc, 579–582
 common emitter, 581
 common emitter ac, 590–592
 edge defects, 773
 edge-emitting diode, 668, 669
 edge-emitting Fabry-Perot laser diode, 684
 edge-emitting laser, 685–686
 edge-emitting LED, 669, 670
 effective channel length, 426–428
 effective density of states, 80–81
 effective electric field, 172–173, 563, 565
 effective mass, 3, 49, 52
 conductivity, 58, 59, 65
 constant, 52
 density-of-states, 58, 59
 direction dependent, 56
 for electrons, 752
 of holes, 59, 70–72, 752
 longitudinal, 73
 negative, 53, 58
 for three-dimensional crystal, 56
 transverse, 73
 tunneling, 201, 270
 efficiency, quantum, 651–652
 eigenfunction, 192
 eigenstate, 192
 eigenvalue, 192
 Einstein coefficients, 675
 Einstein relations, 132
 electric current. *See* current
 electric field:
 effective, 172–173, 563, 565
 in a step junction, 250
 true, 172–173
 electrical neutrality, 162, 165
 electrically neutral regions, 229
 electromagnetic spectrum, 644, 645

electromagnetic waves, 32
 electromigration, 526
 electron affinity, 17, 229, 312
 electron affinity model (EAM), 310
 electron charge, 5
 “electron cloud,” 12–13
 electron collisions, 408
 electron concentration as a function of temperature, 93
 electron conductivity, 117
 electron diffusion, 168, 169
 electron distribution function, 88–89
 electron drift, 168–169
 electron flux, 131, 544, 545. *See also* light flux; optical flux; photon flux
 electron generation rate, 141, 142
 electron-hole pair, 60, 139, 649
 electron injection efficiency. *See* injection efficiency
 electron lifetime, 18, 62, 206. *See also* lifetime
 electron mobility. *See* mobility
 electron momentum, 53
 electron recombination rate. *See* recombination rate
 electron volt (eV), 8
 electron wave function. *See* wave function
 electronic charge, 750
 electronic systems, shielding, 32
 electrons:
 acquiring extra energy, 33
 applying Schroedinger’s equation to, 185–186
 conductivity due to, 116
 continuity equation for, 142, 155
 density of states function for, 73
 diffusion coefficient, 131
 drift velocity, 123
 effective electric field, 172–173
 effective masses of, 752
 finding the values of observable quantities for, 185–186
 free, 6, 23–26, 187–188
 indirect electron transitions, 213–217
 lifetime of, 206
 motion in a crystal, 114
 quasi-electric field for, 174
 quasi-free, 28–32
 tunneling, 32–33
 unopposed, 69
 wavelike properties of, 25

as waves, 180–181
 elements, periodic table of, 753
 emission, 139–141. *See also* optical emission; spontaneous emission; stimulated emission
 emitter, degenerate injection efficiency, 560
 emitter-base current, 580
 emitter-base junctions, recombination in a BJT, 635–636
 emitter resistance, 594, 595
 emitterbase (E-B) junction, 303
 empty states, concentration of, 163
 energies:
 donor, 63
 in the hydrogen atom, 9
 of photons, 34
 of states, 10
 energy band diagrams, 1, 60
 for BJTs, 545, 546
 compared to hybrid diagrams, 522–524
 complemented by *E-K* diagrams, 31
 drawing, 168, 226, 312
 of prototype junctions, 229–236
 energy bands:
 in crystalline solids, 14–15
 occupancy of, 18
 energy gap, 17
 energy “hump,” 526–529
 energy states in a phosphorus ion, 65
 “enhanced” channel, 379
 enhancement-mode FET, 477
 enhancement-mode MESFET, 477
 enhancement MOSFETs, 502
 enhancement-type FET, 380
 enhancement-type NFET, 380, 381
 enhancement-type PFET, 380, 381
 epitaxy, 774–776
 metal-organic vapor-phase (MOCVD), 776
 molecular beam (MBE), 776, 777
 strained-layer, 775
 vapor-phase (VPE), 775–776
 equilibrium, 143
 in an intrinsic semiconductor, 60, 61
 constancy of the Fermi level at, 162–164
 drawing energy band diagrams, 168
 Esaki tunneling, 201
 eutectic bond, 789
 eV (electron volt), 8
 excess carriers, 139

excess majority carriers. *See* majority carriers
 excess minority carriers. *See* minority carriers
 exclusion principle, 13–14
 extraction, 262
 extrinsic semiconductors, 62

F

fabrication techniques for semiconductors, 769–791
 Fabry-Perot cavity, 678–680
 Fabry-Perot laser diode, 684
 fall time, 146, 451
 fast interface states. *See* interface states
 FCC (face-centered cube) lattices, 39
 FD (fully depleted) SOI, 456–457
 feedback in a laser, 677–680
 feedthrough resistance, BJT, 597, 598
 Fermi-Dirac distribution, 76
 Fermi-Dirac distribution function, 75
 Fermi-Dirac probability function, 75, 77
 Fermi-Dirac statistics, 73–77
 Fermi energy, 75
 Fermi function, 165
 Fermi level, 75, 165
 crossing into the conduction band, 98
 at equilibrium, 162–165
 intrinsic materials, 76
 quasi, 152–154
 states of energies above and below, 75–76
 FETs (field-effect transistors), 357–358
 compared to BJTs, 539
 dependence on channel fields, 404
 enhancement, 477
 enhancement-type, 380
 equation for current flow in, 364
 generic, 358–362
 I-V characteristics of, 362–366
 other types of, 468–487
 regions of operation, 360
 resistances of, 420–423
 fiber optics, 664–666
 field plate, 709, 737, 738
 fill factor for a solar cell, 654
 FINFETs, 463–468
 tri-gate, 463–464
 finite potential well, 203–205

first Brillouin zone, 30
 fitting parameters, 118
 fixed oxide charges, 496–497
 flash memory. *See* nonvolatile memory
 flat band, 515–517
 flat band voltage, 499–502, 519
 float-zone process, 772–773
 floating body effect, 456
 “floating” neutral Si region, 456
 flux, 544. *See also* electron flux
 flux density, 131
 forbidden band gap. *See* band gaps
 force on an electron, 53–54
 forward active mode, BJT, 542
 forward bias, 224, 239–240
 forward-biased junction, 661–663
 forward recovery, power diode, 719–720
 four-layer diode, 725–728
 Fourier transforms, 217
 Fowler-Nordheim tunneling, 350
 free electron, 187–188
 free-electron approximation, 23, 49–50
 momentum of, 31–32
 in one dimension, 23–25
 rest mass, 750, 752
 in three dimensions, 27–28
 free-electron model, 28
 free space, permittivity of, 65
 frequency response, BJT, 603–608
 Fresnel reflection, 646, 648, 667
 fully depleted SOI, 456–457

G

GaAs (gallium arsenide):
 conduction band for, 57
 crystal structure of, 16, 40
 donor atom in, 65
 low-field doping dependence, 126, 127
 GaAs-base HBTs, 579
 GaAs-base HFET, 468–472
 GaAs MESFET, 475
 GaAsGe heterojunction, 314, 315, 318
 gain:
 current, common base, 580
 curve, 681
 optical in a laser, 675
 gate-all-around transistor, 464
 gate delay. *See* propagation delay times

gate-source voltage:
 controlling resistance, 385
 MOSFET, 378

gate structure of an FET, 357, 358

gate-substrate capacitance, 516

gate turn-off thyristors (GTOs),
 735–736

Gauss's law, 369

Ge (germanium), 126, 127
 band gap of, 157–158
 conduction band for, 58

generation, 133–135, 141, 142
 optical, 60
 thermal, 60

generation current, 239
 current density, 268
 in a reverse-biased pn junction,
 265–267

generation rate, 266

generation-recombination (G-R)
 current, 264–270

generic FET, 358–362

generic photodetectors, 644–652

germanium. *See* Ge

Giaever tunneling, 199–203

gold in metallization, 782

graded base BJT, 601

graded-base transistor, 565–569

graded doping, 164–169
 combined with graded composition,
 173–174

gradual channel approximation, 405,
 484

grains of polycrystalline
 materials, 20

green gap, 663

GRINSCH laser, 682

ground state of an electron, 11, 12

group velocity, 51
 of an electron in three dimensions,
 56
 of a wave, 25

GTOs. *See* gate turn-off thyristors

guard ring, 660

gull wings, 791

Gummel plot, 635

Gummel-Poon equations, 582

H

h-bar, 7, 750

half-wave rectifier, 225

Hall effect, 70

HBTs (heterojunction bipolar
 transistors), 570–579

double, 577–579, 615–616
 graded-composition, 575–577
 uniformly doped, 571–575

headers, attaching chips to, 787

heavy holes, 58

Heisenberg uncertainty principle,
 205–207

HEMTs. *See* HFETs (heterojunction
 field-effect transistors)

heteroepitaxy, 774

heterojunction bipolar transistors.
See HBTs

heterojunction diodes, *I-V*
 characteristics of, 331–332

heterojunction field-effect transistors
 (HFETs), 468–472
 GaN-based, 473

heterojunctions, 223, 227, 309
 creating, 774
 effects of lattice mismatch on,
 322–323
 GaAs:Ge, 314, 315, 318
 metal-semiconductor, 323
 semiconductor-semiconductor,
 309–323

heterostructure, double, 661,
 662, 682

HFETs (heterojunction field-effect
 transistors), 468–472
 GaN-based, 473

high field effects on conductivity,
 126–130

high-frequency transistors,
 608–611

high injection, 632–633

high temperatures, carrier
 concentrations at, 91–95

higher order terms (HOTs), 50–51

hole conduction, 116

hole current, 132–133

hole-electron pairs, 154

hole mobility, 117, 124

holes, 19, 68–70
 annihilation of, 116
 bound state of, 67–68
 conductivity due to, 116
 conductivity effective mass, 59
 continuity equation for, 142, 155
 density-of-states effective
 mass, 59

density-of-states function for, 73

diffusion current, 132

diffusion length for, 151

effective electric field, 172

effective mass of, 70–72, 752

Fermi-Dirac distribution for, 76

injecting into an n-type
 semiconductor, 341

as particles, 70

quasi-free, 61–62

homogeneous semiconductors, 48,
 113–154, 162

homojunctions, 223, 227, 302–309

hot-carrier-induced degradation,
 526–530

hot carriers, 526

HOTs (higher-order terms), 50–51

hybrid diagrams, 373–375

hybrid-pi model, 594–598

hydrogen:
 extending the Bohr model to, 11
 in passivation, 497

hydrogen atom, 5, 8
 allowed energy levels in, 205
 Bohr energies and orbital radii
 for, 9
 Bohr model of, 5–11

hydrogen-like impurities, 5

hyperabrupt doping profile, 303

hyperabrupt junctions, 309

I

I-E current. *See* injection-extraction
 current

I-V characteristics, basis of derivation,
 362–366

IGBTs. *See* insulated-gate bipolar
 transistors

IGFET (insulated-gate field-effect
 transistor), 367

image, 352

image effect, 351–353

image force effects, 326

image sensors:
 area, 691–692
 charge-coupled devices (CCDs),
 686–688
 linear, 688–691
 semiconductor-based, 686

imager, CMOS, 691

impact ionization, 272

impurity atoms, 62

impurity band, 124, 125

impurity band mobility, 124–126

impurity-induced band-gap narrowing,
 96–98

impurity-induced band-gap
 reduction, 97

indirect gap materials, 137, 138, 140
 indium phosphide (InP), 16
 infinite potential well, 191, 192
 infrared LEDs, 664–671
 injected carrier concentrations, 259
 injection:
 high, 632–633
 low, 338–340, 632
 low-level, 254
 injection current, 241, 635–637
 injection efficiency, 312
 for a BJT, 551
 for a degenerate emitter, 560
 in a prototype npn BJT, 555–556
 injection-extraction (I-E) current, 255
 input impedance, MOSFET compared to BJT, 616
 insulated-gate bipolar transistors (IGBTs), 740–745
 insulated-gate field-effect transistor. *See* IGFET
 insulators, 20–21, 783–785
 integrals, 792
 integrated circuit reliability, MOSFET, 531–532
 intensity, 644
 interface states, 310
 effect on C-V curves, 519–520
 effects of, 318–322
 sources of, 323
 interface trapped charges, 496–498
 interfacial dipole, 316
 internal gain in an APD, 660
 internal reflection loss, 667
 interstitial defects, 773
 intrinsic carrier concentration, 90–92
 intrinsic concentration, silicon (Si), 81
 intrinsic n-type material, 83
 intrinsic semiconductors, 60–62
 intrinsic silicon, equilibrium carrier concentrations for, 119
 “intrinsic” transistor, 421
 inverse mode, operation of a BJT in, 579–581
 inversion layer in an ideal MOS structure, 515
 inverted mode, BJT, 542
 inverted surface region, 370
 inverter circuits, 449–453
 inverters, 444–446
 as signal amplifiers, 511
 ion implantation, 502, 503, 781
 ionic bonding, 16

ionic diffusion, 526
 ionization:
 energy, 16
 impact, 272
 potential, 230, 312
 ionized impurity scattering, 122
 IR drop in junction voltage, 280
 irradiance, 644
 isoelectronic traps, 141
 isolated hydrogen nuclei, 11

J

J leads, 791
 JFETs (junction field-effect transistors), 479–480
 mathematical treatment of, 484
 simple model for, 484–485
 velocity saturation model for, 486–487
 Josephson tunneling, 201
 joules, 8
 junction capacitance, 279–281, 332, 342
 base collector, hybrid pi model, 594–596
 compared to stored-charge capacitance, 283–284
 in a nonuniformly doped junction, 344–345
 in a prototype (step) junction, 342–343
 junction diode, 241
 junction field-effect transistors. *See* JFETs
 junction resistance, 277–278
 junction voltage, 251, 280
 junction widths, 252
 junctions, 223
 drawing energy band diagrams for, 226
 metallurgical, 302
 between a semiconductor and a metal, 323–330
 step (prototype), 228

K

kinetic energy, 12, 25
 for a classical particle, 53
 negative, 53
 kink effect, 456
 Kirchhoff’s current law, 541
 Kirchhoff’s voltage law, 169, 360, 540
 Kirk effect, 633–634

L

lake analogy to FET operation, 361–362
 laser, gain and feedback in, 680–682
 laser diodes, 674–686
 output pattern of, 685
 power-current curve of, 677
 semiconductors important for, 686–687
 structures used to make, 682–686
 lateral bipolar transistors, 637–638
 lattice constants, 39
 of some semiconductors, 666, 775
 lattice matching, 774–775
 lattice mismatch, 322–323
 lattice scattering, 122–123
 LCC (leadless chip carrier), 790
 LDD MOSFET, 530
 lead frame for a DIP, 789
 leaded chip carrier, 791
 leadless chip carrier (LCC), 790
 leakage currents, 239
 in FETs, 359
 subthreshold, FET, 429–432
 LEDs (light-emitting diodes), 661–671
 beams from, 685–686
 blue, 664
 compared to laser diodes, 674
 development of, 663
 edge emitting, 669, 670
 infrared, 664–671
 laser acting as, 677
 OLEDs, 663–664
 physical structure of, 666–668
 typical spectrum of, 667, 668
 visible, 664
 white, 664, 671–674
 lifetime, 675. *See also* electron lifetime
 defined, 62, 142
 minority carrier, 144–146, 148–149
 light-emitting diodes. *See* LEDs
 light energy, absorbing emitting, 34
 light flux, 647–649
 light holes, 58
 light intensity, 644
 lighting, solid-state, 671–674
 lightly doped drain (LDD) MOSFETs, 530
 line defects, 773
 linear image sensors, 688–691
 linear region. *See* sublinear region

linearly graded approximation, 307
 linearly graded junctions, 306–309
 lineshape function, 675, 681
 liquid-phase chemical-vapor deposition (LPCVD)
 process, 784
 lithography, 780–781
 load capacitance, 449–451
 load line, determined by an external circuit, 362, 363
 long-base diode, 255, 347
 long-channel MOSFET model, 389
 with constant mobility, 390–404
 long-channel simple model
 equations for, 433
 revising to account for carrier mobility, 413–415
 with varying mobility, 404–420
 longitudinal effective mass, 57–58
 longitudinal electric field, 365
 effect on the *I-V* characteristics of MOSFETs, 413–415
 producing in the channel, 386
 longitudinal electron mass, 73
 longitudinal field:
 in an NFET device, 389, 390
 dependence of the channel charge on, 493–495
 effect on channel mobility, 411–412
 for various values of drain voltage, 396–397
 longitudinal modes, 678
 Lorentz force, 70, 71–72
 low-field mobility, 127, 404
 effects of the transverse field on, 404–409
 expressing experimentally, 408
 measuring for a MOSFET, 441
 low injection, 338–340, 554
 low-level injection condition, 254, 632
 low-resistance contacts, tunneling in, 351
 LPCVD (liquid-phase chemical-vapor deposition) process, 784

M

macroscopic permittivity, 65
 majority carriers, 117–118
 dielectric relaxation time, 338–340
 drift in an n-type semiconductor, 125
 scattering, 121
 majority electron mobility, 125

manufacture, MOSFET compared to BJT, 617
 maser, 674n6
 mask for a 16-Mbyte memory chip, 782
 mass, effective, 49, 52
 longitudinal, 57–58
 transverse, 57–58
 mass action, law of, 81
 matching to equalize currents, 447
 materials in the operation and design of semiconductor devices, 1
 matter waves, 22
 Matthiessen's rule, 124
 maximum oscillation frequency, 608
 MBE (molecular beam epitaxy), 776, 777
 mean free path, 131
 for bulk Si, 405
 in a MOSFET, 405
 mean free time between collisions, 122, 123, 406
 mean time to failure (MTTF)
 for a thermally activated mechanism, 526
 memory cell, 521
 merged pin-Schottky diode, 723–725
 MESFETs (metal-semiconductor field-effect transistors), 475–479
 enhancement-mode, 477
 mathematical treatment of, 484
 Si-based compared to GaAs-based, 487
 metal-organic chemical vapor-phase (MOCVD) deposition, 776
 metal-oxide-semiconductor field-effect transistors. *See* MESFETs
 metal-semiconductor junctions, 223, 323–330
 metal “vias,” 782
 metallization, 782
 metallurgical junctions, 302
 metal:n-semiconductor Schottky barrier, 327
 metals, 20–21
 minibands, dopants smearing into, 96
 minority carrier lifetime, 144–146, 155
 silicon, empirical expressions for, 148–149
 of a solar cell material, 654–655
 minority carriers, 117, 118
 dielectric relaxation time, 341–342
 diffusion coefficients, 346
 diffusion currents, 239, 552–554
 diffusion lengths, 149–152, 155
 extracted, 262
 hole (electron) mobility, 125
 mobile ion charges, 496, 497
 mobility, 117, 121
 effect of transverse field, 404–405
 low field, 127, 404–405
 majority carrier, 124
 mirror reflectivity, 680
 scattering affecting, 123–124
 temperature dependence of, 126, 128
 transit time (*See* transit time)
 MOCVD (metal-organic chemical vapor-phase) deposition, 776
 modes in a laser, 678
 MODFET. *See* HFETs
 MODFET (modulation doped field-effect transistor), 468n4
 modulation doped field-effect transistor. *See* HFETs
 molecular beam epitaxy (MBE), 776, 777
 momentum:
 conservation of, 53
 crystal, 32, 53
 electron, 53
 monolayers, growing, 776
 MOS capacitors (MOSCs), 367–373, 514–520
 C-V characteristics, 370, 372, 519–520
 charges associated with, 517
 hybrid diagrams, 373–375
 ideal, 515–519
 linear array of, 523
 real, 519–520
 structure of, 514
 MOSFETs (metal-oxide-semiconductor field-effect transistors), 359–360
 analog equivalent circuits, 506–511
 ballistic model for, 423–426
 built-in voltage, 369
 in circuit schematics, 382

compared to BJTs, 616–618
depletion region, 369
energy band diagrams of, 369
enhancement, 502
at equilibrium, 376–377
failure mechanisms in, 526–530
integrated circuit reliability,
 531–532
lightly doped drain (LDD), 530
models compared with experiments,
 421–423
nonvolatile, 465–468
not at equilibrium, 378–389
parameters for typical Si, 415
physics of operation, 378
power, 736–740
qualitative principles of operation,
 367–489
quantitative description, 389–421
scaling, 523–525
small-signal equivalent circuits of,
 507–511
superjunction, 738–740
threshold voltage for, 495–506
trench, 738
MTTF (mean time to failure)
 for a thermally activated
 mechanism, 526
multiple-quantum-well (MQW)
 structure, 683
multiplication current, 272
multiplication factor:
 avalanche, 273, 661
 collection, 551

N

n-channel enhancement MOSFET,
 495–496
n-channel field-effect transistor. *See*
 NFET
n-channel MOSFET, 376–377,
 385–389
n-type semiconductor, 62, 68
n-type silicon, conductivity of, 120
n-well technology, 444
NaCl, 16
native oxide, 320
negative effective mass, 53, 58
negative kinetic energy, 53
net doping concentration, 164
net doping profile, 227
net surface charge, 321
neutral region. *See* quasi-neutral
 region

neutrality, electrical, 162, 165
NFET (n-channel field-effect
 transistor), 358, 359
current flow in, 362
current-voltage characteristics of,
 386–389
depletion-type, 381
electrical characteristics of,
 360–361
enhancement-type, 380, 381
equalizing saturation currents with
 PFETs, 418
inverter circuit, 362, 363
longitudinal field for, 365
matching with PFETs, 447
 simple (inverter) circuit for, 359
nitride, silicon, 784–785
nitrogen in oxide to reduce
 capacitance, 502
NMOS, 420
nn heterojunction, 313
non-neutralized acceptor ions, 231
non-reach-through, 706–708
nondegenerate semiconductors, 79
 quasi Fermi levels and, 152–154
 temperature dependence of, 90–95
nonhomogeneous semiconductors,
 2, 162
nonradiative transmission, 661
nonstep homojunctions, 302–309
nonuniform composition, 170–173
nonuniform doping, 164–167
nonvolatile memory, 465–468
normalized wave function, 23, 182
notch. *See* potential energy well
np junction, 310
npn BJT, 302
 common circuit configurations for,
 541
 current flow in, 544, 546
 energy band diagram for, 598, 599
 I-V characteristics for, 541, 542
 output characteristics of, 548–550
 representing in the Ebers-Moll
 model, 580
npn homojunction transistor, 330
npn transistor, 606
nucleus of an atom, 4

O

observables, 181
occupancy, probability of, 73, 164
OEICs (optoelectronic integrated
 circuits), 685

offset voltage, 636–637
ohmic (low-resistance) contacts, 330
ohmic junctions, 323, 783
Ohm's law, 116, 133, 154
one-dimensional crystals, 49–55
one-dimensional infinite potential
 well, 191
one-sided junctions, 233, 246
 breakdown voltage, 275
 built-in voltage plot, 248
 junction width, 252
one-sided step junction, 254, 280
one-step tunneling process, 270
operators, quantum mechanical,
 181–182
optical absorption, 34, 35
optical communication system, 38
optical emission, 35, 107, 139–141
optical feedback in a laser, 677–680
optical fiber:
 absorption spectrum for, 38
 coupling light to, 668
optical flux, 647–649
optical gain, 675–677
optical generation, 60
optical generation rate, 142, 145
optical loss in optical fibers, 664
optical penetration depth, 658
optical phonons, 128
optical processes, 135–141,
 675–677
optical pumping, 686
optical waveguide, 668, 669, 674
optical window, 312
optoelectronic devices, 644–687
optoelectronic integrated circuits
 (OEICs), 685
organic light-emitting diodes
 (OLEDs), 663–664
output characteristics for the common
 emitter configuration,
 548–550
output conductance, differential, long-
 channel MOSFET, 403
output high-to-low transition time,
 612–614
output low-to-high transition time
 effect on, 613
 increasing the speed of, 614
output resistance in a BJT, 597
oxidation, 783–784
 dry, 787
 thermal, 497
 wet, 783

- oxide:
- breakdown, 526
 - capacitance, 391
 - dielectric constant of, 391
 - layers, 783, 784
 - native, 320
 - trapped charges, 496, 497
- P**
- p-channel field-effect transistor. *See* PFET
- p-type material, 120–121
- p-type semiconductor, 62, 149–150
- packaging of chips, 787–791
- parabolic potential, 11
- parallel plate capacitor:
- capacitance of, 391
 - junction resembling, 280
- parasitic capacitance:
- considering for ac behavior, 590
 - in SOI devices, 454–455
- parasitic effects, 426, 454
- parasitic resistances, minimizing, 598
- parasitic resistances and capacitances, considering for ac behavior, 590
- partially depleted SOI, 456
- particles:
- classical, 21–22
 - holes as, 70
 - phonons as, 122
 - wavelike behavior of, 21
- pass-through current in CMOS switching, 454
- passivation, 321, 497, 784–785
- passivation layer for AlGaAs, 468
- Pauli exclusion principle, 13–14
- PD (partially depleted) SOI, 456
- PECVD (plasma-enhanced CVD), 785
- penetration depth, 658
- periodic table of elements, 753
- permittivity:
- of free space, 5, 65
 - macroscopic, 65
 - of vacuum, 750
- PFET (p-channel field effect transistor), 358, 359
- depletion-type, 381
- electrical characteristics of, 360, 361
- enhancement-type, 380, 381
- equalizing saturation currents with NFETs, 418
- matching with NFETs, 447
- phase velocity, 25
- phonon scattering, 128
- phonons, 21, 33, 60–62, 122–123, 207–211
- carrier scattering by, 211–213
 - excess electron energy released as, 661
 - indirect electron transitions, 213–217
 - optical, 128
- phosphorus as a donor atom, 62–65
- photoconductivity, 144, 145
- photocurrent, 649, 650
- photodetectors, 644
- generic, 644–652
 - photodiode, 644, 650–651
 - photoelectric effect, 4
 - photolithography, 769, 780–782
- photon energy:
- finding the spread of, 206–207
- photon flux:
- density, 649, 655–656
 - variation with distance, 647
- photons, 21, 33, 34
- photosensist, 780, 781
- physical constants, 750
- physical parameters, analyses from *C-V* measurements, 520
- piezoelectric effect, 472–473
- PIN diode, 658–660, 705–706
- pin grid array (PGA), 790
- PIN photodetector, 658–660
- pinch-off voltage. *See* threshold voltage
- Planck's constant, 4, 750
- planes, crystallographic, 40–41
- plasma-enhanced CVD (PECVD), 785
- plastic leaded chip carrier (PLCC), 790
- pn homojunctions, 227
- with bias applied, 234
 - current-voltage characteristics of, 254–274
 - at equilibrium, 232, 236–237
 - fabricating, 227
 - quantitative descriptions of energy band diagrams of, 245–277
 - under reverse bias, 236
 - switching circuit model for, 290
- pn junction diode, 329
- pn junctions, 246
- built-in voltage of, 247
 - capacitance in, 279–285
- small-signal resistance, 277–278
- as switches, 285
- pnp BJT, 541, 542
- point defects, 773
- Poisson's equation, 248, 345
- polarization charge density, 473
- polycrystalline materials, 20
- polycrystalline silicon, 772
- polymers, organic for LEDs, 664
- polysilicon conductors, 783
- polysilicon emitter, 610
- population inversion, 674, 676, 681
- positive charge of holes, 58
- positive effective mass of holes, 58
- potential barrier, reflection and transmission at, 193–195
- potential energy:
- for an electron, 14, 52, 272
 - arbitrariness of, 10
 - describing a physical problem to be solved, 185
 - in a three-dimensional crystal, 31–32
- potential energy well, 189–190, 312. *See also* capacitor well; quantum well
- finite, 203–205
 - infinite, 190–193
 - particle in a one-dimensional, 182–184
- power consumption:
- MOS, 449
- power conversion efficiency of a solar cell, 654
- power dissipation:
- of a CMOS circuit, 451
 - MOSFET compared to BJT, 617
- power semiconductor devices, 699
- power supply voltage, 451
- power transistors, 626
- principal quantum number, 13
- probability:
- density, 22, 182
 - inversion, 674, 676
 - of occupancy, 73, 77
 - quantum mechanical, 22–23
- prompt photocurrent, 650
- propagation constant, 22
- propagation delay times, 451–453
- prototype BJT, 546, 551–562
- prototype homojunctions, 228, 277–285
- prototype pn homojunctions, 245–277
- prototype pn junctions, 229–241

pseudo-classical mechanics, 1, 3, 48–49
 for electrons in crystals, 49–57
 push-out effect. *See* Kirk effect
 pyrolysis process, 783

Q

quality factor, diode, 270
 quanta, 3
 quantized energies, 4, 8
 quantum efficiency, photodiode, 651–652
 quantum mechanical operators, 181–182
 quantum mechanics, 3
 introduction to, 180–210
 old, 4
 results from, 187–207
 quantum numbers, 7, 13–14
 quantum well, 10, 192, 203–205, 682.
See also potential energy well
 quasi-electric field, 174
 quasi-Fermi levels, 152–154
 quasi-free electron model, 28–32, 50
 quasi-free electrons, 188–189
 quasi-free hole, 61–62
 quasi-neutral region, 168, 234, 240
 current flow in, 240
 in a graded-base transistor, 567

R

rad-hard devices, 490
 radiative transition, 661
 RAM (random-access memory), 521
 random telegraph noise (RTN), 531
 random walk, 130
 reach-through, 706–707
 reading a DRAM memory cell, 523
 reclaimable charge, 283, 347, 601
 recombination, 62, 133–135, 141, 142, 144
 recombination and generation:
 under forward bias, 267–270
 under reverse bias, 265–267
 recombination centers, 287
 recombination current, 239, 240, 268, 557
 current density, 268
 in a forward-biased pn junction, 268
 recombination rate, 141, 142, 264
 recoverable charge, 601
 rectifier:
 circuit diagram for, 224
 current-voltage characteristics, 715

half-wave, 225
 power, 700
 rectifying diodes, 700
 rectifying junctions, 323
 reduced zone, 30
 reflection:

of an electron by a barrier, 193–195

Fresnel, 646, 648, 667

total internal, 667

reflectivity mirrors, 680

refreshing stored charges, 523

relative permittivity (dielectric constant) for silicon, 248

relaxation time. *See* dielectric relaxation time

resistances:

channel, 421

differential input, BJT, 596

feedthrough, BJT, 598

hybrid-pi model, 594, 595

output, BJT, 597

parasitic, BJT, 598

series, diode, 277

series, FET, 421

small-signal, pn junction, 277–278

resistivity, 115

resonances of a Fabry-Perot cavity, 679, 680

resonant frequency, 346

resonant wavelengths, 679

responsivity, 651, 652

reverse bias, 224, 238–239

 breakdown, 275–276

 tunneling, 270–272

reverse-biased prototype junction, 262–263

reverse-biased Schottky diode, 350

reverse dark current, 651

reverse recovery, power diode, 720–723

Richardson-Dushman equation, 328

rise time, 146, 451

Rutherford, Ernest, 4

S

sapphire, growing Si onto insulating, 455

saturation:

 BJT, 549, 550

 FET, 402

saturation currents, 392

 equalizing, 418

 matching, 447

saturation mode, BJT, 542

saturation region:

 for a BJT device, 550

 of MOSFET *I*-*V* curves, 360, 361

saturation transconductance, 510

saturation velocity, 128, 392

saturation voltage:

 variation of, 419

scaling:

 constant field, 523–525

 constant voltage, 523

 MOS, 447, 453

 MOSFETs, 523–525

scattering:

 effects on mobility, 123–124

 ionized impurity, 122

 mean free times between, 124

 phonon, 122–123

 phonon or lattice, 126, 128

 physics of, 121–123

Schottky barriers, 331

 barrier lowering due to image effect, 351–353

 tunneling through, 349–351

Schottky-clamped transistor, 614–615

Schottky diodes, 323, 326

 barrier lowering in, 351–353

 current voltage characteristics, 712–714

 first-order model, 348–349

I-*V* characteristics of, 326–329

 junction capacitance, 343, 344

 p-type, 328, 329

 second-order effects in, 348, 349–353

 stored charge capacitance, 332

 tunneling in, 328, 329, 350

 turn-off transient in, 720–723

Schroedinger's equation, 48

 development of, 184–185

 method of applying, 185–186

 solving, 187–190

 time-independent, 23, 24

SCRs. *See* silicon-controlled rectifiers

scrubbing in a wire bond, 788

secondary ion mass spectroscopy. *See* SIMS

seed crystal, extending, 739–740, 770–771

segregation coefficients, 773

self-aligned transistor, 608–611

semiconductor-controlled switch (SCS), 725

semiconductor current, 115

semiconductor diodes. *See* diodes

- semiconductor heterojunctions, 309–323
- semiconductor junctions. *See* junctions
- semiconductor-to-metal barrier, 327
- semiconductors, 1, 20–21
- absorption coefficients, 648
 - band gaps of, 17, 751
 - constants of, 750
 - current in, 2
 - electron affinity, 17
 - energy band diagrams for, 20
 - at equilibrium, 113
 - extrinsic, 62
 - fabrication techniques for, 769–791
 - intrinsic, 60–62
 - laser diodes, 686–687
 - optical emission in, 139–141
 - optical processes in, 135–141
 - tunneling in, 199–203
 - tunneling structure in, 197
 - valence bands of, 751
- sensitivity, capacitance-voltage measurements, 346
- series base resistance, 594, 595
- series resistances, 420–423
- diode, 277
 - FET, 421
- shielding, 32
- short-base diodes, 255, 263–264
- compared to long-base, 291
 - stored-charge capacitance, 347
 - stored charge capacitance of, 346–348
- short-channel effects:
- dependence of effective channel length on drain voltage, 426–428
 - dependence of threshold on drain voltage, 428–429
 - FET, 426–429
- short-circuit current gain, 603–604
- Si. *See* silicon (Si)
- S.I. (International System) units, 8
- Si-base HBTs, 579
- Si-base MESFETs, 487
- Si-Ge alloys:
- characteristics of, 170–171
- Si npn BJT, 552
- sidewall scattering, 407
- SiGe-base HBTs, 579
- Si:Ge Nn heterojunction, 322–323
- signal amplifiers, CMOS inverters as, 511
- silica (SiO_2), 770
- silicide, 455
- silicon (Si), 14–15
- absorption coefficient, 649
 - as an indirect gap material, 137
 - conduction band for, 57–58
 - crystal structure of, 40
 - depletion region, 456
 - doping with boron, 66, 68
 - homojunction diode, 223, 224
 - intrinsic concentration, 81
 - mobility in, 118
 - MOSFETs, 415
 - production of ultrapure, 769–776
 - refining of, 770
 - solar cells, 655, 657–658
 - valence of, 62
- silicon-controlled rectifiers (SCRs), 730–732
- silicon dioxide (SiO_2), 17
- as an electrical insulator, 783–784
 - dielectric constant for, 391
 - as a native oxide, 320
- silicon nitride (Si_3N_4), 784–785
- silicon on insulator (SOI) devices, 454–457
- double-gate, 457–463
- silicon on sapphire (SOS) devices, 455
- SIMOX process (separation by implanted oxygen), 454
- simple cubic, 39
- SIMS (secondary ion mass spectroscopy), 563–565, 779, 780
- sinker, 552
- SiO_2 . *See* silicon dioxide
- skin depth, 32
- small-outline package, 790
- small-signal capacitance, 280
- small-signal equivalent circuit, MOSFET, 507–511
- small-signal impedance of prototype homojunctions, 277–285
- small-signal junction resistance, 277–278
- small-signal mode, diode, 277
- small-signal models:
- BJT, 592–598
- snappiness factor, 720
- Snell's law, 667
- SOI (silicon on insulator) devices, 454–457
- solar cells, 652–658
- solar spectrum, 654, 655
- solid-state lighting, 671–674
- Sommerfeld, 5
- SOS (silicon on sapphire) devices, 455
- source, FET, 357, 358
- space charge neutrality, 92, 95
- space charge region, 232. *See also* transition region
- specific on-resistance, 711–712
- spectrum:
- electromagnetic, 644, 645
 - LED, 667, 668
 - optical fiber, 38
 - solar, 654, 655
- speed, MOSFET compared to BJT, 617
- SPICE:
- determining switching waveforms, 290
 - Level 1 model, 390
 - parameters used in, 581
 - voltage and current response determined by, 290
- SPICE Level 1, channel length modulation parameter, 403–404
- spin of an electron, 13
- spin-orbit coupling, 59
- split-off band, 59
- spontaneous emission, 661, 675
- characteristics of, 666
 - optical, 35
- square law model, 394
- SRAM, 531
- staggered heterojunction, 310
- states:
- degeneracy of, 676
 - energies of, 10
- step function, voltage distribution in, 251
- step junction, 228
- current in, 261
 - electric field in, 250
 - equilibrium energy band diagram for, 257
 - junction capacitance, 280, 344
- step pn junction, built-in voltage of, 246
- stimulated emission, 674, 676–677
- storage time, 286–287
- stored charge:
- BJT, 611–612
 - reclaimable in a step junction, 283

stored-charge capacitance, 278, 281–285, 601, 602–603
 BJT, 598–603
 Schottky diode, 332
 short base diode, 346–348
 straddling heterojunction, 310
 strained-layer epitaxy, 775
 sublinear region:
 FET, 361
 of MOSFET *I-V* curves, 360
 substrate:
 preparation of, 769–776
 subthreshold leakage current, 429–432
 subthreshold region:
 FET, 376
 of MOSFET *I-V* curves, 360
 subthreshold swing, 431, 456
 in double-gate fully depleted SOI
 MOSFET, 461–463
 superjunction MOSFET, 738–740
 surface effects, 318–322
 surface-mount type of packages, 790
 surface potential, 369, 379
 surface states, 319. *See also* traps
 band bending due to, 321
 occupation of, 500
 swing, 431
 switches, transistors operating as, 357
 switching in CMOS inverter circuits, 449–453
 switching time, BJT, 613
 symbols, list of, 754–768

T

temperature dependence:
 of carrier concentrations, 90–95
 of diodes, 291–292
 of mobility, 126, 128
 thermal agitation, 18
 thermal generation, 60
 in an isolated MOS capacitor, 521
 in the region of a capacitor well, 523
 thermal generation rate, 142
 thermal oxidation, 497
 thermionic diode current, 348
 thermionic emission current, 328
 thin outline small package (TSOP), 791
 Thompson, J. J., 4
 Thompson model of an atom, 4
 three-dimensional crystals, 55–57
 threshold, 379
 adjustment of, 496

conditions, 383–385
 control of via ion implantation, 502
 control of via substrate bias, 503
 effect of charges on, 499
 lasing, 677
 measurement of, 441
 occurrence of, 385
 threshold voltage, 496
 calculating for an n-channel
 MOSFET, 500–501
 control, 502–506
 dependence on drain voltage, 428–429
 effect of charges on, 498–499
 measuring MOSFET, 440–442
 MESFET, 478
 through-hole-mount packages, 790
 thyristors, 725–736
 tilt, adjusting, 168
 time-dependent dielectric breakdown (TDDB), 532
 time-dependent wave function, 186
 time-independent Schroedinger's equation, 23, 24
 time-independent wave function, 23, 29, 186
 TO can, 789, 790
 total conductivity, 117
 total energy for an electron, 23–25
 total wave function for a free electron, 25
 transconductance, 509
 BJT, 597
 MOSFET compared to a BJT, 616–617
 transfer characteristics of a CMOS inverter, 446
 transient losses, power devices, 718–719
 transient switching effects, 449–453
 transients, 285
 transistors, 357
 in circuits, 362–363
 classes of, 357
 gate-all-around, 464
 HBTs, 570–579
 high-frequency, 608–611
 lateral bipolar, 637–638
 MODFET, 468n4
 tunnel field-effect, 480–483
 transit time, 291, 347
 transition region, 230. *See also* space charge region
 transition time, 613–614

transmission at a barrier, 193–195
 transport efficiency, BJT, 551
 transverse effective mass, 57–58
 transverse electron mass, 73
 transverse field:
 in an NFET device, 389, 390
 effect of, 434
 effect on mobility, 404–405
 effects on *I-V* characteristics, 410
 effects on low-field mobility, 404–409
 transverse modes, 678
 trap level, 135
 traps, 287. *See also* surface states
 TRIAC, 733–734
 triode region. *See* sublinear region
 true electric field, 172–173
 TSOP (thin small-outline package), 791
 tunnel diodes, 241–245
 tunnel field-effect transistors (TFETs), 480–483
 tunneling, 195–203, 270
 by electrons, 32–33
 under forward bias, 240
 Fowler-Nordheim, 350
 Giaever, 199–203
 probability in a junction, 270
 reverse bias, 270–272
 in a Schottky diode, 328, 329
 through Schottky barriers, 349–351
 tunneling breakdown, 276
 tunneling current, 239
 tunneling distance, 272
 characteristic, 201
 estimating, 276
 tunneling effective mass, 201
 tunneling-induced dipole layer, 317
 tunneling-induced dipoles, 314–318, 325–326
 tunneling process, 32
 turn-off time, 285
 compared to turn-on time, 287–288
 shortening, 287
 turn-off transient, 285–287
 turn-on time, 285
 turn-on transient, 287–291
 two-step generation, 134, 135
 two-step recombination, 134, 135

U

ultrapure silicon, 769–776
 UMOS, 738
 uncertainty principle, 205–207

uncompensated materials, 125

unipolar devices, 539

units:

conversion of, 751

derived, 751

unity current gain frequency, 604–605,
608

unopposed electrons, 69

V

vacancy, defects, 773

vacant states. *See* holes

vacuum level, 6, 312

valence band:

conduction current in, 70

in crystalline Si, 14, 15

heavy hole, 58

light hole, 58

split-off, 59

valence band edge, 16

valence bands, 67

E-K structure of, 58–59

vapor-phase epitaxy (VPE), 775–776

varactors, 345–346

variable capacitance diodes, 345–346

VCSELs (vertical-cavity surface-emitting
lasers), 683, 684

VDMOS, 738

velocity:

electron, 68, 69

group, 25, 51

phase, 25

velocity saturation, 128

effects, 415, 419

JFET, 486–487

long-channel model with, 434

MOSFET simple model, 411

vertical cavity lasers, beams of, 685

vertical-cavity surface-emitting lasers
(VCSELs), 683, 684

vias, metal, 782

visible LEDs, 664

VMOS, 738

voids (open circuits), 782

volatile memory, 521

voltage:

built-in, 168

distribution in a step function, 251

junction, 280

offset, 636–637

VPE (vapor-phase epitaxy), 775–776

W

wafer sizes, 771

wafers, separating, 789

water analogy. *See* lake analogy to
FET operation

wave function, 22–23, 180–181

of an electron, 23, 48

finding from Schroedinger's

equation, 184–185

for a free electron, 187

normalized, 182

probability and, 22–23

wave number, 22

wave-particle duality, 21–22

wave vector, 21, 25

conservation of, 137

relation to the wavelength, 31

in three dimensions, 27

waveguide, optical, 668, 669, 674

wavelengths:

of light, 35

relating to momentum, 26

resonant, 679

waves, 21

wedge bonding, 788

well-defined minority carrier

(electron) lifetime, 142

wet oxidation process, 783

white LEDs, 664, 671–674

wide bandgap, 700

Wilson, 5

Wilson-Sommerfeld model, 4

window effect, 313

wire bonding, 788–789

work functions, 230, 369, 519

writing into a memory cell, 521–523

X

X-ray scattering, 4

Y

yield, 774

Z

Zener diodes, 274

Zener effect, 270

zinc blende structure, 40

zone, Brillouin, 30, 56

