McGRAW-HILL
NANOSCIENCE & TECHNOLOGY
SERIES

Semiconductor Heterojunctions and Nanostructures

Omar Manasreh

Semiconductor Heterojunctions and Nanostructures

Omar Manasreh
University of Arkansas
Fayetteville, Arkansas

McGraw-Hill
New York Chicago San Francisco Lisbon London Madrid
Mexico City Milan New Delhi San Juan Seoul
Singapore Sydney Toronto

The McGraw-Hill Companies

Copyright © 2005 by The McGraw-Hill Companies, Inc. All rights reserved. Manufactured in the United States of America. Except as permitted under the United States Copyright Act of 1976, 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 permission of the publisher.

0-07-146968-0

The material in this eBook also appears in the print version of this title: 0-07-145228-1.

All trademarks are trademarks of their respective owners. Rather than put a trademark symbol after every occurrence of a trademarked name, we use names in an editorial fashion only, and to the benefit of the trademark owner, with no intention of infringement of the trademark. Where such designations appear in this book, they have been printed with initial caps. McGraw-Hill eBooks are available at special quantity discounts to use as premiums and sales promotions, or for use in corporate training programs. For more information, please contact George Hoare, Special Sales, at george_hoare@mcgraw-hill.com or (212) 904-4069.

TERMS OF USE

This is a copyrighted work and The McGraw-Hill Companies, Inc. (“McGraw-Hill”) and its licensors reserve all rights in and to the work. Use of this work is subject to these terms. Except as permitted under the Copyright Act of 1976 and the right to store and retrieve one copy of the work, you may not decompile, disassemble, reverse engineer, reproduce, modify, create derivative works based upon, transmit, distribute, disseminate, sell, publish or sublicense the work or any part of it without McGraw-Hill’s prior consent. You may use the work for your own noncommercial and personal use; any other use of the work is strictly prohibited. Your right to use the work may be terminated if you fail to comply with these terms.

THE WORK IS PROVIDED “AS IS.” McGRAW-HILL AND ITS LICENSORS MAKE NO GUARANTEES OR WARRANTIES AS TO THE ACCURACY, ADEQUACY OR COMPLETENESS OF OR RESULTS TO BE OBTAINED FROM USING THE WORK, INCLUDING ANY INFORMATION THAT CAN BE ACCESSED THROUGH THE WORK VIA HYPERLINK OR OTHERWISE, AND EXPRESSLY DISCLAIM ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. McGraw-Hill and its licensors do not warrant or guarantee that the functions contained in the work will meet your requirements or that its operation will be uninterrupted or error free. Neither McGraw-Hill nor its licensors shall be liable to you or anyone else for any inaccuracy, error or omission, regardless of cause, in the work or for any damages resulting therefrom. McGraw-Hill has no responsibility for the content of any information accessed through the work. Under no circumstances shall McGraw-Hill and/or its licensors be liable for any indirect, incidental, special, punitive, consequential or similar damages that result from the use of or inability to use the work, even if any of them has been advised of the possibility of such damages. This limitation of liability shall apply to any claim or cause whatsoever whether such claim or cause arises in contract, tort or otherwise.

DOI: 10.1036/0071469680

McGraw Hill **Professional**

Want to learn more?

We hope you enjoy this McGraw-Hill eBook! If you'd like more information about this book, its author, or related books and websites, please click here.

To my PhD advisor, Don O. Pederson
and my friend Greg J. Salamo who inspired me in
many ways

ABOUT THE AUTHOR

Omar Manasreh, Ph.D., is a full professor in the Department of Electrical Engineering at the University of Arkansas in Fayetteville and the author of over 150 journal articles. He edited and coedited over ten books and organized over ten symposia. Dr. Manasreh is also a series editor for a technical book series, *Nanoscience and Technology*, published by McGraw-Hill. He has extensive experience in the experimental and theoretical optoelectronic properties of III-V semiconductors, superlattices, nanostructures, and related devices.

Copyright © 2005 by The McGraw-Hill Companies, Inc. Click here for terms of use.

Contents

Preface xi
Acknowledgments xv
List of Symbols and Abbreviations xvi

Chapter 1. Introduction to Quantum Mechanics 1

1.1 Introduction 1
  1.1.1 Blackbody radiation 2
  1.1.2 The specific heat capacity of solids 2
  1.1.3 Photoelectric effect 4
  1.1.4 The Bohr model of the atom 6
1.2 The de Broglie Relation 8
1.3 Wave Functions and the Schrödinger Equation 9
1.4 Wave Packet at a Given Time 10
1.5 Separation of Variables 15
1.6 Dirac Notation 17
1.7 Important Postulates 18
1.8 Important Mathematical Tools 19
  1.8.1 The scalar product 19
  1.8.2 Linear operators 20
  1.8.3 Action of a linear operator on a bra 21
  1.8.4 The adjoint operator $A^\dagger$ of a linear operator A 21
  1.8.5 Eigenvalues and eigenfunctions of an operator 22
  1.8.6 The Dirac δ-function 24
  1.8.7 Fourier series and Fourier transform in quantum mechanics 26
1.9 Variational Method 28
1.10 Perturbation 29
1.11 Angular Momentum 33
Summary 34
Problems 35

Chapter 2. Potential Barriers and Wells 37

2.1 Stationary States of a Particle in a Potential Step 38
2.2 Potential Barrier with a Finite Height 42
2.3 Potential Well with an Infinite Depth 49

2.4 Finite-Depth Potential Well 51
2.5 Unbound Motion of a Particle ($E > V_o$) in a Potential Well with a Finite Depth 56
2.6 Triangular Potential Well 58
2.7 Parabolic Potential Well 61
2.8 Delta-Function Potentials 64
2.9 Transmission in Finite Double-Barrier Potential Wells 70
2.10 Wentzel-Kramers-Brillouin (WKB) Approximation 72
2.11 Energy Levels in Double Quantum Well Structure 78
Summary 80
Problems 81

Chapter 3. Electronic Energy Levels in Periodic Potentials 83

3.1 Bloch's Theorem 84
3.2 The Kronig-Penney Model 85
3.3 Bloch Electron in a Weak Periodic Potential 89
3.4 One-Electron Approximation in a Periodic Dirac δ-functions 93
3.5 Superlattices 95
3.6 Effective Mass 98
3.7 Band Structure Calculation Methods 99
  3.7.1 Tight-binding method 99
  3.7.2 $\mathbf{k} \cdot \mathbf{p}$ method 102
  3.7.3 Envelope function approximation 113
Summary 119
Problems 120

Chapter 4. Tunneling Through Potential Barriers 123

4.1 Transmission Through Potential Barriers 124
4.2 Tunneling Through Pyramidal Potential Barriers 131
4.3 Double-Barrier Potential 135
4.4 The *pn*-Junction Tunneling Diode 139
4.5 Resonant Tunneling Diodes 143
4.6 Coulomb Blockade 147
Summary 149
Problems 150

Chapter 5. Distribution Functions and Density of States 153

5.1 Distribution Functions 156
5.2 Maxwell-Boltzmann Statistic 157
5.3 Fermi-Dirac Statistics 161
5.4 Bose-Einstein Statistics 164
5.5 Density of States 166
5.6 Density of States of Quantum Wells, Wires, and Dots 171
  5.6.1 Quantum wells 171
  5.6.2 Quantum wires 175
  5.6.3 Quantum dots 178

5.7 Density of States of Other Systems 179
  5.7.1 Superlattices 179
  5.7.2 Density of states of bulk electrons in the presence of a magnetic field 181
  5.7.3 Density of states in the presence of an electric field 184
Summary 188
Problems 189

Chapter 6. Optical Properties 191

6.1 Fundamentals 192
6.2 Lorentz and Drude Models 196
6.3 The Optical Absorption Coefficient of the Interband Transition in Direct Bandgap Semiconductors 199
6.4 The Optical Absorption Coefficient of the Interband Transition in Indirect Bandgap Semiconductors 205
6.5 The Optical Absorption Coefficient of the Interband Transition in Quantum Wells 206
6.6 The Optical Absorption Coefficient of the Interband Transition in Type II Superlattices 208
6.7 The Optical Absorption Coefficient of the Intersubband Transition in Multiple Quantum Wells 210
6.8 The Optical Absorption Coefficient of the Intersubband Transition in GaN/AlGaN Multiple Quantum Wells 216
6.9 Electronic Transitions in Multiple Quantum Dots 217
6.10 Selection Rules 222
  6.10.1 Electron-photon coupling of intersubband transitions in multiple quantum wells 222
  6.10.2 Intersubband transition in multiple quantum wells 223
  6.10.3 Interband transition 223
6.11 Excitons 225
  6.11.1 Excitons in bulk semiconductors 225
  6.11.2 Excitons in quantum wells 231
  6.11.3 Excitons in quantum dots 233
6.12 Cyclotron Resonance 234
6.13 Photoluminescence 239
6.14 Lattice Vibrations and Phonons 245
Summary 255
Problems 256

Chapter 7. Electrical and Transport Properties 261

7.1 Introduction 261
7.2 The Hall Effect 265
7.3 Quantum Hall and Shubnikov–de Haas Effects 269
  7.3.1 Shubnikov–de Haas effect 272
  7.3.2 Quantum Hall effect 275
7.4 Charge Carrier Transport in Bulk Semiconductors 277
  7.4.1 Drift current density 278
  7.4.2 Diffusion current density 282
  7.4.3 Generation and recombination 286
  7.4.4 Continuity equation 288

7.5 Boltzmann Transport Equation 294
7.6 Derivation of Transport Coefficients Using the Boltzmann Transport Equation 298
  7.6.1 Electrical conductivity and mobility in *n*-type semiconductors 300
  7.6.2 Hall coefficient R_H 302
7.7 Scattering Mechanisms in Bulk Semiconductors 304
  7.7.1 Scattering from an ionized impurity 306
  7.7.2 Scattering from a neutral impurity 307
  7.7.3 Scattering from acoustic phonons 307
  7.7.4 Optical phonon scattering—polar and nonpolar 309
  7.7.5 Scattering from short-range potentials 310
  7.7.6 Scattering from dipoles 311
7.8 Scattering in a Two-Dimensional Electron Gas 312
  7.8.1 Scattering by remote ionized impurities 314
  7.8.2 Scattering by interface roughness 315
  7.8.3 Electron-electron scattering 317
7.9 Coherence and Mesoscopic Systems 318
Summary 324
Problems 325

Chapter 8. Semiconductor Growth Technologies: Bulk, Thin Films, and Nanostructures 329

8.1 Introduction 329
8.2 Growth of Bulk Semiconductors 334
  8.2.1 Liquid-encapsulated Czochralski (LEC) method 334
  8.2.2 Horizontal Bridgman method 341
  8.2.3 Float-zone growth method 343
  8.2.4 Lely growth method 346
8.3 Growth of Semiconductor Thin Films 347
  8.3.1 Liquid-phase epitaxy method 348
  8.3.2 Vapor-phase epitaxy method 349
  8.3.3 Hydride vapor-phase epitaxial growth of thick GaN layers 352
  8.3.4 Pulsed-laser deposition technique 355
  8.3.5 Molecular beam epitaxy growth technique 357
  8.3.6 Metal-organic chemical vapor deposition growth technique 372
8.4 Fabrication and Growth of Quantum Dots 377
  8.4.1 Nucleation 379
  8.4.2 Fabrications of quantum dots 387
  8.4.3 Epitaxial growth of self-assembly quantum dots 388
Summary 399
Problems 400

Chapter 9. Electronic Devices 403

9.1 Introduction 403
9.2 Schottky Diode 406
9.3 Metal-Semiconductor Field-Effect Transistors 411
9.4 Junction Field-Effect Transistor 420
9.5 Heterojunction Field-Effect Transistors 423
9.6 GaN/AlGaN Heterojunction Field-Effect Transistors 428
9.7 Heterojunction Bipolar Transistors 433

9.8 Tunneling Electron Transistors 435
9.9 Coulomb Blockade and Single Electron Transistor 437
Summary 451
Problems 452

Chapter 10. Optoelectronic Devices 457

10.1 Introduction 457
10.2 Infrared Quantum Detectors 459
  10.2.1 Figures of merit 460
  10.2.2 Basic concepts of photoconductivity 465
  10.2.3 Noise in photodetectors 469
  10.2.4 Multiple-quantum-well infrared photodetectors 472
  10.2.5 Infrared photodetectors based on multiple quantum dots 485
10.3 Light-Emitting Diodes 492
10.4 Semiconductor Lasers 498
  10.4.1 Basic principles 498
  10.4.2 Semiconductor heterojunction lasers 505
  10.4.3 Quantum well edge-emitting lasers 511
  10.4.4 Vertical cavity surface-emitting lasers 515
  10.4.5 Quantum cascade lasers 517
  10.4.6 Quantum dot lasers 521
Summary 526
Problems 527

Appendix: Tables 529

Bibliography 533
Index 539

This page intentionally left blank.

Preface

Applying old concepts to new technology is a very difficult task. Novel and innovative approaches are required for one to reach a scientific understanding of the ever-changing field of semiconductor nanostructures. While the growth of quantum structures, such as quantum dots (also known as atomic designers), has the tendency to be an art rather than a science, the fabrication of these structures changed our way of looking at things. When I decided to create a graduate course on *nanostructures*, I scanned the open literature for a suitable textbook that would cover the topics of interest to students who are eager to learn and understand semiconductor quantum structures at the nanoscale limits. I struggled finding such an ideal textbook due to the fact that the field of semiconductor nanostructures is changing so quickly and the novelty of the field is presented in articles published in technical journals and highly specialized reference books that are directed toward highly specialized researchers and are not suitable as textbooks. This is what motivated me to write this book, which covers various concepts ranging from bulk semiconductor materials to semiconductor quantum dots. To understand quantum wells, wires, and dots, it is imperative to possess a basic knowledge of quantum mechanics and how one can apply Schrödinger's equation to calculate the quantized electronic energy levels in such a tiny structure. This requirement is due to the fact that classical mechanics is limited in providing an explanation of almost all the properties of nanostructures. Quantum mechanics, however, can provide insight and accurate predictions of phenomena observed in cases of semiconductor nanostructures. This textbook is by no means a complete or an ideal textbook, but it is one step in a changing field full of limitless possibilities of innovations and inventions.

The textbook is designed to cover topics in the subject of heterojunctions and nanostructures that are of interest to graduate students in electrical engineering, materials engineering, and applied physics. Advanced undergraduate students as well as researchers in the field of

Copyright © 2005 by The McGraw-Hill Companies, Inc. Click here for terms of use.

semiconductor heterojunctions and nanostructures may benefit from it. I imagine that graduate students who use this textbook in their studies will continue to use it as a reference book after their graduation.

The basic properties of bulk and low-dimensional systems down to quantum structures with zero degrees of freedom are discussed. This book is structured such that the discussion starts with bulk crystalline materials, which is the basis for understanding the basic properties of semiconductors. The discussion then evolves to cover quantum structures, such as single and multiple quantum wells. Then, attempts are made to discuss and explain the properties of even lower-dimensional systems, such as quantum wires and dots. In many cases, the theoretical derivation of various properties is simpler for quantum dots. However, since the field is still in its infancy, there are too many unknowns and many of the properties of the lower-dimensional systems are yet to be understood or have yet to reach their full potential. Thus, the discussion regarding quantum wires and dots is limited to the more mature properties of these quantum structures. Future updates of the discussion will thus be necessary.

The topics covered in this textbook include an introduction to quantum mechanics, quantization of electronic energy levels in periodic potentials, tunneling, distribution functions and density of states, optical and electronic properties, growth issues, and devices. Figure 1 summarizes in a flowchart the major topics discussed. In a nutshell, the chapters are devoted to the introduction of quantum mechanics; calculations of the energy levels in periodic potentials, quantum wells, and quantum dots; derivation of the density of states in bulk materials and quantum wells, wires, and dots, and the density of states under the influence of electric or magnetic fields; growth of the bulk materials and quantum structures; optical properties; electrical and transport properties; electronic devices based on heterojunctions and nanostructures such as ohmic and Schottky contacts, diodes, resonant tunneling diodes, MODFETs, HFETs, Coulomb blockade, and single-electron transistors (known as SETs); and optoelectronic devices such as light-emitting transistors, light-emitting diodes, photodetectors based on quantum wells and quantum dots, edge-emitting lasers, VCSELs, quantum cascade lasers, and laser diodes based on quantum dots. End-of-chapter problems, appendices, tables, and references are included.

Students registering for courses based on this textbook should have a basic knowledge of semiconductor materials and devices. While knowledge in quantum mechanics is not required, it is however recommended that students have taken undergraduate physics courses, such as university physics and modern physics. The first chapter of this book covers the basic formalism of quantum mechanics needed for a student in electrical engineering to grasp the basic idea of how to calculate the

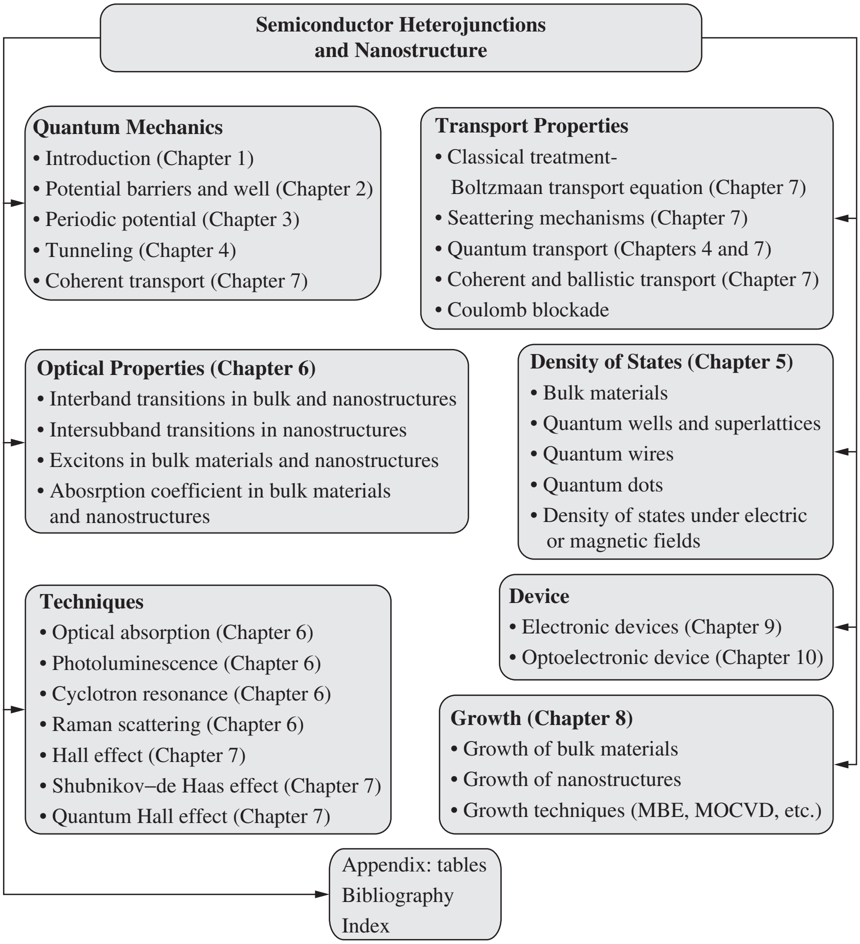

Semiconductor Heterojunctions and Nanostructure
Quantum Mechanics
• Introduction (Chapter 1)
• Potential barriers and well (Chapter 2)
• Periodic potential (Chapter 3)
• Tunneling (Chapter 4)
• Coherent transport (Chapter 7)
Transport Properties
• Classical treatment- Boltzmaan transport equation (Chapter 7)
• Seattering mechanisms (Chapter 7)
• Quantum transport (Chapters 4 and 7)
• Coherent and ballistic transport (Chapter 7)
• Coulomb blockade
Optical Properties (Chapter 6)
• Interband transitions in bulk and nanostructures
• Intersubband transitions in nanostructures
• Excitons in bulk materials and nanostructures
• Abosrption coefficient in bulk materials and nanostructures
Density of States (Chapter 5)
• Bulk materials
• Quantum wells and superlattices
• Quantum wires
• Quantum dots
• Density of states under electric or magnetic fields
Techniques
• Optical absorption (Chapter 6)
• Photoluminescence (Chapter 6)
• Cyclotron resonance (Chapter 6)
• Raman scattering (Chapter 6)
• Hall effect (Chapter 7)
• Shubnikov–de Haas effect (Chapter 7)
• Quantum Hall effect (Chapter 7)
Device
• Electronic devices (Chapter 9)
• Optoelectronic device (Chapter 10)
Growth (Chapter 8)
• Growth of bulk materials
• Growth of nanostructures
• Growth techniques (MBE, MOCVD, etc.)
Appendix: tables
Bibliography
Index

Figure 1 A schematic illustrating the major topics and their locations in the textbook.

energy levels in a simple quantum well. To understand and appreciate the beauty of quantum transport in quantum structure, the readers must have some knowledge in the classical type of transport. This led us to focus on both quantum transport, such as tunneling and coherent transport in mesoscopic systems, and classical transport, such as Boltzmann's transport equation and formalisms.

When an electronic or optoelectronic device is under the influence of an applied electric field and/or photonic excitation, the device is no longer at equilibrium and its transport properties become more complicated. The limits of various transport regimes, which are classified according to the electron phase coherent length and compared to the de Broglie wavelength, are discussed in Chap. 7. Various scattering mechanisms, which dominate the classical regime, are also discussed. When a nanostructure possesses a capacitance on the order of *attofarads*, a new transport phenomenon occurs, known as a Coulomb blockade. This phenomenon, in conjunction with the quantum tunneling effect, forms the basis for single-electron transistors. Discussion regarding this new class of devices is presented in Chaps. 4, 7, and 9. While these devices have the potential to revolutionize the current technology, it should be pointed out that this current technology is still based on carrier-injected and CMOS devices, where the transport is dominated by carrier scattering rather than by ballistic or coherent transports.

In addition to electronic devices, a new generation of optoelectronic devices is under intense research, including long-wavelength infrared detectors based on intersubband transitions, edge-emitting quantum well laser diodes, vertical-cavity surface-emitting lasers, and quantum cascade lasers. All of these devices are discussed in the textbook. Excitons play a major role in optoelectronic and photonic devices. Theoretically, the exciton binding energy in a quantum well is larger than that of excitons in the constituent bulk materials by a factor of 4. Furthermore, it is predicted that the excitons binding energies are even higher in quantum wires and dots. This can be translated to very fast optoelectronic devices that can operate at room temperature. The text presents a detailed discussion and derivation of the exciton binding energies in direct bandgap bulk semiconductors, quantum wells, and quantum dots.

Omar Manasreh
University of Arkansas

Acknowledgments

I would like to thank W. D. Brown, L. W. Schaper, G. J. Salamo, and R. E. Nowlin for proofreading and correcting several chapters. The edge-emitting laser diodes were provided by J. W. Tomm of Max-Born-Institut, and the p-type QWIP devices were provided by G. J. Brown. Many of the graphs reported in the text were provided by my students. Many thanks goes to my Sponsoring Editor Kenneth P. McCombs for his support and encouragement.

Copyright © 2005 by The McGraw-Hill Companies, Inc. Click here for terms of use.

List of Symbols and Abbreviations

0D	Zero dimensional
1D	One dimensional
2D	Two dimensional
2DEG	Two-dimensional electron gas
3D	Three dimensional
AFM	Atomic force microscopy
$\mathrm{Ai}(x)$	Airy function
$\mathbf{B}$	Magnetic field
$\mathrm{Bi}(x)$	Airy function
C	Capacitance
C_{gd}	Gate-drain capacitance
C_{gs}	Gate-source capacitance
C_v	Specific heat capacity
$\mathbf{D}$	Electric displacement vector
D^*	Detectivity
DBR	Distributed Bragg reflector
D_d	Dopant diffusion coefficient
D_e	Electron diffusion coefficient
D_h	Hole diffusion coefficient
D_s	Surface diffusion coefficient
E_c	Conduction band minimum
E_{des}	Desorption activation energy
E_{ex}	Exciton binding energy
E_F	Fermi energy level
E_g	Bandgap
E_{i^*}	Dissociation energy of a cluster made of i^* adatoms
E_n	Quantized energy levels
E_s	Surface diffusion activation energy
E_v	Valence band maximum
f_{BE}	Bose-Einstein distribution function
f_k	Fermi-Dirac distribution function at nonequilibrium

Copyright © 2005 by The McGraw-Hill Companies, Inc. Click here for terms of use.

f_k^o	Fermi-Dirac distribution function at equilibrium
f_{MB}	Maxwell-Boltzmann distribution function
f_n	Electron Fermi-Dirac distribution function
f_p	Hole Fermi-Dirac distribution function
f_T	Cutoff frequency
G	Conductance
$g^{0\mathrm{D}}(E)$	Density of states in quantum dots
$g^{1\mathrm{D}}(E)$	Density of states in quantum wires
$g^{2\mathrm{D}}(E)$	Density of states in quantum wells
HBT	Heterojunction bipolar transistor
$g^{3\mathrm{D}}(E)$	Density of states in bulk semiconductors
g_m	Transconductance
$g^s(E)$	Density of states in superlattices
h or $\hbar$	Planck's constant
$\mathbf{H}$	Hamiltonian
HEMT	High electron mobility transistor
HET	Hot electron transistor
HFET	Heterojunction field-effect transistor
HVPE	Hydride vapor-phase epitaxy
$i_{1/f}$	$1/f$ current noise
I_c	Collector current
I_D	Drain current
i_{gr}	Generation recombination current noise
i_J	Johnson current noise
i_n	Noise current
I_p	Pinch-off current
IR	Infrared
i_s	Shot current noise
J	Electric current density
J_{diff}	Diffusion current density
JFET	Junction field-effect transistor
J_t	Threshold current density
k_B	Boltzmann's constant
$\mathbf{k}$	Wavevector
k_e	Effective segregation coefficient
k_l	Thermal conductivity of the liquid
k_o	Segregation coefficient at equilibrium
k_s	Thermal conductivity of the solid

L	Angular momentum
LA	Longitudinal acoustical phonon mode
Laser	Light amplification by stimulated emission of radiation
L_d	Depletion length
l_e	Elastic collision mean free path
LEC	Liquid-encapsulated Czochralski
LED	Light-emitting diode
l_H	Landau magnetic length or cyclotron radius
l_i	Inelastic collision free path
LO	Longitudinal optical phonon mode
LPE	Liquid-phase epitaxy
l_T	Thermal diffusion length
l_ϕ	Coherence length
MBE	Molecular beam epitaxy
m_e^*	Electron effective mass
MESFET	Metal semiconductor field-effect transistor
m_h^*	Hole effective mass
MOCVD	Metal-organic chemical vapor deposition
MODFET	Modulation-doped field-effect transistor
MOSFET	Metal-oxide semiconductor field-effect transistor
N_d	Donor concentration
NEP	Noise-equivalent power
NETD	Noise-equivalent temperature difference
n_H	Hall concentration
n_i	Intrinsic carrier concentration
N_n	Density of neutral impurity
n_o	Electron density at equilibrium
n_r	Refractive index
N_th	Electron concentration emitted by thermionic mechanism
$\mathbf{P}$	Dipole moment
PL	Photoluminescence
PLD	Pulsed laser deposition
p_o	Hole density at equilibrium
R_abs	Absorption rate
R_c	Specific contact resistance
R_H	Hall coefficient

RHEED	Reflection high-energy electron diffraction
R_i	Current spectral responsivity
R_r or $\mathcal{R}$	Recombination rate
R_{st}	Stimulated emission rate
R_t	Tunnel resistance
R_v	Voltage spectral responsivity
S/N	Signal-to-noise ratio
S_n	Seebeck coefficient
STM	Scanning tunneling microscopy
T	Temperature
TA	Transverse acoustical phonon mode
TEM	Tunneling electron microscopy
THET	Tunneling hot electron transistor
TO	Transverse optical phonon mode
U	Internal energy
UV	Ultraviolet
$V(x)$	Potential energy in one dimension
V_b	Bias voltage
V_{BE}	Base-emitter voltage
V_{bi}	Built-in potential
V_{CE}	Collector-emitter voltage
VCSEL	Vertical cavity surface-emitting laser
V_D	Drain voltage
V_G	Gate voltage
V_H	Hall voltage
V_P	Pinch-off voltage
V_T	Threshold voltage
W	Depletion width
W_B	Base width
WKB	Wentzel-Kramers-Brillouin approximation
X-TEM	Cross-sectional tunneling electron microscopy
ΔG	Gibbs free energy
Δn	Excess electron concentration
Δp	Excess hole concentration
$\mathcal{E}$	Electric field
Φ	Impingement flux
$\Gamma(x)$	Gamma function
Γ_n	Electron generation rate

Γ_p	Hole generation rate
$\alpha(\omega)$	Optical absorption coefficient
α_{ex}	Optical absorption of exciton in bulk semiconductor
α_i	Optical loss due to the laser active region
α_m	Optical loss due to the mirrors or facets
χ	Dielectric susceptibility
δ	Delta function
ϵ_∞	High-frequency dielectric constant
ϵ_o	Permittivity of free space
ϵ_r	Permittivity of a medium
ϵ	Dielectric constant
ϕ_m	Work function for metals
γ_l	Light generation rate
γ_t	Thermal generation rate
η_d	External differential quantum efficiency
η_{ex}	External quantum efficiency
η_{in}	Internal quantum efficiency
η_{inj}	Injection efficiency
η_{pi}	Population inversion efficiency
η_r	Radiative efficiency
λ	Wavelength, de Broglie wavelength
μ_H	Hall mobility
μ_n	Electron mobility
μ_o	Permeability of free space
μ_p	Hole mobility
ρ	Electrical resistivity
σ	Electrical conductivity
τ_n	Electron relaxation time
τ_p	Hole relaxation time
τ_r	Recombination lifetime
υ_d	Drift velocity
υ_g	Group velocity
υ_p	Phase velocity
υ_s	Carrier saturation velocity
ω	Angular frequency
ω_c	Cyclotron frequency
ω_p	Plasmon frequency

Chapter
1

Introduction to Quantum Mechanics

1.1 Introduction

Despite the many successful applications of classical mechanics (based on Newton's famous laws of motion) to a wide range of physical phenomena, it was apparent at the beginning of the last century that many of the known phenomena could not be explained in terms of the concepts of classical mechanics. To meet the challenge of these classical inexplicable observations, a complete new theory, *quantum mechanics*, was developed. The basic underlying assumptions of quantum theory are quite different from those of classical mechanics, and they constitute a fundamentally different way of looking at nature.

Quantum mechanics provides precise answers to many problems, but it tells only the average value of many individual measurements made on a given dynamical system in a certain initial state. One of the fundamental differences between classical mechanics and quantum theory is that in quantum mechanics it is not possible to measure all variables with specific accuracy at the same time, while in classical mechanics it is. Another difference is that classically, the effects of the disturbances due to the measurements can be exactly allowed in predicting the future behavior of the system, whereas quantum mechanically the exact effects of the disturbances accompanying any measurements are inherently unknown. For example, in quantum mechanics the measurement of the position of a particle introduces an unpredictable uncertainty regarding its momentum.

Early examples of observations that required a revision of classical mechanics are numerous. We will discuss a few of them here.

Copyright © 2005 by The McGraw-Hill Companies, Inc. Click here for terms of use.

1.1.1 Blackbody radiation

The attempt of Max Planck to explain blackbody radiation from incandescent hot bodies was actually the first step in developing quantum mechanics. In 1901, Planck described the spectral intensity of blackbody radiation by assuming that the oscillators in equilibrium with radiation can have certain discrete energy E_n, given by

$$E_n = n\hbar\omega_o \qquad \text{for } n = 0, 1, 2, 3, \ldots \tag{1.1}$$

where ω_o is the oscillator frequency and $\hbar$ is Planck's constant. The basic assumption of Planck's work is that for a cavity radiator, the number of internal degrees of freedom (standing waves) can be calculated at a given frequency range per unit volume of the cavity to be $2 \times 4\pi\nu^2/c^3$, where ν is the frequency, c is the speed of light, and 2 is added to account for the fact that each electromagnetic wave can have two orthogonal polarizations. However, none of these standing waves in the cavity can take on all possible energies as Maxwell's equations imply, but can take on only certain integrally related discrete energies, $0, \hbar\omega_o, 2\hbar\omega_o, 3\hbar\omega_o, \ldots$ as shown in Eq. (1.1). Furthermore, it is assumed that the probability that a standing wave has one of these energies associated with it is given by the normal Boltzmann statistical distribution function. With these assumptions, the mean energy of the oscillator can be written as

$$\overline{E} = \frac{\sum_n n\hbar\omega_o e^{(-n\hbar\omega_o/k_BT)}}{\sum_n e^{(-n\hbar\omega_o/k_BT)}} = k_BT\left[\frac{\hbar\omega_o/k_BT}{e^{(\hbar\omega_o/k_BT)} - 1}\right] \tag{1.2}$$

Equation (1.2) differs by the factor in the brackets from the classical calculation of the energy density derived by Rayleigh and Jeans in 1900, which was given by $E = 2\omega_o^2 k_B T/(\pi c^3)$. The final Planck expression of the energy flux W can be written as

$$W = \frac{\hbar\omega_o^3}{2\pi c^2}\frac{1}{e^{(\hbar\omega_o/k_BT)} - 1} \tag{1.3}$$

A plot of both the Rayleigh-Jeans and Planck expressions is shown in Fig. 1.1. The quantity h or $\hbar = h/2\pi$ is known as Planck's constant, which was used as a parameter to fit Eq. (1.3) to the experimental curves of the blackbody radiation. From the fitting procedures, it was determined that $\hbar = 1.0546 \times 10^{-34}$ J·s.

1.1.2 The specific heat capacity of solids

The classical result of the specific heat of solids was derived by Dulong and Petit by assuming that the atoms in the solid crystal are simply

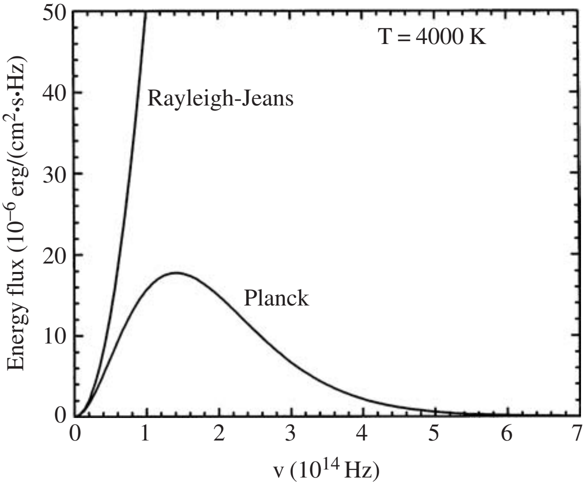


Figure 1.1 Radiation laws for blackbody at $T = 4000$ K ($\omega_o = 2\pi\nu$).

harmonic oscillators with a Maxwell-Boltzmann energy distribution as predicted by the statistical theory. The total vibrational internal energy per oscillator (U) for a system of $3N$ oscillators can be written as

$$U = \frac{\overline{E}}{3N} = \frac{\int_0^\infty E e^{-E/k_B T}\, dE}{\int_0^\infty e^{-E/k_B T}\, dE} = k_B T \tag{1.4}$$

which gives $\overline{E} = 3Nk_B T$. The specific heat capacity C_ν then can be obtained as follows:

$$C_\nu = \left(\frac{\partial \overline{E}}{\partial T}\right)_\nu = 3Nk_B \tag{1.5}$$

This result is in disagreement with the experimental measurements of the specific heat capacity at low temperature. In 1911, Albert Einstein presented a new model for the specific heat capacity of solids based on the assumption that the energies of the harmonic oscillators (atoms in a solid) are restricted to the discrete values given by quantum theory as

$$E_n = \left(n + \frac{1}{2}\right)\hbar\omega \qquad n = 0, 1, 2, 3, \ldots \tag{1.6}$$

The final results of Einstein's model are as follows:

$$E = \frac{3N\hbar\omega}{2} + \frac{3N\hbar\omega}{e^{\hbar\omega/k_BT} - 1}$$

$$C_\nu = 3Nk_B\left(\frac{\hbar\omega}{k_BT}\right)^2 \frac{e^{\hbar\omega/k_BT}}{(e^{\hbar\omega/k_BT} - 1)^2} \tag{1.7}$$

By defining $\Theta_E = \hbar\omega/k_B$, known as the Einstein temperature, the expression for the specific heat of a solid can be rewritten as

$$C_\nu = 3Nk_B\left(\frac{\hbar\omega}{k_BT}\right)^2 \frac{e^{\Theta_E/T}}{(e^{\Theta_E/T} - 1)^2} \tag{1.8}$$

Finally, Debye regarded the atoms of the crystal as harmonic oscillators coupled together by Hooke's law of interatomic forces, which generate acoustic waves that propagate over a range of frequencies from zero to a maximum value given by the dispersion relation. The Debye result for the specific heat capacity can be written as

$$C_\nu = 9Nk\left(\frac{T}{\Theta_D}\right)^3 \int_0^{\Theta_D/T} \frac{x^4e^x}{(e^x - 1)^2}\,dx \tag{1.9}$$

where $\Theta_D = \hbar\omega/k_B$ is defined as the *Debye temperature*. Equation (1.9) cannot be plotted analytically since the integral is very difficult to evaluate. However, for low temperatures, Θ_D/T approaches infinity and the integral can then be evaluated to be $4\pi^4/15$. In the low temperature limits, Eq. (1.9) reduces to $C_\nu = \frac{12\pi^4Nk_B}{5}(\frac{T}{\Theta_D})^3$. The specific heat capacity results obtained from the three approaches discussed are shown in Fig. 1.2.

1.1.3 Photoelectric effect

In 1887, Hertz, while conducting experiments on the generation of electromagnetism, discovered that electrons could be ejected from solids by letting radiation fall onto the solid. Lenard and others found that the maximum energy of these photoejected electrons depended only on the frequency of the light falling on the surface and not on its intensity. It was also found that for shorter wavelengths, the maximum energy of the electrons was greater than for longer wavelengths. In 1905, Einstein explained the photoelectric effect in a satisfactory way by making use of Planck's ideas. He assumed that radiation exists in the form of quanta of definite size; that is, light consists of packets of energy of size $\hbar\omega$. He also assumed that when light falls on a surface, individual electrons in a solid can absorb these energy quanta. Therefore, the energy received

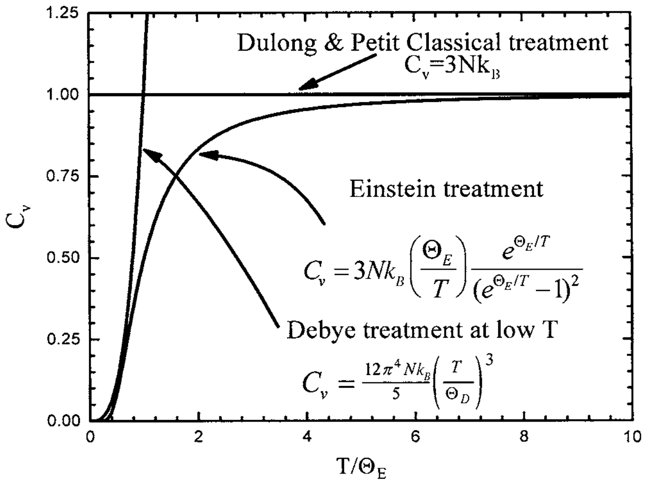


Figure 1.2 The specific heat capacity of a single crystal plotted as a function of temperature. The three curves represent the Dulong and Petit, Einstein, and Debye models.

by an electron depends only on the frequency of the light and is independent of its intensity. The intensity merely determines how many photoelectrons will leave the surface per second. Thus, the maximum kinetic energy of an electron excited by such light can be expressed as

$$E_m = \hbar(\omega - \omega_o) = \hbar\omega - q\varphi_o \tag{1.10}$$

where $\hbar\omega_o = q\varphi_o$ or $\omega_o = q\varphi_o/\hbar$, which is known as the threshold frequency, and q is the charge of the electron. Above this threshold frequency, the light quanta has more than enough energy to excite the electrons into the vacuum. The quantity φ_o is a characteristic property of the metal called the work function. The electron must obtain energy $q\varphi_o$ from the incident light to be emitted as a photoelectron. Einstein's analysis of the photoemission phenomenon assumes that it can be considered as a two-body collision in which the light is giving up all its energy to a single electron.

Other early experiments that were found difficult to explain in terms of classical mechanics were Compton scattering, electron diffraction from solid crystals, and emission and absorption spectra of atoms and molecules. The failure of classical mechanics was associated with two general types of effects. The first one is that physical quantities such as the energies of the electromagnetic waves and of lattice vibrations of a given frequency, or the energies and angular momenta associated with electronic orbits in the hydrogen atoms, which in classical theory can take on a continuous range of values, were found to take on discrete values instead. The second type of effect is called wave-particle duality, where both the wave nature of light as shown by diffraction and interference effects, and the particle nature of light, as shown by the photoelectric and Compton effects, are exhibited. Particle parameters (the energy E and momentum $\mathbf{p}$ of a photon) and wave parameters (angular frequency $\omega = 2\pi\nu$ and wave vector $\mathbf{k}$, where $\mathbf{k} = 2\pi/\lambda$, $\nu =$ frequency, and $\lambda =$ wavelength) are linked by the fundamental relations $E = \hbar\omega$ and $\mathbf{p} = \hbar\mathbf{k}$, known as the Planck-Einstein relations. During each elementary process, total energy and momentum must be conserved.

1.1.4 The Bohr model of the atom

Classical mechanics failed to explain the sharply defined spectral lines observed in the optical emission spectra of the elements. In 1913, Neils Bohr found a way of quantizing the hydrogen atom that described the spectrum of the element with impressive accuracy. This delay was partly due to the fact that the atomic nucleus was not discovered until 1910 when Rutherford's scattering experiments were performed. It was only then that the concept of an atom as a point nucleus surrounded by a swarm of electrons emerged. In Bohr's model, a single electron of mass m and charge q is assumed to move in a circular orbit around the nucleus with a positive charge of qz, where z is an integer. In classical electrodynamics, accelerated charges like the orbiting electron always radiate energy in the form of electromagnetic waves. Classically, one would expect the electron to continually lose energy, spiraling inward toward the nucleus as its energy is depleted by radiation. Bohr suggested that stable nonradiative states of the atom can exist, corresponding to circular electron orbits whose angular momentum L is quantized in integral multiples of $\hbar$ so that

$$L_n = mr_n^2\omega_n = mr_n v = n\hbar \qquad (n = 0, 1, 2, 3, \ldots) \tag{1.11}$$

where v is the electron velocity in its orbit and r_n is the orbit radius. This quantization of the angular momentum also quantizes the orbit radii and angular velocities as indicated in Eq. (1.11). The allowed total energies, kinetic plus potential, can be written as

$$E_n = \frac{1}{2}mr_n^2\omega_n^2 - \frac{zq^2}{4\pi\epsilon_0 r_n} \tag{1.12}$$

where ϵ_0 is the permittivity of free space.

By equating the Coulomb force to the centripetal force, one can write

$$mr_n\omega_n^2 = \frac{zq^2}{4\pi\epsilon_0 r_n^2} \tag{1.13}$$

where $r_n\omega_n^2$ is the centripetal acceleration. From Eqs. (1.11) to (1.13), one can obtain r_n, ω_n, and E_n according to the following relationships:

$$r_n = \frac{4\pi\epsilon_0 n^2\hbar^2}{mzq^2} \tag{1.14}$$

$$\omega_n = \frac{mz^2q^4}{(4\pi\epsilon_0)^2 n^3\hbar^3} \tag{1.15}$$

$$E_n = -\frac{mz^2q^4}{2(4\pi\epsilon_0)^2 n^2\hbar^2} \tag{1.16}$$

By introducing the dimensionless unit α, known as the "fine structure constant,"

$$\alpha = \frac{q^2}{(4\pi\epsilon_0)\hbar c} = \frac{1}{137.036} \approx \frac{1}{137} \tag{1.17}$$

the quantities in Eqs. (1.14) to (1.16) can be rewritten in much simpler expressions such as

$$r_n = \frac{n^2\hbar}{mzc\alpha} \qquad \omega_n = \frac{m(zc\alpha)^2}{n^3\hbar} \qquad \text{and} \qquad E_n = -\frac{m(z\alpha c)^2}{2n^2} \tag{1.18}$$

The allowed energies are negative, corresponding to stable bound states of the electron. For $z = 1$ (hydrogen atom) and $n = 1$, one can find that $r_1 = 0.5293$ Å and $E_1 = -13.62$ eV. Thus, one can rewrite the orbit radii and the energy as $r_n = n^2 r_1$ and $E_n = -E_1/n^2$. The energy levels of the hydrogen atom are illustrated in Fig. 1.3.

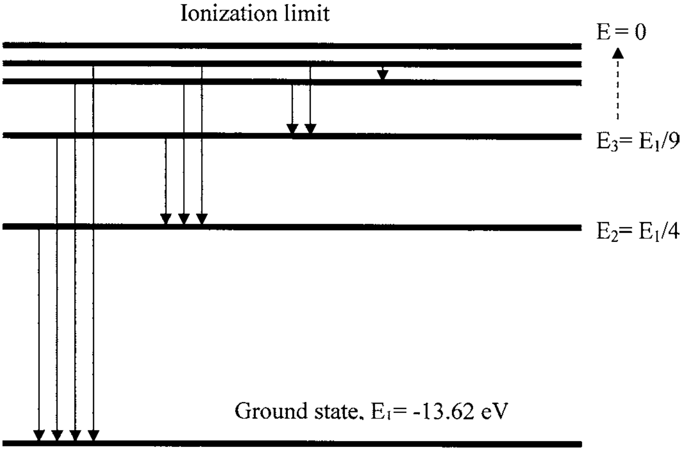


Figure 1.3 Energy level diagram for the hydrogen atom derived from Bohr's model.

Bohr's model was significant in the development of quantum theory because it showed the potential usefulness of its concepts in describing the structure of atoms and molecules. However, attempts to extend Bohr's model to helium and more complex atoms were not very successful. These problems were not fully solved until after 1930, and in order to work them out, a completely new and much more general theory of quantum mechanics had to be developed. Still, Bohr's atomic system provided a simple picture of the structure of the one-electron atomic system.

1.2 The de Broglie Relation

Classical mechanics failed to explain the narrow lines that composed the atomic emission and absorption spectra. In other words, a given atom emits or absorbs only photons having well-defined frequencies. This can be easily understood if one accepts the fact that the energy of the atom is quantized. The emission or absorption of a photon is then accompanied by a jump in the energy of the atom from one permitted value (E_i) to another (E_f). Conservation of energy implies that the photon has a frequency ν_{if} such that $h\nu_{if} = |E_i - E_f|$. In 1923, de Broglie presented the following hypothesis: *Material particles, just like photons, can have a wavelike aspect*. He then derived the Bohr-Sommerfeld quantization rules as a consequence of this hypothesis. The various permitted energy levels appeared analogous to the normal modes of a vibrating string. The electron diffraction experiment by Davisson and Germer in 1927 strikingly confirmed the existence of the wavelike aspect of matter by showing that interference patterns could be obtained with material particles such as electrons. One therefore associates with a material particle of an energy E and momentum $\mathbf{p}$, a wave whose angular frequency is $\omega = 2\pi\nu$ and a wave vector $\mathbf{k}$ given by the same relations presented by the Planck-Einstein relations ($E = \hbar\omega$ and $\mathbf{p} = \hbar\mathbf{k}$). The corresponding wavelength is

$$\lambda = \frac{2\pi}{|\mathbf{k}|} = \frac{h}{|\mathbf{p}|} \tag{1.19}$$

The small value of h explains why the wavelike nature of matter is very difficult to demonstrate on a macroscopic scale.

Example Consider a dust particle of diameter $r = 1\ \mu$m and mass $m = 10^{-15}$ kg. For such a particle of small mass and a speed of $v = 10^{-3}$ m/s, the de Broglie wavelength is $\lambda = 6.6 \times 10^{-34}/(10^{-15} \times 10^{-3}) = 6.6 \times 10^{-16}$ m $= 6.6 \times 10^{-6}$ Å. This wavelength is negligible on the scale of the dust particle. Let us now consider a thermal neutron ($m_n \approx 1.67 \times 10^{-27}$ kg) of energy $1.5k_BT$. Hence, $\frac{1}{2}m_nv^2 = p^2/(2m_n)$, where $k_B = 1.38 \times 10^{-23}$ J/K. This gives

$\lambda = h/(3m_n k_B T)^{0.5} \approx 1.4$ Å at $T = 300$ K. This wavelength is on the order of the lattice constant in crystalline solid. A beam of thermal neutrons falling on a crystal therefore gives rise to diffraction phenomena analogous to those observed using x rays.

Example Let us now examine the de Broglie wavelengths associated with electrons ($m_e \approx 0.9 \times 10^{-30}$ kg). If the electron is accelerated by a potential difference V, then the electron kinetic energy is $E = qV = 1.6 \times 10^{-19} V$ joules. Since $E = p^2/(2m_e)$, then $\lambda = h/(2m_e E)^{0.5} = 12.3/(V)^{0.5}$ Å. With a potential difference of several hundred volts, one can obtain a wavelength comparable to those of x rays. Thus, electron diffraction phenomena can be observed in crystals or crystalline powders.

1.3 Wave Functions and the Schrödinger Equation

By considering the de Broglie hypothesis, one can apply the wave properties for the case of photons to all material particles. Thus, for the classical concept of a trajectory, the time-varying state is substituted by the quantum state characterized by a *wave function* $\psi(\mathbf{r}, t)$, which contains all the information (in terms of space $\mathbf{r}$ and time t) that is possible to obtain about the particle. The wave function $\psi(\mathbf{r}, t)$ can be thought of as a *probability amplitude* of the particle's presence. The measurements of an arbitrary physical quantity must belong to a set of eigenvalues. Each eigenvalue is associated with an eigenstate. The equation describing the evolution of the wave function $\psi(\mathbf{r}, t)$ remains to be written. The wave equation can be introduced by using the Planck and de Broglie relations to yield the fundamental equation known as the Schrödinger equation. The form of this equation for a particle of mass m, which is subject to the influence of a potential $V(\mathbf{r}, t)$, takes the following form:

$$i\hbar \frac{\partial}{\partial t}\psi(\mathbf{r}, t) = -\frac{\hbar^2}{2m}\Delta\psi(\mathbf{r}, t) + V(\mathbf{r}, t)\psi(\mathbf{r}, t) \tag{1.20}$$

where Δ is the Laplacian operator given by $\Delta = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$. This equation is linear and homogeneous in $\psi(\mathbf{r}, t)$. Consequently, for material particles, there exists a superposition principle.

When $V(\mathbf{r}, t) = 0$, the Schrödinger equation is reduced to

$$i\hbar \frac{\partial}{\partial t}\psi(\mathbf{r}, t) = -\frac{\hbar^2}{2m}\Delta\psi(\mathbf{r}, t) \tag{1.21}$$

which is the wave equation for a free particle. A solution of this equation has the form

$$\psi(\mathbf{r}, t) = Ae^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)} \tag{1.22}$$

where A is a constant and the dispersion relation obtained by substituting Eq. (1.22) into (1.21) is

$$\omega = \frac{\hbar k^2}{2m} \tag{1.23}$$

Equations (1.19) and (1.23) give the relation between energy E and momentum $\mathbf{p}$ as $E = p^2/(2m)$, where $\mathbf{p} = \hbar\mathbf{k}$.

The constant A in Eq. (1.22) can be obtained by normalization. Using the wave function form given by Eq. (1.22), one can write

$$|\psi(\mathbf{r},\, t)\psi^*(\mathbf{r},\, t)| = |A|^2 \qquad \text{where } \psi^*(\mathbf{r},\, t) = A^* e^{-i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \tag{1.24}$$

This equation tells us that a plane wave of this type represents a particle whose probability of presence is uniform throughout all space.

The principle of superposition tells us that every linear combination of plane waves satisfying the dispersion relation given by Eq. (1.23) will also be a solution of Eq. (1.21). This superposition can be written as

$$\psi(\mathbf{r},\, t) = \frac{1}{(2\pi)^{3/2}} \int g(\mathbf{k}) e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}\, d^3\mathbf{k} \tag{1.25}$$

where $d^3\mathbf{k}$ is the infinitesimal volume element in $\mathbf{k}$-space and $g(\mathbf{k})$ can be complex but must be sufficiently regular to allow differentiation inside the integral. A wave function such as Eq. (1.25) is called a three-dimensional *wave packet*.

1.4 Wave Packet at a Given Time

In this section, we follow Cohen-Tannoudji formalism, but the subject of packets has been discussed in almost every quantum mechanics textbook. The reason for discussing the wave packet is to demonstrate the duality concepts and show how the uncertainty principle is obtained. For the sake of simplicity, we will discuss the case of a one-dimensional wave packet where the wave function depends only on x and t as

$$\psi(x,\, t) = \frac{1}{\sqrt{2\pi}} \int g(\mathbf{k}) e^{i(kx-\omega t)}\, d\mathbf{k} \tag{1.26}$$

For $t = 0$ we have

$$\psi(x, 0) = \frac{1}{\sqrt{2\pi}} \int g(\mathbf{k}) e^{ikx}\, d\mathbf{k} \tag{1.27a}$$

The Fourier transform of this equation can be obtained as

$$g(\mathbf{k}) = \frac{1}{\sqrt{2\pi}} \int \psi(x, 0) e^{-ikx}\, dx \tag{1.27b}$$

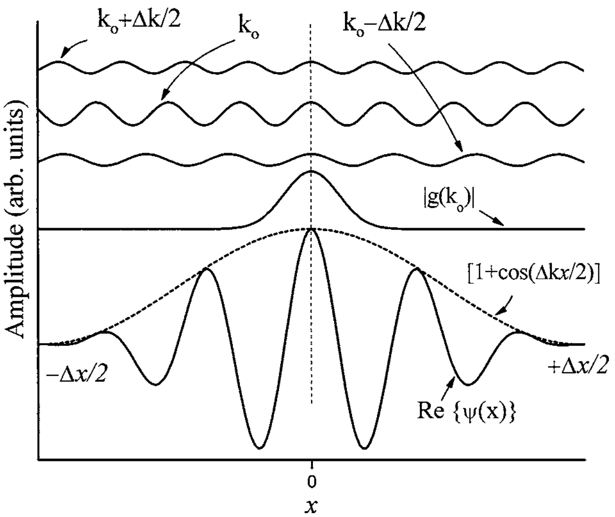


Figure 1.4 The shape of the function $|g(\mathbf{k})|$ is plotted along with the real parts of the three functions whose sum gives the function $\psi(x)$ of Eq. (1.28). The real part of $\psi(x)$ is also shown. The dashed-line curve corresponds to the function $[1 + \cos(\Delta\mathbf{k}x/2)]$ which represents the form of the wave packet.

Thus, $g(\mathbf{k})$ is the Fourier transform of $\psi(\mathbf{r}, 0)$. The wave packet is given by the x-dependent wave function expressed in Eq. (1.27a). If $|g(\mathbf{k})|$ has the shape depicted in Fig. 1.4 and $\psi(x)$, instead of having the form shown in Eq. (1.27a), is composed of three plane waves with wave vectors of $\mathbf{k}_o$, $\mathbf{k}_o + \Delta\mathbf{k}/2$ and $\mathbf{k}_o - \Delta\mathbf{k}/2$ and amplitudes proportional to 1, $\frac{1}{2}$, and $\frac{1}{2}$, respectively, then one can write the new wave packet as

$$\psi(x) = \frac{g(\mathbf{k}_o)}{\sqrt{2\pi}}\left[e^{i\mathbf{k}_o x} + \frac{1}{2}e^{i(\mathbf{k}_o-\Delta\mathbf{k}/2)x} + \frac{1}{2}e^{i(\mathbf{k}_o+\Delta\mathbf{k}/2)x}\right]$$
$$= \frac{g(\mathbf{k}_o)}{\sqrt{2\pi}}e^{i\mathbf{k}_o x}\left[1 + \cos\left(\frac{\Delta\mathbf{k}}{2}x\right)\right] \tag{1.28}$$

From Fig. 1.4, $|\psi(x)|$ is maximum at $x = 0$. This result is due to the fact that when x is zero, the three waves are in phase and interfere constructively as shown in the figure. As one moves away from the value $x = 0$, the waves become more and more out of phase and $|\psi(x)|$ decreases. The interference becomes completely destructive when the phase shift between $e^{i\mathbf{k}_o x}$ and $e^{i(\mathbf{k}_o\pm\Delta\mathbf{k}/2)x}$ is equal to $\pm\pi$ and $|\psi(x)| = 0$ when $x = \pm\,\Delta x/2$. In other words, $|\psi(x)| = 0$ when $\cos(\Delta\mathbf{k}\ \Delta x/4) = -1$. This leads to the following equation:

$$\Delta\mathbf{k}\ \Delta x = 4\pi \tag{1.29}$$

This equation shows that the larger the width of $|\psi(x)|$, the smaller the width of $g(\mathbf{k})$. Equation (1.28), however, shows that $|\psi(x)|$ is periodic in x and therefore has a series of maxima and minima. This arises from the fact that $\psi(x)$ is the superposition of a finite number of waves. For a continuous superposition of an infinite number of waves as shown in Eq. (1.27*a*), such a phenomenon does not exist and $|\psi(x, 0)|$ can have only one maxima.

Let us return to the general wave packet formula shown in Eq. (1.27*a*). Its form results from an interference phenomenon. Let $\alpha(\mathbf{k})$ be the argument of the function $g(\mathbf{k})$, which yields

$$g(\mathbf{k}) = |g(\mathbf{k})|e^{i\alpha(\mathbf{k})} \tag{1.30}$$

Assume that $\alpha(\mathbf{k})$ varies sufficiently smoothly within the interval $[\mathbf{k}_o - \Delta\mathbf{k}/2, \mathbf{k}_o + \Delta\mathbf{k}/2]$, where $|g(\mathbf{k})|$ is appreciable. Hence, when $\Delta\mathbf{k}$ is small enough, one can expand $\alpha(\mathbf{k})$ around $\mathbf{k} \approx \mathbf{k}_o$ such that $\alpha(\mathbf{k}) \approx \alpha(\mathbf{k}_o) + (\mathbf{k} - \mathbf{k}_o) d\alpha/d\mathbf{k}|_{\mathbf{k}=\mathbf{k}_o}$, which enables us to rewrite Eq. (1.27*a*) in the following form:

$$\psi(x, 0) \approx \frac{e^{i[\mathbf{k}_o x + \alpha(\mathbf{k}_o)]}}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} |g(\mathbf{k})|e^{i(\mathbf{k}-\mathbf{k}_o)(x-x_o)}\, d\mathbf{k} \tag{1.31}$$

where $x_o = -[d\alpha/d\mathbf{k}]_{\mathbf{k}=\mathbf{k}_o}$. Equation (1.31) is very useful for studying the variation of $|\psi(x)|$ in terms of x. When $|x - x_o|$ is large as compared to $1/(\Delta\mathbf{k})$, the wave function oscillates rapidly within the interval $\Delta\mathbf{k}$ as shown in Fig. 1.5. On the other hand, when $|x - x_o|$ is small as compared to $1/(\Delta\mathbf{k})$, the wave function oscillates only once as shown in the figure. Thus, when x moves away from x_o, $|\psi(x)|$ decreases. The decrease becomes appreciable if $e^{i(\mathbf{k}-\mathbf{k}_o)(x-x_o)}$ oscillates approximately once. That is when

$$\Delta\mathbf{k}\,(x - x_o) \approx 1 \tag{1.32}$$

If Δx is the approximate width of the wave packet, then one can write

$$\Delta\mathbf{k}\ \Delta x \geq 1 \tag{1.33}$$

This classical relation relates the widths of two functions that are Fourier transforms of each other. The important fact is that the product $\Delta\mathbf{k}\ \Delta x$ has a lower bound that depends on the precise definition of each width. With the help of the relation $\Delta\mathbf{p} = \hbar\ \Delta\mathbf{k}$, Eq. (1.33) can be rewritten as

$$\Delta p\ \Delta x \geq \hbar \tag{1.34}$$

This relationship is called Heisenberg's uncertainty principle.

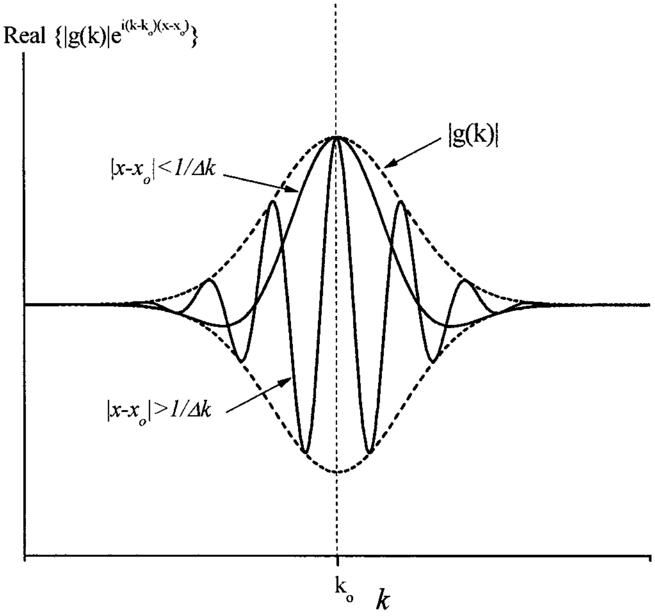


Figure 1.5 Variation of the wave function versus **k**. When $|x - x_o| > 1/\Delta\mathbf{k}$, we see several oscillations, but when $|x - x_o| < 1/\Delta\mathbf{k}$, we see only one oscillation.

The same procedure can be repeated by assuming

$$\psi(\mathbf{r}, t) = Ae^{i\omega t} \tag{1.35}$$

to obtain a wave packet that is localized in time and frequency with their widths being related by

$$\Delta\omega\ \Delta t \approx 1 \tag{1.36}$$

With the aid of the relation $\Delta E = \hbar\,\Delta\omega$, the uncertainty principle becomes

$$\Delta E\ \Delta t \geq \hbar \tag{1.37}$$

The inequalities shown in Eqs. (1.34) and (1.37) are introduced to show that $\hbar$ is the lower limit. It is possible that one can construct wave packets for which the products of the quantities in these equations are larger than $\hbar$.

Let us consider the time evolution of the wave packet where it is given in one dimension by Eq. (1.26). If ω has a simple dependence on **k** (nondispersive media), that is, $\omega = v_p k$, where v_p is known as the phase velocity, then Eq. (1.26) becomes

$$\psi(x,\,t)=\frac{1}{\sqrt{2\pi}}\int g(\mathbf{k}-\mathbf{k}_o)e^{i\mathbf{k}(x-v_p t)}\,d\mathbf{k} \tag{1.38}$$

This means that the wave packet moves with its center at $x = v_p t$ and its shape is unchanged with time. If we have a dispersive media and $\omega(\mathbf{k})$ is more complex, we can use Taylor's expansion to write

$$\omega(\mathbf{k})=\omega(\mathbf{k}_o)+\frac{\partial\omega}{\partial\mathbf{k}}|_{\mathbf{k}=\mathbf{k}_o}(\mathbf{k}-\mathbf{k}_o)+\frac{1}{2}\frac{\partial^2\omega}{\partial\mathbf{k}^2}|_{\mathbf{k}=\mathbf{k}_o}(\mathbf{k}-\mathbf{k}_o)^2+\cdots \tag{1.39}$$

Assuming $\omega(\mathbf{k}_o)=\omega_o$, $\frac{\partial\omega}{\partial\mathbf{k}}|_{\mathbf{k}=\mathbf{k}_o}=v_g$, and $\frac{\partial^2\omega}{\partial\mathbf{k}^2}|_{\mathbf{k}=\mathbf{k}_o}=\alpha$, the wave packet takes the following form:

$$\psi(x,\,t)=\frac{e^{i(\mathbf{k}_o x-\omega_o t)}}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} g(\mathbf{k}-\mathbf{k}_o)e^{i[(\mathbf{k}-\mathbf{k}_o)(x-v_g t)-\frac{\alpha}{2}(\mathbf{k}-\mathbf{k}_o)^2 t]}\,d\mathbf{k} \tag{1.40}$$

If $\alpha = 0$, the wave packet would move with its peak centered at $x = v_g t$, where $v_g=\frac{\partial\omega}{\partial\mathbf{k}}|_{\mathbf{k}=\mathbf{k}_o}$ is known as the group velocity. If α is not zero, then the wave packet will change. Let us assume that $g(\mathbf{k}-\mathbf{k}_o)$ has a Gaussian form

$$g(\mathbf{k}-\mathbf{k}_o)=e^{-(\mathbf{k}-\mathbf{k}_o)^2/(2\Delta\mathbf{k}^2)} \tag{1.41}$$

then the wave packet becomes

$$\psi(x,\,t)=\frac{e^{i(\mathbf{k}_o x-\omega_o)t}}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{i[(\mathbf{k}-\mathbf{k}_o)(x-v_g t)-\frac{1}{2}(\mathbf{k}-\mathbf{k}_o)^2(\alpha t-i/\Delta\mathbf{k}^2)]}\,d\mathbf{k} \tag{1.42}$$

Thus, the probability $|\psi(\mathbf{r},\,t)|^2$ depends on space and time and can be expressed in the following form:

$$|\psi(x,\,t)|^2=\frac{B}{2\pi}\,\exp\left[-\frac{(\Delta\mathbf{k})^2(x-v_g t)^2}{1+\alpha^2 t^2(\Delta\mathbf{k})^4}\right] \tag{1.43}$$

where B is a constant that may depend on time. For simplicity, we assume that B is time-independent. This is a Gaussian distribution centered at $x = v_g t$, and the mean width in real space is given by

$$\delta x(t)=\delta x(0)\sqrt{1+\frac{\alpha^2 t^2}{\delta x(0)^2}} \tag{1.44}$$

where $\delta x(0) = 1/\Delta\mathbf{k}$. For short times such as $\alpha^2 t^2(\Delta\mathbf{k})^4 \ll 1$, the width does not change appreciably from its starting value, but as time passes,

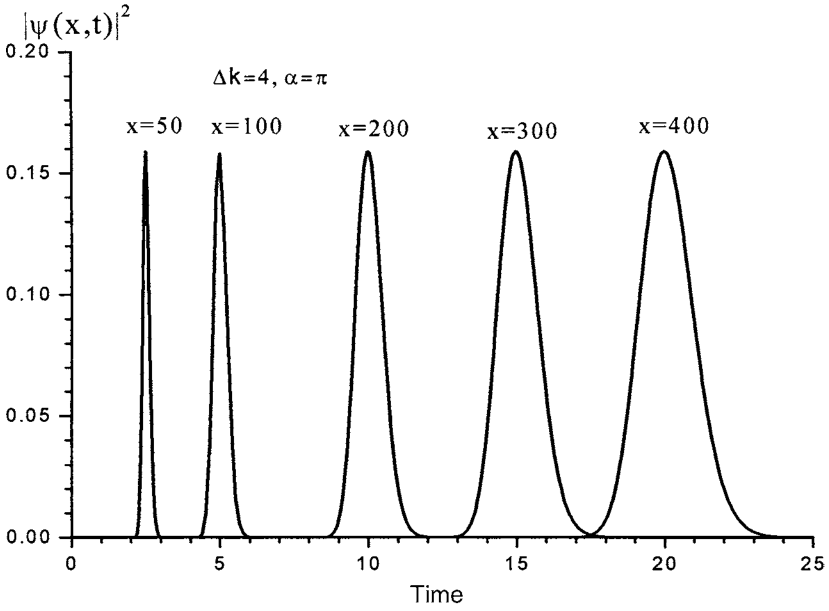


Figure 1.6 Plot of $|\psi(x, t)|^2$, given by Eq. (1.43), as a function of time for different values of x, where $\Delta\mathbf{k} = 4$ and $\alpha = 0.314$. It is clear that this probability spreads in time and space.

the wave packet will start spreading for $\alpha \neq 0$. As an example, we plotted Eq. (1.43) as a function of time for different values of x, as shown in Fig. 1.6. It is clear that this function is spread out as x increases for a fixed value of α. The probability function presented in Fig. 1.6 spreads even faster for larger values of α and becomes localized (unchanged) for $\alpha = 0$.

1.5 Separation of Variables

The wave function of a particle whose potential energy is time-independent must satisfy the Schrödinger equation [Eq. (1.20)]. Consider the following wave function:

$$\psi(\mathbf{r}, t) = \varphi(\mathbf{r})\chi(t) \tag{1.45}$$

Substituting Eq. (1.45) into Eq. (1.20), we obtain

$$i\hbar\varphi(\mathbf{r})\frac{\partial}{\partial t}\chi(t) = \chi(t)\left[-\frac{\hbar^2}{2m}\Delta\varphi(\mathbf{r})\right] + V(\mathbf{r}, t)\varphi(\mathbf{r})\chi(t) \tag{1.46}$$

Dividing both sides by $\varphi(\mathbf{r})\chi(t)$, we get

$$\frac{i\hbar}{\chi(t)}\frac{\partial}{\partial t}\chi(t) = \frac{1}{\varphi(\mathbf{r})}\left[-\frac{\hbar^2}{2m}\Delta\varphi(\mathbf{r})\right] + V(\mathbf{r},\, t) \tag{1.47}$$

This equation equates a function of t only and a function of $\mathbf{r}$ only. This equality is only possible if each of these functions is in fact a constant, which is set to be $\hbar\omega$. Thus, the left-hand side takes the following form:

$$i\hbar\frac{\partial}{\partial t}\chi(t) = \hbar\omega\chi(t) \qquad \text{where } \chi(t) = Ae^{-i\omega t} \tag{1.48}$$

Similarly, the right-hand side of Eq. (1.47) can be written as

$$\left[-\frac{\hbar^2}{2m}\Delta\varphi(\mathbf{r})\right] + V(\mathbf{r},\, t)\varphi(\mathbf{r}) = \hbar\omega\varphi(\mathbf{r}) \tag{1.49}$$

Finally, the wave function can be written as

$$\psi(\mathbf{r},\, t) = \varphi(\mathbf{r})e^{-i\omega t} \tag{1.50}$$

where the prefactor A in $\chi(t)$ is incorporated in $\varphi(\mathbf{r})$. The wave function presented in Eq. (1.50) is the solution for the Schrödinger equation. The time and space variables are said to have been separated. A wave function of the form (1.50) is called a stationary solution of the Schrödinger equation. This is because it leads to a time-independent probability density $|\psi(\mathbf{r},\, t)|^2 = |\varphi(\mathbf{r})|^2$. In a stationary function, only one angular frequency ω appears. According to the Planck-Einstein relations, a stationary state is a state with a well-defined energy $E = \hbar\omega$ (energy eigenvalue). Equation (1.49) can be rewritten as

$$\left[-\frac{\hbar^2}{2m}\Delta + V(\mathbf{r},\, t)\right]\varphi(\mathbf{r}) = \hbar\omega\varphi(\mathbf{r}) \tag{1.51}$$

or

$$\mathbf{H}\varphi(\mathbf{r}) = E\varphi(\mathbf{r}) \tag{1.52}$$

where $\mathbf{H}$ is the differential operator known as the Hamiltonian operator:

$$\mathbf{H} = -\frac{\hbar^2}{2m}\Delta + V(\mathbf{r},\, t) \tag{1.53}$$

The operator $\mathbf{H}$ is a linear operator, since if α_1 and α_2 are constants, we have

$$\mathbf{H}[\alpha_1\varphi_1(\mathbf{r}) + \alpha_2\varphi_2(\mathbf{r})] = \mathbf{H}\alpha_1\varphi_1(\mathbf{r}) + \mathbf{H}\alpha_2\varphi_2(\mathbf{r}) \tag{1.54}$$

Equation (1.41) is the eigenvalue equation of the linear operator **H**. As we shall see in the following chapters, Eq. (1.41) has a solution only for certain values of E. This is the origin of energy quantization.

1.6 Dirac Notation

Each quantum state of a particle is characterized by a state vector belonging to an abstract space $\mathcal{S}$, called the state space of the particle. Any element or vector of $\mathcal{S}$-space is called a *ket vector* or simply a *ket*, which is represented by the symbol $|\rangle$. Inside this ket we can place a quantity, which distinguishes it from all others, for example, $|\psi\rangle$. Also we can define a *bra vector* with every ket $|\psi\rangle \in \mathcal{S}$, which is denoted $\langle\psi| \in \mathcal{S}^*$, where $\mathcal{S}^*$ is the complex conjugate of $\mathcal{S}$. The origin of this terminology is based on the word *bracket* used to denote the symbol $\langle|\rangle$, hence the name *bra* for the left-hand side and the name *ket* for the right-hand side of this symbol. Thus, the notation $\langle\varphi|\psi\rangle$ is identical to the familiar wave mechanics expression

$$\langle\varphi|\psi\rangle = \int_{-\infty}^{+\infty} \varphi^*(x)\psi(x)\,dx \tag{1.55}$$

The bra and ket satisfy the scalar products defined as

$$\langle\varphi|\psi\rangle = (|\varphi\rangle,\ |\psi\rangle) \tag{1.56a}$$

$$\langle\varphi|\psi\rangle^* = \langle\psi|\varphi\rangle \tag{1.56b}$$

The product of two linear operators **A** and **B** is defined as $(\mathbf{AB})\,|\psi\rangle = \mathbf{A}(\mathbf{B}|\psi\rangle)$. In general $\mathbf{AB} \neq \mathbf{BA}$. The commutator $[\mathbf{A}, \mathbf{B}]$ is by definition given as $[\mathbf{A}, \mathbf{B}] = \mathbf{AB} - \mathbf{BA}$. Now let $|\varphi\rangle$ and $|\psi\rangle$ be two kets; we define $\langle\varphi|A|\psi\rangle$ as the matrix element of **A**. Now assume that $|\varphi\rangle$ and $|\psi\rangle$ are written in opposite order: $|\varphi\rangle\langle\psi|$. This is actually an operator since applying it to an arbitrary ket $|\chi\rangle$ yields $|\varphi\rangle\langle\psi|\chi\rangle = \alpha|\varphi\rangle$, where α is a real constant. Thus, applying $|\varphi\rangle\langle\psi|$ to an arbitrary ket gives another ket, which is the definition of an operator.

The order of symbols is very important in Dirac notation. The following discussion is focused on the properties of bra and ket functions. If λ is a complex number and $|\psi\rangle$ is a ket, then $\lambda|\psi\rangle$ is a ket, which can be presented as

$$\lambda|\psi\rangle = |\lambda\psi\rangle \tag{1.57}$$

One then must remember that $\langle\lambda\psi| = \lambda^*\langle\psi|$ is the bra associated with the ket $|\lambda\psi\rangle$. Additionally,

$$\langle\varphi|\lambda_1\psi_1 + \lambda_2\psi_2\rangle = \lambda_1\langle\varphi|\psi_1\rangle + \lambda_2\langle\varphi|\psi_2\rangle \tag{1.58}$$

$$\langle\lambda_1\varphi_1 + \lambda_2\varphi_2|\psi\rangle = \lambda_1^*\langle\varphi_1|\psi\rangle + \lambda_2^*\langle\varphi_2|\psi\rangle \tag{1.59}$$

$$\langle\psi|\psi\rangle = \begin{cases} A & \text{if } |\psi\rangle \neq 0 \\ 0 & \text{if } |\psi\rangle = 0 \end{cases} \qquad A = \text{real positive number} \tag{1.60}$$

In Dirac notation, the wave function can be written as

$$|\psi\rangle = \sum_i C_i|u_i\rangle \tag{1.61}$$

where $\{|u_i\rangle\}$ is a discrete set for the basis of the ket $|\psi\rangle$.

1.7 Important Postulates

The discussion in this section is developed to help in answering the following questions. How can the state of a quantum system at a given time be described mathematically? Given this state, how can one predict the results of the measurements of various physical quantities? How can the state of the system at an arbitrary time t be found when the state at time t_o is known?

First postulate. At a fixed time t_o the state of a physical system is defined by specifying a ket $|\psi(t_o)\rangle$ belonging to the state space $\mathcal{S}$. This postulate implies that a linear combination of state vectors is a state vector. It should be emphasized here that the ket is not a statistical mixture of states.

Second postulate. Every measurable physical quantity is described by an operator in $\mathcal{S}$-space. This operator is an observable. Unlike classical mechanics, quantum mechanics describes, in a fundamentally different manner, a system and the associated physical quantities: A state is represented by a vector and a physical quantity by an operator.

Third postulate. The only possible result of the measurement of a physical quantity is one of the eigenvalues of the corresponding observable.

Fourth postulate. When a physical quantity is measured for a system in the normalized state $|\psi\rangle$, the probability P of obtaining a nondegenerate eigenvalue of the corresponding observable is

$$P = |\langle u_n|\psi\rangle|^2 \tag{1.62}$$

where $|u_n\rangle$ is the normalized eigenvector of the observable associated with the eigenvalue. If the eigenvalues are degenerate, several orthonormal eigenvectors $|u_n\rangle$ correspond to them. The probability then can be rewritten as

$$P = \sum_{i=1}^{g_n} |\langle u_n|\psi\rangle|^2 \tag{1.63}$$

where g_n is the degree of degeneracy. However, for continuous nondegenerate systems the probability of obtaining a result between α and $\alpha + d\alpha$ is equal to

$$dP = |\langle \upsilon_n|\psi\rangle|^2 \, d\alpha \tag{1.64}$$

where $|\upsilon_n\rangle$ is the eigenvector corresponding to the eigenvalue α of the observable associated with the physical quantity.

Fifth postulate. If the measurement of a physical quantity of a system in the state $|\psi\rangle$ gives the result a_n, the state of the system immediately after the measurement is the normalized projection $P_n|\psi\rangle/\sqrt{\langle\psi|P_n|\psi\rangle}$ of $|\psi\rangle$ onto the eigensubspace associated with a_n.

Sixth postulate. The time evolution of the state vector $|\psi(t)\rangle$ is governed by the Schrödinger equation:

$$i\hbar \frac{\partial}{\partial t}|\psi(t)\rangle = \mathbf{H}(t)|\psi(t)\rangle \tag{1.65}$$

where $\mathbf{H}(t)$ is the observable associated with the total energy of the system and, as stated before, is called the Hamiltonian operator of the system.

1.8 Important Mathematical Tools

This section is intended to provide the needed basic mathematical tools used in quantum mechanics without going through the rigorous proofs that are required by mathematicians. Let us first introduce the terms *wave function space* $\mathcal{F}$ and the *state space* $\mathcal{E}$. The wave function introduced earlier belongs to $\mathcal{F}$, and the state vector belongs to $\mathcal{E}$. $\mathcal{F}$ satisfies all criteria of a vector space.

1.8.1 The scalar product

For each pair of elements of $\mathcal{F}$, $|\varphi\rangle$ and $|\psi\rangle$, we associate a complex number denoted $(|\varphi\rangle, |\psi\rangle)$, which by definition is equal to $\langle\varphi|\psi\rangle = (|\varphi\rangle, |\psi\rangle)$ [see Eq. (1.56a)]. The quantity $\langle\varphi|\psi\rangle$ always converges so long as both wave functions belong to $\mathcal{F}$. Based on this definition we have

$$\begin{aligned} &(\langle\varphi|\psi\rangle) = (\langle\psi|\varphi\rangle)^* && \\ &\langle\varphi|\lambda_1\psi_1 + \lambda_2\psi_1\rangle = \lambda_1(\langle\varphi|\psi_1\rangle) + \lambda_2(\langle\varphi|\psi_2\rangle) && \text{Called linear} \qquad (1.66) \\ &\langle\lambda_1\varphi_1 + \lambda_2\varphi_2|\psi\rangle = \lambda_1^*\langle\varphi|\psi_1\rangle + \lambda_2^*\langle\varphi|\psi\rangle && \text{Called antilinear} \end{aligned}$$

If $\langle\varphi|\psi\rangle = 0$, then $|\varphi\rangle$ and $|\psi\rangle$ are said to be orthogonal. Furthermore, $|\varphi\rangle$ and $|\psi\rangle$ must satisfy Eq. (1.60).

1.8.2 Linear operators

Equation (1.57) is a simple definition of a linear operator. Let **A** and **B** be two linear operators. Their product is defined as

$$(\mathbf{AB})|\psi\rangle = \mathbf{A}(\mathbf{B}|\psi\rangle) \tag{1.67}$$

B is first to operate on $|\psi\rangle$, and then **A** operates on the new product. In general $\mathbf{AB} \neq \mathbf{BA}$. We call the commutator of **A** and **B** the operator $[\mathbf{A}, \mathbf{B}]$ defined as $[\mathbf{A}, \mathbf{B}] = \mathbf{AB} - \mathbf{BA}$. An example is the operator X and $\partial/\partial x$ that are operating on an arbitrary function $|\psi\rangle$.

$$\begin{aligned} \left[X, \frac{\partial}{\partial x}\right]|\psi\rangle &= \left(X\frac{\partial}{\partial x} - \frac{\partial}{\partial x}X\right)|\psi\rangle \\ &= X\frac{\partial}{\partial x}|\psi\rangle - \frac{\partial}{\partial x}X|\psi\rangle = X\frac{\partial}{\partial x}|\psi\rangle - \frac{\partial}{\partial x}(X|\psi\rangle) \\ &= X\frac{\partial}{\partial x}|\psi\rangle - |\psi\rangle - X\frac{\partial}{\partial x}|\psi\rangle \\ &= -|\psi\rangle \end{aligned} \tag{1.68a}$$

Thus

$$\left[X, \frac{\partial}{\partial x}\right] = -1 \tag{1.68b}$$

Example

(a) Show that $[\mathbf{A}, \mathbf{B}] = -[\mathbf{B}, \mathbf{A}]$.

(b) Expand $[\mathbf{A}, (\mathbf{B} + \mathbf{C})]$.

(c) Expand $[\mathbf{A}, \mathbf{BC}]$.

(d) Show that $[\mathbf{X}, \mathbf{P_x}] = i\hbar$, where $\mathbf{P_x} = \frac{\hbar}{i}\frac{\partial}{\partial x}$.

Solution

(a) $[\mathbf{A}, \mathbf{B}] = \mathbf{AB} - \mathbf{BA} = -(\mathbf{BA} - \mathbf{AB}) = -[\mathbf{B}, \mathbf{A}]$.

(b) $[\mathbf{A}, (\mathbf{B} + \mathbf{C})] = \mathbf{A}(\mathbf{B} + \mathbf{C}) - (\mathbf{B} + \mathbf{C})\mathbf{A} = \mathbf{AB} + \mathbf{AC} - \mathbf{BA} - \mathbf{CA} = \mathbf{AB} - \mathbf{BA} + \mathbf{AC} - \mathbf{CA} = [\mathbf{A}, \mathbf{B}] + [\mathbf{A}, \mathbf{C}]$.

(c) $[\mathbf{A}, \mathbf{BC}] = \mathbf{ABC} - \mathbf{BCA}$. Adding $\mathbf{BAC}$ and subtracting $\mathbf{BAC}$, we get $\mathbf{ABC} - \mathbf{BCA} + \mathbf{BAC} - \mathbf{BAC} = \mathbf{ABC} - \mathbf{BAC} + \mathbf{BAC} - \mathbf{BCA} = [\mathbf{A}, \mathbf{B}]\mathbf{C} + \mathbf{B}[\mathbf{A}, \mathbf{C}]$.

(d)
$$\begin{aligned}\langle\varphi|[\mathbf{X}, \mathbf{P_x}]|\psi\rangle &= \langle\varphi|[\mathbf{XP_x} - \mathbf{P_xX}]|\psi\rangle = x\langle\varphi|\mathbf{P_x}|\psi\rangle - \frac{\hbar}{i}\frac{\partial}{\partial x}\langle\varphi|\mathbf{X}|\psi\rangle\\ &= x\langle\varphi|\mathbf{P_x}|\psi\rangle - \frac{\hbar}{i}\frac{\partial}{\partial x}(x\langle\varphi|\psi\rangle)\\ &= x\langle\varphi|\mathbf{P_x}|\psi\rangle - x\frac{\hbar}{i}\frac{\partial}{\partial x}\langle\varphi|\psi\rangle - \frac{\hbar}{i}\langle\varphi|\psi\rangle\\ &= x\frac{\hbar}{i}\frac{\partial}{\partial x}\langle\varphi|\psi\rangle - x\frac{\hbar}{i}\frac{\partial}{\partial x}\langle\varphi|\psi\rangle - \frac{\hbar}{i}\langle\varphi|\psi\rangle = i\hbar\langle\varphi|\psi\rangle\end{aligned}$$

Then

$$[\mathbf{X}, \mathbf{P_x}] = i\hbar$$

Similarly

$$[\mathbf{R_i}, \mathbf{P_j}] = i\hbar\delta_{ij} \qquad \text{where } \delta_{ij} = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases}$$

1.8.3 Action of a linear operator on a bra

Let $\langle\varphi|$ be a well-defined bra, and consider the set of all kets $|\psi\rangle$. With each of these kets can be associated the complex number $\langle\varphi|\mathbf{A}|\psi\rangle$, which is defined as the matrix element of $\mathbf{A}$ between $\langle\varphi|$ and $|\psi\rangle$. Since $\mathbf{A}$ is linear and the scalar product depends linearly on the ket, the matrix element depends linearly on $|\psi\rangle$. Thus, for fixed $\langle\varphi|$ and $\mathbf{A}$, we can associate with every ket $|\psi\rangle$ a number that depends on $|\psi\rangle$. The specification of $\langle\varphi|$ and $\mathbf{A}$ therefore defines a new linear function, that is, a new bra belonging to the conjugate state space $\mathcal{E}^*$. The new bra is denoted $\langle\varphi|\mathbf{A}$. The relation that can define $\langle\varphi|\mathbf{A}$ can be written as

$$(\langle\varphi|\mathbf{A})|\psi\rangle = \langle\varphi|(\mathbf{A}|\psi\rangle) \tag{1.69}$$

This equation defines the linear operation on bras.

1.8.4 The adjoint operator $\mathbf{A}^\dagger$ of a linear operator A

For every linear operator $\mathbf{A}$, there is another linear operator $\mathbf{A}^\dagger$, called the adjoint operator or Hermitian conjugate. This could be clearly understood by examining Fig. 1.7. According to this figure, $\mathbf{A}$ is a linear operator defined by the formula

$$|\psi'\rangle = \mathbf{A}|\psi\rangle \Leftrightarrow \langle\psi'| = \langle\psi|\mathbf{A}^\dagger \tag{1.70}$$

$$\begin{tikzcd}
{|\psi\rangle} \arrow[rr, Rightarrow, "\mathbf{A}"] \arrow[d, leftrightarrow] & {} \arrow[d, leftrightarrow] & {|\psi'\rangle = \mathbf{A}|\psi\rangle} \arrow[d, leftrightarrow] \\
{\langle\psi|} \arrow[rr, Rightarrow, "\mathbf{A}^{\dagger}"'] & {} & {\langle\psi| = \langle\psi| \ \mathbf{A}^{\dagger}}
\end{tikzcd}$$

Figure 1.7 Definition of the adjoint operator $\mathbf{A}^{\dagger}$ of an operator $\mathbf{A}$ using the correspondence between kets and bras.

An operator $\mathbf{A}$ is a Hermitian if it is equal to its adjoint, that is, if $\mathbf{A} = \mathbf{A}^{\dagger}$, and the product of two Hermitian operators $\mathbf{A}$ and $\mathbf{B}$ is Hermitian if and only if $[\mathbf{A}, \mathbf{B}] = 0$. Another important quantity is the trace of an operator. Tr $\mathbf{A}$ is the trace of operator $\mathbf{A}$, and it is defined as the sum of the diagonal matrix elements of $\mathbf{A}$. Thus, Tr $\mathbf{A} = \sum_j \langle u_j|\mathbf{A}|u_j\rangle$. Tr $\mathbf{A}$ is invariant and Tr $\mathbf{AB}$ = Tr $\mathbf{BA}$. Also, Tr $\mathbf{ABC}$ = Tr $\mathbf{BCA}$ = Tr $\mathbf{CAB}$.

1.8.5 Eigenvalues and eigenfunctions of an operator

The ket $|\psi\rangle$ is said to be an eigenvector or eigenket of the linear operator $\mathbf{A}$ if

$$\mathbf{A}|\psi\rangle = \lambda|\psi\rangle \tag{1.71}$$

where λ is a complex number. This equation is called the eigenvalue equation of the linear operator $\mathbf{A}$. In general, this equation has a solution only when λ takes on certain values called eigenvalues. The set of the eigenvalues are called the spectrum of $\mathbf{A}$. If $|\psi\rangle$ is an eigenvector for $\mathbf{A}$, then $\alpha|\psi\rangle$ is also an eigenvector, where α is an arbitrary complex number.

$$\mathbf{A}(\alpha|\psi\rangle) = \alpha\mathbf{A}|\psi\rangle = \alpha\lambda|\psi\rangle = \lambda(\alpha|\psi\rangle) \tag{1.72}$$

To get rid of α, the eigenvectors are usually normalized to 1:

$$\langle\psi|\psi\rangle = 1 \tag{1.73}$$

The eigenvalue λ is called nondegenerate when its corresponding eigenvector is unique to within a constant factor, that is, when all its associated eigenkets are collinear. On the other hand, if there exists at least two linearly independent kets that are eigenvectors of $\mathbf{A}$ with the same eigenvalue, this eigenvalue is said to be degenerate.

To find the eigenvalues and eigenvectors of an operator, we shall consider the case where the state space is of a finite dimension. If $\{|u_i\rangle\}$ is the base for all the state vectors in the state space and if we project Eq. (1.71) on the basis vector $|u_i\rangle$, we obtain

$$\langle u_i|\mathbf{A}|\psi\rangle = \lambda\langle u_i|\psi\rangle \tag{1.74}$$

Inserting what is called a closure relation, $\sum_j |u_j\rangle\langle u_j| = 1$ between $\mathbf{A}$ and $|\psi\rangle$, we obtain

$$\sum_j \langle u_i|\mathbf{A}|u_j\rangle\langle u_j|\psi\rangle = \lambda\langle u_i|\psi\rangle \tag{1.75}$$

With the aid of the following relations

$$\langle u_i|\psi\rangle = C_i \tag{1.76}$$

$$\langle u_i|\mathbf{A}|u_j\rangle = A_{ij} \tag{1.77}$$

we can rewrite Eq. (1.74) as

$$\sum_j \langle u_i|\mathbf{A}|u_j\rangle\langle u_j|\psi\rangle = \lambda C_i \qquad \text{or} \qquad \sum_j [\mathbf{A}_{ij} - \lambda\delta_{ij}]C_j = 0 \tag{1.78}$$

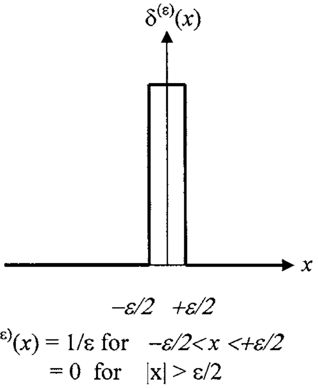
This equation can be considered to be a system of equations where the unknowns are C_j, which are the components of the eigenvector in the chosen representation. This system is linear and homogeneous. It is composed of N equations ($j = 1, 2, 3, \ldots, N$) with N unknowns (C_j). It has a nontrivial solution if and only if the determinant of the coefficients is zero (the trivial solution is $C_j = 0$). This condition can be written as

$$\text{Det}\ [\mathbf{A} - \lambda\mathbf{I}] = 0 \tag{1.79}$$

where $\mathbf{A}$ is an $N \times N$ matrix of A_{ij} elements and $\mathbf{I}$ is the unit matrix. This equation is called the *characteristic equation*, or *secular equation*, and enables us to determine all the eigenvalues of the operator $\mathbf{A}$. The spectrum of the operator can be written in the following form:

$$\begin{vmatrix} A_{11}-\lambda & A_{12} & A_{13} & \cdots & A_{1N} \\ A_{21} & A_{21}-\lambda & A_{23} & \cdots & A_{2N} \\ \vdots & \vdots & \vdots & & \vdots \\ A_{N1} & A_{N2} & A_{N3} & \cdots & A_{NN}-\lambda \end{vmatrix} = 0 \tag{1.80}$$

This is the Nth-order equation in λ; consequently, it has N roots (real or imaginary). This characteristic equation is independent of the



$\delta^{(\varepsilon)}(x)$
x
$-\varepsilon/2$ $+\varepsilon/2$
$\delta^{(\varepsilon)}(x) = 1/\varepsilon$ for $-\varepsilon/2 < x < +\varepsilon/2$
$= 0$ for $|x| > \varepsilon/2$

Figure 1.8 The δ-function: a square function of width ε and height $1/\varepsilon$ centered at $x = 0$.

representation chosen; therefore, the eigenvalues of an operator are the roots of its characteristic equation.

The eigenvectors are then determined by choosing an eigenvalue λ_o, which is a solution of the characteristic equation. Then look for the corresponding eigenvector.

1.8.6 The Dirac δ-function

The δ-function is a distribution, but it is usually treated as an ordinary function. Consider a δ-function as shown in Fig. 1.8 with a width of ε and a height of $1/\varepsilon$. By definition we have

$$\int_{-\infty}^{+\infty} \delta(x)\, dx = 1 \tag{1.81}$$

Let us evaluate the following integral, where $f(x)$ is an arbitrary function. $\int_{-\infty}^{+\infty} \delta^{(\varepsilon)}(x) f(x)\, dx$. If ε is sufficiently small, the variation of $f(x)$ over the effective interval $[-\varepsilon/2 + \varepsilon/2]$ is negligible and $f(x)$ remains practically equal to $f(0)$; therefore,

$$\int_{-\infty}^{+\infty} \delta^{(\varepsilon)}(x) f(x)\, dx \approx f(0) \int_{-\infty}^{+\infty} \delta^{(\varepsilon)}(x)\, dx = f(0) \tag{1.82}$$

The smaller ε is, the better the approximation, and for the limit $\varepsilon = 0$, we define the δ-function as

$$\int_{a}^{b} \delta(x) f(x)\, dx = \begin{cases} f(0) & \text{for } 0 \in [a, b] \\ 0 & \text{for } 0 \notin [a, b] \end{cases} \tag{1.83}$$

For a more general form of $\delta(x - x_o)$, we have

$$\int_{-\infty}^{+\infty} \delta(x - x_o) f(x)\, dx = f(x_o) \tag{1.84}$$

Other properties of the δ-function are

$$\delta(-x) = \delta(x) \tag{1.85a}$$

$$\delta(cx) = \frac{1}{|c|}\delta(x) \tag{1.85b}$$

$$\delta[g(x)] = \sum_j \frac{1}{\left|\frac{\partial g_j(x)}{\partial x}\right|}\delta(x - x_j) \tag{1.85c}$$

$$x\delta(x - x_o) = x_o\delta(x - x_o) \tag{1.85d}$$

$$g(x)\delta(x - x_o) = g(x_o)\delta(x - x_o) \tag{1.85e}$$

$$\int_{-\infty}^{+\infty} \delta(x - y)\delta(x - z)\, dx = \delta(y - z) \tag{1.85f}$$

The Fourier transform $\bar{\delta}_{x_o}(p)$ of $\delta(x - x_o)$ is

$$\bar{\delta}_{x_o}(p) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{+\infty} dx\, e^{ipx/\hbar}\ \delta(x - x_o) = \frac{1}{\sqrt{2\pi\hbar}} e^{ipx_o/\hbar} \tag{1.86a}$$

and

$$\bar{\delta}_{x_o}(p) = \frac{1}{\sqrt{2\pi\hbar}} = \text{Fourier transform of } \delta(x) \tag{1.86b}$$

The inverse Fourier transform is

$$\delta(x - x_o) = \frac{1}{2\pi\hbar} \int_{-\infty}^{+\infty} dp\, e^{ip(x - x_o)/\hbar} = \frac{1}{2\pi} \int_{-\infty}^{+\infty} dk\, e^{ip(x - x_o)} \tag{1.87}$$

Additionally, $\delta(x)$ is a derivative of a unit step function $\theta(x)$, i.e.,

$$\frac{d}{dx}\theta(x) = \delta(x) \tag{1.88a}$$

and

$$\theta(x) = \frac{1}{2\pi\hbar} \int_{-\infty}^{x} \delta^{(\varepsilon)}(y)\, dy \tag{1.88b}$$

Other properties of $\delta(x)$ include the following:

$$\int_{-\infty}^{+\infty} \delta'(x) f(x)\, dx = -\int_{-\infty}^{+\infty} \delta(x) f'(x)\, dx = -f'(0) \tag{1.89}$$

and

$$\delta'(-x) = -\delta'(x) \tag{1.90a}$$

$$x\delta'(x) = -\delta(x) \tag{1.90b}$$

where the prime indicates the first derivative and Eq. (1.87) allows us to write

$$\delta'(x) = \frac{i}{2\pi} \int_{-\infty}^{+\infty} kd\, ke^{ip(x-x_o)} \tag{1.91}$$

The nth-order derivative (n) can be defined in the same way:

$$\int_{-\infty}^{+\infty} \delta^{(n)}(x) f(x)\, dx = (-1)^n f^{(n)}(0) \tag{1.92}$$

Equation (1.90) can then be generalized to the following:

$$\delta^{(n)}(-x) = (-1)^n \delta^{(n)}(x) \tag{1.93a}$$

$$x\delta^{(n)}(x) = -n\delta^{(n-1)}(x) \tag{1.93b}$$

The δ-function is very useful in quantum mechanics as we will see in subsequent chapters.

1.8.7 Fourier series and Fourier transform in quantum mechanics

In this section we will review a few definitions that are important in quantum mechanics. A function $f(x)$ is said to be periodic if there exists a real nonzero number L such that for all x: $f(x+L) = f(x)$, where L is called the period of the function. If $f(x)$ is periodic with a period L, then all numbers nL, where n is an integer, are also periods of $f(x)$. Another important example of periodic functions is the periodic exponential. For an exponential $e^{\alpha L}$ to have a period L, it is necessary to have $e^{\alpha L} = 1$, that is, $\alpha L = i2n\pi$, where n is an integer. Thus, if $f(x)$ is a periodic function with a fundamental period of L, one can expand this function in the following form, known as the Fourier series:

$$f(x) = \sum_{n=-\infty}^{\infty} c_n e^{ik_n x} \qquad \text{with}\, k_n = n\frac{2\pi}{L} \tag{1.94}$$

The coefficients c_n are given by the following formula:

$$c_n = \frac{1}{L} \int_{x_o}^{x_o+L} dx\, e^{-ik_n x} f(x) \tag{1.95}$$

where x_o is an arbitrary number. The coefficients c_n are called the Fourier spectrum of $f(x)$. Another useful relation, known as the Bessel-Parseval relation, is

$$\frac{1}{L} \int_{x_o}^{x_o+L} dx |f(x)|^2 = \sum_{n=-\infty}^{\infty} |c_n|^2 \tag{1.96}$$

Now assume that we have two functions, $g(x)$ and $f(x)$ with the same period and having Fourier coefficients d_n and c_n, respectively. We can generalize Eq. (1.96) according to the following relation:

$$\frac{1}{L} \int_{x_o}^{x_o+L} dx\, f(x) g(x) = \sum_{n=-\infty}^{\infty} c_n d_n \tag{1.97}$$

If $\psi(x)$ is a one-dimensional wave function, its Fourier transform $\bar{\psi}(p)$ is defined as

$$\bar{\psi}(p) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} dx\, e^{-ipx/\hbar} \psi(x) \tag{1.98}$$

and the inverse formula is

$$\psi(x) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} dp\, e^{ipx/\hbar} \bar{\psi}(p) \tag{1.99}$$

Another useful relationship in quantum mechanics is the Parseval-Plancherel formula, which has the following general form:

$$\int_{-\infty}^{\infty} \varphi^*(x) \psi(x)\, dx = \int_{-\infty}^{\infty} \bar{\varphi}^*(p) \bar{\psi}(p)\, dp \tag{1.100}$$

1.9 Variational Method

There are several well-known approximations used to solve quantum mechanical problems. One approximation is the variational method, which has numerous applications in solid-state physics where the exact solution requires extensive computational analysis. Let us consider a nondegenerate arbitrary physical system with a time-independent Hamiltonian **H**. The general solution of the Schrödinger equation can be written as

$$\mathbf{H}|\varphi_n\rangle = E_n|\varphi_n\rangle \tag{1.101}$$

where E_n are the discrete eigenvalues for $n = 0, 1, 2, \ldots$ While the Hamiltonian is known, the eigenvalues are not necessarily known. In this case, the variational method can be used to obtain an approximate expression for the eigenvalues. This method is very useful for cases where **H** cannot be exactly diagonalized. To proceed, let us choose an arbitrary ket where the mean value of the Hamiltonian can be expressed as

$$\langle\mathbf{H}\rangle = \frac{\langle\psi|\mathbf{H}|\psi\rangle}{\langle\psi|\psi\rangle} \geq E_o \tag{1.102}$$

where E_o is the ground-state eigenvalue, and the inequality is valid if $|\psi\rangle$ is the eigenvector of **H** with an eigenvalue of E_o. Without going through the derivation, we simply state the final result as

$$\mathbf{H}|\psi\rangle = \langle\mathbf{H}\rangle|\psi\rangle \tag{1.103}$$

This method can be generalized and applied to the approximate determination of the eigenvalues of the Hamiltonian. Equation (1.103) tells us that if the function $\langle\mathbf{H}\rangle(\alpha)$ obtained from the trial kets $|\psi(\alpha)\rangle$ has several extrema, the extrema give the approximate values of the function's energies E_n.

As an example, let us find the first energy level of the simple harmonic oscillator with the following Hamiltonian:

$$\mathbf{H} = -\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + \frac{1}{2}m\omega^2x^2 \tag{1.104}$$

and the following trial function

$$\psi(x) = e^{-\alpha x^2} \qquad \text{where } \alpha > 0 \tag{1.105}$$

The objective here is to evaluate $\langle\psi|\mathbf{H}|\psi\rangle$ and $\langle\psi|\psi\rangle$, which are

$$\langle\psi|\mathbf{H}|\psi\rangle = \sqrt{\frac{\pi}{2}}\left[\frac{\hbar^2}{2m}\alpha^{1/2} + \frac{1}{8}m\omega^2\alpha^{-3/2}\right]$$

and (1.106)

$$\langle\psi|\psi\rangle = \sqrt{\frac{\pi}{2\alpha}}$$

Divide the two expressions to obtain

$$\langle\mathbf{H}\rangle(\alpha) = \frac{\hbar^2}{2m}\alpha + \frac{1}{8}m\omega^2\alpha^{-1} \tag{1.107}$$

We assume at the beginning that the wave function has extrema such that the derivative of Eq. (1.107) is zero. This yields $\alpha = m\omega/(2\hbar)$, which can be substituted back into Eq. (1.107) to give $\langle\mathbf{H}\rangle(\alpha) = \frac{1}{2}\hbar\omega$. This is exactly the ground-state energy obtained from the exact solution (see Chap. 2).

1.10 Perturbation

Simple physical systems such as simple harmonic oscillators and hydrogen atoms can be solved exactly where the Hamiltonian is simple enough to generate exact eigenvalues. In general, the Hamiltonian is very complicated for most systems, such as many electron atoms, semiconductor heterostructures, multiple quantum wells, and quantum dots. Solving the Schrödinger equation for such complicated systems is difficult, and therefore one needs to make several approximations to reach a reasonable answer. One of these approximations is the perturbation theory. In this section we will treat the time-independent perturbation (stationary) approximation, which is widely used in many systems such as solid-state physics. To understand this approximation, one needs to define a physical system and isolate the main effects that are responsible for the main features of the system. Once these features are understood, then the finer details could be discussed by considering the less important effects that were neglected in the first approximation. Treating these secondary effects can be performed using the perturbation theory. Thus, the Hamiltonian of the system can be presented in the following form:

$$H = H_o + H_1 \tag{1.108}$$

where H_o is the unperturbed Hamiltonian with known eigenvectors and eigenvalues, and H_1 is the perturbation that describes the secondary effects in the system. The problem now is to find the modification of the eigenvalues produced by adding H_1. The matrix elements of H_1 are assumed to be much smaller that those of H_o. Let us now define a very small real number $\lambda \ll 1$ such as

$$H_1 = \lambda \hat{H}_1 \tag{1.109}$$

where $\hat{H}_1$ is an operator whose matrix elements are comparable to those of H_o. Perturbation theory deals with expanding the eigenvalues and eigenvectors of H in terms of powers of λ with a finite number of terms.

Let us assume that the discrete eigenvalues (E_u^o) and eigenvectors ($|\varphi_u^i\rangle$) are known for H_o. The subscript u indicates the unperturbed terms, and the superscript i is added in cases where we have degenerate states. For the first approximation (unperturbed system) we have

$$H_o|\varphi_u^i\rangle = E_u^o|\varphi_u^i\rangle \tag{1.110}$$

where the set of vectors $|\varphi_u^i\rangle$ forms an orthogonal basis such that $\langle\varphi_u^i|\varphi_{u'}^{i'}\rangle = \delta_{ii'}\delta_{uu'}$ and $\sum_u \sum_i |\varphi_u^i\rangle\langle\varphi_u^i| = 1$. Now we can consider the system Hamiltonian that depends on the parameter λ by substituting Eq. (1.109) into (1.108):

$$H(\lambda) = H_o + \lambda \hat{H}_1 \tag{1.111}$$

For $\lambda = 0$ we have only the unperturbed Hamiltonian and the eigenvalues of $H(\lambda)$ that depend on λ. To find the approximate solution of the Schrödinger equation, one needs to find $E(\lambda)$ and $|\psi(\lambda)\rangle$ of $H(\lambda)$:

$$H(\lambda)|\psi(\lambda)\rangle = E(\lambda)|\psi(\lambda)\rangle \tag{1.112}$$

Let us assume that $E(\lambda)$ and $|\psi(\lambda)\rangle$ can be expanded in powers of λ such as

$$E(\lambda) = E_o + \lambda E_1 + \lambda^2 E_2 + \cdots + \lambda^n E_n \tag{1.113a}$$

and

$$|\psi(\lambda)\rangle = |0\rangle + \lambda|1\rangle + \lambda^2|2\rangle + \cdots + \lambda^n|n\rangle \tag{1.113b}$$

Substituting these expansions into Eq. (1.112), we obtain

$$(H_o + \lambda\hat{H}_1)\sum_n \lambda^n|n\rangle = \sum_{n'} \lambda^{n'} E_{n'} \sum_n \lambda^n|n\rangle \tag{1.114}$$

From this equation we obtain the following relations. For the zeroth order of λ we have

$$H_o|0\rangle = E_o|0\rangle \tag{1.115}$$

For the first order we have

$$(H_o - E_o)|1\rangle + (\hat{H}_1 - E_1)|0\rangle = 0 \tag{1.116}$$

For the second order we have

$$(H_o - E_o)|2\rangle + (\hat{H}_1 - E_1)|1\rangle - E_2|0\rangle = 0 \tag{1.117}$$

and so on.

From the orthogonal property we have for the zeroth order

$$\langle 0|0\rangle = 1 \tag{1.118}$$

For the first order we have

$$\langle \psi(\lambda)|\psi(\lambda)\rangle = \langle 0|0\rangle + \lambda[\langle 1|0\rangle + \langle 0|1\rangle] + O(\lambda^2) \tag{1.119}$$

where $O(\lambda^2)$ is a term of the second order. Since λ is a real number, then $\langle 1|0\rangle = \langle 0|1\rangle = 0$. Similarly, one can find from the second order in λ that

$$\langle 2|0\rangle = \langle 0|2\rangle = -\frac{1}{2}\langle 1|1\rangle \tag{1.120}$$

Let us consider the modification to the unperturbed E_u^o defined in Eq. (1.106). Consider first the zeroth perturbation for $\lambda \to 0$. By comparing Eqs. (1.110) and (1.115), we have $E_o = E_u^o$, and $|0\rangle = |\varphi_n\rangle$. This simple result demonstrates how to obtain the eigenvalues and eigenvectors of the H_o. For the first-order correction, we need to determine E_1 and $|1\rangle$ from Eq. (1.116). Let us project Eq. (1.116) onto the eigenvector $|\varphi_n\rangle$ to obtain

$$\langle \varphi_n|(H_o - E_o)|1\rangle + \langle \varphi_n|(\hat{H}_1 - E_1)|0\rangle = 0 \tag{1.121}$$

By letting H_o operate on the bra $\langle \varphi_n|$, we find that the first term in this equation is zero since $|0\rangle = |\varphi_n\rangle$. Hence, Eq. (1.121) is reduced to

$$E_1 = \langle \varphi_n|\hat{H}_1|0\rangle \tag{1.122}$$

Substituting Eq. (1.122) into (1.113a), we have

$$\begin{aligned} E_n(\lambda) &= E_u^o + \lambda\langle \varphi_n|\hat{H}_1|0\rangle + O(\lambda^2) \\ &= E_u^o + \langle \varphi_n|H_1|0\rangle + O(\lambda^2) \end{aligned} \tag{1.123}$$

Thus, the first-order correction to the unperturbed eigenvalue is the mean value of the perturbed term H_1.

For the eigenvector correction, let us project Eq. (1.116) onto the eigenvectors $|\varphi_p^i\rangle \neq |\varphi_n\rangle$ to obtain

$$\left\langle \varphi_p^i|(H_o - E_o)|1\right\rangle + \left\langle \varphi_p^i|(\hat{H}_1 - E_1)|\varphi_n\right\rangle = 0 \tag{1.124}$$

Recall that $|0\rangle = |\varphi_n\rangle$, the index p is different than n, and i is the degeneracy index. Since the eigenvectors of H_o associated with different eigenvalues are orthogonal, then $E_1\langle \varphi_p^i|\varphi_n\rangle = 0$. Recall that $E_o = E_u^o$

and let H_o act on the bra $\langle\varphi_p^i|$ to give E_p^o. Equation (1.124) can then be written as

$$\left(E_p^o - E_u^o\right)\left\langle\varphi_p^i|1\right\rangle + \left\langle\varphi_p^i|\hat{H}_1|\varphi_n\right\rangle = 0 \tag{1.125}$$

This equation gives the coefficients of the desired expansion of the eigenvector $|1\rangle$ on all the unperturbed basis states except $|\varphi_n\rangle$:

$$\left\langle\varphi_p^i|1\right\rangle = \frac{1}{E_u^o - E_p^o}\left\langle\varphi_p^i|\hat{H}_1|\varphi_n\right\rangle \qquad p \neq n \tag{1.126}$$

The last coefficient $\langle\varphi_n|1\rangle = 0$ according to Eq. (1.119). Finally, the eigenvector $|1\rangle$ can be written as

$$|1\rangle = \sum_{p\neq n}\sum_{i}\frac{\left\langle\varphi_p^i|\hat{H}_1|\varphi_n\right\rangle}{E_u^o - E_p^o}\left|\varphi_p^i\right\rangle \tag{1.127}$$

Consequently, the eigenvector $|\psi_n(\lambda)\rangle$ has the following form:

$$|\psi_n(\lambda)\rangle = |\varphi_n\rangle + \sum_{p\neq n}\sum_{i}\frac{\left\langle\varphi_p^i|H_1|\varphi_n\right\rangle}{E_u^o - E_p^o}\left|\varphi_p^i\right\rangle + O(\lambda^2) \tag{1.128}$$

For the second-order perturbation theory, the energy correction is obtained by projecting Eq. (1.117) onto the vector $|\varphi_n\rangle$:

$$\langle\varphi_n|(H_o - E_o)|2\rangle + \langle\varphi_n|(\hat{H}_1 - E_1)|1\rangle - \langle\varphi_n|E_2|0\rangle = 0 \tag{1.129}$$

By letting H_o operate on the bra $\langle\varphi_n|$ and knowing that $E_o = E_u^o$, the first term is zero. The E_1 term is also zero since $\langle\varphi_n|1\rangle = 0$, and hence Eq. (1.129) is reduced to

$$E_2 = \langle\varphi_n|\hat{H}_1|1\rangle \tag{1.130}$$

By substituting the expression of $|1\rangle$ as shown in Eq. (1.127) into (1.130), the second-order corrections to the eigenvalue can be written as

$$E_2 = \sum_{p\neq n}\sum_{i}\frac{\left|\left\langle\varphi_p^i|\hat{H}_1|\varphi_n\right\rangle\right|^2}{E_u^o - E_p^o} \tag{1.131}$$

The final expression for $E_n(\lambda)$ to the second-order perturbation takes the following form:

$$\begin{aligned} E_n(\lambda) &= E_u^o + \lambda\langle\varphi_n|\hat{H}_1|0\rangle + \lambda^2\sum_{p\neq n}\sum_{i}\frac{|\langle\varphi_p^i|\hat{H}_1|\varphi_n\rangle|^2}{E_u^o - E_p^o} + O(\lambda^3) \\ &= E_u^o + \langle\varphi_n|H_1|0\rangle + \sum_{p\neq n}\sum_{i}\frac{|\langle\varphi_p^i|H_1|\varphi_n\rangle|^2}{E_u^o - E_p^o} + O(\lambda^3) \end{aligned} \tag{1.132}$$

For the eigenvector $|\psi_n(\lambda)\rangle$ corrections, one can project Eq. (1.129) onto $|\varphi_u^i\rangle$ to obtain the following result:

$$|\psi_n(\lambda)\rangle = |\varphi_n\rangle + \sum_{p\neq n}\sum_i \frac{\langle\varphi_p^i|H_1|\varphi_n\rangle}{E_u^o - E_p^o}|\varphi_p^i\rangle + \sum_{p\neq n}\sum_i \frac{|\langle\varphi_p^i|H_1|\varphi_n\rangle|^2}{(E_u^o - E_p^o)^2}|\varphi_p^i\rangle + O(\lambda^3) \tag{1.133}$$

For additional details on the perturbation of the degenerate states, see for example Merzbacher and Cohen-Tannoudji et al.

1.11 Angular Momentum

Angular momentum is a very important problem in many fields including semiconductor materials. One may encounter angular momentum when dealing with doping in semiconductors, solving the Schrödinger equation for semiconductor energy bands, and in many other cases. In this section, we will present the most important properties of angular momentum without going through derivations. In dealing with angular momentum, one needs to distinguish between the spin, the orbital angular momentum, and the total angular momentum. For the orbital angular momentum of a spinless particle, we have three observables L_x, L_y, and L_z that are the components of the orbital angular momentum operator **L**. The three components can be written as

$$\begin{aligned} L_x &= YP_z - ZP_y \\ L_y &= ZP_x - XP_z \\ L_z &= XP_y - YP_x \end{aligned} \tag{1.134}$$

where X, Y, and Z are the position observables and P_x, P_y, and P_z are the momentum observables. It was shown in Sec. 1.8.2 that $[X, P_x] = [Y, P_y] = [Z, P_z] = i\hbar$. Using this relation, we can write

$$\begin{aligned} [L_x, L_y] &= i\hbar L_z \\ [L_y, L_z] &= i\hbar L_x \\ [L_z, L_x] &= i\hbar L_y \end{aligned} \tag{1.135}$$

Similarly, the components of the total angular momentum J can be expressed as

$$\begin{aligned} [J_x, J_y] &= i\hbar J_z \\ [J_y, J_z] &= i\hbar J_x \\ [J_z, J_x] &= i\hbar J_y \end{aligned} \tag{1.136}$$

It can be shown that $[J^2, J] = 0$. It is customary to use the following raising and lowering operators:

$$J_\pm = J_x \pm J_y \tag{1.137}$$

The eigenvalues of J^2 usually take the form $j(j+1)\hbar$ where $j \geq 0$. It can be shown that for orthogonal wave vectors $|k, j, m\rangle$, the corresponding eigenvalues are

$$\begin{aligned} J_z|k,\ j,\ m\rangle &= m\hbar|k,\ j,\ m\rangle \\ J_\pm|k,\ j,\ m\rangle &= \hbar\sqrt{j(j+1) - m(m\pm 1)}|k,\ j,\ m\pm 1\rangle \end{aligned} \tag{1.138}$$

where m is the quantum number that indicates the projections of the angular momentum on the z axis. The addition of two angular momenta or spin-orbit coupling are usually discussed thoroughly in quantum mechanics textbooks. We may revisit the spin-orbit coupling when dealing with quantization of energy levels in heterostructures and quantum wells.

Summary

In this chapter we reviewed the basic concepts of quantum mechanics needed for the analysis of bulk and low-dimensional semiconductor systems. Several examples, such as blackbody radiation, specific heat capacity of solids, and photoelectric effects, showing the limitation of classical mechanics were presented. The concept of duality and the de Broglie relation were briefly discussed. The Schrödinger equation and the concept of wave functions were presented, from which a spectrum of energy levels can be obtained. The concept of energy quantization, probabilities, wave packets, the Heisenberg uncertainty principle, Dirac notations, and the most important postulates of quantum mechanics were discussed. Quantum mechanics models and theories require the knowledge of mathematical tools. Thus, we briefly discussed the separation of variables, scalar product, linear operators, adjoint operators, eigenfunction operators, the Dirac δ-function, and the Fourier transform. There are several approximations in quantum mechanics that one needs to understand at early stages. The variational method and perturbation are among them. These two approximations are encountered in many quantum mechanical treatments of solids in general and of semiconductors in particular. The derivation of the first- and second-order perturbations is presented toward the end of the chapter. Finally, the most important properties of the angular momentum were discussed briefly at the end of this chapter.

Problems

1.1 Show that the total internal energy per oscillator for a system of $3N$ oscillators is $U = kT$. Start with the Dulong and Petit model.

1.2 Derive Einstein's expression for the specific heat capacity of a solid given by Eq. (1.8).

1.3 The work function of a material is the minimum energy required to remove an electron from the surface of the material. Calculate the maximum wavelength of light for the photoelectric emission from gold ($\varphi_o = 4.90$ V) and cesium ($\varphi_o = 1.90$ V).

1.4 Use the uncertainty relation to evaluate the ground state of the hydrogen atom.

1.5 Calculate the de Broglie wavelength for (a) an electron with a kinetic energy of 10^4 eV, (b) a proton of kinetic energy of 10^2 eV, and (c) a (150 kg) man running at a speed of 0.25 m/s.

1.6 Starting from Eq. (1.35), write the Fourier transform of this wave function in terms of $g(\omega)$. Then assume that $g(\omega) = |g(\omega)|e^{i\beta(\omega)}$. Derive the uncertainty principle presented in Eq. (1.37).

1.7 Show that $\Delta p\ \Delta x \approx \Delta E\ \Delta t$.

1.8 If $|\psi\rangle$ can be normalized to unity and assuming that an operator $\mathbf{A} = |\psi\rangle\langle\psi|$, show that $\mathbf{A}^2 = \mathbf{A}$.

1.9 Assume that $[X, P] = i\hbar$. Show that $[X, P^2] = 2i\hbar P$, and then show that $[X, P^n] = i\hbar n P^{n-1}$.

1.10 Care must be taken when working with operators. The order of the operators is very important. Assume that $\mathbf{A}$ and $\mathbf{B}$ are operators that do not commute. Show that $e^{\mathbf{A}}e^{\mathbf{B}}$, $e^{\mathbf{B}}e^{\mathbf{A}}$, and $e^{\mathbf{A}+\mathbf{B}}$ are not equal.

1.11 A series of lines in hydrogen correspond to transitions to a final state characterized by some quantum number n. If the wavelength of the radiation giving rise to the first line is 657 nm, what are the wavelengths corresponding to the next two lines? Assume that $\Delta n = 1$.

1.12 Show that the integration of a δ-function is a step function.

1.13 Derive the expression of the Fourier transform function shown in Eq. (1.97).

1.14 If $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$, write an expression for $e^{\mathbf{A}}$ in matrix form.

1.15 Show that $\sum_j |u_j\rangle\langle u_j| = 1$. This is called the closure relation.

1.16 Find the Fourier transform of the following functions:

(a)

$$\bar{\psi}(x) = \begin{cases} \frac{1}{a} & \text{for } -\frac{a}{2} < x < \frac{a}{2} \\ 0 & \text{for } |x| > \frac{a}{2} \end{cases}$$

(b)

$$\psi(x) = \begin{cases} e^{-ax} & \text{for } x > 0 \\ 0 & \text{for } x < 0 \end{cases}$$

(c)

$$\psi(x) = e^{-x^2/a^2}$$

1.17 Show that the second-order perturbation Schrödinger equation is given by Eq. (1.117). Project this equation onto the wave vectors $|\varphi_n\rangle$ to obtain the final expression $E_n(\lambda)$ to the second order as shown in Eq. (1.132).

1.18 Use the following trial function $\psi(x) = (x^2 + a)^{-1}$, where a is a positive number, to calculate $\langle H \rangle$ for a simple harmonic oscillator as described in the variational method approach.

Chapter

2

Potential Barriers and Wells

This chapter deals with particles in time-independent potential barriers and wells. The quantum effects such as transmission through barriers (tunneling) and energy quantization should increase when the potential barrier varies over a distance shorter than the wavelength of the quantum particle (either photon or electron). The time-independent Schrödinger equation with an arbitrary potential was discussed briefly in Sec. 1.5. In order to distinguish between the various possible values of the energy and the corresponding eigenfunctions, we label them with a quantum number n such that the Schrödinger equation can be written as

$$\mathbf{H}\varphi_n(\mathbf{r}) = E_n\varphi_n(\mathbf{r}) \tag{2.1}$$

and the stationary state of the particle has a wave function with the form

$$\psi_n(\mathbf{r}, t) = \varphi_n(\mathbf{r})e^{-iE_nt/\hbar} \tag{2.2}$$

where $\psi_n(\mathbf{r}, t)$ is a solution to the Schrödinger equation [Eq. (2.1)]. The exponential $e^{-iE_nt/\hbar}$ is factored out in the Schrödinger equation, and Eq. (2.2) is still called a time-independent wave function. Since Eq. (2.1) is linear, it has other solutions of the form

$$\psi(\mathbf{r}, t) = \sum_n C_n\varphi_n(\mathbf{r})e^{-iE_nt/\hbar} \tag{2.3}$$

where C_n are arbitrary complex numbers. In this chapter, we consider only one-dimensional systems where the potentials are presented by functions that make discontinuities along the x coordinate. These functions may or may not represent real physical potentials, but we shall use them for illustration on how to obtain the eigenvalues and eigenfunctions.

Copyright © 2005 by The McGraw-Hill Companies, Inc. Click here for terms of use.

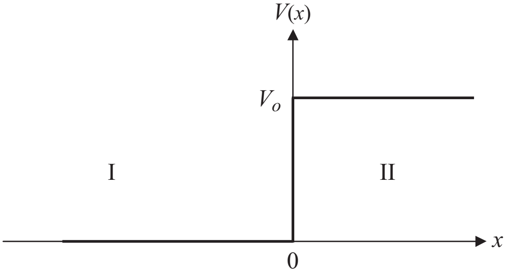


Figure 2.1 Step potential where $V(x) = 0$ for $x < 0$ (region I) and $V(x) = V_o$ for $X > 0$ (region II).

2.1 Stationary States of a Particle in a Potential Step

Consider the potential step shown in Fig. 2.1 and consider that a particle with a mass m is traveling from left to right with an energy $E > V_o$, where V_o is the height of the potential step. The Schrödinger equation for this potential can be written as

$$-\frac{\hbar^2}{2m}\frac{d^2}{dx^2}\varphi(x) + V(x)\varphi(x) = E\varphi(x) \tag{2.4}$$

which can be rearranged in the following form:

$$\frac{d^2}{dx^2}\varphi(x) + \frac{2m}{\hbar^2}(E - V_o)\varphi(x) = 0 \tag{2.5}$$

The solution for Eq. (2.5) has the form of Eq. (2.3) in both region I ($x < 0$) and region II ($x > 0$). Let us introduce the following positive numbers, known as propagation vectors, k_1 *and* k_2, such that

$$k_1^2 = \frac{2m(E)}{\hbar^2} \quad \text{for region I} \qquad \text{and} \qquad k_2^2 = \frac{2m(E - V_o)}{\hbar^2} \quad \text{for region II} \tag{2.6}$$

Thus, the solutions of Eq. (2.5) for both regions can be expressed as

$$\varphi_{\mathrm{I}}(x) = Ae^{ik_1x} + Be^{-ik_1x} \tag{2.7a}$$

$$\varphi_{\mathrm{II}}(x) = Ce^{ik_2x} + De^{-ik_2x} \tag{2.7b}$$

where A, B, C, and D are complex constants and are equivalent to the constants C_n shown in Eq. (2.3). In quantum mechanics, the wave functions in both regions must be matched. This requires the introduction of boundary conditions, which are stated as follows:

1. The wave functions at the boundaries must be continuous. Thus $\varphi_{\mathrm{I}}(x = 0) = \varphi_{\mathrm{II}}(x = 0)$.

2. The first derivative of both wave functions must also be continuous, that is, $\frac{d}{dx}\varphi_{\mathrm{I}}(x=0) = \frac{d}{dx}\varphi_{\mathrm{II}}(x=0)$, or $\varphi'_{\mathrm{I}}(x=0) = \varphi'_{\mathrm{II}}(x=0)$, where the prime stands for the first derivative.

By applying these boundary conditions to Eq. (2.7), we obtain

$$\begin{aligned} A + B &= C + D \\ ik_1A - ik_1B &= ik_2C - ik_2D \end{aligned} \tag{2.8}$$

Thus, the boundary conditions produces two equations with four unknowns. If we assume that the particle is coming from $x = -\infty$, then we can choose $D = 0$ or $A + B = D$ and $k_1A - k_1B = k_2C$. Even with this simplification, we can only determine the ratios B/A and C/A, which are shown as

$$\frac{B}{A} = \frac{k_1 - k_2}{k_1 + k_2} \qquad \text{and} \qquad \frac{C}{A} = \frac{2k_1}{k_1 + k_2} \tag{2.9}$$

Thus far we have $\varphi_{\mathrm{I}}(x)$ composed of two waves or two parts; one represents the particle coming from $x = -\infty$, and the other represents the particle as being reflected from the potential step. Since we have chosen $D = 0, \varphi_{\mathrm{II}}(x)$ is composed of only one wave representing the particle as being transmitted above the potential step with an energy $E > V_o$.

The concepts of transmissions and reflections of particles based on the ratios shown in Eq. (2.9) can be understood in terms of a *probability current*, which can be discussed as follows. Let us consider a system composed of only a single spinless particle with a normalized wave function of $\psi(\mathbf{r}, t)$. A quantity known as a *probability density* is defined as the probability $dp(\mathbf{r}, t)$ of finding the particle at time t in an infinitesimal volume $d^3\mathbf{r}$ located at the point $\mathbf{r}$ in the system and is defined as

$$dp(\mathbf{r}, t) = \rho(\mathbf{r}, t)d^3\mathbf{r} \tag{2.10}$$

where

$$\rho(\mathbf{r}, t) = |\psi(\mathbf{r}, t)|^2 \tag{2.11}$$

The *probability density* is analogous to an isolated physical system with a volume charge density distribution in space of $\rho(\mathbf{r}, t)$. The total charge in this case is conserved over time. But the spatial charge distribution may vary within the system, giving rise to electric currents. More precisely, the variation of the charge, dQ, during a time interval dt contained within the volume V is given by $-I\,dt$, where I is the current.

The current density $\mathbf{J}(\mathbf{r}, t)$ according to the classical vector analysis can be written as

$$\frac{\partial}{\partial t}\rho(\mathbf{r}, t) + \operatorname{div} \mathbf{J}(\mathbf{r}, t) = 0 \tag{2.12}$$

The objective here is to show that it is possible to find $\mathbf{J}(\mathbf{r}, t)$, known as the *probability current*, which satisfies an equation identical to Eq. (2.12). Let us first assume that the particle under study is subject to a potential $V(\mathbf{r}, t)$, and thus the Hamiltonian of the particle is

$$\mathbf{H} = \frac{p^2}{2m} + V(\mathbf{r}, t) \tag{2.13}$$

The corresponding Schrödinger equation is

$$i\hbar\frac{\partial}{\partial t}\psi(\mathbf{r}, t) = -\frac{\hbar^2}{2m}\Delta\psi(\mathbf{r}, t) + V(\mathbf{r}, t)\psi(\mathbf{r}, t) \tag{2.14}$$

and the complex conjugate equation is

$$i\hbar\frac{\partial}{\partial t}\psi^*(\mathbf{r}, t) = -\frac{\hbar^2}{2m}\Delta\psi^*(\mathbf{r}, t) + V(\mathbf{r}, t)\psi^*(\mathbf{r}, t) \tag{2.15}$$

where $V(\mathbf{r}, t)$ is real and $\mathbf{H}$ is Hermitian. Multiply both sides of Eq. (2.14) by $\psi^*(\mathbf{r}, t)$ and both sides of Eq. (2.15) by $-\psi(\mathbf{r}, t)$, and then add both equations to obtain the following:

$$i\hbar\frac{\partial}{\partial t}\psi^*(\mathbf{r}, t)\psi(\mathbf{r}, t) = -\frac{\hbar^2}{2m}[\psi^*(\mathbf{r}, t)\Delta\psi(\mathbf{r}, t) - \psi(\mathbf{r}, t)\Delta\psi^*(\mathbf{r}, t)] \tag{2.16}$$

By substituting Eq. (2.11) into (2.16), we obtain

$$\frac{\partial}{\partial t}\rho(\mathbf{r}, t) + \frac{\hbar}{2mi}[\psi^*(\mathbf{r}, t)\Delta\psi(\mathbf{r}, t) - \psi(\mathbf{r}, t)\Delta\psi^*(\mathbf{r}, t)] = 0 \tag{2.17}$$

If we set

$$\mathbf{J}(\mathbf{r}, t) = \frac{\hbar}{2mi}[\psi^*(\mathbf{r}, t)\nabla\psi(\mathbf{r}, t) - \psi(\mathbf{r}, t)\nabla\psi^*(\mathbf{r}, t)] \tag{2.18}$$

then Eq. (2.17) can be written in the form of Eq. (2.12) since

$$\begin{aligned}
\operatorname{div} \mathbf{J}(\mathbf{r}, t) &= \boldsymbol{\nabla} \cdot \mathbf{J}(\mathbf{r}, t) \\
&= \frac{\hbar}{2mi}\begin{bmatrix} \nabla\psi^*(\mathbf{r}, t) \cdot \nabla\psi(\mathbf{r}, t) + \psi^*(\mathbf{r}, t)\nabla^2\psi(\mathbf{r}, t) \\ -\nabla\psi(\mathbf{r}, t) \cdot \nabla\psi^*(\mathbf{r}, t) - \psi(\mathbf{r}, t)\nabla^2\psi^*(\mathbf{r}, t) \end{bmatrix} \\
&= \frac{\hbar}{2mi}\left[\psi^*(\mathbf{r}, t)\nabla^2\psi(\mathbf{r}, t) - \psi(\mathbf{r}, t)\nabla^2\psi^*(\mathbf{r}, t)\right] \\
&\equiv \text{Second term of Eq. (2.16)}
\end{aligned} \tag{2.19}$$

This proved the equation of local conservation of probabilities, and we have found the expression for the *probability current* in terms of the normalized wave function $\psi(\mathbf{r}, t)$. Hence, if we have a plane wave $\psi(\mathbf{r}, t) = Ae^{ik\cdot\mathbf{x}}$, we can calculate the probability density $\rho(\mathbf{r}, t)$ and the probability current $\mathbf{J}(\mathbf{r}, t)$ such that

$$\rho(\mathbf{r}, t) = |\psi(\mathbf{r}, t)|^2 = |A|^2 \tag{2.20a}$$

and

$$\mathbf{J}(\mathbf{r}, t) = |A|^2 \frac{\hbar k}{m} = \rho(\mathbf{r}, t)\mathbf{V}_g \tag{2.20b}$$

where $\mathbf{V}_g$ is the group velocity obtained with the help of $\hbar\omega = \hbar^2 k^2/2m$.

The objective from the preceding derivation is to show that the *probability current* is proportional to $|A|^2$, and thus the definition of the transmission T and reflection R coefficients depend on the squares of the ratios of Eq. (2.9). In other words, the reflection coefficient can be shown to be

$$R = \left|\frac{B}{A}\right|^2 = \left|\frac{k_1 - k_2}{k_1 + k_2}\right|^2 = 1 - \frac{4k_1k_2}{(k_1 + k_2)^2} \tag{2.21}$$

The transmission coefficient can thus be obtained from $T + R = 1$ or from $T = (k_2/k_1)|C/A|^2$.

The reflection and transmission coefficients are plotted in Fig. 2.2 as a function of the particle energy with a fixed potential height of $V_o = 0.6$ eV. It is clear from this figure that the transmission coefficient

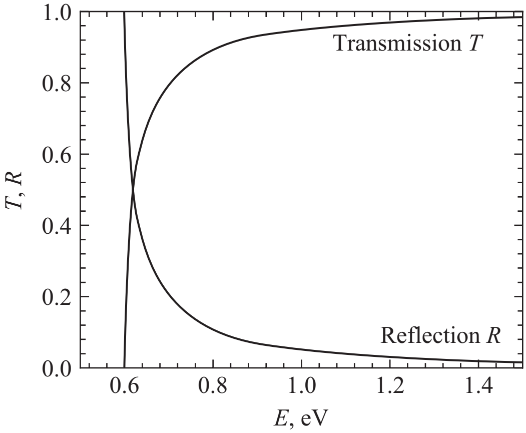


Figure 2.2 The transmission and reflection coefficients plotted as a function of the particle energy E for a barrier height of $V_o = 0.6$ eV.

T is zero for $E < V_o$ and starts to increase as the energy of the particle increases above V_o. The particle is completely transmitted when the energy is way above the barrier. The reflection coefficient R exhibits an opposite behavior since the sum of T and R is unity.

For the case when $E < V_o$, the problem is quite different. Let us assume that the propagation vectors are defined in a similar fashion as in Eq. (2.6). The parameter k_2 is a complex quantity since $E < V_o$ and can be replaced by $k_2 = i\rho_2 = i\sqrt{2m(V_o - E)/\hbar^2}$. The wave function in region II becomes

$$\varphi_{\mathrm{II}}(x) = Ce^{\rho_2 x} + De^{-\rho_2 x} \tag{2.22}$$

For the solution to remain bounded when $x \to \infty$, it is necessary to have the coefficient $C = 0$, reducing the wave function to $\varphi_{\mathrm{II}}(x) = De^{-\rho_2 x}$ while $\varphi_{\mathrm{I}}(x)$ remains the same as in Eq. (2.7*a*). The boundary conditions at $x = 0$ give

$$\frac{B}{A} = \frac{k_1 - i\rho_2}{k_1 + i\rho_2} \qquad \text{and} \qquad \frac{D}{A} = \frac{2k_1}{k_1 + i\rho_2} \tag{2.23}$$

The reflection coefficient can then be given as

$$R = \left| \frac{B}{A}\frac{B^*}{A^*} \right| = \frac{k_1 - i\rho_2}{k_1 + i\rho_2}\frac{k_1 + i\rho_2}{k_1 - i\rho_2} = 1 \qquad \text{and} \qquad T = 0 \tag{2.24}$$

Equation (2.24) shows that we have a total reflection. This effect is demonstrated in Fig. 2.2 where $R \to 1$ and $T \to 0$ when $E \ll V_o$. This is similar to classical mechanics where the particle is always reflected. In quantum mechanics, however, the wave function in region II of Fig. 1.1 is an evanescent wave with the form $e^{-\rho_2 x}$. Thus the particle has a nonzero probability of being in region II, which is decreased as x is increased.

The ratio B/A is complex, and a certain phase shift appears upon reflection, which is due to the fact that the particle is delayed when it penetrates the region $x > 0$. This effect is analogous to the phase shift appearing when light is reflected from a metallic material. The delay time τ will be discussed later in Sec. 2.2.

2.2 Potential Barrier with a Finite Height

Let us now derive the transmission and reflection coefficient for a particle with an energy E larger than the potential barrier height V_o and a width L as shown in Fig. 2.3. First we assume that the particle coming from $x = -\infty$; thus the particle cannot be reflected back in region III and there is only one wavevector associated with the particle in this

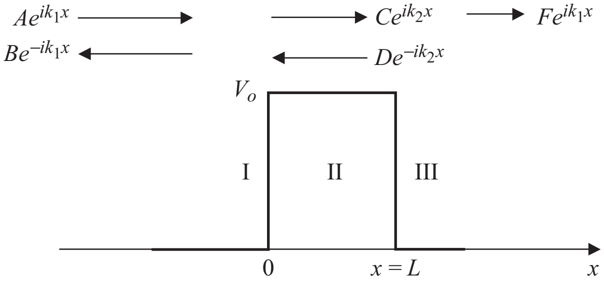


Figure 2.3 The wave function of a particle with energy $E > V_o$ is sketched with a potential barrier of width L and a height V_o.

region. The wave functions for the three regions can be written as

$$\phi_{\mathrm{I}} = Ae^{ik_1x} + Be^{-ik_1x} \tag{2.25a}$$

$$\phi_{\mathrm{II}} = Ce^{ik_2x} + De^{-ik_2x} \tag{2.25b}$$

$$\phi_{\mathrm{III}} = Fe^{ik_1x} \tag{2.25c}$$

where A, B, C, D, and F are complex numbers. By applying the boundary conditions at $x = 0$ and $x = L$, one can obtain the ratios A/F and B/A. The best approach to solve this problem is by finding the coefficients A and B in terms of C and D using the boundary conditions at $x = 0$, finding C and D in terms of F using the boundary conditions at $x = L$, and then relating A and B to F.

Let us start with the boundary conditions at $x = L$:

$$Ce^{ik_2L} + De^{-ik_2L} = Fe^{ik_1L} \tag{2.26a}$$

$$Ck_2e^{ik_2L} - Dk_2e^{-ik_2L} = Fk_1e^{ik_1L} \tag{2.26b}$$

Multiply Eq. (2.26a) by k_2 and add Eq. (2.26a) to (2.26b); then subtract the same two equations, to obtain the following two relations:

$$\frac{C}{F} = \frac{k_1 + k_2}{2k_2}e^{i(k_1-k_2)L} \tag{2.27a}$$

$$\frac{D}{F} = \frac{k_1 - k_2}{2k_2}e^{i(k_1+k_2)L} \tag{2.27b}$$

Similarly, the boundary conditions at $x = 0$ gives

$$A = \frac{k_1 + k_2}{2k_1}C + \frac{k_1 - k_2}{2k_1}D \tag{2.28a}$$

$$B = \frac{k_1 - k_2}{2k_1}C + \frac{k_1 + k_2}{2k_1}D \tag{2.28b}$$

By substituting Eqs. (2.27) into (2.28) and with rearrangement, we obtain

$$\frac{A}{F} = e^{ik_1L}\left[\cos(k_2L) - i\frac{k_1^2 + k_2^2}{2k_1k_2}\sin(k_2L)\right] \tag{2.29a}$$

$$\frac{B}{F} = ie^{ik_1L}\left(\frac{k_2^2 - k_1^2}{2k_1k_2}\right)\sin(k_2L) \tag{2.29b}$$

The transmission coefficient T can be obtained from Eq. (2.29a) as

$$T = \left|\frac{F}{A}\frac{F^*}{A^*}\right| = \frac{4k_1^2k_2^2}{4k_1^2k_2^2 + \left(k_1^2 - k_2^2\right)^2\sin^2(k_2L)} \tag{2.30}$$

Substitute the expressions

$$\begin{aligned} k_1^2 &= \frac{2m(E)}{\hbar^2} \quad \text{for regions I and III,} \qquad \text{and} \\ k_2^2 &= \frac{2m(E - V_o)}{\hbar^2} \quad \text{for region II} \end{aligned} \tag{2.31}$$

and insert in Eq. (2.30); then the transmission coefficient can be rewritten as

$$T = \left|\frac{F}{A}\frac{F^*}{A^*}\right| = \frac{4E(E - V_o)}{4E(E - V_o) + V_0^2\sin^2(\sqrt{(2m(E - V_o)/\hbar^2}L)} \tag{2.32}$$

The reflection coefficient R can be obtained from the relation $T + R = 1$. A plot of both R and T are shown in Fig. 2.4 as a function of the barrier width. This transmission coefficient is also plotted as a function of the particle energy as shown in Fig. 2.5. The transmission coefficient plotted in this figure exhibits oscillations for $E > V_o$. When the sin term is zero, we have $kL = n\pi$, which gives $E = (n\hbar\pi)^2/(2mL^2) + V_o$. The displayed curves for the transmission coefficient as a function of energy and a fixed barrier width show that the number of oscillations is increased as the barrier height is increased. Additionally, the transmission probability is higher for thinner wells.

For the case of $E < V_o$, one can go through the same analysis shown previously to obtain an expression for the transmission coefficient. The final expression of the transmission coefficient is identical to Eq. (2.32) except for the fact that the sin argument is $[-2m(V_o - E)/h^2]^{1/2}L$. This quantity is a complex number. By using the trigonometry relation

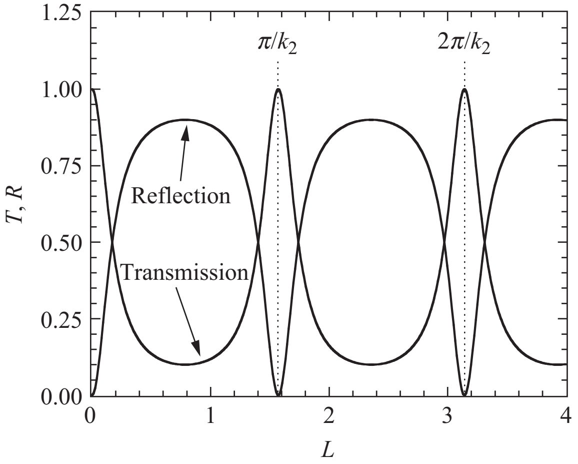


Figure 2.4 The transmission [Eq. (2.32)] and reflection ($R = 1 - T$) coefficients plotted as a function of the barrier width L.

$\sin(ix) = i\sinh(x)$ in Eq. (2.32) we can write the transmission coefficient as follows:

$$T = \left|\frac{F}{A}\frac{F^*}{A^*}\right| = \frac{4E(V_o - E)}{4E(V_o - E) + V_0^2 \sinh^2(\sqrt{2m(V_o - E)/\hbar^2}L)} \qquad (2.33)$$

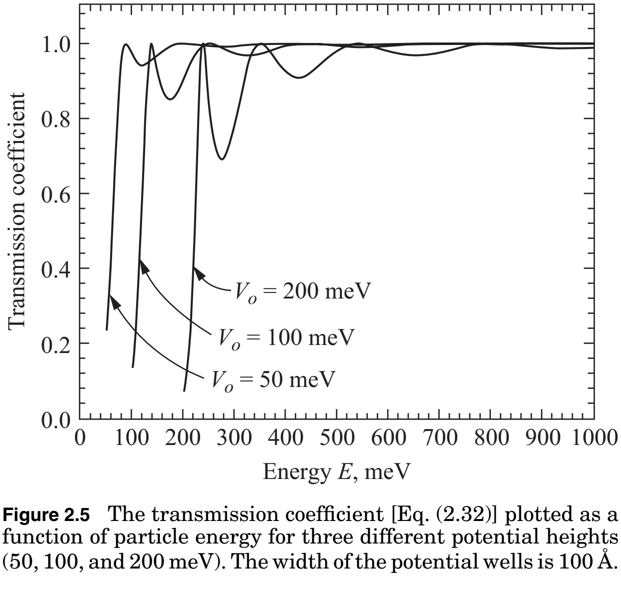


Figure 2.5 The transmission coefficient [Eq. (2.32)] plotted as a function of particle energy for three different potential heights (50, 100, and 200 meV). The width of the potential wells is 100 Å.

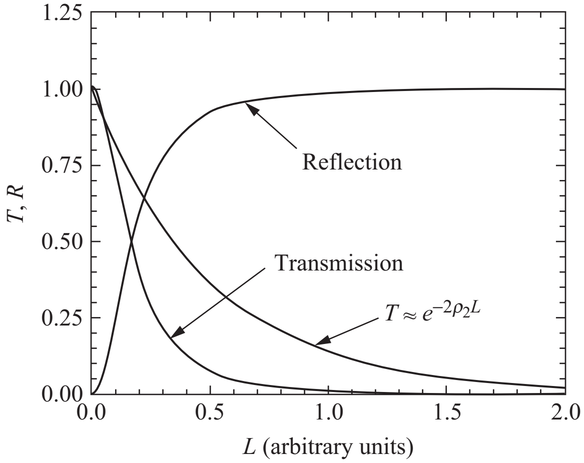


Figure 2.6 The reflection and transmission coefficients plotted as a function of the barrier width L.

A plot of T and R as a function of the barrier width is shown in Fig. 2.6. For $\rho_2 L \gg 1$, where $\rho_2 = [2m(V_o - E)/\hbar^2]^{1/2}$, we have $4E(V_o - E) \ll V_0^2 \sinh^2(\sqrt{2m(V_o - E)/\hbar^2}L)$ and $\sinh^2(\sqrt{2m(V_{-o}E)/\hbar^2}L) \approx \frac{1}{4}e^{2\rho_2 L}$. This leads to the following expression for the transmission coefficient:

$$T \approx \frac{16E(V_o - E)}{V_o^2} e^{-2\rho_2 L} \tag{2.34}$$

This equation is also plotted in Fig. 2.6, where we have shown the transmission coefficient is plotted as $T \approx \exp(-2\rho_2 L)$. It is clear from this figure that the particle penetrates the barrier and the probability of finding the particle at $x > 0$ does exist. This behavior cannot be explained in terms of classical mechanics. The particle has a considerable probability of crossing the barrier by the *tunneling effect*. The evanescent wave has a range of $1/\rho_2$. For a free electron of mass m, this range is $(1/\rho_2) \approx 1.95/\sqrt{V_o - E}$ Å, and for the conduction electron in GaAs with an effective mass of $m^* = 0.067$ m, we have $(1/\rho_2) \approx 7.55/\sqrt{V_o - E}$ Å.

The tunneling of the particle through the barrier is shown schematically in Fig. 2.7 where the wave functions for the regions are shown assuming the particle is traveling from the left to right. The reflected wave functions in regions I and II are not shown. The wave function inside the barrier is a decaying function with the width of the barrier. This effect is due to the fact that $E < V_o$ and the propagation vector $\mathbf{k}_2$ is a complex quantity. Thus the wave function takes the following form: $\phi_{\mathrm{II}} \sim e^{i\mathbf{k}_2 x} = e^{-\rho_2 x}$, where $\mathbf{k}_2 = i\rho_2$.

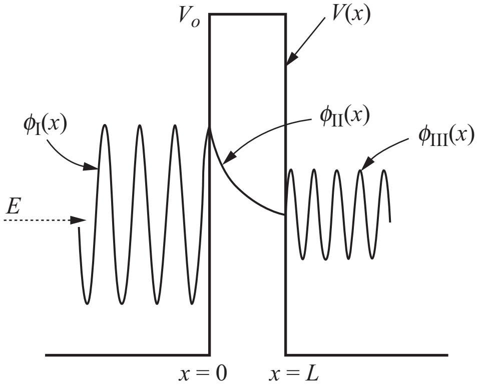


Figure 2.7 A schematic diagram showing the tunneling of a particle through a potential barrier. The wave function inside the barrier is a decaying function.

Since the ratio F/A [see Eq. 2.29a] is a complex number, a certain phase shift appears upon reflection, which is physically due to the fact that the particle is delayed when it penetrates the $x > 0$ region. To demonstrate this kind of delay, we will revisit the step potential presented in Sec. 2.1 to estimate the delay time τ needed for the particle to penetrate the potential and be reflected back to region I. When $E < V_o$, we obtain relations between the wave function coefficients shown in Eq. (2.23). The delay time τ can be obtained from constructing the wave packet of the particle as follows. First, let us introduce the parameter

$$k_o = \sqrt{\frac{2mV_o}{\hbar^2}} = \sqrt{k_1^2 + \rho_2^2} \tag{2.35}$$

We will choose the value of k_1 to be smaller than that of k_o such that the wave packet is formed for the total reflection case. As discussed in Chap. 1, a function $g(k_1)$ is chosen to contain the wave packet characteristics. Thus, this function is zero for $k_1 > k_o$. Our attention now is focused on region I in Fig. 2.1. Let us set $B/A = e^{-i2\theta}$ with $\tan(\theta) = \sqrt{k_o^2 - k_1^2}/k_1$. The wave packet at $t = 0$, for negative x, can be written as

$$\psi(x, 0) = \frac{1}{\sqrt{2\pi}} \int_0^{k_o} g(k)[Ae^{ik_1x} + Be^{ik_1x}]\,dk \tag{2.36}$$

The wave function in the brackets is that presented in Eq. (2.7a). Since the coefficients A and B have the same modulus, we can rewrite

Eq. (2.36) with the help of $B/A = e^{-i2\theta}$ as

$$\psi(x,0) = \frac{1}{\sqrt{2\pi}} \int_{0}^{k_o} g(k)[e^{ik_1x} + e^{ik_1x}e^{-i2\theta}]\,dk \tag{2.37}$$

As discussed in Chap. 1, we assumed that $|g(k)|$ is real and has a pronounced peak with a width of Δk about the value $k = k_1 < k_o$. For $\psi(x,t)$, we use the following general form:

$$\psi(x,t) = \frac{1}{\sqrt{2\pi}} \int_{0}^{k_o} dk\, g(k) e^{i(kx-\omega t)} + \frac{1}{\sqrt{2\pi}} \int_{0}^{k_o} dk\, g(k) e^{-i(kx+\omega t+2\theta)} \tag{2.38}$$

where $\omega = \hbar k_1^2/(2m)$. The first term of Eq. (2.38) represents the incident wave packet, and the second term represents the reflected wave packet. From the argument of the first term and a constant phase condition $[(kx - \omega t) = \text{constant phase}]$, we have after differentiating

$$x_i = t\left[\frac{d\omega}{dk}\right]_{k=k_1} = \frac{\hbar k_1}{m}t \tag{2.39}$$

where x_i is the center of the incident wave packet. Similarly, the center of the reflected wave packet x_r can be obtained by differentiating the argument of the second term in Eq. (2.38) to give

$$x_r = -t\left[\frac{d\omega}{dk} + 2\frac{d\theta}{dk}\right]_{k=k_1} = -\frac{\hbar k_1}{m}t + \frac{2}{\sqrt{k_o^2 - k_1^2}} \tag{2.40}$$

The results in Eq. (2.40) were obtained by differentiating $\tan(\theta) = \sqrt{k_o^2 - k_1^2}/k_1$, where we used the following procedure:

$$\begin{aligned}
[1 + \tan^2(\theta)]\,d\theta &= \left[1 + \frac{k_o^2 - k_1^2}{k_1}\right] d\theta \\
&= -\frac{dk_1}{k_1^2}\sqrt{k_o^2 - k_1^2} - \frac{dk_1}{\sqrt{k_o^2 - k_1^2}}
\end{aligned}$$

or (2.41)

$$\frac{k_0^2}{k_1^2}\,d\theta = -\frac{k_0^2}{k_1^2}\frac{dk_1}{\sqrt{k_o^2 - k_1^2}}$$

At $x_r = 0$, Eq. (2.40) gives the following expression for the delay time τ:

$$\tau = -2\left[\frac{d\theta/dk}{d\omega/dk}\right]_{k=k_1} = \frac{2m}{\hbar k_1\sqrt{k_o^2 - k_1^2}} \tag{2.42}$$

This equation tells us that the particle spends a time on the order of τ in the region $x > 0$ before retracting its steps. For Eq. (2.42) to be valid, k_1 must be larger than zero; i.e., the particle is not at rest, and $k_1 < k_o$, or the particle energy is smaller than the barrier height ($E < V_o$).

2.3 Potential Well with an Infinite Depth

Let us consider a potential well with an infinite depth as shown in Fig. 2.8 where $V(x)$ is zero for $0 < x < L$ and infinite everywhere else. Inside the quantum well, the Schrödinger equation is

$$\frac{d^2}{dx^2}\varphi(x) + \frac{2mE}{\hbar^2}\varphi(x) = 0 \tag{2.43}$$

By setting the propagation vector as $k = \sqrt{2mE/\hbar^2}$, where $E > 0$, the Schrödinger equation can be rewritten as

$$\frac{d^2}{dx^2}\varphi(x) + k^2\varphi(x) = 0 \tag{2.44}$$

The general solution of this equation is

$$\varphi(x) = A\sin(kx) + B\cos(kx) \tag{2.45}$$

The boundary conditions in this case are $\varphi(0) = \varphi(L) = 0$. Thus, B in Eq. (2.45) must be zero, reducing the wave function to $\varphi(x) = A\sin(kx)$.

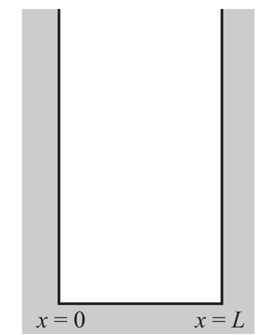


Figure 2.8 A schematic of an infinite potential well.

To have a nontrivial solution at $x = L$, A cannot equal 0, which implies that $\sin(kL) = 0$. Hence,

$$kL = n\pi \qquad n = 1, 2, 3, \ldots \tag{2.46}$$

By substituting the expression of k into Eq. (2.46), one can obtain the eigenvalues as

$$E_n = \frac{\hbar^2 k^2}{2m} = \frac{\hbar^2 \pi^2 n^2}{2mL^2} \qquad n = 1, 2, 3, \ldots \tag{2.47}$$

By normalizing the wave function and substituting for $k = n\pi/L$, we finally can write the wave function as

$$\varphi(x) = \sqrt{\frac{2}{L}} \sin\left(\frac{n\pi}{L}x\right) \tag{2.48}$$

The example of the potential well, when it is defined such that $V_o = 0$ for $-L/2 \le x \le +L/2$, can be solved by performing the transformation $x \to x - L/2$. The wave function becomes

$$\varphi(x) = \sqrt{\frac{2}{L}} \sin\left[\frac{n\pi}{L}\left(x - \frac{L}{2}\right)\right] = \sqrt{\frac{2}{L}} \sin\left(\frac{n\pi}{L}x - \frac{n\pi}{2}\right) \tag{2.49}$$

By expanding the sin function we obtain the following:

$$\sin\left(\frac{n\pi}{L}x - \frac{n\pi}{2}\right) = \sin\left(\frac{n\pi}{L}x\right)\cos\left(\frac{n\pi}{2}\right) - \cos\left(\frac{n\pi}{L}x\right)\sin\left(\frac{n\pi}{2}\right) \tag{2.50}$$

For $n = 1, 3, 5, \ldots$, the wave function is proportional to $\cos\left(\frac{n\pi}{L}x\right)$, which is an even function. For $n = 2, 4, 6, \ldots$, the wave function is proportional to $\sin\left(\frac{n\pi}{L}x\right)$, which is an odd function. Thus, for $-L/2 < x < +L/2$, the wave function is given as

$$\varphi(x) = \begin{cases} \sqrt{\frac{2}{L}} \sin\left(\frac{n\pi}{L}x\right) & \text{for } n = 2, 4, 6, \ldots \\ \sqrt{\frac{2}{L}} \cos\left(\frac{n\pi}{L}x\right) & \text{for } n = 1, 3, 5, \ldots \end{cases} \tag{2.51}$$

The energy levels and the probability functions of a particle in an infinite potential well are shown in Fig. 2.9. The energy levels are proportional to n^2 as indicated in Eq. (2.47).

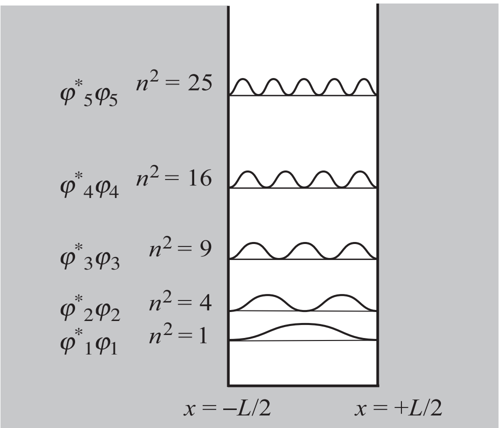


Figure 2.9 Energy levels $E_n = \hbar^2\pi^2 n^2/(2mL^2)$ of a particle in an infinite potential well plotted along with the probability $\phi^*(x)\phi(x)$ using the expressions shown in Eq. (2.51).

2.4 Finite-Depth Potential Well

A one-dimensional finite-depth potential well is illustrated in Fig. 2.10 where the potential is defined as

$$V(x) = \begin{cases} V_o & \text{for } -\dfrac{L}{2} < x < +\dfrac{L}{2} \\ 0 & \text{for } x < -\dfrac{L}{2} \quad \text{and} \quad x > +\dfrac{L}{2} \end{cases} \tag{2.52}$$

The wave functions for the case where $-V_o < E < 0$ can be chosen as

$$\phi_{\mathrm{I}} = Ae^{\rho x} + Be^{-\rho x} \tag{2.53a}$$

$$\phi_{\mathrm{II}} = Ce^{ikx} + De^{-ikx} \tag{2.53b}$$

$$\phi_{\mathrm{III}} = Fe^{\rho x} + Ge^{-\rho x} \tag{2.53c}$$

where $k = \sqrt{2mE/\hbar^2}$ and $\rho = \sqrt{2m(V_o - E)/\hbar^2}$. To simplify the analysis, we can assume that the particle is bound in the well so that $B = 0$.

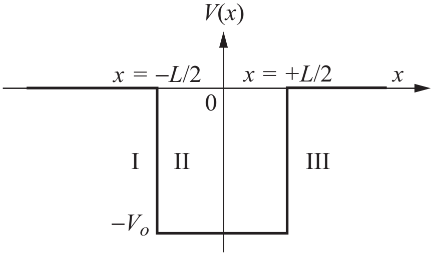


Figure 2.10 Finite-depth potential well plotted for a particle with $-V_o < E < 0$.

By imposing the boundary conditions at $x = -L/2$, the relations between A, C, and D are obtained as follows:

$$\begin{aligned} C &= e^{(-\rho+ik)L/2}\frac{\rho+ik}{2ik}A \\ D &= e^{-(\rho+ik)L/2}\frac{\rho-ik}{2ik}A \end{aligned} \tag{2.54}$$

With the help of Eq. (2.54) and the matching conditions at $x = +L/2$ we can obtain the relations between A, F, and G according to the following:

$$\begin{aligned} \frac{F}{A} &= e^{-\rho L/2}[(\rho+ik)^2 e^{ikL} - (\rho-ik)^2 e^{-ikL}] \\ \frac{G}{A} &= \frac{\rho^2+k^2}{2k\rho}\sin(kL) \end{aligned} \tag{2.55}$$

We still cannot obtain meaningful results without an additional assumption. Since the particle is bound in the well, one would find it necessary to set $F = 0$. Thus, the first equality in Eq. (2.55) leads to the following relation:

$$\left(\frac{\rho-ik}{\rho+ik}\right)^2 = e^{2ikL} \tag{2.56}$$

Since ρ and k depend on E, Eq. (2.56) can only be satisfied for certain values of E. In solving this problem, we consider the two possible cases for the following relation:

$$\frac{\rho-ik}{\rho+ik} = \pm e^{ikL} \tag{2.57}$$

The first case is when $(\rho-ik)/(\rho+ik) = -e^{ikL}$, which yields

$$\frac{\rho}{k} = \tan\left(\frac{kL}{2}\right) \tag{2.58a}$$

Let us define k_o such as $k_o^2 = \rho^2 + k^2 = 2mV_o/\hbar^2$, which leads to

$$\frac{1}{\cos^2(kL/2)} = 1 + \tan^2\left(\frac{kL}{2}\right) = \frac{\rho^2+k^2}{k^2} = \frac{k_o^2}{k^2} \tag{2.58b}$$

Equation (2.58*b*) is equivalent to the following set of solutions:

$$\left|\cos\left(\frac{kL}{2}\right)\right| = \frac{k}{k_o} \tag{2.59}$$

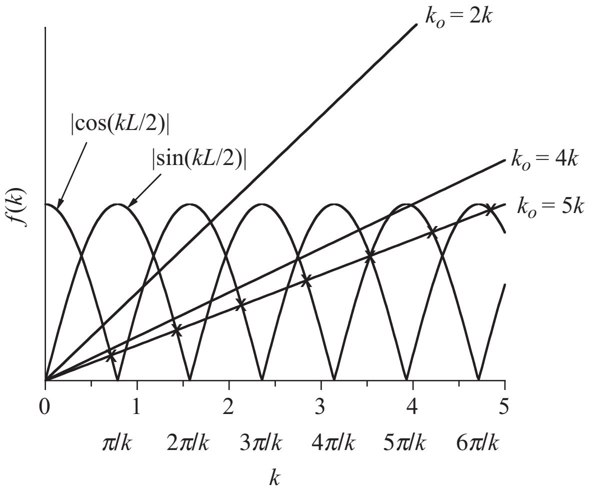


Figure 2.11 Graphic solutions of Eqs. (2.59) and (2.60) giving the bound states of a particle in a finite-depth potential well.

The second case is when $(\rho - ik)/(\rho + ik) = +e^{ikL}$. By following this procedure, one can reach the following equation:

$$\left|\sin\left(\frac{kL}{2}\right)\right| = \frac{k}{k_o} \tag{2.60}$$

It is difficult to solve Eqs. (2.59) and (2.60); however, a graphic solution is possible. This solution is shown in Fig. 2.11, where the values of k_o are taken in units of k. From this figure one can calculate the values of the energy levels from the intersections of k_o and the sin and cos curves. The intersections give the values of k from which the energy level values are calculated. The number of confined states in the well can be obtained from the number of intersections that the straight line makes with the curves. For example, when k_o is $5k$ we have four even states and three odd states. For the line marked $k_o = 2k$, we have three states (two even and one odd). This figure indicates that there will be at least one bound state in the potential well.

The finite-depth well potential problems can be solved in a different way, as discussed in many textbooks. Let us assume that the wave functions for the three regions in Fig. 2.12 have the following forms:

$$\phi_{\mathrm{I}} = Ae^{\rho x} \tag{2.61a}$$

$$\phi_{\mathrm{II}} = B\sin(kx) + C\cos(kx) \tag{2.61b}$$

$$\phi_{\mathrm{III}} = De^{-\rho x} \tag{2.61c}$$

where $k = \sqrt{2mE/\hbar^2}$ and $\rho = \sqrt{2m(V_o - E)/\hbar^2}$.

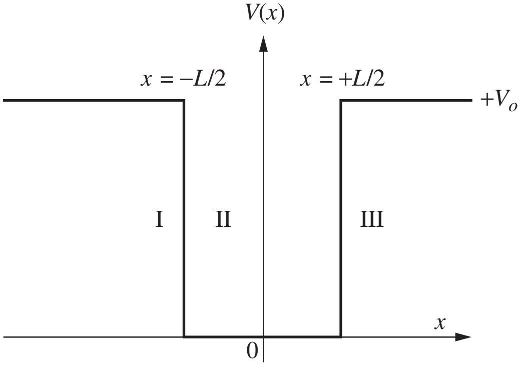


Figure 2.12 A sketch of a finite-depth potential well plotted for a particle with $0 < E < V_o$.

From the boundary conditions at $x = -L/2$ and $x = +L/2$ we have the following set of equations:

$$Ae^{-\rho L/2} + B\sin\left(\frac{kL}{2}\right) - C\cos\left(\frac{kL}{2}\right) = 0 \tag{2.62a}$$

$$A\rho e^{-\rho L/2} - Bk\cos\left(\frac{kL}{2}\right) - Ck\sin\left(\frac{kL}{2}\right) = 0 \tag{2.62b}$$

$$B\sin\left(\frac{kL}{2}\right) + C\cos\left(\frac{kL}{2}\right) - De^{-\rho L/2} = 0 \tag{2.62c}$$

$$Bk\cos\left(\frac{kL}{2}\right) - Ck\sin\left(\frac{kL}{2}\right) + D\rho e^{-\rho L/2} = 0 \tag{2.62d}$$

This is a system of four homogeneous linear equations for the coefficients A, B, C, and D. For the nontrivial solution we must set the determinant of the coefficients to zero. While the values of coefficients are arbitrary, one can solve for their ratios. The determinant of the coefficients can now be written as

$$\begin{vmatrix} 1 & \sin\left(\frac{kL}{2}\right) & \cos\left(\frac{kL}{2}\right) & 0 \\ \rho & -k\cos\left(\frac{kL}{2}\right) & -k\sin\left(\frac{kL}{2}\right) & 0 \\ 0 & \sin\left(\frac{kL}{2}\right) & \cos\left(\frac{kL}{2}\right) & -1 \\ & k\cos\left(\frac{kL}{2}\right) & -k\sin\left(\frac{kL}{2}\right) & \rho \end{vmatrix} e^{-\rho L/2} = 0 \tag{2.63}$$

This determinant can be expanded in minors to give

$$\left[k\sin\left(\frac{kL}{2}\right)-\rho\cos\left(\frac{kL}{2}\right)\right]\left[k\cos\left(\frac{kL}{2}\right)+\rho\sin\left(\frac{kL}{2}\right)\right]=0 \qquad (2.64)$$

Divide by $\cos^2(kL/2)$ to obtain

$$\left[k\tan\left(\frac{kL}{2}\right)-\rho\right]\left[\rho\tan\left(\frac{kL}{2}\right)+k\right]=0 \qquad (2.65)$$

By introducing $k_o^2=\rho^2+k^2=2mV_o/\hbar^2$ and by knowing that Eq. (2.65) can vanish by setting either of the quantities in the parentheses to zero, one can obtain

$$k\tan\left(\frac{kL}{2}\right)=\sqrt{k_o^2-k^2} \quad \text{and} \quad -k\cot\left(\frac{kL}{2}\right)=\sqrt{k_o^2-k^2} \qquad (2.66)$$

Once again, these equations can be solved graphically to obtain k values from which the energy eigenvalues can be determined. To find the values of k, Eq. (2.66) is plotted in Fig. 2.13 for both expressions along with the propagation vector $\rho=\sqrt{k_o^2-k^2}$. The intersections of ρ with the functions $\tan(kL/2)$ and $\cot(kL/2)$ are shown as crosses for three

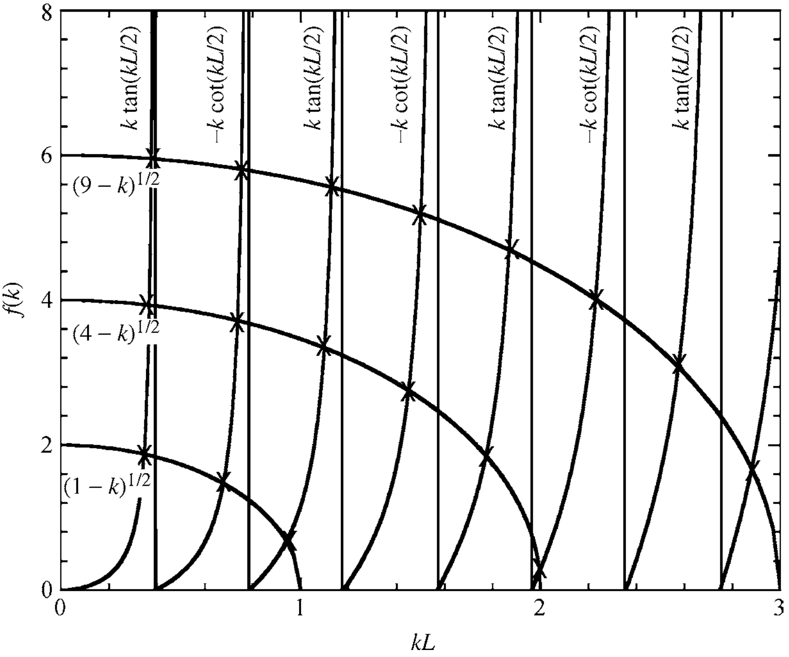


Figure 2.13 A plot of the expressions given in Eq. (2.66) as a function of kL. The intersections of $(k_o - k)^{1/2}$ with the tan and cot functions determine the values of k and the number of bound states.

values of ρ. While Figs. 2.11 and 2.13 are the graphical solutions for a finite-depth potential well, we noted that the number of bound states in Fig. 2.13 is twice as many as the number of bound states obtained from Fig. 2.11. This is due to the fact that the x axis is plotted as kL instead of $kL/2$. But the number of the bound energy levels in the quantum well is the same for both of the cases discussed.

2.5 Unbound Motion of a Particle ($E > V_o$) in a Potential Well with a Finite Depth

Consider the potential well shown in Fig. 2.14 and consider that a particle is traveling from the left ($x = -\infty$) to the right ($x = +\infty$) with energy $E > V_o$. The propagation vectors are given by

$$\rho = \sqrt{\frac{2mE}{\hbar^2}} \qquad \text{and} \qquad k = \sqrt{\frac{2m(V_o + E)}{\hbar^2}} \tag{2.67}$$

and the wave functions are constructed for the three regions according to Bastard (1988) and given as

$$\phi_{\rm I} = e^{i\rho(x+L/2)} + r e^{-i\rho(x+L/2)} \tag{2.68a}$$

$$\phi_{\rm II} = \alpha e^{ikx} + \beta e^{-ikx} \tag{2.68b}$$

$$\phi_{\rm III} = t e^{\rho(x-L/2)} \tag{2.68c}$$

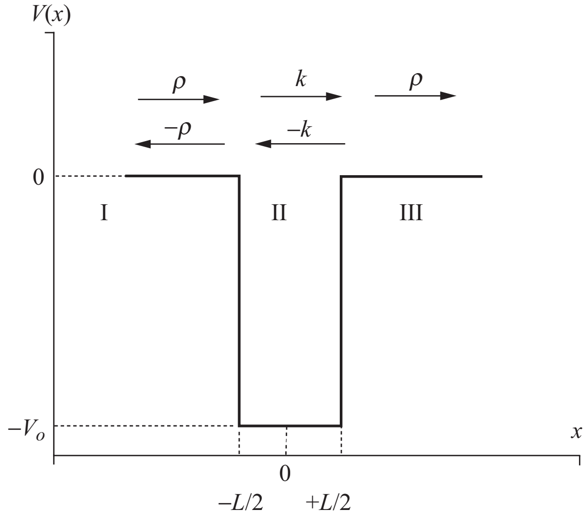


Figure 2.14 A schematic presentation of a potential well with a particle with energy $E > 0$ plotted along the x axis. The potential well $V(x) = -V_o$ for $-L/2 < x < +L/2$.

From the boundary conditions at $x = -L/2$ we have

$$1 + r = \alpha e^{-ikL/2} + \beta e^{ikL/2} \tag{2.69a}$$

$$\rho - \rho r = \alpha k e^{-ikL/2} - \beta k e^{ikL/2} \tag{2.69b}$$

and the boundary conditions at $x = +L/2$ give

$$t = \alpha e^{ikL/2} + \beta e^{-ikL/2} \tag{2.70a}$$

$$\rho t = \alpha k e^{ikL/2} - \beta k e^{-ikL/2} \tag{2.70b}$$

By multiplying Eq. (2.69*a*) by ρ and adding to Eq. (2.69*b*) we obtain

$$2\rho = \alpha(k + \rho)e^{-ikL/2} - \beta(k - \rho)e^{ikL/2} \tag{2.71a}$$

and multiplying Eq. (2.70*a*) by ρ and subtracting it from Eq. (2.70*b*) we have

$$0 = \alpha(k - \rho)e^{ikL/2} - \beta(k + \rho)e^{-ikL/2} \qquad \text{or} \qquad \beta = \alpha\frac{k - \rho}{k + \rho}e^{ikL} \tag{2.71b}$$

By dividing Eq. (2.71*a*) by $(k - \rho)e^{ikL/2}$ and utilizing Eq. (2.71*b*), we can obtain the following:

$$\begin{aligned} \alpha &= \frac{\rho(k + \rho)e^{-ikL/2}}{2k\rho\cos(kL) - i(k^2 + \rho^2)\sin(kL)} \\ \beta &= \frac{\rho(k - \rho)e^{ikL/2}}{2k\rho\cos(kL) - i(k^2 + \rho^2)\sin(kL)} \end{aligned} \tag{2.72a}$$

Substituting Eq. (2.72*a*) into (2.70*a*) and (2.69*a*) yields

$$t = \frac{1}{\cos(kL) - \frac{1}{2}i(k/\rho + \rho/k)\sin(kL)}$$

and (2.72*b*)

$$r = \frac{i/2(k/\rho - \rho/k)\sin(kL)}{\cos(kL) - \frac{1}{2}i(k/\rho + \rho/k)\sin(kL)}$$

Let the transmission coefficient be $T(E) = |t(E)|^2$ and the reflection coefficient be $R(E) = |r(E)|^2$, where $T(E) + R(E) = 1$, then

$$T(E) = \frac{1}{1 + \frac{1}{4}(k/\rho - \rho/k)^2\sin^2(kL)}$$

and (2.73)

$$R(E) = \frac{(k/\rho - \rho/k)^2\sin^2(kL)}{4 + (k/\rho - \rho/k)^2\sin^2(kL)}$$

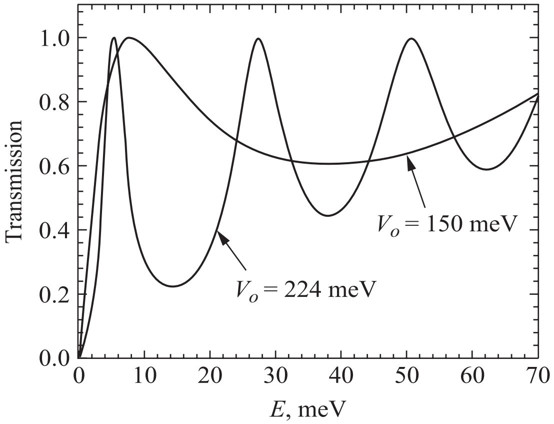


Figure 2.15 The transmission coefficient $[T(E)]$ of an unbound particle $(E > 0)$ plotted as a function of energy for a optional well with a height of 224 meV and thickness of 250 Å. $T(E)$ is also plotted for a similar potential well of height 150 meV.

The transmission coefficient in this equation is a function of energy as plotted in Fig. 2.15. It exhibits an oscillatory behavior as the energy of the particle increases. It reaches unity when $\sin(kL) = 0$ or $kL = n\pi$, where n is an integer. The form of $T(E)$ corresponds to constructive interference inside the potential well. The discrete energies that fulfill the condition $kL = n\pi$ are called transmission resonances (see Bastard 1988). They correspond to an enhanced probability of finding the particle inside the quantum well.

2.6 Triangular Potential Well

Another important potential well is the triangular quantum well. This type of well is common at the semiconductor interfaces such as the GaAs/AlGaAs heterojunction. In particular, the high electron mobility transistor (HEMT) is based on the energy quantization in the triangular well formed at the heterojunction interface. A schematic representation of the conduction band edge of the GaAs/AlGaAs HEMT structure is shown in Fig. 2.16. The space W is an undoped barrier region. If N is the number of electrons transferred to the well per unit area (known as the two-dimensional electron gas), the electric field $\mathcal{E}_s$ is given by Gauss's law as

$$\mathcal{E}_s = \frac{eN}{\epsilon_o \epsilon} \tag{2.74a}$$

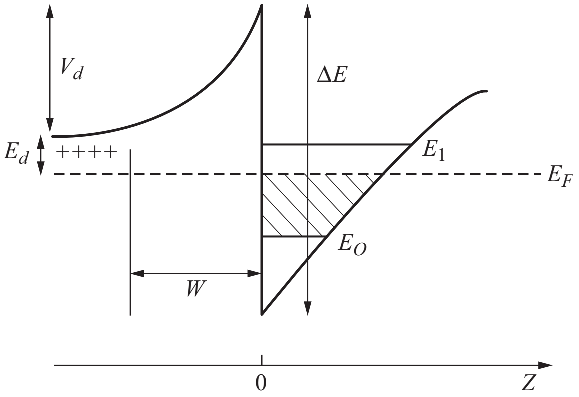


Figure 2.16 A schematic plot of the conduction band of a HEMT structure.

where ϵ is the dielectric constant of the well material. For a triangular well, as shown in Fig. 2.16, the electrostatic potential $\varphi(z)$ is linear in the $z > 0$ region and is given by

$$\varphi(z) = -\mathcal{E}_s z \tag{2.74b}$$

The Hamiltonian for an electron in the triangular well, assuming that the potential barrier is infinite at $z = 0$, can be written as

$$\mathbf{H} = -\frac{\hbar^2}{2m}\frac{d^2}{dz^2} + V_p(z) + e\varphi(z) \tag{2.75}$$

where $V_p(z)$ is a periodic potential energy. Using the envelop function approximation (this approximation will be discussed in more detail in upcoming chapters) for a one-dimensional system, the wave function can be written as

$$\psi(z) = F(z)U(z) \tag{2.76}$$

where $U(z)$ is the conduction band Bloch function for a zero wavevector and $F(z)$ is the envelop function that satisfies the effective mass Schrödinger equation

$$\left[-\frac{\hbar^2}{2m^*}\frac{d^2}{dz^2} + e\varphi(z)\right]F(z) = E_n F(z) \tag{2.77}$$

The index n identifies the eigenvalues, and m^* is the conduction electron effective mass of the well material. The wave function $F(z)$ can be further written as

$$F(z) = e^{ik_z z}\chi_n(z) \tag{2.78}$$

where k_z is the two-dimensional wavevector perpendicular to the surface normal, and $\chi(z)$ satisfies the equation

$$\left[-\frac{\hbar^2}{2m^*}\frac{d^2}{dz^2} + e\varphi(z)\right]\chi(z) = E_n\chi(z) \tag{2.79}$$

and

$$E_n = E - \frac{\hbar^2 k_z^2}{2m^*} \tag{2.80}$$

where E is the total energy eigenvalues of the carriers. The boundary conditions to be satisfied by $\chi(z)$ are $\chi_n(0) = \chi_n(\infty) = 0$. A solution that satisfies the boundary condition at infinity is the Airy function (see Stern 1972, Balanski and Wallis 2000, and Ferry 2001) given by $\mathrm{Ai}[(2m^*/\hbar^2 e^2 \mathcal{E}_s^2)^{1/3}(e\mathcal{E}_s z - En)]$. The boundary condition at $z = 0$ determines the allowed values of E_n as

$$E_n = -\left(\frac{\hbar^2 e^2 \mathcal{E}_s^2}{2m^*}\right)^{1/3} a_n \tag{2.81}$$

The quantity a_n is the zero of the Airy function and is approximated as (see Stern 1972)

$$a_n \approx -\left[\frac{3\pi}{2}\left(n + \frac{3}{4}\right)\right]^{2/3} \tag{2.82}$$

where $n = 0, 1, 2, \ldots$ The values of E_n are then

$$E_n = \left(\frac{\hbar^2}{2m^*}\right)^{1/3}\left[\frac{3\pi e\mathcal{E}_s}{2}\left(n + \frac{3}{4}\right)\right]^{2/3} \quad \text{with}$$

$$E_o \approx \left(\frac{\hbar^2}{2m^*}\right)^{1/3}\left(\frac{9\pi e^2 N}{8\epsilon_o \epsilon}\right)^{2/3} \tag{2.83}$$

The triangular potential is a very good approximation for the potential distribution near the semiconductor interfaces. The quantity E_n in Eq. (2.83) is obtained as a function of the quantum number n, which represents the energy levels in the approximation of an infinite triangular quantum well.

2.7 Parabolic Potential Well

Another example of this extremely important class of one-dimensional bound-states in quantum mechanics is the simple harmonic oscillator where the potential can be written as

$$V(x) = \frac{1}{2}Kx^2 \tag{2.84}$$

where K is the force constant of the oscillator. The Hamiltonian for this potential is given by

$$\mathbf{H} = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + \frac{1}{2}Kx^2 \tag{2.85}$$

The Schrödinger equation which gives the possible energies of the oscillator is

$$-\frac{\hbar^2}{2m}\frac{\partial^2\varphi(x)}{\partial x^2} + \frac{1}{2}Kx^2\varphi(x) = E_n\varphi(x) \tag{2.86}$$

This equation can be simplified by choosing a new measure length and a new measure of energy, each of which is dimensionless. $\zeta \equiv (mK/\hbar^2)^{1/4}x$ and $\eta = 2E_n/(\hbar\omega)$, where $\omega = \sqrt{K/m}$. With these substitutions, Eq. (2.86) becomes

$$\frac{d^2\varphi(\zeta)}{dx^2} + (\eta - \zeta^2)\,\varphi(\zeta) = 0 \tag{2.87}$$

In looking for bounded solutions, one can notice that as ζ approaches infinity, η gecomes too small compared to ζ^2. The resulting differential equation can thus be easily solved to yield an asymptotic solution in the following form:

$$\varphi(\zeta) \approx e^{\pm(1/2)\zeta^2} \tag{2.88}$$

This expression for the asymptotic dependence is suitable only for the negative sign in the exponent. It is clear that because of the very rapid decay of the resulting gaussian function as ζ goes to infinity, the function will still have the same asymptotic dependence multiplied by any finite polynomial in ζ (see, for example, Dicke and Wittke 1960):

$$\varphi(\zeta) = H(\zeta)e^{-(1/2)\zeta^2} \tag{2.89}$$

where $H(\zeta)$ is a finite polynomial. By substituting Eq. (2.89) into (2.87) one can obtain

$$\frac{d^2H(\zeta)}{dx^2} - 2\zeta\frac{dH(\zeta)}{dx} + (\eta - 1)H(\zeta) = 0 \tag{2.90}$$

If we assume a solution to this equation in the form of a finite polynomial such that

$$H(\zeta) = A_0 + A_1\zeta + A_2\zeta^2 + \cdots + A_n\zeta^n \tag{2.91}$$

a recursion formula connecting the coefficients can be obtained in the following form:

$$A_{n+2} = \frac{2n+1-\eta}{(n+2)(n+1)}A_n \qquad \text{for} \qquad n \geq 0 \tag{2.92}$$

For an upper cutoff to the coefficients so that the polynomial $H(\zeta)$ remains finite, the condition

$$\eta = 2n + 1 \tag{2.93}$$

must be satisfied. Substitute $\eta = 2E_n/(\hbar\omega)$ into Eq. (2.93), we obtain

$$E_n = \left(n + \frac{1}{2}\right)\hbar\omega \tag{2.94}$$

The energy levels described by this equation and the parabolic potential are shown in Fig. 2.17, where the energy levels are evenly spaced by the amount of $\hbar\omega$.

The polynomial solutions lead to wave functions that approach zero at $x = \pm\infty$, which can all be normalized. These polynomials are called *Hermite* polynomials, and they are the acceptable solutions as wave functions. The Hermite polynomial is defined as follows:

$$H_n(\zeta) = (-1)^n e^{\zeta^2}\frac{d^n}{dx^n}(e^{-\zeta^2}) \tag{2.95}$$

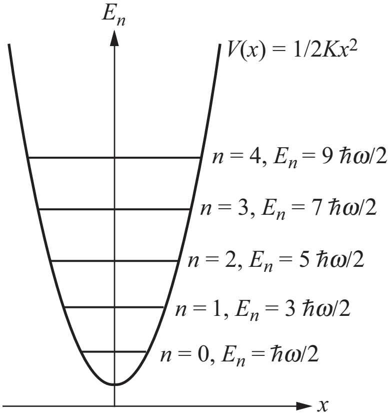

E_n
$V(x) = 1/2Kx^2$
$n = 4, E_n = 9\,\hbar\omega/2$
$n = 3, E_n = 7\,\hbar\omega/2$
$n = 2, E_n = 5\,\hbar\omega/2$
$n = 1, E_n = 3\,\hbar\omega/2$
$n = 0, E_n = \hbar\omega/2$
x

Figure 2.17 A parabolic one-dimensional potential well with a few of the allowed energy levels shown.

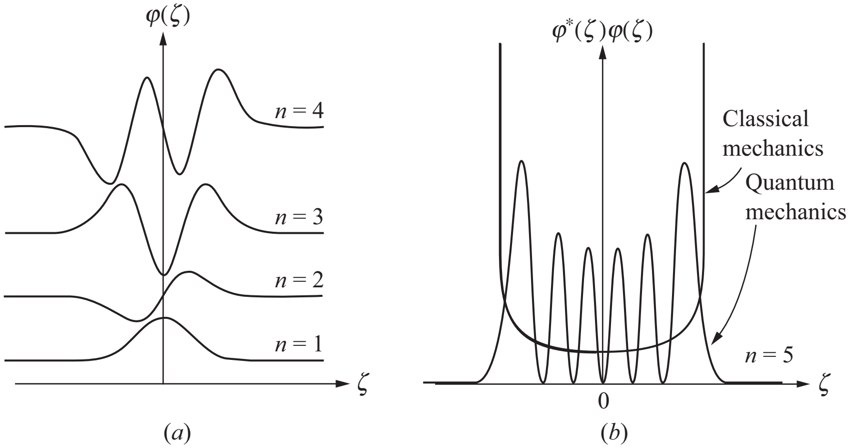


Figure 2.18 (a) The lowest four wave functions of the simple harmonic oscillator are plotted as a function of coordinate ζ. (b) The probability amplitude for $n = 5$ is shown along the classical probability density of the simple harmonic oscillator.

Finally, the wave function can be written as

$$\varphi(\zeta) = N_n H_n(\zeta) e^{-(1/2)\zeta^2} = N_n(-1)^n e^{(1/2)\zeta^2} \frac{d^n}{dx^n}(e^{-\zeta^2}) \tag{2.96}$$

The normalization factor N_n can be found to be

$$N_n = \frac{1}{2^n n!}\sqrt{\frac{\alpha}{\pi}} \qquad \text{where } \alpha = \left(\frac{mK}{\hbar^2}\right)^{1/2} \tag{2.97}$$

The lowest four wave functions are illustrated in Fig. 2.18a. The probability amplitude for the $n = 5$ eigenstate is shown in Fig. 2.18b along with the classical probability density. The first few Hermite polynomial functions are shown in the following table.

N	$H_n(\zeta)$
0	1
1	2ζ
2	$4\zeta^2 - 2$
3	$8\zeta^3 - 12\zeta$
4	$16\zeta^4 - 48\zeta^2 + 12$
5	$32\zeta^5 - 160\zeta^3 + 120\zeta$
6	$64\zeta^6 - 480\zeta^4 + 720\zeta^2 - 120$

Figure 2.18b shows several oscillations in the $\varphi^*(\zeta)\varphi(\zeta)$ curve with their amplitudes fairly small near the origin and considerably larger near the end of the curve. As n increases, the probability density is becoming

larger and larger near the end of the curves and smaller and smaller near the center of the curve approaching the classical limit. According to classical mechanics (see, for example, McKelvey 1993), the probability of finding a particle in an interval of dx is proportional to the time dt it spends in the interval. This is in turn directly related to the velocity by $dx = \upsilon_x dt$. More precisely, if $T/2$ is the half-period of oscillation, the function of time spent in dx is $dt/(T/2)$ or $2dx/(T\upsilon_x)$. This fraction is the classical analog of the probability density $\varphi^*(\zeta)\varphi(\zeta)$. For a classical harmonic oscillator, conservation of energy requires that the total energy E is

$$E = \frac{1}{2}m\upsilon_x^2 + \frac{1}{2}m\omega^2 x^2 = \frac{1}{2}m\omega^2 A^2 \tag{2.98}$$

Solving for υ_x we obtain $\upsilon_x = \omega(A^2 - x^2)^{1/2}$. The classical analog to $\varphi^*(\zeta)\varphi(\zeta)$ can then be written as $p(x)\,dx$, where

$$p(x)\,dx = \frac{2dx}{T\upsilon_x} = \frac{dx}{\pi\sqrt{A^2 - x^2}} \tag{2.99}$$

If the quantum energy is given by Eq. (2.94), then with the help of Eq. (2.98) one can write the classical probability as

$$p(x) = \frac{2dx}{T\upsilon_x} = \frac{dx}{\pi\sqrt{(2n+1)/\alpha - x^2}} \tag{2.100}$$

where α is defined in Eq. (2.97). The classical probability is plotted in Fig. 2.18*b* along with the quantum probability.

2.8 Delta-Function Potentials

The δ-function problem will be discussed in this section for one particular reason. The current technology in optoelectronics is gravitating toward semiconductor nanostructures. The recent research is focused on the use of quantum dots for lasers and detectors. Quantum dots are a small collection of semiconductor atoms such as InAs sandwiched between GaAs barrier materials. The quantum dots are sometimes called the designer atoms. In other words, the quantum dots could be represented or approximated by δ-function wells. A few δ-function characteristics were discussed in Sec. 1.8.6. First let us find the Fourier transform of a step function similar to the potential step shown in Fig. 2.1. Let us define a potential $V(x)$ such as

$$V(x) = -\frac{\hbar^2\lambda}{2ma}\delta(x) \tag{2.101}$$

where a is a quantity with a dimension of length and λ is a dimensionless quantity introduced to characterize the strength of the δ-function

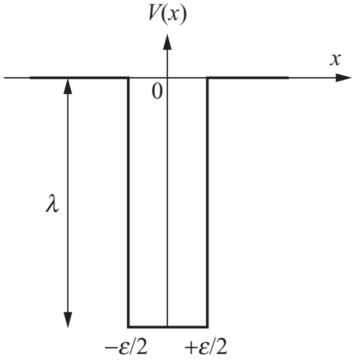


Figure 2.19 A sketch of a δ-function with a width of a small quantity ε and a strength of λ.

(see, for example, Gasiorowicz 2003). The δ-function well is shown in Fig. 2.19. The Schrödinger equation can be written as

$$-\frac{\hbar^2}{2m}\frac{d^2u(x)}{dx^2}-\frac{\hbar^2\lambda}{2ma}\delta(x)u(x)=-\left|E_n\right|u(x) \tag{2.102}$$

By integrating Eq. (2.102) we can obtain the condition at $x = 0$ such as

$$-\int_{-\varepsilon}^{+\varepsilon}\frac{d^2u(x)}{dx^2}\,dx-\frac{\lambda}{a}\int_{-\varepsilon}^{+\varepsilon}\delta(x)u(x)\,dx=-\frac{2m}{\hbar^2}\left|E_n\right|\int_{-\varepsilon}^{+\varepsilon}u(x)\,dx \tag{2.103}$$

The right-hand side of this equation is zero, since if we choose as an example $u(x) = Ae^{\mathbf{k}x}$, the right-hand side can then be proportional to $\sinh(\varepsilon)$ and $\lim_{\varepsilon\to 0}\sinh(\varepsilon) \to 0$. With the help of Eq. (1.83), one can rewrite Eq. (2.103) as

$$\left.\frac{du(x)}{dx}\right|_{-\varepsilon}^{+\varepsilon}=-\frac{\lambda}{a}u(o)$$

or (2.104)

$$\left.\frac{du(x)}{dx}\right|_{x=+\varepsilon}-\left.\frac{du(x)}{dx}\right|_{x=-\varepsilon}=-\frac{\lambda}{a}u(o)$$

For the solution of Eq. (2.102) at $x \neq 0$, we have

$$\frac{d^2u(x)}{dx^2}-\frac{2m}{\hbar^2}\left|E_n\right|u(x)=0$$

or (2.105)

$$\frac{d^2u(x)}{dx^2}-k^2u(x)=0$$

where $k = \sqrt{2m\,|E_n|/\hbar^2}$. The solution that satisfies Eq. (2.104) at all x values except for $x = 0$ and vanishes at $x = \pm\infty$ is

$$u(x) = \begin{cases} e^{-\mathbf{k}x} & \text{for } x > 0 \\ e^{\mathbf{k}x} & \text{for } x < 0 \end{cases} \tag{2.106}$$

The amplitude of $u(x)$ is the same by symmetry for $x > 0$ and $x < 0$, and for simplicity, it was chosen as unity. From Eqs. (2.104) and (2.106) one can find that

$$-k - k = -\frac{\lambda}{a} \qquad \text{or} \qquad 2k = \frac{\lambda}{a} \tag{2.107}$$

Substituting the value of **k** into Eq. (2.107), we find the energy to be

$$E = \frac{\hbar^2\lambda^2}{8ma^2} \tag{2.108}$$

which means that there is only one bound state in the δ-function potential well. This is in many ways similar to small-size semiconductor quantum dots, where each quantum dot has only one bound state. The situation is different as the quantum dot size is increased beyond approximately five monolayers.

A more interesting problem is the double narrow-deep δ-function potential well shown in Fig. 2.20. The potential can be written as

$$V(x) = -\frac{\hbar^2\lambda}{2ma}\,[\delta(x-a) + \delta(x+a)] \tag{2.109}$$

The potential is symmetric under the interchange $x \to -x$; therefore, the solutions have definite parity. Let us consider both even- and odd-parity solutions.

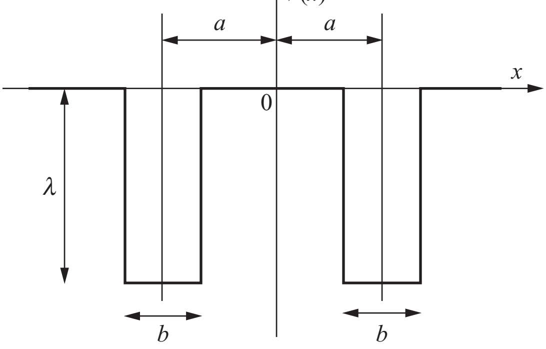


Figure 2.20 Double narrow-deep δ-function potential.

Even-parity solution. Let us consider the following wave function that satisfies the even parity:

$$u(x) = \begin{cases} e^{-kx} & \text{for } x > a \\ A\cosh(kx) & \text{for } a < x < -a \\ e^{kx} & \text{for } x < -a \end{cases} \tag{2.110}$$

By applying the boundary conditions at $x = a$ and with the help of Eq. (2.104), we obtain the following relations:

$$e^{-ka} = A\cosh(ka) \tag{2.111}$$

and

$$-ke^{-ka} - kA\sinh(ka) = -\frac{\lambda}{a}e^{-ka} \tag{2.112}$$

The constant A can be eliminated by combining Eqs. (2.111) and (2.112) to yield

$$\tanh(ka) = \frac{\lambda}{ka} - 1 \tag{2.113}$$

Equation (2.113) can be rewritten as

$$\frac{e^{ka} - e^{-ka}}{e^{ka} + e^{-ka}} = \frac{\lambda}{ka} - 1 \tag{2.114a}$$

or

$$e^{-2ka} = \frac{2ka}{\lambda} - 1 \tag{2.114b}$$

Equation (2.113) can be used to obtain the eigenvalues graphically in a manner similar to the finite-depth potential well. The result is shown in Fig. 2.21. It is clear from this figure that there is only one solution corresponding to an eigenvalue. Additionally, since $\tanh(ka) < 1$ as shown in the figure, it follows from Eq. (2.113) that

$$k > \frac{\lambda}{2a} \qquad \text{or} \qquad E > \frac{\hbar^2\lambda^2}{8ma^2} \tag{2.115}$$

By comparing Eq. (2.115) to Eq. (2.107), it implies that the energy level in the double δ-function potential well is smaller (larger negative

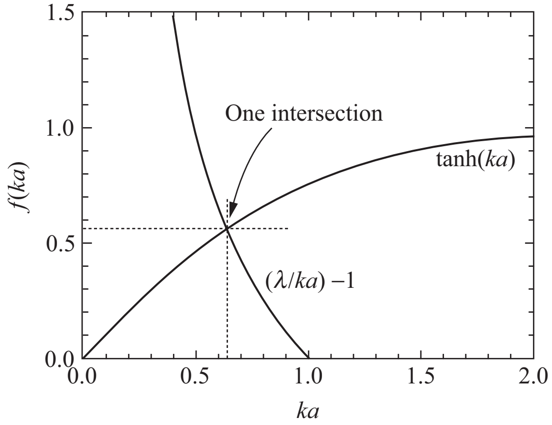


Figure 2.21 A graphical solution for the eigenvalue of a double δ-function potential well.

number) than the energy level in a single δ-function. The wave function of the double δ-function potential well [see Eq. (2.110)] is plotted in Fig. 2.22 to show the singularities at $+a$ and $-a$. The reduction of the energy level in the double δ-function as compared to that of the single δ-function barrier is difficult to explain, but such an effect is observed experimentally in multiple quantum wells. It was observed that the confined energy levels are reduced as the number of quantum wells in the structure is increased. This, however, could be a coincidence.

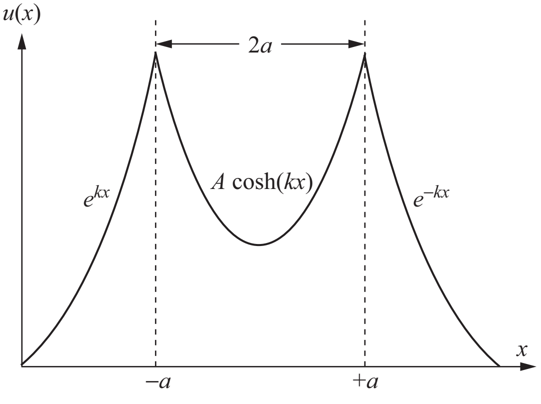


Figure 2.22 A plot of the wave function [see Eq. (2.110)] of a double δ-function potential well as a function of x.

Odd-parity solution. For the odd-parity solution, consider the following wave function:

$$u(x) = \begin{cases} e^{-kx} & \text{for } x > a \\ A\sinh(kx) & \text{for } a < x < -a \\ -e^{kx} & \text{for } x < -a \end{cases} \tag{2.116}$$

The analysis here is similar to the analysis followed in the even-parity solution case. By taking the boundary conditions at $x = +a$, and by the help of Eq. (2.104), we have

$$A \sinh(ka) = e^{-ka} \tag{2.117a}$$

and

$$-ke^{-ka} - kA\cosh(ka) = -\frac{\lambda}{a}e^{-ka} \tag{2.117b}$$

Substituting Eq. (2.117*a*) into (2.117*b*), we obtain

$$\coth(ka) = \frac{\lambda}{ka} - 1 \qquad \text{or} \qquad \tanh(ka) = \left(\frac{\lambda}{ka} - 1\right)^{-1} \tag{2.118}$$

This equation can be solved graphically as was the case with the even-parity solution discussed previously. The results are shown in Fig. 2.23.

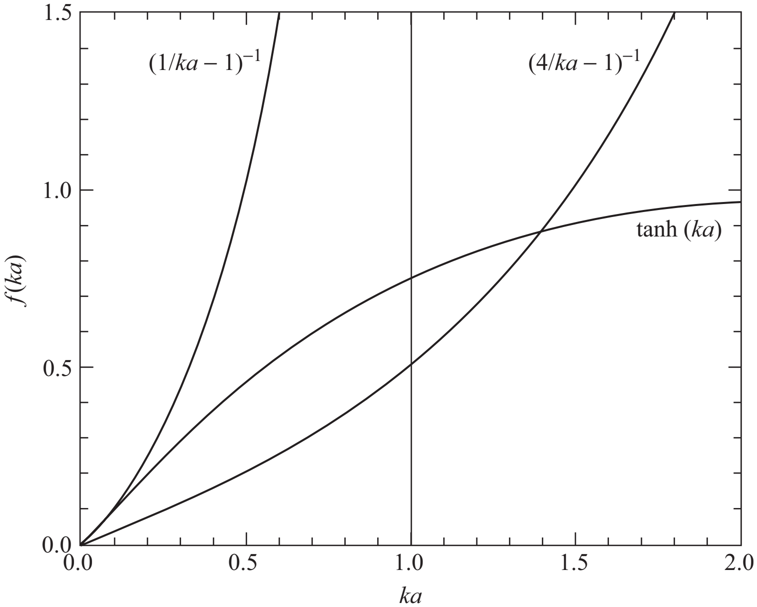


Figure 2.23 The graphical representation of the odd-parity solution of the double δ-function potential well.

Equation (2.118) indicates that there is a singularity when $\lambda = ka$. The vertical line in Fig. 2.23 is due to the singularity when λ is chosen to be unity. When λ is larger than one, for example, $\lambda = 4$, the singularity occurs at 4, which is not shown in the figure. As indicated in this figure, there is only one bound state when $\lambda \geq 1$. However, when λ is less than unity, we may or may not have a bound state due to the odd-parity solution.

2.9 Transmission in Finite Double-Barrier Potential Wells

This is a more complicated problem, and we will follow Harrison's assessment of the solution without giving explicit expressions to the transfer matrix elements. Consider the double barrier potential shown in Fig. 2.24. Harrison (2000) considered the case where $L_1 \neq L_2 \neq L_3$. The aim here is to obtain an expression for the transmission coefficient of a particle traveling from $z = -\infty$ to $z = +\infty$, assuming that the particle mass does not change when it travels through the barriers. The structure consists of five regions, and the solutions to the Schrödinger equation within each region for $E < V_o$ are

$$\begin{aligned}
&\text{Region 1: } \psi_1(z) = Ae^{ikz} + Be^{-ikz}\\
&\text{Region 2: } \psi_2(z) = Ce^{\rho z} + De^{-\rho z}\\
&\text{Region 3: } \psi_3(z) = Fe^{ikz} + Ge^{-ikz}\\
&\text{Region 4: } \psi_4(z) = He^{\rho z} + Je^{-\rho z}\\
&\text{Region 5: } \psi_5(z) = Ke^{ikz} + Le^{-ikz}
\end{aligned} \tag{2.119}$$

where $k = \sqrt{2mE/\hbar^2}$ and $\rho = \sqrt{2m(V_o - E)/\hbar^2}$. The boundary conditions at $z = 0, L_1, L_1 + L_2$, and $L_1 + L_2 + L_3$ give the following relations:

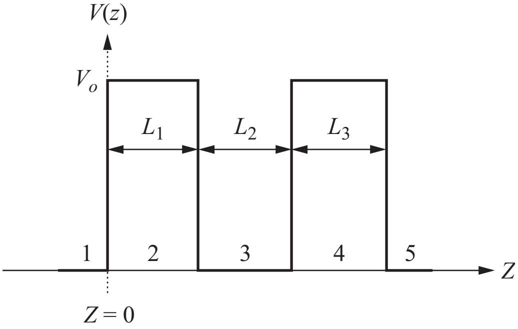


Figure 2.24 A sketch of a double-barrier potential well.

$$z = I_1 = 0: \qquad A + B = C + D \tag{2.120}$$

$$ikA - ikB = \rho C - \rho D \tag{2.121}$$

$$z = I_2 = L_1: \quad Ce^{\rho I_2} + De^{-\rho I_2} = Fe^{ikI_2} + Ge^{-ikI_2} \tag{2.122}$$

$$C\rho e^{\rho I_2} - D\rho e^{-\rho I_2} = Fike^{ikI_2} - Gike^{-ikI_2} \tag{2.123}$$

$$z = I_3 = L_1 + L_2: \quad Fe^{ikI_3} + Ge^{-ikI_3} = He^{\rho I_3} + Je^{-\rho I_3} \tag{2.124}$$

$$Fike^{ikI_3} - Gike^{-ikI_3} = H\rho e^{\rho I_3} - J\rho e^{-\rho I_3} \tag{2.125}$$

$$z = I_4 = L_1 + L_2 + L_3: \quad He^{\rho I_4} + Je^{-\rho I_4} = Ke^{ikI_4} + Le^{-ikI_4} \tag{2.126}$$

$$H\rho e^{\rho I_4} - J\rho e^{-\rho I_4} = Kike^{ikI_4} - Like^{-ikI_4} \tag{2.127}$$

The best method to proceed from here is to put the results of Eqs. (2.120) to (2.127) in matrix form. This method is known as the *transfer matrix technique*, and it yields

$$\mathbf{M}_1\begin{pmatrix} A \\ B \end{pmatrix} = \mathbf{M}_2\begin{pmatrix} C \\ D \end{pmatrix} \tag{2.128a}$$

$$\mathbf{M}_3\begin{pmatrix} C \\ D \end{pmatrix} = \mathbf{M}_4\begin{pmatrix} F \\ G \end{pmatrix} \tag{2.128b}$$

$$\mathbf{M}_5\begin{pmatrix} F \\ G \end{pmatrix} = \mathbf{M}_6\begin{pmatrix} H \\ J \end{pmatrix} \tag{2.128c}$$

$$\mathbf{M}_7\begin{pmatrix} H \\ J \end{pmatrix} = \mathbf{M}_8\begin{pmatrix} K \\ L \end{pmatrix} \tag{2.128d}$$

The coefficients of the outer regions can be linked by forming the *transfer matrix* such as

$$\begin{pmatrix} A \\ B \end{pmatrix} = \mathbf{M}_1^{-1}\mathbf{M}_2\mathbf{M}_3^{-1}\mathbf{M}_4\mathbf{M}_5^{-1}\mathbf{M}_6\mathbf{M}_7^{-1}\mathbf{M}_8\begin{pmatrix} K \\ L \end{pmatrix} \tag{2.129}$$

Since we assumed that the particle is traveling from $z = -\infty$ to $z = +\infty$, then the coefficient L can be set to zero. Furthermore, if the 2×2 matrix is written as $\mathbf{M}$, we obtain

$$\begin{pmatrix} A \\ B \end{pmatrix} = \mathbf{M}\begin{pmatrix} K \\ 0 \end{pmatrix} \tag{2.130}$$

Thus, we have $A = M_{11}K$, and the transmission coefficient can be written as

$$T(E) = \left|\frac{K \cdot K^*}{A \cdot A^*}\right| = \frac{1}{|M_{11} \cdot M_{11}^*|} \tag{2.131}$$

The matrix multiplication needed to obtain M_{11} using hand analysis is tedious and time-consuming. However, this problem can be easily solved using computer programs, such as Mathematica or MathLab.

2.10 Wentzel-Kramers-Brillouin (WKB) Approximation

The potential barriers and wells considered thus far are geometrically simple. If the barrier hight is an arbitrary function of the position, the solution of the Schrödinger equation becomes very complicated. A simple example where the barrier is a function of the distance is the triangular potential well that is usually encountered at the semiconductor heterojunction interfaces. This problem was discussed briefly in Sec. 2.6, where the solution was expressed in terms of Airy functions. Another example is the simple harmonic oscillator where the potential is parabolic in distance (see Sec. 2.7). The solution of this problem is expressed in terms of Hermite polynomials. For an arbitrary potential barrier as shown in Fig. 2.25, one can follow the WKB approximation discussed in most quantum mechanics textbooks. In this instance, we will follow the Merzbacher treatment to summarize the WKB approximation. Let us consider Fig. 2.25*a*, where we show an arbitrary spatially

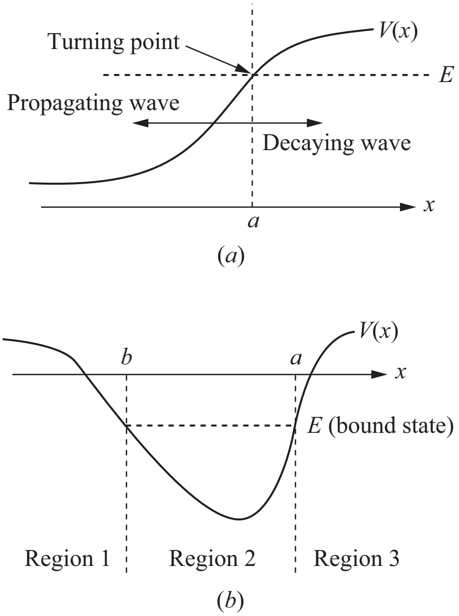


Figure 2.25 (*a*) Variation of a potential barrier as a function of the distance showing the corresponding energy level. (*b*) An arbitrary potential well used for the WKB approximation.

varying potential. The position a is called the turning point at which the wave function changes from propagating to decaying. The propagation vectors of both wave functions are given by

$$k(x) = \sqrt{\frac{2m}{\hbar^2}\,[E - V(x)]} \qquad \text{for} \quad E > V(x) \tag{2.132}$$

and

$$\rho(x) = \sqrt{\frac{2m}{\hbar^2}\,[V(x) - E]} \qquad \text{for} \quad E < V(x) \tag{2.133}$$

The WKB approximation suggests that the wave functions, either decaying or propagating, are wave-type functions generally defined as

$$\psi(x) \approx e^{i\varphi(x)} \tag{2.134}$$

By applying the Schrödinger equation to the propagating wave function, we have

$$\frac{\partial^2\psi(x)}{\partial x^2} + k^2(x)\psi(x) = 0 \tag{2.135}$$

Assuming the proportionality constant of Eq. (2.135) is spatially invariant, the Schrödinger equation becomes

$$\frac{\partial^2 e^{i\varphi(x)}}{\partial x^2} + k^2(x)e^{i\varphi(x)} = 0 \tag{2.136}$$

which can be reduced to the differential equation of $\varphi(x)$ as

$$i\frac{\partial^2\varphi(x)}{\partial x^2} - \left(\frac{\partial\varphi}{\partial x}\right)^2 + k^2(x) = 0 \tag{2.137}$$

This equation is equivalent to the Schrödinger equation except that it is nonlinear whereas the Schrödinger equation is linear. One, however, can take advantage of the nonlinearity to solve Eq. (2.137). If we have a true free particle, then the second derivative is very small, assuming that the potential does not vary too much.

$$i\frac{\partial^2\varphi(x)}{\partial x^2} = 0 \tag{2.138}$$

When Eq. (2.138) is omitted from Eq. (2.137), we obtain the first crude approximation by replacing φ with φ_o:

$$\left(\frac{\partial\varphi_o}{\partial x}\right)^2 = k^2(x) \qquad \text{or} \qquad \varphi_o = \pm\int^x k(x)\,dx + C \tag{2.139}$$

The next approximation is to set Eq. (2.137) in the following form:

$$\left(\frac{\partial \varphi}{\partial x}\right)^2 = +k^2(x) + i\frac{\partial^2 \varphi(x)}{\partial x^2} \tag{2.140}$$

The nth approximation can be set for the right-hand side of this equation to obtain the $(n+1)$th approximation as follows:

$$\varphi_{n+1}(x) = \pm \int^x \sqrt{k^2(x) + i\varphi_n''(x)}\, dx + C_{n+1} \tag{2.141}$$

For $n = 0$, we have

$$\varphi_1(x) = \pm \int^x \sqrt{k^2(x) + i\varphi_o''(x)}\, dx + C_1 = \pm \int^x \sqrt{k^2(x) \mp ik'(x)}\, dx + C_1 \tag{2.142}$$

The correct $\varphi(x)$ is baseless unless $\varphi_1(x)$ is close to $\varphi_o(x)$, which means

$$|k'(x)| \ll |k^2(x)| \tag{2.143}$$

If this condition holds, then the integrand can be expanded to obtain

$$\varphi_1(x) \approx \int^x \left[\pm k(x) + \frac{i}{2}\frac{k'(x)}{k(x)}\right] dx + C_1 = \pm \int^x k(x)\, dx + \frac{i}{2}\ln[k(x)] + C_1 \tag{2.144}$$

All $\simeq$ the above approximations are known as *WKB approximations*, which leads to writing the wave function as

$$\psi(x) \approx \frac{1}{\sqrt{k(x)}} e^{\pm i \int^x k(x) dx} \tag{2.145}$$

The equivalent solution for the decaying wave is

$$\psi(x) \approx \frac{1}{\sqrt{k(x)}} e^{\pm \int^x \rho(x) dx} \tag{2.146}$$

If $k(x)$ is regarded as the effective wave number, then for the propagating wave function we have $\lambda(x) = 2\pi/k(x)$. If condition (1.143) holds, then we have

$$\lambda(x)\left|\frac{dp}{dx}\right| \ll |p(x)| \tag{2.147}$$

where $p(x) = \pm\hbar k(x)$ is the momentum that the particle would possess at point x. Condition (2.147) implies that the change of the momentum over a wavelength must be small compared to the momentum itself. This condition breaks down if $k(x)$ is zero or varies violently, such as in a sharp corner. The entire approach breaks down if the energy of the particle is close in value to the potential extremum because proceeding from left to right, the turning point a is reached before the particle gets sufficiently far away from the turning point b (see Fig. 2.25*b*) for the WKB approximation to hold.

To connect the waves from one type to another (decaying and propagating) at the turning point requires mathematical details, which we will not consider here, but the reader can find these details in other textbooks such as in that by Merzbacher (1970). The connecting formulas are written in terms of sin and cos and are given as follows:

For $x = a$:

$$\begin{aligned} &\frac{2}{\sqrt{k}}\cos\left(\int_x^a k\,dx - \frac{\pi}{4}\right) \Leftrightarrow \frac{2}{\sqrt{k}}e^{-\int_a^x kdx} \\ &\frac{2}{\sqrt{k}}\sin\left(\int_x^a kdx - \frac{\pi}{4}\right) \Leftrightarrow -\frac{2}{\sqrt{k}}e^{\int_a^x kdx} \end{aligned} \tag{2.148}$$

For $x = b$:

$$\begin{aligned} &\frac{2}{\sqrt{k}}e^{-\int_x^b kdx} \Leftrightarrow \frac{2}{\sqrt{k}}\cos\left(\int_b^x k\,dx - \frac{\pi}{4}\right) \\ &-\frac{2}{\sqrt{k}}e^{\int_x^b kdx} \Leftrightarrow \frac{2}{\sqrt{k}}\sin\left(\int_b^x k\,dx - \frac{\pi}{4}\right) \end{aligned} \tag{2.149}$$

Let us now consider the WKB approximation to solve the bound states in an arbitrary potential well. Consider three regions in Fig. 2.25*b* where the potential is arbitrary. The WKB approximation will be used in regions 1, 2, and 3 away from the turning points. The connection formulas (2.148) and (2.149) will be used near $x = a$ and $x = b$. The requirement is to have $\psi(x)$ be finite, and the solution to the Schrödinger equation must vanish as the particle moves outward from the turning points.

The wave function can be written as

$$
\begin{aligned}
\psi_1(x) &\approx \frac{1}{\sqrt{\rho}} e^{-\int_x^b \rho(x)dx} && \text{for } x < b \\
\psi_2(x) &\approx \frac{2}{\sqrt{k}} \cos\left(\int_b^x k\,dx - \frac{\pi}{4}\right) && \text{for } b < x < a \\
&\approx \frac{2}{\sqrt{k}} \cos\left(\int_b^a k\,dx - \int_x^a k\,dx - \frac{\pi}{4}\right) \\
&\approx -\frac{2}{\sqrt{k}} \cos\left(\int_b^a k\,dx\right) \sin\left(\int_x^a k\,dx - \frac{\pi}{4}\right) \\
&\quad +\frac{2}{\sqrt{k}} \sin\left(\int_b^a k\,dx\right) \cos\left(\int_x^a k\,dx - \frac{\pi}{4}\right) \\
\psi_3(x) &\approx \frac{1}{\sqrt{\rho}} e^{\int_a^x \rho dx} && \text{for } x > a
\end{aligned}
\tag{2.150}
$$

From the boundary condition at the turning point a, given by Eq. (2.148), only the second term of $\psi_2(x)$ gives rise to a decreasing exponential satisfying the boundary conditions at infinity. Thus, the first term of $\psi_2(x)$ must be zero which leads to the following relation:

$$
\cos\left(\int_b^a k dx\right) = \left(n + \frac{1}{2}\right)\pi \qquad n = 0, 1, 2, 3, \ldots \tag{2.151}
$$

This equation determined the possible discrete values of E. The energy E appears in the integrand as well as in the limits of integration, since the turning points a and b are determined such that $V(a) = V(b) = E$.

For example, let us consider the triangular potential well shown in Fig. 2.16. Since the potential is sharp at $x = 0$, the WKB approximation cannot be used at this point. The energy E_n can be related to the turning points x_n such that $E_n(x_n) = V(x_n) = e\mathcal{E}_s x_n$, where $\mathcal{E}_s$ is the electric field. For the turning point of the decaying function we have

$$
\psi_1(x) \approx \frac{1}{\sqrt{\rho}} e^{-\int_{x_n}^b \rho(x)dx} \qquad \text{for } x > x_n \tag{2.152}
$$

This wave function must be connected to the cosine function according to Eq. (2.148) such that

$$\psi_1(x) \approx \frac{2}{\sqrt{k}} \cos\left(\int_x^{x_n} k\, dx - \frac{\pi}{4} \right) \tag{2.153}$$

This equation must vanish at $x = 0$, so the bound states are found from the following relation:

$$\cos\left(\int_x^{x_n} k\, dx - \frac{\pi}{4} \right) = 0 \tag{2.154}$$

Then the propagation vector can now be written as

$$k(x) = \sqrt{\frac{2m}{\hbar^2}[E_n - V(x)]} = \sqrt{\frac{2m}{\hbar^2}[e\mathcal{E}_s x_n - e\mathcal{E}_s x]} = \sqrt{\frac{2me\mathcal{E}_s}{\hbar^2}(x_n - x)} \tag{2.155}$$

Furthermore, the condition in Eq. (2.154) yields

$$\int_x^{x_n} k\, dx - \frac{\pi}{4} = (2n+1)\frac{\pi}{2} \tag{2.156}$$

Combine Eqs. (2.155) and (2.156) to obtain

$$\begin{gathered} \int_0^{x_n} k\, dx = (2n+1)\frac{\pi}{2} + \frac{\pi}{4} \\ \sqrt{\frac{2me\mathcal{E}_s}{\hbar^2}} \int_0^{x_n} \sqrt{x_n - x}\, dx = \frac{\pi}{2}\left(2n + \frac{3}{2}\right) \end{gathered} \tag{2.157}$$

The integral can now be evaluated to give

$$\sqrt{\frac{2me\mathcal{E}_s}{\hbar^2}} \frac{2}{3} x_n^{3/2} = \frac{\pi}{2}\left(2n + \frac{3}{2}\right) \tag{2.158a}$$

or

$$x_n = \left[\frac{3\pi}{4}\left(2n + \frac{3}{2}\right)\right]^{2/3} \left(\frac{\hbar^2}{2me\mathcal{E}_s}\right)^{1/3} \tag{2.158b}$$

Finally, substitute $E_n = e\mathcal{E}_s x_n$ in Eq. (2.158*b*) to obtain the quantized energy levels as

$$E_n = \left[\frac{3\pi}{4}\left(2n + \frac{3}{2}\right)\right]^{2/3} \left(\frac{e^2\mathcal{E}_s^2\hbar^2}{2m}\right)^{1/3} \qquad \text{for } n = 0, 2, 3, \ldots \tag{2.159}$$

The results shown in this equation are in exactly the same form obtained using the Airy function approach as shown in Eq. (2.83). Notice that the particle mass in the WKB approximation is assumed to be m, while the electron mass in Eq. (2.83) is assumed to be the effective mass m^*.

2.11 Energy Levels in Double Quantum Well Structure

This is a typical example discussed by others (see, for example, Bastard 1988, Balkanski and Wallis 2000, and Singh 2003), but we will discuss it briefly. Consider the two wells shown in Fig. 2.26 that are separated by a potential barrier of width h. Assume that each well contains $n_{\max} \gg 1$ bound states when they are isolated. The localized wave function decays exponentially far away from the well. In the limit of infinite h, the bound states $(0 \geq E \geq -V_o)$ are twofold degenerate. This means that the particle can be found in either one well or the other. At a finite value of h, the wave functions that describe the isolated wells are no longer valid for the coupled well Hamiltonian of

$$\mathbf{H} = T + V_1(z) + V_2(z) \tag{2.160}$$

where $V_1(z)$ and $V_2(z)$ are the potential energies associated with wells 1 and 2, respectively. Let $\chi_1(z)$ and $\chi_2(z)$ be the ground-state wave function for the isolated wells. The Schrödinger equations for the two wells can be written as

$$\begin{aligned} [T + V_1(z)]\chi_1(z) &= E_1\chi_1(z) \\ [T + V_2(z)]\chi_2(z) &= E_1\chi_2(z) \end{aligned} \tag{2.161}$$

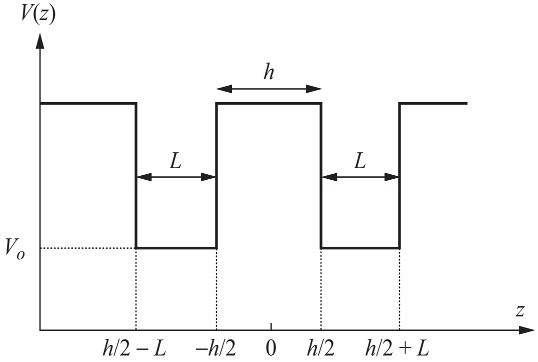


Figure 2.26 Identical double potential wells separated by a potential barrier of width h.

where E_1 is the ground state in the isolated wells. The wave function of the coupled wells can be expressed as

$$\psi(z) = A_1\chi_1(z) + A_2\chi_2(z) \tag{2.162}$$

and the Schrödinger equation can be written as

$$\mathbf{H}\psi(z) = E\psi(z) \tag{2.163}$$

Using Eqs. (2.162) and (2.163), the matrix elements $\langle\psi|\mathbf{H}|\psi\rangle$ can be obtained as follows:

$$\begin{aligned}(E_1 + \overline{V}_1 - E)A_1 + (SE_1 + \overline{V}_{12} - SE)A_2 = 0\\ (SE_1 + \overline{V}_{12} - SE)A_1 + (E_1 + \overline{V}_1 - E)A_2 = 0\end{aligned} \tag{2.164}$$

where

$$\begin{aligned} S &= \langle\chi_1 \mid \chi_2\rangle\\ \overline{V}_1 &= \langle\chi_1|\, V_2(z)\,|\chi_1\rangle = \langle\chi_2|\, V_1(z)\,|\chi_2\rangle\\ \overline{V}_{12} &= \langle\chi_1|\, V_1(z)\,|\chi_2\rangle = \langle\chi_2|\, V_2(z)\,|\chi_1\rangle\end{aligned} \tag{2.165}$$

For no-trivial solutions, the determinant of the coefficients of A_1 and A_2 must be set to zero such as

$$\begin{vmatrix} E_1 + \overline{V}_1 - E & SE_1 + \overline{V}_{12} - SE\\ SE_1 + \overline{V}_{12} - SE & E_1 + \overline{V}_1 - E\end{vmatrix} = 0 \tag{2.166}$$

The solution of this determinant is obtained as

$$E = E_1 + \frac{\overline{V}_1 \pm \overline{V}_{12}}{1 \pm S} \tag{2.167}$$

For $S \ll 1$, Eq. (2.167) is reduced to $E = E_1 + \overline{V}_1 \pm \overline{V}_{12}$. The coupling of the two wells produces a splitting of their ground-state level by $\sim 2\overline{V}_{12}$. The quantities, $S, \overline{V}_1$, and $\overline{V}_{12}$ are the overlap, shift, and transfer integrals, respectively, as illustrated in Fig. 2.27. The exact

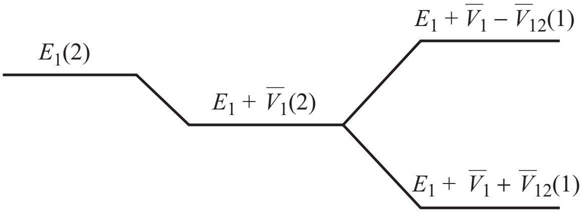

$E_1(2)$
$E_1 + \overline{V}_1(2)$
$E_1 + \overline{V}_1 - \overline{V}_{12}(1)$
$E_1 + \overline{V}_1 + \overline{V}_{12}(1)$

Figure 2.27 Shifting and lifting the degeneracy of the two ground-state isolated quantum wells due to the coupling between the wells. The numbers in parentheses reflect the degeneracy [see Bastard (1988) for additional details].

solution of the double symmetric quantum well is given by Bastard (1988) as

$$\frac{2}{\tan(kL)} + \left(\frac{\rho}{k} - \frac{k}{\rho}\right) = \mp\left(\frac{\rho}{k} + \frac{k}{\rho}\right)e^{-\rho h} \tag{2.168}$$

where k and ρ are the propagating decaying wave vectors, respectively. The $-$ and $+$ signs are for the symmetric and asymmetric solutions, respectively. This equation shows that as h is approaching infinity, there is at least one eigenvalue in each of the isolated quantum wells. As h is decreased to zero, the ground state evolves from the ground state in a well of width L to a ground state in a well of width $2L$.

Summary

The Schrödinger equation was constructed for several potential wells and barriers including a step potential, single rectangular barrier, single rectangular well, parabolic potential, single δ-function, and double δ-functions. The quantized energy levels were derived for all these systems. The quantum transmission and reflection coefficients were derived for a particle traveling with energy above or below the potential barriers, bearing in mind that the sum of the two coefficients is unity. One striking feature of the transmission coefficient of a particle with an energy larger than the potential barrier is that the transmission coefficient exhibits interference resonance. Finally, the transfer matrix technique was introduced to derive the transmission coefficients for more complex potential barrier systems, such as a double-barrier potential.

The WKB semiclassical approximation was introduced to obtain the quantized energy levels in an arbitrary smooth potential barrier. For this approximation to work, the potential barriers should not exhibit a large variation such as sharp corners or abrupt spatial variations. As an example of how to apply this approximation, the energy levels in a triangular quantum well, which commonly exists at semiconductor heterojunction interfaces, were derived and compared to the Airy function solution. The energy levels of a simple harmonic oscillator were derived for a single parabolic potential. In addition to the geometrically shaped potentials, δ-function potentials were also considered. The energy levels for single and double identical δ-function potentials were obtained using graphical solutions.

The analysis and derivation of energy levels and transmission coefficients in this chapter were first presented for simple cases, such as the step function and infinitely deep potentials. The derivation becomes more complicated as the potential barriers and wells start taking on complex structures, such as double barriers or double wells. For even further complex structures, the analysis can be obtained with the aid of computer programs.

Problems

2.1 Derive an expression for the transmission coefficient for the potential barrier shown in Fig. P2.1. Simplify your answer for the case of $k_2a = n\pi$, where n is an even integer.

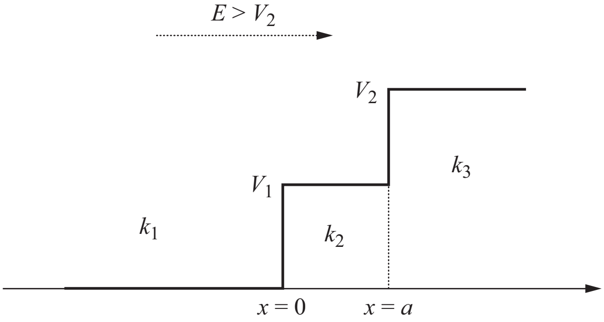


Figure P2.1

2.2 Consider the potential well shown in Fig. P2.2. Derive an expression for the energy levels in the potential well.

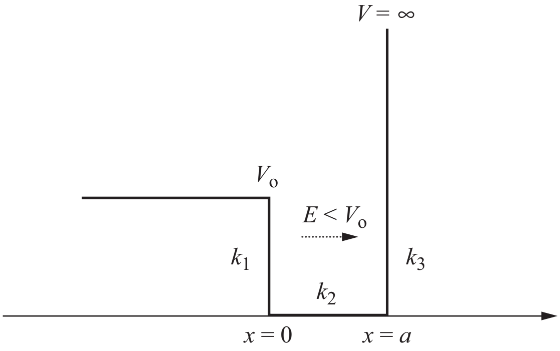


Figure P2.2

2.3 Consider an infinite three-dimensional cubic potential well with a side a where $V(0) = 0$ for $0 < x < a$ and infinity everywhere else. Derive an expression for the eigenvalues.

2.4 Consider the step potential barrier shown in Fig. 2.1 where an electron is traveling from left to right with an energy of 2.0 eV and the potential height is

2.2 eV. Determine the relative probability of finding the electron at 10 and 30 beyond the barrier.

2.5 Derive the normalization factor N_n of the simple harmonic oscillator as shown in Eq. (2.97).

2.6 Derive Eqs. (2.59) and (2.60).

2.7 The eigenvalues of the Schrödinger equation for a finite well can be obtained graphically as shown in Fig. 2.13. Start from the eigenfunctions shown in Eq. (2.61) and use the boundary conditions at $x = -L/2$ and $x = +L/2$ to derive Eq. (2.66).

2.8 For the WKB approximation to work the following inequality should be valid: $|k'(x)| \ll |k^2(x)|$. Show that if this condition holds, then $\lambda(x)|dp/dx| \ll |p(x)|$ is true. Explain the meaning of your results.

2.9 Use the WKB approximation to derive an expression for the potential $V(x) = \frac{1}{2}\beta x^2$, where β is a constant. Assume that $\beta = m\omega^2$, where m is the mass of the oscillator. Compare your results to the simple harmonic oscillator results shown in Sec. 2.7.

2.10 Derive the recursion formula: $A_{n+2} = \frac{2n+1-\eta}{(n+2)(n+1)} A_n$, for $n \leq 0$.

2.11 Derive Eq. (2.90) from Eq. (2.87).

2.12 Derive the transmission coefficient for the step potential well shown in Fig. 2.1 assuming that the particle is traveling from $x = +\infty$ to $x = -\infty$ with energy $E > V_o$.

2.13 Use the WKB approximation method to calculate the energy levels in a spherical potential well with a radius R such as $V(x) = 0$ for $x < R$ and $V(x) = \infty$ for $x > R$. Compare your results to the results of Prob. 2.3 assuming that the volume of the cube of side a is the same as the volume of the sphere with radius R.

2.14 The δ-function is very useful in solving many mathematical problems. Show that the following properties of the δ-function are true.

$$\begin{aligned} f(x)\delta(x) &= f(0)\delta(x) \\ x\delta(x) &= 0 \\ \delta(ax) &= \frac{1}{|a|}\delta(x) \\ \delta(-x) &= \delta(x) \end{aligned}$$

Chapter 3

Electronic Energy Levels in Periodic Potentials

This chapter focuses on the discussion of quantum mechanics of a single electron in a periodic potential. It is difficult to find such a system, but the closest example is that of a free electron in a solid single crystal. The free electron here means that there is only one electron in the conduction band of the crystal. This simplistic example requires that the atoms of the single crystal be perfectly arranged in a single lattice and the electron-electron interactions be ignored. Such a *one-electron single-crystal approximation* leads to a description of allowed electronic energy levels in the crystal under the constraints of the Pauli exclusion principle and Fermi-Dirac statistics. This approximation is actually the foundation of most theoretical analyses of crystalline solids. Based on this foundation, there are other approximations such as the absence of imperfections in the single crystal, the tight-binding method, and the effective mass approximation. For the *one-electron single-crystal approximation* to work, the periodic potential must satisfy the following relation assuming a one-dimensional crystal:

$$V(x) = V(x + L) \tag{3.1}$$

where L is the period of the potential. The periodic potential could be square-shaped, a δ-function, or any arbitrary shape that repeats itself in a periodic fashion and has the same periodicity of the lattice. The Schrödinger equation of the one-electron single crystal can be written as

$$\frac{\partial^2 \psi(x)}{\partial x^2} + \left\{ \frac{2m}{\hbar^2}[E_n - V(x)] \right\} \psi(x) = 0 \tag{3.2}$$

If $V(x)$ is a periodic function, then $(2m/\hbar^2)[E_n - V(x)]$ must be periodic. A typical periodic potential is shown in Fig. 3.1. We plot a square

Copyright © 2005 by The McGraw-Hill Companies, Inc. Click here for terms of use.

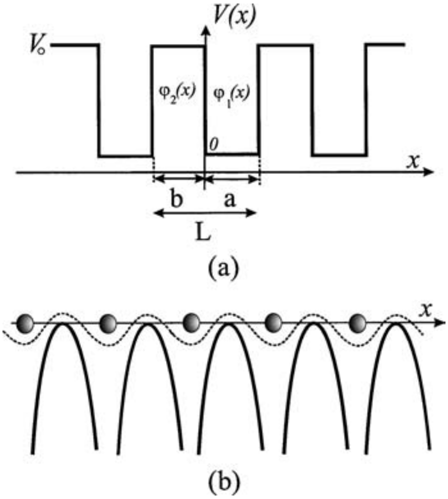


Figure 3.1 (*a*) Square periodic potential wells and (*b*) a typical crystalline periodic potential plotted along a line of ions. The solid lines are the potential along the line of ions, and the dashed line is the potential along the line between planes of ions.

periodic potential in Fig. 3.1*a* and a period potential due to a line of atoms in a crystal in Fig. 3.1*b*.

Independent electrons in a crystalline solid that each obeys a one-dimensional Schrödinger equation with a periodic potential are commonly called *Bloch electrons*. A Bloch electron reduces to a free electron when the periodic potential is zero. The discussion in this chapter starts by introducing Bloch's theorem. A simple model known as the Kronig-Penney model will be presented in which the allowed and forbidden energy bands are obtained for an electron in a periodic potential. The discussion covers other approximations, such as a Bloch electron in a weak periodic potential and an electron in a periodic δ-function potential. Superlattice system will be briefly discussed as an example of a periodic structure. Additionally, the most widely used theories employed to calculate the bandgaps of bulk semiconductor quantum structures such as quantum wells and quantum dots will be briefly discussed.

3.1 Bloch's Theorem

The Bloch theorem was derived in 1928 and was based on the nineteenth century result of Floquet. This theorem states that the eigenstates of the one-electron Hamiltonian in one dimension can be written as $H = -\hbar^2\Delta/2m + V(x)$, where $V(x + L) = V(x)$ and L is the period of the periodic potential. The wave function can be chosen to have the form of a plane wave e^{ikx} times a function of periodicity $\varphi_k(x)$ of the primitive lattice cell such that

$$\psi_k(x) = e^{ikx}\,\varphi_k(x) \tag{3.3}$$

where $\varphi_k(x)$ satisfies the following condition:

$$\varphi_k(x+L) = \varphi_k(x) \tag{3.4}$$

The one-dimensional propagation vector k is introduced as a subscript. Each k may have several eigenvalues. For the three-dimensional case, x is replaced by $\mathbf{r}$. By combining Eqs. (3.3) and (3.4) we obtain

$$\psi_k(x+L) = e^{ikL}\,\psi_k(x) \tag{3.5}$$

This equation states that the eigenstates of **H** can be chosen so that associated with each ψ is a wavevector k such that the following condition is satisfied for every L in the lattice:

$$\psi(x+L) = e^{ikL}\,\psi(x) \tag{3.6}$$

The proof of the Bloch theorem is left as an exercise. The Bloch theorem is the key to answering many of the unresolved questions posed by the free-electron theory and serves as the starting point for most of the more detailed calculations of wave functions and energy levels in crystalline solids including semiconductors and insulators.

3.2 The Kronig-Penney Model

Let us consider a periodic rectangular well potential as shown in Fig. 3.1*a*. The Schrödinger equation for this periodic potential was first solved by R. de L. Kronig and G. Penney in 1931, which led to the well-known Kronig-Penny model. This model allows one to reach an exact solution to the Schrödinger equation. While the model is a crude approximation of real crystal potentials, it illustrates explicitly most of the important characteristics of the quantum behavior of electrons in real crystalline solids such as semiconductors. Using the one-electron approximation, the wave function of the Schrödinger equation can be obtained by assuming that the net force acting on the electron is regarded as derivable from the periodic potential. The Schrödinger equation has the familiar form given by Eq. (3.2). The periodic potential in this equation satisfies the following conditions: $V(x) = 0$ for $0 < x < a$ and $V(x) = V_o$ for $-b < x < 0$. Thus, the lattice constant can be considered as $L = a+b$, which is the potential period. The wave functions of the Schrödinger equation are given by Eq. (3.3). Substituting the wave functions into the Schrödinger equation gives

$$\frac{\partial^2 e^{ikx}\,\varphi_k(x)}{dx^2} + \left\{\frac{2m}{\hbar^2}\,[E_n - V(x)]\right\} e^{ikx}\,\varphi_k(x) = 0 \tag{3.7}$$

By performing the second derivative on $e^{ikx}\,\varphi(x)$, this equation can be rewritten as

$$\frac{\partial^2\varphi_k(x)}{\partial x^2} + 2ik\frac{d\varphi_k(x)}{dx} - \left\{k^2 - \frac{2m}{\hbar^2}\left[E_n - V(x)\right]\right\}\varphi_k(x) = 0 \tag{3.8}$$

The square periodic potential shown in Fig. 3.1a requires that two equations for $\varphi_1(x)$ and $\varphi_2(x)$ be written such that

$$\begin{aligned}
&\frac{\partial^2\varphi_1(x)}{dx^2} + 2ik\frac{d\varphi_1(x)}{dx} - (k^2-\alpha^2)\,\varphi_1(x) = 0 \qquad \text{for } 0 < x < a\\
&\frac{\partial^2\varphi_2(x)}{dx^2} + 2ik\frac{d\varphi_2(x)}{dx} - (k^2-\beta^2)\,\varphi_2(x) = 0 \qquad \text{for } -b < x < 0
\end{aligned} \tag{3.9}$$

where $\alpha^2 = 2mE_n/\hbar^2$ and $\beta^2 = 2m(E_n - V_o)/\hbar^2$. The solutions to these two linear differential equations are taken as

$$\begin{aligned}
\varphi_1(x) &= Ae^{i(\alpha-k)x} + Be^{-i(\alpha+k)x} \qquad \text{for } 0 < x < a\\
\varphi_2(x) &= Ce^{i(\beta-k)x} + De^{-i(\beta+k)x} \qquad \text{for } -b < x < 0
\end{aligned} \tag{3.10}$$

where $A, B, C,$ and D are arbitrary constants. Using the continuous boundary conditions (i.e., the wave functions and their first derivatives are continuous at the boundaries, at $x = 0$ and $x = -b$) one can obtain the following four equations:

$$\begin{aligned}
A + B &= C + D\\
i(\alpha-k)A - i(\alpha+k)B &= i(\beta-k)C - i(\beta+k)D\\
e^{i(\alpha-k)a}A + e^{-i(\alpha+k)a}B &= e^{-i(\beta-k)b}C + e^{i(\beta+k)b}D\\
i(\alpha-k)e^{i(\alpha-k)a}A - i(\alpha+k)e^{-i(\alpha+k)a}B &= i(\beta-k)e^{-i(\beta-k)b}C\\
&\quad -i(\beta+k)e^{i(\beta+k)b}D
\end{aligned} \tag{3.11}$$

Notice that the periodic function at $x = a$ is the same as at $x = -b$. A trivial solution of Eq. (3.11) would be to set $A = B = C = D = 0$. However, a nontrivial solution is to set the determinant of the coefficients to zero such as the following:

$$\begin{vmatrix}
1 & 1 & -1 & -1\\
\alpha-k & -(\alpha+k) & -(\beta-k) & \beta+k\\
e^{i(\alpha-k)a} & e^{-i(\alpha+k)a} & -e^{-i(\beta-k)b} & -e^{i(\beta+k)b}\\
(\alpha-k)e^{i(\alpha-k)a} & -(\alpha+k)e^{-i(\alpha+k)a} & -(\beta-k)e^{-i(\beta-k)b} & (\beta+k)e^{i(\beta+k)b}
\end{vmatrix} = 0 \tag{3.12}$$

Using the determinant minor technique and very tedious algebra one can reach the following well-known result:

$$\frac{-(\alpha^2+\beta^2)}{2\alpha\beta}\sin(\alpha a)\,\sin(\beta b)+\cos(\alpha a)\cos(\beta b)=\cos(ka+kb)=\cos(kL) \tag{3.13}$$

This equation is derived for the case where the electron energy E is larger than the potential barrier height ($E_n > V_o$). This means that β^2 is a positive real quantity. For the case where $0 < E_n < V_o$, β is a pure imaginary number. By letting $\beta = i\gamma$, we have $\beta^2 = -\gamma^2$. From the trigonometry relations, we have $\cos(ix) = \cosh(x)$ and $\sin(ix) = i\sinh(x)$. Substitute these relations into Eq. (3.13) to obtain

$$\frac{-(\alpha^2-\gamma^2)}{2\alpha\gamma}\sin(\alpha a)\,\sinh(\gamma b)+\cos(\alpha a)\cosh(\gamma b)=\cos(ka+kb) \\ =\cos(kL) \tag{3.14}$$

Additional approximations (see, for example, Kittle 1996) can be made to Eq. (3.14). One of these approximations is based on the assumption that the periodic square wells can be replaced by δ-functions such that the product of the width and height of the δ-function remains finite. Incorporating this simplified approximation reduces Eq. (3.14) to

$$P\frac{\sin(\alpha a)}{\alpha a}+\cos(\alpha a)=\cos(ka+kb)=\cos(kL) \qquad \text{for}\, E < V_o \tag{3.15}$$

where $P = mbaV_o/\hbar^2$. The left-hand side of Eq. (3.15) is plotted as a function of αa in Fig. 3.2. Notice that the function $\cos(kL)$ on the right-hand side of Eq. (3.15) is always within the interval $-1 \le \cos(kL) \le +1$ for all real values of kL. For the nonzero imaginary part of kL, we have wave functions that diverge at $\pm\infty$, which is not an acceptable solution for the one-electron approximation in a periodic potential. Thus, there are ranges of energy in which no quantum states can exist. These bands are shown as the shaded regions in Fig. 3.2. The unshaded bands between ± 1 are the allowed energy bands in which energy states exist. Equation (3.15) is the dispersion relation, which gives the relation between the propagation vector k and the energy E_n for which the Schrödinger equation has a solution.

To understand Fig. 3.2, one may consider the case where the potential height is zero, which is the case of a free electron. Equation (3.15) is then reduced to

$$\cos(\alpha a)=\cos(kL) \qquad \text{or} \qquad \alpha = k \tag{3.16}$$

where $E = \hbar^2K^2/(2m)$ is the free-electron energy, which is shown as the dashed parabola in Fig. 3.3. Notice that the propagation vector for the

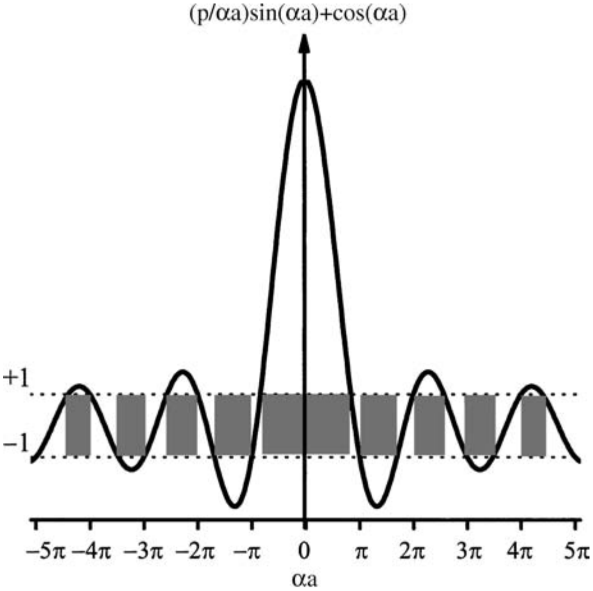


Figure 3.2 A plot of the left-hand side of Eq. (3.15) with $P = 10$. The allowed energy bands are shown as the unshaded bands for which the function lies between ±1. The forbidden bands are shown as the gray bands.

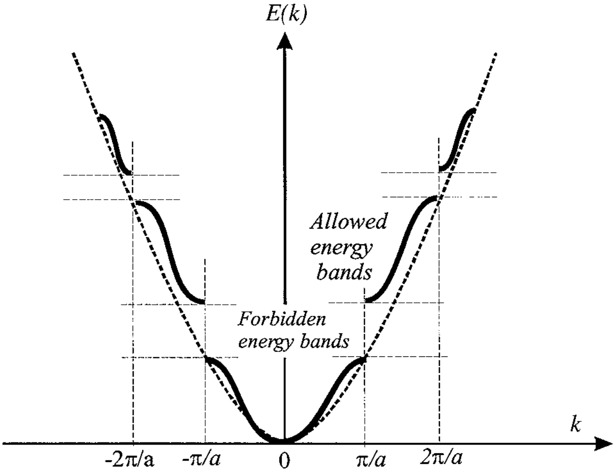


Figure 3.3 The electron energy, $E(k)$ versus k for both the Bloch electron (segments) showing the allowed and forbidden bands according to the Kronig-Penney model and the free-electron energy (dashed parabola).

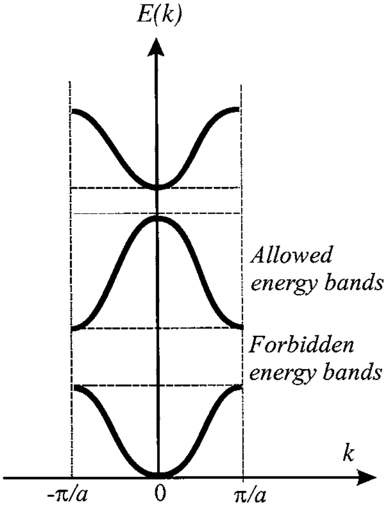


Figure 3.4 Reduced-zone representation of the allowed and forbidden bands. The curve segments of the Bloch electron where displaced by $\pm 2n\pi$ and mirror reflected at each Bragg plane.

free electron is written as K to distinguish it from the propagation vector k of the Bloch electron. The solid segment lines in Fig. 3.3 represent the allowed energy bands where the energy is a continuous function of k. This figure illustrates the concept of allowed and forbidden bands in solids such as semiconductor materials. Notice that k is continuous in the allowed band. The discontinuities at $n\pi/a$ in Fig. 3.3, where n is a positive or negative integer, show where Bragg reflections take place. At these places, the slope of the $E(k)$ should be zero. The Bloch electron energy bands can be presented by folding the curve segment as shown in Fig. 3.4. This is called a reduced-zone representation. To understand this representation, let us consider the right-hand side of Eq. (3.15) which is a periodic function and satisfies the following condition:

$$\cos(kL) = \cos(kL + 2n\pi) = \cos(kL - 2n\pi) \tag{3.17}$$

where n is a positive integer. This equation is still satisfied by adding or subtracting $2n\pi$ from the cosine argument; hence, one can displace the curve segments as shown in the figure.

3.3 Bloch Electron in a Weak Periodic Potential

There are two fundamental reasons why the strong interactions of the conduction electrons with each other in solids and with the positive ions can have the net effect of a very weak potential (see, for example, Ashcroft and Mermin 1976). First, the electron-ion interaction is strongest at small separations, but the conduction electrons (by the

Pauli principle) are forbidden from entering the immediate neighborhood of the ions because this region is already occupied by the core electrons. Second, in the region in which the conduction electrons are allowed, their mobility further diminishes the net potential on any single electron, for they can screen the fields of positively charged ions, diminishing the total effective potential. Thus the one-electron approximation in a weak periodic potential has extensive practical applications.

The Bloch wave function, with a crystal momentum, of an electron can be written as

$$\psi_{\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{K}} C_{\mathbf{k}-\mathbf{K}} e^{i(\mathbf{k}-\mathbf{K})\cdot\mathbf{r}} \tag{3.18}$$

where $\mathbf{k}$ is the crystal momentum of the electron, or simply the propagation vector of the electron in a periodic potential, and $|\mathbf{K}| = 2\pi/a$, where a is the lattice constant. $\mathbf{K}$ is thus known as the reciprocal lattice vector. The coefficients $C_{\mathbf{k}-\mathbf{K}}$ and the energy E can be determined from solving the Schrödinger equation using the wave function described in Eq. (3.18), which gives

$$\left[\frac{\hbar^2}{2m}(\mathbf{k}-\mathbf{K})^2 - E\right] C_{\mathbf{k}-\mathbf{K}} + \sum_{\mathbf{K}'} C_{\mathbf{k}-\mathbf{K}'} V_{\mathbf{K}'-\mathbf{K}} = 0 \tag{3.19}$$

where $V_{\mathbf{K}'-\mathbf{K}}$ is the Fourier transform of the periodic potential $V(x)$. Equation (3.19) is sometimes called the central equation. For the free-electron case, $V_{\mathbf{K}'-\mathbf{K}}$ is zero and Eq. (3.19) is reduced to

$$\left(E^o_{\mathbf{k}-\mathbf{K}} - E\right) C_{\mathbf{k}-\mathbf{K}} = 0 \tag{3.20}$$

where $E^o_{\mathbf{k}-\mathbf{K}} = (\hbar^2/2m)(\mathbf{k}-\mathbf{K})^2$. When $V_{\mathbf{K}'-\mathbf{K}}$ is not zero, but very small, the analysis can be made for degenerate and nondegenerate cases of free electrons. These two cases are discussed in more detail by Ashcroft and Mermin (1976). For our purpose here, we will consider the energy levels near a single Bragg plane discussed in the Kronig-Penney model. Let us assume that two free electrons are within an order of V of each other, but far from all other electrons. Equation (3.19) is then reduced to a set of two equations:

$$\begin{aligned}\left(E - E^o_{\mathbf{k}-\mathbf{K}_1}\right) C_{\mathbf{k}-\mathbf{K}_1} &= C_{\mathbf{k}-\mathbf{K}_1} V_{\mathbf{K}_2-\mathbf{K}_1} \\ \left(E - E^o_{\mathbf{k}-\mathbf{K}_2}\right) C_{\mathbf{k}-\mathbf{K}_2} &= C_{\mathbf{k}-\mathbf{K}_1} V_{\mathbf{K}_1-\mathbf{K}_2}\end{aligned} \tag{3.21}$$

where $E^o_{\mathbf{k}-\mathbf{K}}$ is as previously defined. Since there are only two electrons with two energy levels, i.e., a two-energy-level problem, the following notations are introduced: $\mathbf{K} = \mathbf{K}_2 - \mathbf{K}_1$ and $\mathbf{q} = \mathbf{k} - \mathbf{K}_1$. Hence, Eq. (3.21) is simplified as

$$\begin{aligned}(E - E^o_{\mathbf{q}})C_{\mathbf{q}} &= C_{\mathbf{q}-\mathbf{K}}V_{\mathbf{K}} \\ (E - E^o_{\mathbf{q}-\mathbf{K}})C_{\mathbf{q}-\mathbf{K}} &= C_{\mathbf{q}}V_{-\mathbf{K}} = C_{\mathbf{q}}V^*_{\mathbf{K}}\end{aligned} \tag{3.22}$$

Since we assumed that there is only a single Bragg plane, Eq. (3.22) has solutions only if the determinant of the coefficient is zero such that

$$\begin{vmatrix} E - E^o_{\mathbf{q}} & -V_{\mathbf{K}} \\ -V^*_{\mathbf{K}} & E - E^o_{\mathbf{q}-\mathbf{K}} \end{vmatrix} = 0 \tag{3.23}$$

which leads to a quadratic solution

$$\left(E - E^o_{\mathbf{q}}\right)\left(E - E^o_{\mathbf{q}-\mathbf{K}}\right) = |V_{\mathbf{K}}|^2 \tag{3.24}$$

with the following roots:

$$E = \frac{1}{2}\left(E^o_{\mathbf{q}} + E^o_{\mathbf{q}-\mathbf{K}}\right) \pm \sqrt{\frac{1}{4}\left(E^o_{\mathbf{q}} - E^o_{\mathbf{q}-\mathbf{K}}\right)^2 + |V_{\mathbf{k}}|^2} \tag{3.25}$$

Equation (3.25) shows the effect of the weak periodic potential on the nearly free electron eigenvalues $E^o_{\mathbf{q}}$ and $E^o_{\mathbf{q}-\mathbf{K}}$ when **q** is very close to the Bragg plane, determined by **K** as shown in Fig. 3.5. When the electrons possess energy close to the Bragg plane, we have $E^o_{\mathbf{q}} = E^o_{\mathbf{q}-\mathbf{K}}$, which reduces Eq. (3.25) to

$$E = E^o_{\mathbf{q}} \pm |V_{\mathbf{K}}| \tag{3.26}$$

This relation shows that at the Bragg plane one level is raised by $|V_{\mathbf{K}}|$ and the other is lowered by the same amount. The result is shown in Fig. 3.6. Another important result is that if $E^o_{\mathbf{q}} = E^o_{\mathbf{q}-\mathbf{K}}$, then

$$\frac{\partial E}{\partial \mathbf{q}} = \frac{\hbar^2}{m}\left(\mathbf{q} - \frac{1}{2}\mathbf{K}\right) \tag{3.27}$$

From Fig. 3.6 and Eq. (3.27) one can conclude that if **q** is on the Bragg plane, then the gradient of E is parallel to the plane. Since from a mathematical point of view the gradient is perpendicular to the surfaces on

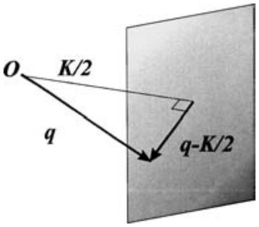


Figure 3.5 Bragg plane is defined by **K**. If a point **q** is defined on the plane, then **q** − **K**/2 is parallel to the plane.

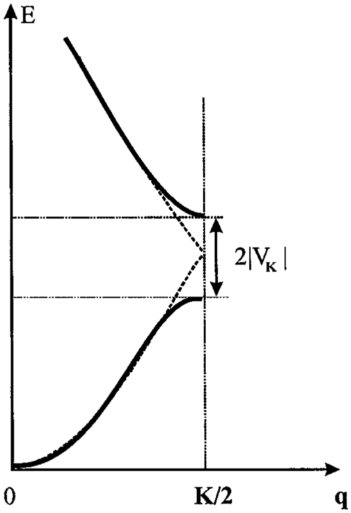


Figure 3.6 Plot of the energy bands given by Eq. (3.25) for $\mathbf{q}$ parallel to $\mathbf{K}$. The lower band corresponds to the minus sign, and the upper band corresponds to the plus sign in the equation. The dotted line represents the free-electron energy.

which a function is constant, the constant energy surfaces at the Bragg plane are perpendicular to the plane. This conclusion is mostly valid at high-symmetry points in the Brillouin zones, and it is illustrated in Figs. 3.3, 3.4, and 3.6.

The wave functions for the case presented in Eq. (3.26) can be written (see Ashcroft and Mermin 1976) for electrons in a weak periodic potential as

$$|\psi(\mathbf{r})|^2 \propto \begin{cases} \cos\left(\dfrac{1}{2}\mathbf{K}\cdot\mathbf{r}\right)^2 & \text{for } E = E^o_{\mathbf{q}} + |V_{\mathbf{K}}|,\ V_{\mathbf{K}} > 0 \\ \sin\left(\dfrac{1}{2}\mathbf{K}\cdot\mathbf{r}\right)^2 & \text{for } E = E^o_{\mathbf{q}} - |V_{\mathbf{K}}|,\ V_{\mathbf{K}} > 0 \\ \sin\left(\dfrac{1}{2}\mathbf{K}\cdot\mathbf{r}\right)^2 & \text{for } E = E^o_{\mathbf{q}} + |V_{\mathbf{K}}|,\ V_{\mathbf{K}} < 0 \\ \cos\left(\dfrac{1}{2}\mathbf{K}\cdot\mathbf{r}\right)^2 & \text{for } E = E^o_{\mathbf{q}} - |V_{\mathbf{K}}|,\ V_{\mathbf{K}} < 0 \end{cases} \tag{3.28}$$

Sometimes the two types of linear combination for $V_{\mathbf{k}} < 0$ are called "p-like" [$|\psi(\mathbf{r})|^2 \propto \sin(\frac{1}{2}\mathbf{K}\cdot\mathbf{r})^2$] and "$s$-like" [$|\psi(\mathbf{r})|^2 \propto \cos(\frac{1}{2}\mathbf{K}\cdot\mathbf{r})^2$] wave functions. The s-like combination does not vanish at the ion, while in the p-like combination the charge density vanishes as the square of the distance from the ion for small distances.

3.4 One-Electron Approximation in a Periodic Dirac δ-functions

It was mentioned in Sec. 3.2 that the Kronig-Penney model can be simplified by assuming that the periodic potential can be approximated as δ-functions with a finite product of the width and height. In fact, this assumption is quite feasible since the atoms in single crystals can be considered as periodic δ-function potentials in many theoretical models. This problem is treated by Mihály and Martin (1996). Let us assume that the atoms are arranged in a one-dimensional crystal with a lattice constant of a. Each atom is thus represented by the potential $V(x) = aV_o\delta(x)$ where V_o is the height of the δ-function. Assume that the atoms are placed at $x = na$ where n is an integer. The Schrödinger equation between the atoms for the range $0 < x < a$ is

$$-\frac{\hbar^2}{2m}\frac{d^2\psi(x)}{dx^2} = E\psi(x) \tag{3.29}$$

and the wave function is

$$\psi(x) = Ae^{iKx} + Be^{-iKx} \tag{3.30}$$

where $K = \sqrt{2mE/\hbar^2}$. For the wave function in the full range $-\infty < x < +\infty$, the Bloch wave function, Eq. (3.3), will be adopted and combined with Eq. (3.30) to give

$$\psi(x) = Ae^{iKx} + Be^{-iKx} = e^{ikx}\varphi_k(x) \tag{3.31a}$$

$$\varphi_k(x) = Ae^{i(K-k)x} + Be^{-i(K+k)x} \qquad \text{for } 0 < x < a \tag{3.31b}$$

where $\varphi_k(x)$ is a periodic function and can be generated for the whole crystal by setting $x = na$. This function is continuous, but its derivative is not, as shown in Sec. 2.8. The jump (discontinuity) in the derivative of the δ-function can be found by integrating the Schrödinger equation over a small range around a such that $a - \varepsilon < x < a + \varepsilon$:

$$\psi(x)|_{x=a+\varepsilon} = \psi(x)|_{x=a-\varepsilon} \tag{3.32a}$$

$$\left.\frac{d}{dx}\psi(x)\right|_{x=a+\varepsilon} - \left.\frac{d}{dx}\psi(x)\right|_{x=a-\varepsilon} = \frac{2maV_o}{\hbar^2}\psi(a) \tag{3.32b}$$

Since $\varphi_k(x)$ is periodic, then $\varphi_k(0) = \varphi_k(a)$, which in the limit of $\varepsilon \to 0$ leads to

$$A + B = Ae^{i(K-k)a} + Be^{-i(K+k)a} \tag{3.33}$$

The derivative of the Bloch wave function gives

$$\frac{d}{dx}\psi(x) = ik\,e^{ikx}\,\varphi_k(x) + e^{ikx}\,\frac{d\varphi_k(x)}{dx} \tag{3.34}$$

Combining Eqs. (3.32*b*) and (3.34), we have

$$e^{ik(a+\varepsilon)}\left.\frac{d\varphi_k(x)}{dx}\right|_{x=a+\varepsilon}-e^{ik(a-\varepsilon)}\left.\frac{d\varphi_k(x)}{dx}\right|_{x=a-\varepsilon}=\frac{2maV_o}{\hbar^2}e^{ika}\varphi_k(x)$$

or (3.35)

$$\left.\frac{d\varphi_k(x)}{dx}\right|_0-\left.\frac{d\varphi_k(x)}{dx}\right|_a=\frac{2maV_o}{\hbar^2}\varphi_k(0)$$

By using the explicit form of $\varphi_k(x)$ in Eq. (3.31*a*), the first derivative yields

$$\frac{d\varphi_k(x)}{dx}=i(K-k)Ae^{i(K-k)x}-i(K+k)Be^{-i(K+k)x} \tag{3.36}$$

Combining Eqs. (3.35) and (3.36) to obtain

$$\begin{aligned}&i(K-k)A-i(K+k)B-i(K-k)Ae^{i(K-k)a}+i(K+k)Be^{-i(K+k)a}\\&=\frac{2maV_o}{\hbar^2}(A+B)\end{aligned} \tag{3.37}$$

One can now solve Eqs. (3.33) and (3.37) such that

$$\begin{bmatrix}1-e^{i(K-k)a} & 1-e^{-i(K+k)a}\\ i(K-k)(1-e^{i(K-k)a})-\gamma & i(K+k)(1-e^{-i(K+k)a})-\gamma\end{bmatrix}\begin{pmatrix}A\\B\end{pmatrix}=0 \tag{3.38}$$

where $\gamma=2maV_o/\hbar^2$. For a nontrivial solution, the determinant of the coefficient should be zero:

$$\begin{vmatrix}1-e^{i(K-k)a} & 1-e^{-i(K+k)a}\\ i(K-k)(1-e^{i(K-k)a})-\gamma & i(K+k)(1-e^{-i(K+k)a})-\gamma\end{vmatrix}=0 \tag{3.39}$$

With simple algebra, the determinant gives the following solution:

$$\cos(ka)=\frac{ma^2V_o}{\hbar^2}\frac{\sin(Ka)}{Ka}+\cos(Ka) \tag{3.40}$$

which is the same form obtained for the Kronig-Penney model as illustrated in Eq. (3.15). The plot of this equation is similar to the plot shown in Fig. 3.2. Equation (3.40) is derived for $V_o>0$, where $E=\hbar^2K^2/(2m)$, but when $V_o<0$, we have negative energy. For this reason we can define $K=i\rho$, which gives $E=-\hbar^2\rho^2/(2m)$, and Eq. (3.40) is changed to

$$\cos(ka)=\frac{ma^2V_o}{\hbar^2}\frac{\sinh(\rho a)}{\rho a}+\cosh(\rho a) \tag{3.41}$$

The graphical solution of Eq. (3.41) shows that at least one bound state exists for $V_o<0$. The proof of this case is left as an exercise.

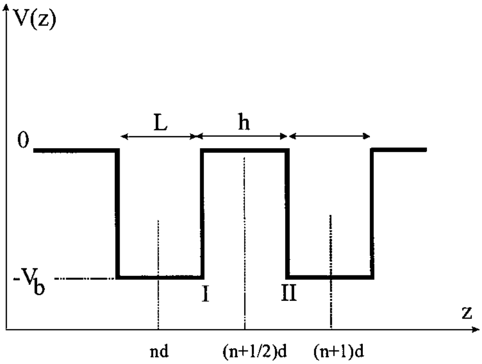


Figure 3.7 A segment of the potential energy profile of a superlattice plotted along the z direction.

3.5 Superlattices

A semiconductor superlattice is a periodic structure that can be used to illustrate the behavior of a periodic potential. This class of systems is composed of a number of semiconductor quantum wells of thickness L separated by barriers of thickness h as shown in Fig. 3.7. Generally speaking, the number of periods ranges between 10 and 50, but for the analysis here we assume that the number of periods is approaching infinity. The superlattice here means that the quantum wells are close to each other such that an electron can tunnel through the barriers and exists in any of the wells with a nonzero probability. The potential energy $V(z)$ is a periodic function of z, where z is the growth direction, with a period of $d = L + h$. Thus it can be written as

$$V(z) = \sum_{n=-\infty}^{+\infty} V(z - nd) \tag{3.42}$$

where

$$V(z - nd) = \begin{cases} -V_b & \text{if } |z - nd| \leq \dfrac{L}{2} \\ 0 & \text{if } |z - nd| > \dfrac{L}{2} \end{cases} \tag{3.43}$$

Following Bastard formalisms, the form of the wave function solutions can be chosen as

$$\psi(z) = \begin{cases} Ae^{ik(z-nd)} + Be^{-ik(z-nd)} & \text{for the well, i.e., } |z - nd| \leq \dfrac{L}{2} \\ Ce^{i\rho(z-nd-d/2)} + De^{-i\rho(z-nd-d/2)} & \text{for the barrier, i.e., } \left|z - nd - \dfrac{d}{2}\right| \leq \dfrac{h}{2}, \end{cases} \tag{3.44}$$

where k is the propagation vector in the well and ρ is the propagation vector in the barrier. Vectors k and ρ are related through the energy

$$E = \frac{\hbar^2 \rho^2}{2m^*} = -V_b + \frac{\hbar^2 k^2}{2m^*} \tag{3.45}$$

The parameter m^* is the effective mass of the electron in the superlattice. Since the potential function of the superlattice is periodic, the wave function expressed in Eq. (3.44) must satisfy the Bloch theorem such that $\psi_q(z + nd) = \psi_q(z)$, where we introduced the subscript q to indicate the function is a Bloch function. Again, the q-space is called the reciprocal or momentum space. Thus, the solution of the Schrödinger equation can be limited to the first Brillouin zone (see Ashcroft and Mermin 1976 or Kittel 1996 for further discussions on the Brillouin zones). The continuity conditions at the interfaces labeled I and II in Fig. 3.7 give the following results for the case of $E > 0$:

$$\begin{aligned}
e^{ikL/2}A + e^{-ikL/2}B &= e^{-i\rho h/2}C + e^{i\rho h/2}D \\
ke^{ikL/2}A - ke^{-ikL/2}B &= \rho e^{-i\rho h/2}C - \rho e^{i\rho h/2}D \\
e^{i(k-q)L/2}A + e^{-i(k+q)L/2}B &= e^{-i(\rho-q)h/2}C + e^{i(\rho+q)h/2}D \\
(k-q)e^{i(k-q)L/2}A - (k+q)e^{-i(k+q)L/2}B &= (\beta - k)e^{-i(\rho-q)h/2} \\
&\quad C - (\rho+q)e^{i(\rho+q)h/2}D
\end{aligned} \tag{3.46}$$

We used the Bloch theorem at interface II in a fashion similar to that of the Kronig-Penney model discussed in Sec. 3.2. To solve these four equations with four unknowns, we rely on the determinant method, which gives

$$\begin{vmatrix}
e^{ikL/2} & e^{-ikL/2} & -e^{-i\rho h/2} & -e^{i\rho h/2} \\
ke^{ikL/2} & -ke^{-ikL/2} & -\rho e^{-i\rho h/2} & \rho e^{i\rho h/2} \\
e^{i(k-q)L/2} & e^{-i(k+q)L/2} & -e^{-i(\rho-q)h/2} & -e^{i(\rho+q)h/2} \\
(k-q)e^{i(k-q)L/2} & -(k+q)e^{-i(k+q)L/2} & -(\beta-k)e^{-i(\rho-q)h/2} & (\rho+q)e^{i(\rho+q)h/2}
\end{vmatrix} = 0 \tag{3.47}$$

The solution of the determinant is similar to that of Eq. (3.13), which yields

$$\frac{-(k^2+\rho^2)}{2k\rho}\sin(kL)\sin(\rho h) + \cos(kL)\cos(\rho h) = \cos(qL+qh) = \cos(qd) \tag{3.48}$$

Similarly, for $-V_b \leq E < 0$ we have

$$\frac{-(k^2-\gamma^2)}{2k\gamma}\sin(kL)\sinh(\gamma h) + \cos(kL)\cosh(\gamma h) = \cos(qL+qh) = \cos(qd) \tag{3.49}$$

where $\rho = i\gamma$. For the case where h is very large, Eq. (3.49) diverges unless the multiplication coefficients were set to zero such that

$$\cos(kL) - \frac{k^2 - \gamma^2}{2k\gamma}\sin(kL) = 0 \tag{3.50}$$

The form of this equation is familiar to us; it is the solution of isolated quantum wells (see Chap. 2).

Equation (3.49) tells us that the energy is a strong function of γ. However, in order to obtain approximate subband energy, one may use the Taylor's series to expand the left-hand side of Eq. (3.49). Thus, let $F(E)$ be the left-hand side of this equation, expanding this function around E for the jth subband such that $E - E_j$ is very small, we obtain

$$F(E) = F(E)|_{E=E_j} + \left.\frac{\partial F(E)}{\partial E}\right|_{E=E_j}(E - E_j) + \cdots \tag{3.51}$$

Retaining only the first-order terms, we have

$$E(q) = E_j - \frac{F(E)|_{E=E_j}}{\left.\dfrac{\partial F(E)}{\partial E}\right|_{E=E_j}} + \frac{F(E)}{\left.\dfrac{\partial F(E)}{\partial E}\right|_{E=E_j}} \tag{3.52}$$

Since the right-hand side of Eq. (3.49) is equal to the left-hand side, i.e., $F(E) = \cos(qd)$, then we can write Eq. (3.52) as

$$E_j(q) = E_j + S_j + 2T_j\cos(qd) \tag{3.53}$$

where

$$S_j = -\frac{F(E)|_{E=E_j}}{\left.\dfrac{\partial F(E)}{\partial E}\right|_{E=E_j}} \quad \text{and} \quad 2T_j = \frac{1}{\left.\dfrac{\partial F(E)}{\partial E}\right|_{E=E_j}} \tag{3.54}$$

A typical plot of the energy band is shown in Fig. 3.8, where the bandwidth is $4T_j$. The parameters S_j and T_j defined in Eq. (3.54) can be

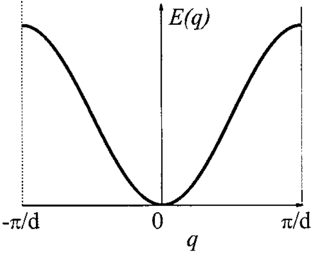


Figure 3.8 A plot of the energy band [Eq. (3.53)] as a function of the crystal momentum q for a superlattice of period d is shown for the first Brillouin zone.

evaluated by assuming that the wave functions satisfy the Bloch theorem, and by considering the nearest-neighbor interactions. The final results are given as

$$T_j = \langle j, z|\, V_b(z)\, |j, z - d\,\rangle \tag{3.55}$$

and

$$S_j = \sum_{n \neq 0} \langle j, z|\, V_b(z - nd\,)\, |j, z\rangle \tag{3.56}$$

3.6 Effective Mass

When the periodic potential is zero, we have a free electron with mass m. But when the periodic potential is nonzero, the electron will move in the periodic crystal with a different mass known as the "effective mass." It is usually denoted as m^*. To obtain an expression for the effective mass, one can start from the free-electron case and work the problem by exerting a force on the electron. We know from previous discussions that the energy of a free electron can be written as $E = \hbar\omega = \hbar^2\mathbf{k}^2/(2m)$. By using the duality concept, the wave packet of the electron is assumed to be moving with a group velocity $\boldsymbol{\upsilon}_g = \partial\omega/\partial\mathbf{k}$. When a force $\mathbf{F}$ is applied to an electron, the electron is accelerated and the motion of the electron is given by the classical relation

$$\frac{\partial E}{\partial t} = \mathbf{F} \cdot \boldsymbol{\upsilon}_g \tag{3.57}$$

On the other hand, if the energy band of the electron $E(\mathbf{k})$ is peaking at $\mathbf{k_o}$, then one can expand $E(\mathbf{k})$ about $\mathbf{k_o}$, assuming that $\mathbf{k}$ is very close to $\mathbf{k_o}$. The linear term in $(\mathbf{k} - \mathbf{k_o})$ vanishes at $\mathbf{k} = \mathbf{k_o}$ and the quadratic term will be proportional to $(\mathbf{k} - \mathbf{k_o})^2$ according to the following relation, where $\mathbf{k_o}$ is assumed to be a point of high symmetry:

$$E(\mathbf{k}) \approx E(\mathbf{k_o}) + A(\mathbf{k} - \mathbf{k_o})^2 \tag{3.58}$$

where A is a positive quantity since E is maximum at $\mathbf{k_o}$. It is obvious that one can easily guess that $A = \hbar^2/(2m^*)$. Furthermore, for energy levels with wavevectors near $\mathbf{k_o}$, we have

$$\boldsymbol{\upsilon}_g = \frac{1}{\hbar}\frac{\partial E}{\partial \mathbf{k}} \approx \frac{\hbar(\mathbf{k} - \mathbf{k_o})}{m^*} \tag{3.59}$$

The acceleration α of the electron in the applied force is thus given by

$$\alpha = \frac{\partial \boldsymbol{\upsilon}_g}{\partial t} = \frac{\hbar}{m^*}\frac{d\mathbf{k}}{dt} = \frac{1}{m^*}\frac{d\hbar\mathbf{k}}{dt} = \frac{1}{m^*}\frac{d\mathbf{p}}{dt} = \frac{1}{m^*}\mathbf{F} \tag{3.60}$$

Furthermore, the first derivative of the group velocity with respect to time can be expressed as

$$\frac{\partial \boldsymbol{\upsilon}_g}{\partial t} = \frac{1}{\hbar}\frac{\partial}{\partial t}\frac{\partial E}{\partial \mathbf{k}} = \frac{1}{\hbar}\frac{\partial}{\partial \mathbf{k}}\frac{\partial E}{\partial t} = \frac{1}{\hbar}\frac{\partial}{\partial \mathbf{k}}\mathbf{F}\cdot\boldsymbol{\upsilon}_g = \frac{1}{\hbar}\frac{\partial}{\partial \mathbf{k}}\frac{1}{\hbar}\frac{\partial E}{\partial \mathbf{k}}\mathbf{F} = \frac{1}{\hbar^2}\frac{\partial^2 E}{\partial \mathbf{k}^2}\mathbf{F} \tag{3.61}$$

By equating Eqs. (3.60) and (3.61) one can find that

$$m^* = \left(\frac{1}{\hbar^2}\frac{\partial^2 E}{\partial \mathbf{k}^2}\right)^{-1} \tag{3.62}$$

which can be written in the tensor form as

$$M_{ij}^{-1} = \frac{1}{\hbar^2}\frac{\partial^2 E}{\partial \mathbf{k}_i \partial \mathbf{k}_j} = \frac{1}{\hbar^2}\nabla_{\mathbf{k}_i}\nabla_{\mathbf{k}_j} E \tag{3.63}$$

M^{-1} is called the inverse of the effective mass tensor. This derivation is made for the electron in the conduction band. The same procedure could be followed for the hole in the valence band, and the result is similar to Eq. (3.63) except the mass tensor has a minus sign from the second derivative of the energy. From Eq. (3.63), one can conclude that the curvature of the energy band is proportional to the inverse effective mass tensor. This means that the smaller the effect mass, the larger the band curvature.

3.7 Band Structure Calculation Methods

This section is concerned with the most common theoretical models used to calculate the band structure of semiconductors, quantum wells, and quantum dots. A brief discussion is presented for the tight-binding method, the Kane theory also known as $\mathbf{k}\cdot\mathbf{p}$ theory, and the envelop function approximation. The theoretical presentation here is very general, and in most cases we simply show the results without going through the derivation, which can be very extensive and complicated.

3.7.1 Tight-binding method

The calculation of energy bands in solids is a very difficult task, and as mentioned in previous sections, there are many approximations that one has to take into account to obtain reasonable answers. In addition to the approximation, discussed in the previous sections, there are many other methods that are used to calculate the dispersion relation $E(\mathbf{k})$ in solids. Instead of considering a one-electron approximation in a weak periodic potential where the energy of the electrons is perturbed slightly, one may consider the electron-ion interaction to be very strong such that the electron is localized and tightly bound to the positive ion.

This approximation is called the *tight-binding method*, where the electron wave function is expanded in a series of functions that are localized about the atoms in the crystal. Thus, if one takes a linear combination of Bloch functions for one atom per unit cell, the resulting function is called the *Wannier function* defined as

$$w_n(\mathbf{r}-\mathbf{R}) = N^{-1/2}\sum_{\mathbf{K}} e^{-i\mathbf{k}\cdot\mathbf{R}}\psi_{n\mathbf{k}}(\mathbf{r}) \tag{3.64}$$

where N is the number of atoms in the crystal and $\mathbf{R}$ is the interatomic separation within the unit cell. Substitute the Bloch function Eq. (3.3) in the Wannier function to obtain

$$\begin{aligned} w_n(\mathbf{r}-\mathbf{R}) &= N^{-1/2}\sum_{\mathbf{K}} e^{-i\mathbf{k}\cdot\mathbf{R}}e^{i\mathbf{k}\cdot\mathbf{r}}\varphi_{n\mathbf{k}}(\mathbf{r}) \\ &= N^{-1/2}\sum_{\mathbf{K}} e^{i\mathbf{k}\cdot(\mathbf{r}-\mathbf{R})}\varphi_{n\mathbf{k}}(\mathbf{r}-\mathbf{R}) \end{aligned} \tag{3.65}$$

If $\varphi_{n\mathbf{k}}$ is periodic and independent of $\mathbf{k}$ as is the case for a cubic lattice, we have $\varphi_{n\mathbf{k}}(\mathbf{r}-\mathbf{R}) = \varphi_{n\mathbf{k}}(\mathbf{r}) = \varphi_{n0}(\mathbf{r})$ and Eq. (3.65) becomes

$$\begin{aligned} w_n(\mathbf{r}-\mathbf{R}) &= N^{-1/2}\varphi_{n0}(\mathbf{r})\sum_{\mathbf{K}} e^{i\mathbf{k}\cdot(\mathbf{r}-\mathbf{R})} \\ &= N^{-1/2}\varphi_{n0}(\mathbf{r})\frac{\sin[\pi(\mathbf{r}-\mathbf{R})/a]}{\pi(\mathbf{r}-\mathbf{R})/a} \end{aligned} \tag{3.66}$$

where a is the lattice constant. This equation is identical to the function $f(x) = \sin(x)/x$, which is a localized function with a maximum at $x = 0$ and decays in an oscillatory fashion as $x \to \pm\infty$. The expression $\sin(x)/x$ is plotted as a function of x in Fig. 3.9 to illustrate the localization of the Wannier wave function near an atom in a single crystal. The Wannier functions, while useful for producing localized wave functions at the lattice sites, are limited in the energy band calculations.

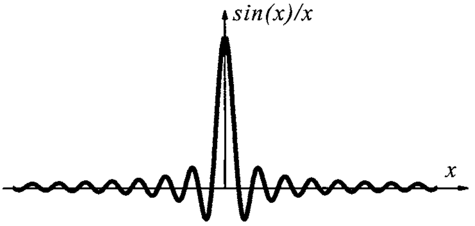


Figure 3.9 The expression $\sin(x)/x$ is plotted as a function of x to illustrate how the Wannier function is localized around the ions in the crystal.

The first-order energy can be obtained (see Kittle 1996) by calculating the diagonal matrix element of the Hamiltonian of the crystal

$$\langle \mathbf{k}|\mathbf{H}|\mathbf{k}\rangle = N^{-1}\sum_j \sum_m e^{i\mathbf{k}\cdot(\mathbf{r}_j-\mathbf{r}_m)}\langle \varphi_m|\mathbf{H}|\varphi_j\rangle \tag{3.67}$$

where $\varphi_m = \varphi_m(\mathbf{r}-\mathbf{r}_m)$. By letting $\mathbf{R}_m = \mathbf{r}_m - \mathbf{r}_j$ this equation can be rewritten as

$$\langle \mathbf{k}|\mathbf{H}|\mathbf{k}\rangle = N^{-1}\sum_m e^{-i\mathbf{k}\cdot\mathbf{R}_m}\int dV\, \varphi^*(\mathbf{r}-\mathbf{R}_m)H\varphi(\mathbf{r}) \tag{3.68}$$

By neglecting all the integrals in Eq. (3.68) except for those on the same atom and those between the nearest neighbor connected by **R**, we have

$$\langle \mathbf{k}|\mathbf{H}|\mathbf{k}\rangle = -\alpha - \gamma\sum_m e^{-i\mathbf{k}\cdot\mathbf{R}_m} = E(\mathbf{k}) \tag{3.69}$$

where

$$\alpha = -\int dV\ \varphi^*(\mathbf{r})\mathbf{H}\varphi(\mathbf{r})$$

and (3.70)

$$\gamma = -\int dV\ \varphi^*(\mathbf{r}-\mathbf{R})\mathbf{H}\varphi(\mathbf{r})$$

The overlap integral γ can be evaluated in Rydberg (Ry$= me^4/2\hbar^2$) for two hydrogen atoms in the 1s state as $\gamma = 2(1+\mathbf{R}/a_o)e^{-R/a_o}$, where $a_o = \hbar^2/me^2$. The overlap integral indicates that the overlap energy is decreasing exponentially with the interatomic separation **R**. For a cubic case where the atomic s-level as an example is given by α, the energy is given by

$$E(k) = -\alpha - 2\gamma(\cos k_x a + \cos k_y a + \cos k_z a) \tag{3.71}$$

For the case where $ka \ll 1$, Eq. (3.71) is reduced to

$$E(k) = -\alpha - 6\gamma + \gamma k^2 a^2 \tag{3.72}$$

By using the definition of the effective mass (3.63), one can obtain $m^* = \hbar^2/2\gamma a^2$. Thus, for a smaller energy overlap, we have a larger effective mass and a narrower energy band. For a body-centered cubic (bcc) structure with eight nearest neighbors we have

$$E(k) = -\alpha - 8\gamma\cos\frac{1}{2}k_x a\cos\frac{1}{2}k_y a\cos\frac{1}{2}k_z a \tag{3.73}$$

and for the face-centered cubic (fcc) structure with 12 nearest neighbor we have

$$E(k) = -\alpha - 4\gamma\left(\cos\frac{1}{2}k_y a\cos\frac{1}{2}k_z a + \cos\frac{1}{2}k_z a\cos\frac{1}{2}k_x a + \cos\frac{1}{2}k_x a\cos\frac{1}{2}k_y a\right) \tag{3.74}$$

For $ka \ll 1$, Eq. (3.73) is reduced to $E(k) = -\alpha - 12\gamma + \gamma k^2 a^2$. Equations (3.72) to (3.74) reveal the characteristic feature of tight-binding energy bands, where the bandwidth (the spread between the minimum and maximum energies in the band) is proportional to the small overlap integral γ. The narrower the bands, the smaller the overlap, and in the limit of vanishing overlap, the bandwidth also vanishes and the band becomes N-fold degenerate. This implies that the electron becomes bound to any one of the N isolated atoms.

3.7.2 k · p method

The advantage of using the $\mathbf{k}\cdot\mathbf{p}$ method is that the optical matrix elements can be used as inputs in the band structure calculations. With this method the calculations are made near $\mathbf{k} = 0$ (the first Brillouin zone center), but it can be extrapolated over the entire Brillouin zone. Interpretation of the optical measurements, analytical expressions for band dispersion, and effective masses can all be easily obtained around high-symmetry points in the Brillouin zone. Before continuing the discussion, it is worth pointing out a few important aspects of the symmetry points in the Brillouin zone. Group theory has been used extensively in investigating crystalline materials, and there are several textbooks on the subject (see, for example, Falicov 1966 and Koster 1957). In this chapter we limit our discussion on the zinc-blende structure, which is a fcc structure. Figure 3.10*a* is a sketch of the first Brillouin zone showing a few of the high-symmetry points. For example, the [100] direction is $\Gamma \to \Delta \to X$, the [110] direction is $\Gamma \to \Sigma \to K$, and the [111] direction is $\Gamma \to \Lambda \to$ L. Each atom in the Brillouin zone has tetrahedral point group symmetry, denoted T_d. The point group symmetry is defined with respect to the three perpendicular crystallographic axes with the origin placed at one of the two atoms in the primitive unit cell. This symmetry has 24 operations commonly known as follows:

E. Identity

C_3 operations. Clockwise and counterclockwise rotation of 120° about the following axes: [111], [$\bar{1}$11], [1$\bar{1}$1], and [11$\bar{1}$] axes (total of eight operations).

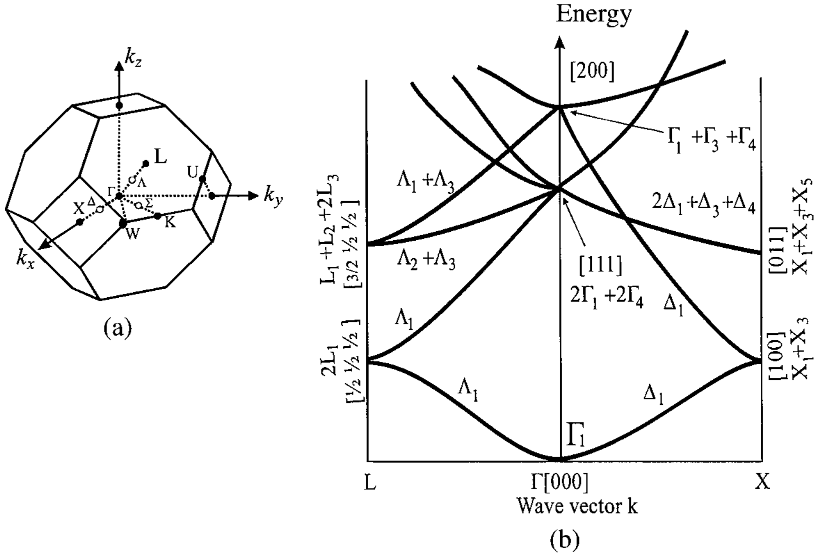


Figure 3.10 (*a*) A sketch of the first Brillouin zone of a face-centered cubic structure with a few high-symmetry points as indicated by the letters. (*b*) The band structure of nearly free electrons in a zinc-blende-type crystal in the reduced zone scheme. (*After Yu and Cardona 2003*).

C_2 operations. Rotations of 180° about the [100], [010], and [001] axes (total of three operations).

S_4 operations. Clockwise and counterclockwise rotations of 90^o followed by a reflection on the plane perpendicular to the rotation about the [100], [010], and [001] axes (total of six operations).

S operations. Reflections with respect to the $(110), (1\bar{1}0), (101), (10\bar{1}), (011)$, and $(01\bar{1})$ planes (total of six operations).

Another notation of the irreducible T_d group is Γ_1, Γ_2, Γ_3, Γ_4, and Γ_5, which is mostly known in semiconductor physics. This notation is reached by the fact that the wave function of a wavevector **k** at the Γ point in the Brillouin zone (the center of the zone) always transforms like the irreducible representation of the point group of the crystal. Using these group theory notations, Yu and Cardona (2003) constructed the nearly free electron band structure in a zinc-blende crystal assuming that the crystal potential is vanishingly small as shown in Fig. 3.10*b*. This figure resembles in many ways the band structures obtained by more complex theory such as $\mathbf{k}\cdot\mathbf{p}$ and pseudopotential methods. A generic band structure for GaAs is sketched in Fig. 3.11 showing the point group symmetries at L, Γ, K and X extrema in the first Brillouin zone.

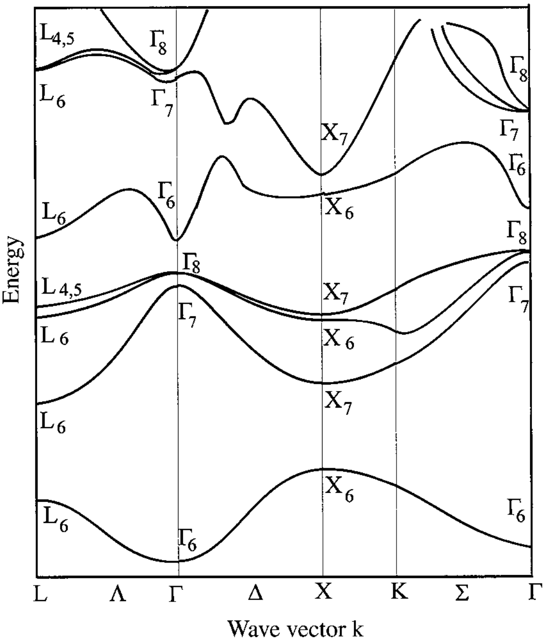


Figure 3.11 A generic energy band structure of GaAs plotted as function of a wavevector. The point group symmetries at L, Γ, and X points in the first Brillouin zone are shown.

The $\mathbf{k}\cdot\mathbf{p}$ method is an approximation method that is most valid in the vicinity of the band extrema of electrons and holes in single-crystal solids such as semiconductors. This method can be derived from the one-electron Schrödinger equation using Bloch wave functions in the reduced zone scheme. The wave function for an extrema at the wavevector $\mathbf{k}_o$ can be written as

$$\begin{aligned}\psi_{n\mathbf{k}}(\mathbf{r}) &= e^{i\mathbf{k}\cdot\mathbf{r}}u_{n\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}}e^{i\mathbf{k}_o\cdot\mathbf{r}}e^{-i\mathbf{k}_o\cdot\mathbf{r}}u_{n\mathbf{k}}(\mathbf{r})\\ &= e^{i\mathbf{k}_o\cdot\mathbf{r}}e^{i(\mathbf{k}-\mathbf{k}_o)\cdot\mathbf{r}}u_{n\mathbf{k}}(\mathbf{r})\\ &= e^{i\mathbf{k}_o\cdot\mathbf{r}}e^{i\Delta\mathbf{k}\cdot\mathbf{r}}u_{n\mathbf{k}_o+\Delta\mathbf{k}}(\mathbf{r})\end{aligned} \tag{3.75}$$

where $\Delta\mathbf{k} = \mathbf{k} - \mathbf{k}_o$. Substitute the wave function into the Schrödinger equation

$$\left[\frac{p^2}{2m^*} + V(\mathbf{r})\right]\psi_{n\mathbf{k}}(\mathbf{r}) = E_{n\mathbf{k}}\psi_{n\mathbf{k}}(\mathbf{r}) \tag{3.76}$$

where $\mathbf{p} = -i\hbar\nabla$ is the momentum operator and $V(\mathbf{r})$ is the periodic potential, to obtain

$$\left[\frac{\mathbf{p}^2}{2m} + \frac{\hbar\mathbf{k}_o\cdot\mathbf{p}}{m} + \frac{\hbar\Delta\mathbf{k}\cdot\mathbf{p}}{m} + \frac{\hbar^2(\mathbf{k}_o+\Delta\mathbf{k})^2}{2m} + V(\mathbf{r})\right] u_{n\mathbf{k}_o+\Delta\mathbf{k}}(\mathbf{r})$$
$$= E_{n\mathbf{k}_o+\Delta\mathbf{k}} u_{n\mathbf{k}_o+\Delta\mathbf{k}}(\mathbf{r}) \tag{3.77}$$

The subscript n indicates the energy level of interest. If one assumes that the wave functions $u_{n\mathbf{k}_o}(\mathbf{r})$ and the eigenvalues $E_{n\mathbf{k}_o}$are solved for the case of $\mathbf{k} = \mathbf{k}_o$, the term involving $\Delta\mathbf{k}$ can be treated as a perturbation term. Additionally, the terms $\hbar^2\mathbf{k}_o\cdot\Delta\mathbf{k}/m$ and $\hbar^2(\Delta\mathbf{K})^2/2m$ are constants and can be combined with$E_{n\mathbf{k}_o}$. If the energy $E_{n\mathbf{k}}$ has an extrema at $\mathbf{k} = \mathbf{k}_o$, the terms linear in $\Delta\mathbf{k}$ must vanish due to the fact that the first derivative of the energy is zero at the extrema. Putting all these together, Eq. (3.77) can be reduced to

$$\left[\frac{\mathbf{p}^2}{2m} + \frac{\hbar\mathbf{k}\cdot\mathbf{p}}{m} + \frac{\hbar^2\mathbf{k}^2}{2m} + V(\mathbf{r})\right] u_{n\mathbf{k}}(\mathbf{r}) = E_{n\mathbf{k}} u_{n\mathbf{k}}(\mathbf{r}) \tag{3.78}$$

Notice, that we replaced $\mathbf{k}_o + \Delta\mathbf{k}$ with $\mathbf{k}$ in Eq. (3.78).

For a nondegenerate band such as the conduction band (Γ_1 symmetry), one can obtain the effective mass as follows. Using standard nondegenerate perturbation theory, the wave function $u_{n\mathbf{k}}$ and the eigenvalues $E_{n\mathbf{k}}$ can be expanded to second-order $\mathbf{k}$ in terms of the unperturbed wave functions $u_{n\mathbf{k}_o}$ and eigenvalues $E_{n\mathbf{k}_o}$such as

$$u_{n\mathbf{k}} = u_{n\mathbf{k}_o} + \frac{\hbar}{m}\sum_{n'\neq n} \frac{\langle u_{n\mathbf{k}_o}|\mathbf{k}\cdot\mathbf{p}|u_{n'\mathbf{k}_o}\rangle}{E_{n\mathbf{k}_o} - E_{n'\mathbf{k}_o}} u_{n'\mathbf{k}_o} \tag{3.79}$$

and

$$E_{n\mathbf{k}} = E_{n\mathbf{k}_o} + \frac{\hbar^2\mathbf{k}^2}{2m} + \frac{\hbar^2}{m^2}\sum_{n'\neq n} \frac{|\langle u_{n\mathbf{k}_o}|\mathbf{k}\cdot\mathbf{p}|u_{n'\mathbf{k}_o}\rangle|^2}{E_{n\mathbf{k}_o} - E_{n'\mathbf{k}_o}} \tag{3.80}$$

On the other hand, Eq. (3.80) is usually expressed as

$$E_{n\mathbf{k}} = E_{n\mathbf{k}_o} + \frac{\hbar^2\mathbf{k}^2}{2m^*} \tag{3.81}$$

where m^* is defined as the effective mass of the band. It is thus clear from Eqs. (3.80) and (3.81) that the effective mass can be written as

$$\frac{1}{m^*} = \frac{1}{m} + \frac{2}{m^2\mathbf{k}^2}\sum_{n'\neq n} \frac{|\langle u_{n\mathbf{k}_o}|\mathbf{k}\cdot\mathbf{p}|u_{n'\mathbf{k}_o}\rangle|^2}{E_{n\mathbf{k}_o} - E_{n'\mathbf{k}_o}} \tag{3.82}$$

This equation shows that the electron mass in a period potential is different from that of the free electron due to the coupling of electronic states in different bands through the $\mathbf{k}\cdot\mathbf{p}$ term. The matrix elements $\langle u_{n\mathbf{k}_o}|\mathbf{p}|u_{n'\mathbf{k}_o}\rangle$ are nonzero for the conduction band of symmetry Γ_1 and the valence band of symmetry Γ_4 of zinc-blende structures such as GaAs. Equation (3.82) also shows that the energy separation $(E_{n\mathbf{k}_o} - E_{n'\mathbf{k}_o})$ between the two bands n and n' determine the relative importance of the contribution of n' to the effective mass of n. Thus, if $E_{n\mathbf{k}_o} > E_{n'\mathbf{k}_o}$, the n' bands will contribute a positive term to $1/m^*$ as is the case for the conduction band in zinc-blende structures and if $E_{n\mathbf{k}_o} < E_{n'\mathbf{k}_o}$, the n' bands will contribute a negative term to $1/m^*$ as is the case for the top of the valence bands.

The effective mass in the conduction band in direct semiconductors is determined mainly by the coupling of the Γ_1 conduction band and the Γ_4 valence band via the $\mathbf{k}\cdot\mathbf{p}$ term as follows:

$$\frac{1}{m_c^*} = \frac{1}{m_o} + \frac{2|\langle\Gamma_{1c}|\mathbf{k}\cdot\mathbf{p}|\Gamma_{4v}\rangle|^2}{m_o^2\mathbf{k}^2E_g} \tag{3.83}$$

where m_c^* = conduction band effective mass

m_o = free-electron mass

$\Gamma_{1c} = \Gamma_1$ conduction band

$\Gamma_{4v} = \Gamma_4$ valence band

$E_g = E_{n\mathbf{k}_o} - E_{n'\mathbf{k}_o}$ = band gap

Moreover, it is customary to represent Γ_4 wave functions as $|X\rangle$, $|Y\rangle$, and $|Z\rangle$. From the T_d symmetry, it can be shown that the only nonzero elements of $\langle\Gamma_{1c}|\,\widehat{\mathbf{k}}\cdot\mathbf{p}\,|\Gamma_{4v}\rangle$ are

$$\langle X|\,\mathbf{p}_x\,|\Gamma_1\rangle = \langle Y|\,\mathbf{p}_y\,|\Gamma_1\rangle = \langle Z|\,\mathbf{p}_z\,|\Gamma_1\rangle = iP \tag{3.84}$$

Substituting Eq. (3.84) into (3.83), one can write the effective mass as

$$\frac{m_o}{m_c^*} \approx 1 + \frac{2P^2}{m_oE_g} \tag{3.85}$$

It was found that $2P^2/m_o \approx 20$ eV since P for most zinc-blende semiconductors is close to those calculated for the nearly free electron $P = 2\pi\hbar/a_o$. The bandgap E_g is typically less than 2 eV, which leads to $2P^2/(m_oE_g) \gg 1$, and Eq. (3.85) is reduced to

$$\frac{m_o}{m_c^*} \approx \frac{2P^2}{m_oE_g} \tag{3.86}$$

For GaAs, $E_g = 1.52$ eV at 4.2 K, and this gives $m_c^* \approx 0.076m_o$. This value is in good agreement with the effective mass of $0.067m_o$ measured by cyclotron resonance technique.

The analysis is more complicated for degenerate bands such as the top of the valence band in zinc-blende semiconductors. To apply the $\mathbf{k} \cdot \mathbf{p}$ method to calculate the dispersion near the top of the valence band in direct zinc-blende semiconductors, one needs to consider the highest Γ_4 valence band at the center of the Brillouin zone. The wave functions of the valence bands are p-like and are denoted by $|X\rangle$, $|Y\rangle$, and $|Z\rangle$. The electron spin is $\frac{1}{2}$, and the wave functions are denoted by ↑ and ↓ for spin-up and spin-down, respectively. For semiconductors with heavier atoms such as Ga, As, and Sb, one expects that spin-orbit coupling to be significant, and it must be included in the unperturbed Hamiltonian ($\mathbf{H}_{so}$) for states near $\mathbf{k} = 0$. This spin-orbit Hamiltonian can be written as

$$\mathbf{H}_{so} = \lambda \, \boldsymbol{l} \cdot \mathbf{s} \tag{3.87}$$

where $\lambda =$ spin-orbit coupling constant
$\boldsymbol{l} =$ angular momentum
$\mathbf{s} =$ spin

The total angular momentum can be defined as $\mathbf{j} = \boldsymbol{l} + \mathbf{s}$ and the z-component of $\mathbf{j}$ is $m_j = \pm j, \pm(j-1), \ldots$ The wave functions of j and m_j are expressed as linear combinations of the wave functions of the orbital angular momentum and spin ↑ and ↓ as follows:

$$|j m_j\rangle = \begin{cases} \left|\frac{3}{2}, \frac{3}{2}\right\rangle = |1,1\rangle \uparrow \\ \left|\frac{3}{2}, \frac{1}{2}\right\rangle = \frac{1}{\sqrt{3}}(|1,1\rangle \downarrow + \sqrt{2}|1,0\rangle \uparrow) \\ \left|\frac{3}{2}, -\frac{1}{2}\right\rangle = \frac{1}{\sqrt{3}}(|1,-1\rangle \uparrow + \sqrt{2}|1,0\rangle \downarrow) \\ \left|\frac{3}{2}, -\frac{3}{2}\right\rangle = |1,-1\rangle \downarrow \\ \left|\frac{1}{2}, \frac{1}{2}\right\rangle = \frac{1}{\sqrt{3}}(|1,0\rangle \uparrow - \sqrt{2}|1,1\rangle \downarrow) \\ \left|\frac{1}{2}, -\frac{1}{2}\right\rangle = \frac{1}{\sqrt{3}}(|1,0\rangle \downarrow - \sqrt{2}|1,-1\rangle \uparrow) \end{cases} \tag{3.88}$$

The p-like Γ_4 states in a zinc-blende crystal can be compared to the atomic p-wave functions to define the "$l = 1$"-like state such as

$$\begin{aligned} |1,1\rangle &= -(|X\rangle + iY\rangle)/\sqrt{2} \\ |1,0\rangle &= |Z\rangle \\ |1,-1\rangle &= (|X\rangle - i|Y\rangle)/\sqrt{2} \end{aligned} \tag{3.89}$$

where the $j = \frac{3}{2}$ and $j = \frac{1}{2}$ states can be obtained by substituting Eq. (3.89) into (3.88).

The nonzero momentum matrix elements due to coupling Γ_{1c} and Γ_{4v} are given by Eq. (3.84), while the nonzero matrix elements between Γ_{4v}

and the sixfold degenerate Γ_{4c} band states are

$$\begin{aligned}\langle X|\,\mathbf{p}_y\,|\Gamma_{4c}(z)\rangle &= \langle Y|\,\mathbf{p}_z\,|\Gamma_{4c}(x)\rangle = \langle Z|\,\mathbf{p}_x\,|\Gamma_{4c}(y)\rangle = iQ\\ \langle X|\,\mathbf{p}_z\,|\Gamma_{4c}(y)\rangle &= \langle Y|\,\mathbf{p}_x\,|\Gamma_{4c}(z)\rangle = \langle Z|\,\mathbf{p}_y\,|\Gamma_{4c}(x)\rangle = iQ\end{aligned} \tag{3.90}$$

where $|X\rangle$, $|Y\rangle$, and $|Z\rangle$, as mentioned earlier, are the wave functions for Γ_{4v}.

The effective Hamiltonian can now be obtained by calculating the 6×6 matrix elements from the following relation:

$$H'_{ij} = H_{ij} + \sum_{\mathbf{k}\neq\Gamma_{4v}} \frac{H_{i\mathbf{k}}H_{\mathbf{k}j}}{E_i - E_{\mathbf{k}}} \tag{3.91}$$

To simplify the notation, the six Γ_{4v} wave functions can be written as

$$\begin{aligned}\phi_1 &= |3/2, 3/2\rangle & \phi_2 &= |3/2, 1/2\rangle\\ \phi_3 &= |3/2, -1/2\rangle & \phi_4 &= |3/2, -3/2\rangle\\ \phi_5 &= |1/2, 1/2\rangle & \phi_6 &= |1/2, -1/2\rangle\end{aligned} \tag{3.92}$$

and the doubly degenerate Γ_{1c} and sixfold degenerate Γ_{4c} conduction band wave functions as ϕ_7 to ϕ_{14}. The H'_{11} can be calculated as follows:

$$\begin{aligned}H'_{11} &= \langle\phi_1|\,\frac{\hbar^2\mathbf{k}^2}{2m_o} + \frac{\hbar\mathbf{k}\cdot\mathbf{p}}{m_o}\,|\phi_1\rangle + \sum_j \left|\langle\phi_1|\,\frac{\hbar^2\mathbf{k}^2}{2m_o} + \frac{\hbar\mathbf{k}\cdot\mathbf{p}}{m_o}|\phi_j\rangle\right|^2 \frac{1}{E_1 - E_j}\\ &= \frac{\hbar^2\mathbf{k}^2}{2m_o} + \langle\phi_1|\,\frac{\hbar\mathbf{k}\cdot\mathbf{p}}{m_o}\,|\phi_1\rangle - \left(\left|\langle\phi_1|\,\frac{\hbar\mathbf{k}\cdot\mathbf{p}}{m_o}\,|\Gamma_{1c}\rangle\right|^2 \frac{1}{E_o}\right)\\ &\quad - \left(\left|\langle\phi_1|\,\frac{\hbar\mathbf{k}\cdot\mathbf{p}}{m_o}\,|\Gamma_{4c}\rangle\right|^2 \frac{1}{E'_o}\right)\end{aligned} \tag{3.93}$$

where E_o is the bandgap between Γ_{1c} and the $j = \frac{3}{2}$ valence band, E'_0 is the bandgap between Γ_{4c} and $j = \frac{3}{2}$ valence band, and

$$\begin{aligned}\left(\left|\langle\phi_1|\,\frac{\hbar\mathbf{k}\cdot\mathbf{p}}{m_o}\,|\Gamma_{1c}\rangle\right|^2 \frac{1}{E_o}\right) &= \frac{1}{2}L\left(k_x^2 + k_y^2\right)\\ \left(\left|\langle\phi_1|\,\frac{\hbar\mathbf{k}\cdot\mathbf{p}}{m_o}\,|\Gamma_{4c}\rangle\right|^2 \frac{1}{E'_o}\right) &= \frac{1}{2}M\left(k_x^2 + k_y^2 + 2k_z^2\right)\end{aligned} \tag{3.94}$$

Here, the following terms are introduced for simplicity:

$$\begin{aligned}
L &= \frac{-\hbar^2 P^2}{m_o^2 E_o} \\
M &= \frac{-\hbar^2 Q^2}{m_o^2 E_o'} \\
N &= L + M \\
L' &= \frac{-\hbar^2 P^2}{m_o^2 (E_o + \Delta_0)} \\
M' &= \frac{-\hbar^2 Q^2}{m_o^2 (E_o' + \Delta_0)}
\end{aligned} \tag{3.95}$$

The final result forH'_{11} is

$$H'_{11} = \frac{\hbar^2 \mathbf{k}^2}{2m_o} + \frac{1}{2}\left(k_x^2 + k_y^2\right) + M k_z^2 \tag{3.96}$$

Similar procedures can be used to obtain the rest of the matrix elements. These matrix elements are as follows:

$$\begin{aligned}
H'_{12} &= \frac{N}{\sqrt{3}}(k_x k_z - i k_y k_z) & H'_{13} &= \frac{1}{2\sqrt{3}}\left[(L - M)\left(k_x^2 - k_y^2\right) - 2iN k_x k_y\right] \\
H'_{14} &= 0 & H'_{15} &= \frac{1}{\sqrt{2}} H'_{12} \\
H'_{16} &= \sqrt{2} H'_{13} & H'_{22} &= \frac{\hbar^2 \mathbf{k}^2}{2m_o} + \frac{1}{3}(M + 2L)\mathbf{k}^2 \\
& & &\quad - \frac{1}{2}(L - M)\left(k_x^2 + k_y^2\right) \\
H'_{23} &= 0 & H'_{24} &= H'_{13} \\
H'_{25} &= \frac{1}{\sqrt{2}}(H'_{22} - H'_{11}) & H'_{26} &= \sqrt{\frac{3}{2}} H'_{12} \\
H'_{33} &= H'_{22} & H'_{34} &= -H'_{12} \\
H'_{35} &= -(H'_{26})^* & H'_{36} &= H'_{25} \\
H'_{44} &= H'_{11} & H'_{45} &= -\sqrt{2}(H'_{13})^* \\
H'_{46} &= -(H'_{15})^* & H'_{55} &= \frac{\hbar^2 \mathbf{k}^2}{2m_o} + \frac{1}{3}\Bigg[(2M' + L')\mathbf{k}^2 \\
& & &\quad - \frac{1}{2}(L - M)\Bigg] - \Delta_o \\
H'_{56} &= 0 & H'_{66} &= H'_{55}
\end{aligned} \tag{3.97}$$

The matrix $\{H'_{ij}\}$ is hermitian, which means $H'_{ij} = [H'_{ji}]^*$, and the matrix can be written as

$$H'_{ij} = \begin{bmatrix} H'_{11} & H'_{12} & H'_{13} & 0 & H'_{15} & H'_{16} \\ (H'_{12})^* & H'_{22} & 0 & H'_{13} & \sqrt{\frac{1}{2}}(H'_{22} - H'_{11}) & \sqrt{\frac{3}{2}}H'_{12} \\ (H'_{13})^* & 0 & H'_{22} & -H'_{12} & -(H'_{26})^* & H'_{25} \\ 0 & (H'_{13})^* & (-H'_{12})^* & H'_{11} & -\sqrt{2}(H'_{13})^* & -(H'_{15})^* \\ (H'_{15})^* & [\sqrt{\frac{1}{2}}(H'_{22} - H'_{11})]^* & -(H'_{26}) & -\sqrt{2}(H'_{13}) & H'_{55} & 0 \\ (H'_{16})^* & (\sqrt{\frac{3}{2}}H'_{12})^* & (H'_{25})^* & -(H'_{15}) & 0 & H'_{66} \end{bmatrix} \tag{3.98}$$

This matrix can be diagonalized with some approximation, such as for small **k**, the matrix elements H'_{15}, H'_{16}, and H'_{25} are zero and by limiting the eigenvalues to $\mathbf{k}^2$ terms only, the 6 × 6 matrix reduces to 4 × 4 and 2 × 2 matrices.

The 2 × 2 matrix gives the doubly degenerate $j = \frac{1}{2}\Gamma_7$ band as

$$\begin{aligned} E_{so} = H'_{55} &= \frac{\hbar^2\mathbf{k}^2}{2m_o} + \frac{1}{3}\left[(2M' + L')\mathbf{k}^2 - \frac{1}{2}(L - M)\right] - \Delta_o \\ &= -\Delta_o + \frac{\hbar^2\mathbf{k}^2}{2m_o}\left\{1 - \frac{2}{3}\left[\frac{P^2}{m\,(E_o + \Delta_o)} + \frac{Q^2}{m\big(E'_o + \Delta_o\big)}\right]\right\} \end{aligned} \tag{3.99}$$

which yields a constant spherical surface for the spin-orbit valence band and an effective mass given by

$$\frac{m_o}{m_{so}} = 1 - \frac{2}{3}\left[\frac{P^2}{m_o(E_o + \Delta_o)} + \frac{Q^2}{m_o(E'_o + \Delta_o)}\right] \tag{3.100}$$

The dispersion for the $j = \frac{3}{2}$ bands was first obtained by Dresselhaus et al. (1955) by diagonalizing the following 4 × 4 matrix:

$$H'_{ij} = \begin{bmatrix} H'_{11} & H'_{12} & H'_{13} & 0 \\ (H'_{12})^* & H'_{22} & 0 & H'_{13} \\ (H'_{13})^* & 0 & H'_{22} & -H'_{12} \\ 0 & (H'_{13})^* & (-H'_{12})^* & H'_{11} \end{bmatrix} \tag{3.101}$$

The secular equation for this matrix reduces to two identical equations of the form

$$(H'_{11} - E)(H'_{22} - E) = |H'_{12}|^2 + |H'_{13}|^2 \tag{3.102}$$

The solutions for this equation are

$$E_{\pm} = \frac{1}{2}(H'_{11} + H'_{22}) \pm \frac{1}{2}\sqrt{(H'_{11} + H'_{22})^2 - 4(H'_{11}H'_{22} - |H'_{12}|^2 - |H'_{13}|^2)}$$

$$= A\mathbf{k}^2 \pm \sqrt{B^2\mathbf{k}^2 + C^2\left(k_x^2k_y^2 + k_y^2k_z^2 + k_z^2k_x^2\right)} \tag{3.103}$$

where A, B, and $|C|^2$ are called the valence band parameters and are related to the momentum matrix elements and the energy gaps by

$$\frac{2m_o}{\hbar^2}A = 1 - \frac{2}{3}\left(\frac{P^2}{m_oE_o} + \frac{2Q^2}{m_oE'_o}\right) \tag{3.104a}$$

$$\frac{2m_o}{\hbar^2}B = \frac{2}{3}\left(\frac{-P^2}{m_oE_o} + \frac{Q^2}{m_oE'_o}\right) \tag{3.104b}$$

$$\left(\frac{2m_o}{\hbar^2}C\right)^2 = \frac{16P^2Q^2}{3m_o^2E_oE'_o} \tag{3.104c}$$

They are given in the units of $(\hbar^2/2m_o)^2$, and their values for GaAs are -7.0, -4.5, and 38, respectively. The dispersion of Γ_8 $\left(j = \frac{3}{2}\right)$ bands near the zone center is given by Eq. (3.103), which was derived after much simplification, and it is valid only for energies smaller than the spin-orbit splitting. Moreover, the values of A and B are negative due to the dominant $2P^2/(3m_oE_o) \gg 1$ term in Eq. (3.104). This implies that the effective masses of these bands are negative. The concept of heavy and light holes can be introduced here as them being particles with negative masses. Their energies are given by

$$E_{\text{hh}} = A\mathbf{k}^2 - \sqrt{B^2\mathbf{k}^2 + C^2\left(k_x^2k_y^2 + k_y^2k_z^2 + k_z^2k_x^2\right)} \tag{3.105a}$$

$$E_{\text{lh}} = A\mathbf{k}^2 + \sqrt{B^2\mathbf{k}^2 + C^2\left(k_x^2k_y^2 + k_y^2k_z^2 + k_z^2k_x^2\right)} \tag{3.105b}$$

where hh stands for "heavy hole" and lh stands for "light hole." The hole band dispersions along the [100] and [111] directions are parabolic, but the hole masses are different along the two directions according to the following:

$\mathbf{k}||(100)$:

$$\frac{1}{m_{\text{hh}}} = \frac{2}{\hbar^2}(-A + B) \tag{3.106a}$$

$$\frac{1}{m_{\text{lh}}} = \frac{2}{\hbar^2}(-A - B) \tag{3.106b}$$

and

$$\mathbf{k}||(111): \qquad \frac{1}{m_{\mathrm{hh}}} = \frac{2}{\hbar^2}\left[-A + B\left(1 + \frac{|C|^2}{3B^2}\right)^{1/2}\right] \qquad (3.107a)$$

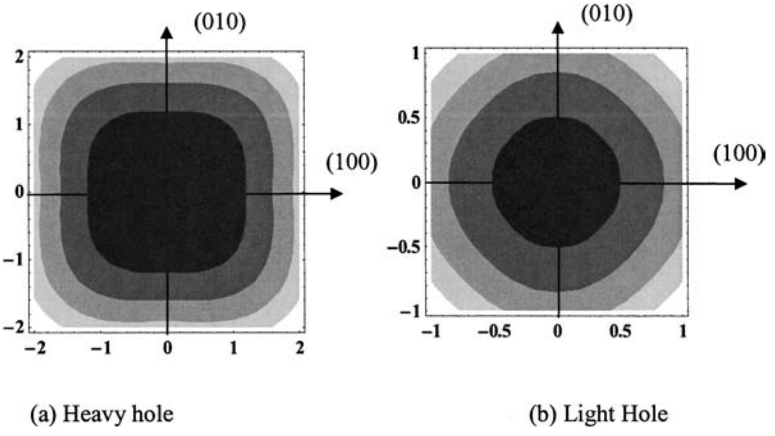
$$\frac{1}{m_{\mathrm{lh}}} = \frac{2}{\hbar^2}\left[-A - B\left(1 + \frac{|C|^2}{3B^2}\right)^{1/2}\right] \qquad (3.107b)$$

Taking the average of Eqs. (3.106) and (3.107) and expanding the square root, we have

$$\frac{1}{m^*_{\mathrm{hh}}} = \frac{1}{\hbar^2}\left[-2A + 2B\left(1 + \frac{|C|^2}{12B^2}\right)\right] \qquad (3.108a)$$

$$\frac{1}{m^*_{\mathrm{lh}}} = \frac{1}{\hbar^2}\left[-2A - 2B\left(1 + \frac{|C|^2}{12B^2}\right)\right] \qquad (3.108b)$$

Notice that in the limit of $C = 0$, Eq. (3.108) is reduced to Eq. (3.106). By plugging in the values of $A = -7$, $B = -4.5$, and $|C|^2 = 38$ for GaAs into Eq. (3.108), we find that $m^*_{\mathrm{hh}} = 0.556m_o$ and $m^*_{\mathrm{lh}} = 0.079m_o$, which are in good agreement with the experimental values of $m^*_{\mathrm{hh}} = 0.53m_o$ and $m^*_{\mathrm{lh}} = 0.08m_o$. The energy contours of the heavy and light holes are shown in Fig. 3.12. These contours are called *warped* spheres.



(a) Heavy hole (b) Light Hole

Figure 3.12 Constant energy surfaces of the $J = \frac{3}{2}(\Gamma_8)$ bands in zinc-blende semiconductors.

Another note is that the parameters A, B, and C are related to the Kohn-Luttinger parameters, γ_1, γ_2, and γ_3 according to the following relations:

$$\begin{aligned}\frac{\hbar^2}{2m_o}\gamma_1 &= -A\\ \frac{\hbar^2}{2m_o}\gamma_2 &= -\frac{\mathrm{B}}{2}\\ \frac{\hbar^2}{2m_o}\gamma_3 &= \sqrt{\frac{B^2}{4}+\frac{C^2}{12}}\end{aligned} \tag{3.109}$$

The Kohn-Luttinger parameters for GaAs are $\gamma_1 = 6.790, \gamma_2 = 1.924$, and $\gamma_3 = 2.681$. For $\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{As}$, these parameters are $\gamma_1 = 6.790 - 3.0x, \gamma_2 = 1.924 - 0.694x$, and $\gamma_3 = 2.681 - 1.286x$. The Kohn-Luttinger parameters are introduced in the following $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian, which was derived for $\Gamma_4$valence bands:

$$\begin{aligned}\mathbf{H}_L = \frac{\hbar^2}{2m_o}\Bigg[\left(\gamma_1+\frac{5}{2}\gamma_2\right)\nabla^2 - 2\gamma_2(\nabla\cdot\mathbf{J})^2 + 2\,(\gamma_3-\gamma_2)\\ \times\left(\nabla_x^2 J_x^2 + \nabla_y^2 J_y^2 + \nabla_z^2 J_z^2\right)\Bigg]\end{aligned} \tag{3.110}$$

with the following eigenvalues:

$$E_\pm = \frac{\hbar^2}{2m_o}\left\{\gamma_1\mathbf{k}^2 \pm \sqrt{4\gamma_2^2\mathbf{k}^4 + 12\left(\gamma_3^2-\gamma_2^2\right)\left(k_x^2k_y^2+k_y^2k_z^2+k_z^2k_x^2\right)}\right\} \tag{3.111}$$

where $\mathbf{J} = (J_x, J_y, J_z)$ is an operator whose effects on the Γ_4 valence bands are identical to those of the angular momentum operator on the $j = \frac{3}{2}$ atomic states.

3.7.3 Envelope function approximation

The theoretical calculations of the energy states in the semiconductor heterostructures are very complicated and require computer analysis. One approach is to find the boundary conditions at the heterojunction interfaces where the wave functions are almost invariant. This approach is called envelope function approximation and has been used successfully (see, for example, Bastard 1988 and Bastard et al. 1991) in determining the energy states in quantum wells, superlattices, and heterojunctions. Following Bastard's formalism, the envelope function approximation assumes that the constituents of the quantum wells are lattice-matched with abrupt interfaces such as GaAs/AlGaAs multiple quantum wells, as shown in Fig. 3.13. This figure shows two types of multiple quantum wells. Type I is illustrated in Fig. 3.13*a*, where the

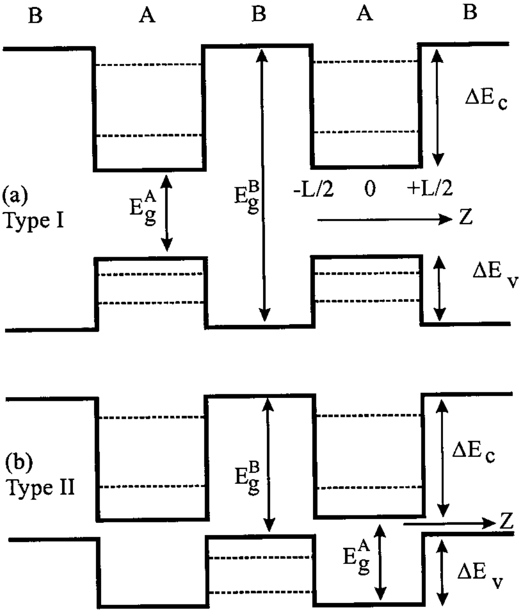


Figure 3.13 A sketch of the band alignment of the quantum wells for (*a*) type I such as GaAs/AlGaAs quantum wells and (*b*) type II such as InAs/InGaSb superlattices. The dashed lines represent the confined energy levels in both the conduction and valence quantum wells. The layers *A* and *B* are designated as the wells and barriers, respectively.

electrons and holes exist in the same layer. An example of this type is GaAs/AlGaAs multiple quantum wells. The second type of band alignment is called type II, which is illustrated in Fig. 3.13*b* where the electrons are located in one layer and the holes are located in the adjacent layer. A typical example of this band alignment is found in InAs/ InGaSb superlattices.

Inside each layer of the multiple quantum wells, the wave function is expanded in the periodic part of the Bloch functions such as

$$\psi(\mathbf{r}) = \sum_n f_n^A(\mathbf{r}) u_{n\mathbf{k}_o}^A \qquad \text{for } -\frac{L}{2} \le z \le \frac{L}{2} \text{ (well)} \tag{3.112a}$$

and

$$\psi(\mathbf{r}) = \sum_n f_n^B(\mathbf{r}) u_{n\mathbf{k}_o}^B \qquad \text{for } z > \frac{L}{2} \quad \text{or} \quad z < -\frac{L}{2} \text{ (barrier)} \tag{3.112b}$$

where $f_n^A(\mathbf{r})$ is the envelope function and $u_{n\mathbf{k}_o}^A(\mathbf{r})$ is the Bloch function in the well. Equation (3.112) stands for the barrier, and the summation is over all the finite energy bands. If one assumes that the Bloch function is the same for the well and the barrier, then Eq. (3.112) becomes

$$\psi(\mathbf{r}) = \sum_n f_n^{A,B}(\mathbf{r}) u_{n\mathbf{k}_o} \tag{3.113}$$

It is thus required to determine the function $f_n^{A,B}(\mathbf{r})$, where $f_n^A(\mathbf{r})$ stands for the wave function in the well and $f_n^B(\mathbf{r})$ stands for the wave function in the barrier. The assumption of identical Bloch functions in the well and the barrier implies that the interband matrix element is the same for the barrier and the well, which yields

$$f_n^A(\mathbf{r}_\perp, z_o) = f_n^B(\mathbf{r}_\perp, z_o) \tag{3.114}$$

where $\mathbf{r}_\perp$ is the xy plane and the subject interface along the growth axis, which is z, occurs at $z = z_o$. Since the lattice constants are assumed to be the same for both the well and the barrier, the envelope wave function can be factorized as

$$f_n^{A,B}(\mathbf{r}_\perp, z) = \frac{1}{\sqrt{a}} e^{i(\mathbf{k}_\perp \cdot \mathbf{r}_\perp)} \chi_n^{A,B}(z) \tag{3.115}$$

where a is the sample area and $\mathbf{k}_\perp$ is the (k_x, k_y) bi-dimensional wavevector, which is assumed to be the same for both A and B, and $\chi_n^{A,B}(z)$ is a slowly varying function with respect to the host's unit cells. To summarize, the heterojunction wave function $\psi(r)$ is composed of the sum of the product of the rapidly varying Bloch function $u_{n\mathbf{k}_o}(r)$ and the slowly varying envelop function $f_n^{A,B}$.

Since the effective masses in the A and B layers are different, the equation of motion of the envelope functions inside the well and the barrier are

$$\left[E_c + V_c(z) - \frac{\hbar^2}{2\mu(z)}\left(\frac{d^2}{dz^2} + k_x^2 + k_y^2\right)\right]\chi_{A,B}(z) = E\chi_{A,B}(z) \tag{3.116}$$

where E_c is the conduction band edge in the well, which can be set to zero, $V_c(z)$ is zero in the well and equal to the conduction band offset in the barrier, and $\mu(z)$ is either m_A^* in the well or m_B^* in the barrier. Notice that $\mathbf{k}_\perp = \sqrt{k_x^2 + k_y^2}$. We assume that Eq. (3.116) is written for the conduction band, but the case of the valence band is more complicated. The boundary conditions are those of the BenDaniel-Duke conditions such that

$$\chi_A\left(\pm\frac{L}{2}\right) = \chi_B\left(\pm\frac{L}{2}\right)$$

and (3.117)

$$\left.\frac{1}{m_A^*}\frac{d\chi_A(z)}{dz}\right|_{z=\pm L/2} = \left.\frac{1}{m_B^*}\frac{d\chi_B(z)}{dz}\right|_{z=\pm L/2}$$

The effective mass mismatch leads to a discontinuity in the derivative of the envelope function at the interfaces and $\mathbf{k}_\perp$adds a steplike variation to the barrier $V_c(z)$ since the effective potential is now $V_c(z) + \hbar^2\mathbf{k}_\perp^2/[2\mu(z)]$. The wave functions of the bound states can be chosen for the even state as

$$\begin{aligned} \chi_{\text{even}}(z) &= A\cos(\mathbf{k}_A z) && \text{for the well} \\ \chi_{\text{even}}(z) &= Be^{[-\mathbf{k}_B(z-L/2)]} && \text{for the barrier} \end{aligned} \tag{3.118}$$

and

$$\chi_{\text{even}}(-z) = \chi_{\text{even}}(z)$$

and for the odd state as

$$\begin{aligned} \chi_{\text{odd}}(z) &= A\sin(\mathbf{k}_A z) && \text{for the well} \\ \chi_{\text{odd}}(z) &= Be^{[-\mathbf{k}_B(z-L/2)]} && \text{for the barrier} \end{aligned} \tag{3.119}$$

and

$$\chi_{\text{odd}}(-z) = -\chi_{\text{odd}}(z)$$

where the even and odd states are shown schematically in Fig. 3.14. The wavevectors are given as follows

$$\begin{aligned} \mathbf{k}_A &= \sqrt{\frac{2m_A^* E}{\hbar^2} - \mathbf{k}_\perp^2} \\ \mathbf{k}_B &= \sqrt{\frac{2m_B^*(V_c - E)}{\hbar^2} - \mathbf{k}_\perp^2}, && \text{for } E < V_c \end{aligned} \tag{3.120}$$

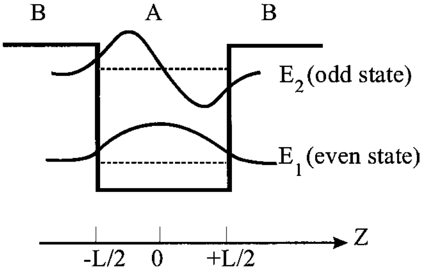


Figure 3.14 A sketch of the conduction quantum well plotted with the ground state (even) and the excited state (odd).

The BenDaniel-Duke boundary conditions give the following straightforward relations from which the energy levels are obtained:

$$\tan(\mathbf{k}_A L/2) = \frac{m_A^* \mathbf{k}_B}{m_B^* \mathbf{k}_A} \qquad \text{for the even state} \tag{3.121}$$

and

$$-\cot(\mathbf{k}_A L/2) = \frac{m_A^* \mathbf{k}_B}{m_B^* \mathbf{k}_A} \qquad \text{for the odd state} \tag{3.122}$$

The left-hand sides of Eqs. (3.121) and (3.122) are plotted as a function of $(\mathbf{k}_A L/2)$ in unit of $\mathbf{k}_B$ for several values of m_B^*/m_A^* as shown in Fig. 3.15 with $\mathbf{k}_\perp = 0$. The nodes indicate the intersections that correspond to the values of $(\mathbf{k}_A L/2)$ from which the bound states can be obtained. It is clear from this figure that the ratio of the effective masses plays a major role in determining the eigenvalues. As m_B^*/m_A^* is increased, the values of the bound energy levels are decreased.

Another note from Fig. 3.15 is that the energy levels seem to reach constant values for higher m_B^*/m_A^* ratios, which implies that $\mathbf{k}_A L = n\pi$ for $n = 0, 1, 2, \ldots$ This result resembles the infinite-depth potential well where the wave function is zero at the boundary conditions. This is actually the essence of the envelope function approximation where the wave function vanishes at the interfaces. By varying $L/2$ and reploting

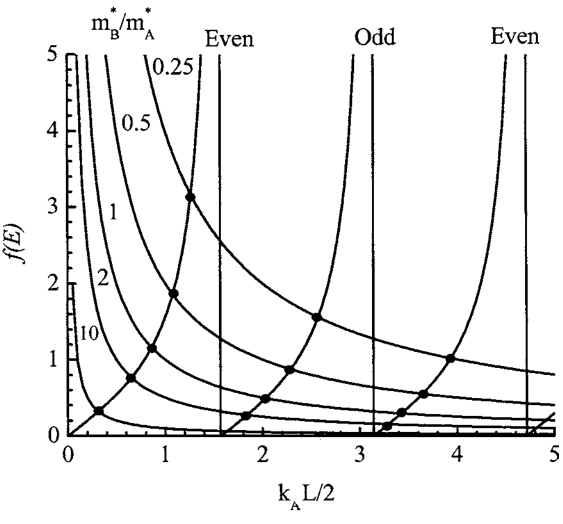


Figure 3.15 The left hand-sides of Eqs. (3.121) and (3.122) plotted as a function of $(\mathbf{k}_A L/2)$. The right-hand side of either equation is plotted for several values of m_B^*/m_A^* in units of $\mathbf{k}_B$. The nodes indicate the intersections from which the energy levels are obtained.

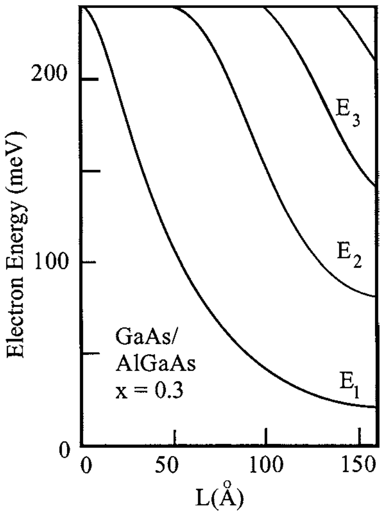


Figure 3.16 The energy-level diagram for electrons in GaAs/AlGAs quantum wells versus the GaAs well thickness. The conduction band offset was taken as 0.245 eV_o. (*After Bastard 1988*)

Fig. 3.15, one can obtain the confined energy levels as a function of the well width. A typical example is shown in Fig. 3.16.

The energy levels of electrons in superlattices can be calculated analytically or by the aid of a computer using Eq. (3.47) for the simple case of $k_x = k_y = 0$ and for the approximate isotropic effective masses ($m_B^* = m_A^*$). As a comparison between the quantum wells and superlattices, the energy levels are sketched as a function of the barrier width h as shown in Fig. 3.17. The illustration in this figure is that when the barrier width is large enough, the energy levels resemble those of

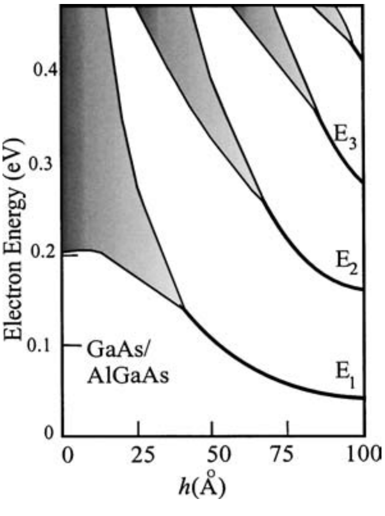


Figure 3.17 The bound energy levels in quantum wells are transformed into minibands as the barrier width h is decreased.

quantum wells. As the barrier width decreases, the neighboring quantum wells interact with each other and the energy levels are broadened to form minibands. It should be pointed out that the envelope function approximation may not yield satisfactory results for short-period superlattices.

The calculation of the holes' energy bands in quantum wells and superlattices is complicated and will not be discussed here. The $6 \times 6\, \mathbf{k} \cdot \mathbf{p}$ matrix of the $J = \frac{3}{2}$ and $J = \frac{1}{2}$ valence bands were briefly discussed in Sec. 3.7.2. For additional reading materials on this subject, the readers are encouraged to search the open literature. The hole effective masses of the $J = \frac{3}{2}$ and $J = \frac{1}{2}$ states, in the framework of the envelope function approximation, are obtained (see Yu and Cardona 2003) as follows:

$$\begin{aligned}
\frac{m_{hz}^*}{m_o} &= \frac{1}{\gamma_1 - 2\gamma_2}, \quad \text{for } z\text{-direction} \qquad \text{and} \\
\frac{m_{hy}^*}{m_o} &= \frac{1}{\gamma_1 + \gamma_2} \quad \text{for } y\text{-direction} \\
\frac{m_{lz}^*}{m_o} &= \frac{1}{\gamma_1 + 2\gamma_2}, \quad \text{for } z\text{-direction} \qquad \text{and} \\
\frac{m_{ly}^*}{m_o} &= \frac{1}{\gamma_1 - \gamma_2} \quad \text{for } y\text{-direction}
\end{aligned} \tag{3.123}$$

where the subscript h stands for the $J = \frac{3}{2}$ state (heavy hole) and l stands for $J = \frac{1}{2}$ state (light hole).

Summary

In this chapter, we introduced the periodic potentials and various approximations used to calculate the single-electron energy levels in these type of potentials. The periodic potential was considered here because it can represent a real single-crystal solid. The Bloch theorem was briefly discussed. It provides a means to construct the wave function of a single electron in a periodic potential. Once the forms of the wave functions and the periodic potentials are known, then the Schrödinger equation can be solved. The solution of the Schrödinger equation, however, can be complicated, and in fact is impossible to obtain exactly for a real crystal. Therefore, several approximations were introduced to construct the dispersion relations. The Kronig-Penney theory was introduced here because it provides the concept of energy bands in solids.

A few examples of periodic potentials were introduced in this chapter, which include a weak periodic potential, periodic δ-function potential, and semiconductor superlattices. The concept of the effective mass of a charge carrier, such as an electron or hole, was discussed. Expressions for the electron, heavy hole, and light hole effective masses were derived.

The tight-binding method, $\mathbf{k} \cdot \mathbf{p}$ model, and the envelope function approximations were discussed because of their applicability to constructing the band structure of bulk materials as well as that of quantum structures, such as quantum wells, superlattices, and quantum dots. The quantization and calculations of the energy levels in quantum wells were discussed, and the BenDaniel-Duke boundary conditions were introduced. These boundary conditions are very helpful in calculating the discrete energy levels in a quantum well.

Problems

3.1 Show that if $V(x)$ is a periodic function, then $f(x) = (2m/\hbar^2)[E_n - V(x)]$ is also a periodic function and that $f(x + L) = f(x)$.

3.2 Provide a proof of the Bloch theorem.

3.3 Derive the Kronig-Penney model in the momentum space (reciprocal space).

3.4 Solve the determinant of the one-electron approximation with the periodic δ-function potential. Start from Eq. (3.39) and obtain the relationship shown in Eq. (3.40).

3.5 Show graphically that Eq. (3.41) has at least one bound state. When do you expect to see more than one bound state? Show your results.

3.6 Use the wave packet analysis to show that the group velocity is the gradient of the electron energy such that $\upsilon_g = \frac{1}{\hbar}\frac{\partial E}{\partial \mathbf{k}}$. Sketch E and υ_g in the first Brillouin zone.

3.7 Assume that an electronic energy band has an extremum at $\mathbf{k}_o = 0$. Use the Bloch wave function in the Schrödinger equation to derive Eq. (3.78).

3.8 Reproduce Fig 3.14 for a GaAs/AlGaAs isolated quantum well, determine the energy values for the first three bound states as a function of m_B^*/m_A^*, and then plot E as a function of the effective masses ratio. Assume that the conduction band offset is 0.3 eV and that $m_A^* = 0.067m_o$, where m_o is the free-electron mass. When does the ground state becomes zero?

3.9 Calculate the first and second energy levels in a GaAs/AlGaAs quantum well with a thickness of 100 Å and a conduction band offset of 0.30 eV. The effective masses are $m_A^* = 0.067m_o$ and $m_B^* = 0.092m_o$.

3.10 Consider the following periodic δ-function's potential $V(x) = \frac{\hbar^2\lambda}{2m_o a}\sum_n \delta(x+na)$, where λ is a positive dimensionless constant, m_o is the mass of the electron, and a is the lattice constant. Use Eq. (3.41) to derive an expression for the lowest energy level at $k=0$ and then find the bandgap at $\mathbf{k} = \pi/a$.

3.11 Use the BenDaniel-Duke boundary conditions to obtain Eqs. (3.121) and (3.122).

This page intentionally left blank.

Chapter

4

Tunneling Through Potential Barriers

The tunneling phenomenon was briefly discussed in Chap. 2 for a single potential barrier, single potential well, and a double-barrier structure. This effect was first reported by Esaki in a narrow germanium *pn* junction (see Esaki 1959). Currently, there are many devices based on the tunneling effect such as resonant tunneling diodes, point contact diodes, Schottky diodes, bipolar transistors, and field-effect transistors. One important aspect of the tunneling effect is that the tunneling time of carriers is proportional to the function $\exp(-2\rho L)$, where ρ is the decaying wave vector inside the barrier and L is the width of the potential barrier. The wave functions of the tunneling carriers are characterized as propagating waves in the wells and evanescent waves inside the barriers. In this chapter, we will discuss examples that are relevant to semiconductor heterojunctions and nanostructures.

In Sec. 2.2 we showed the form of the transmission coefficient for a particle tunneling through a rectangular barrier. When the product of the decaying wavevector and the width of the barrier is much larger than unity, the transmission coefficient is approximated by Eq. (2.34). For a potential barrier with an arbitrary shape as depicted in Fig. 4.1 (solid line), the exact derivation of the transmission coefficient becomes more complicated. It can be obtained with the help of approximation methods. For example, the WKB method becomes very handy in obtaining an approximate form of the tunneling probability.

To obtain an approximate expression for the transmission coefficient in case of an arbitrary potential barrier, as shown in Fig. 4.1, as opposed to the rectangular potential barriers shown as the dashed line in the figure, one can consider the barrier as being composed of small rectangular segments as shown in the figure. The form of the transmission

Copyright © 2005 by The McGraw-Hill Companies, Inc. Click here for terms of use.

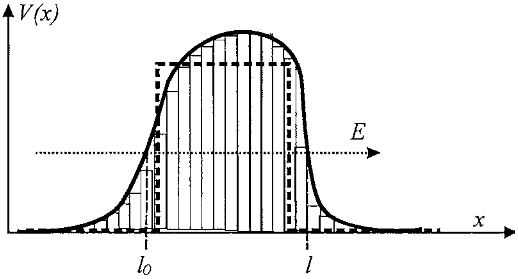


Figure 4.1 An arbitrary potential barrier divided in many small rectangular segments and overlapped on a rectangular potential barrier (dashed line) for comparison.

coefficient for each segment is identical to the expression presented in Eq. (2.34). The total transmission coefficient can now be approximated by the product of the transmission coefficient of the segments such as

$$\begin{aligned} T(E) &= T_1(E)T_2(E)T_3(E)\cdots T_n(E) \\ &= T_o e^{\sum\limits_n -2\rho d_n} \end{aligned} \tag{4.1}$$

where the transmission through the nth segment of a width d_n is $T_n(E) \propto e^{-2\rho d_n}$. The number of segments can be made large enough so that the integration can be used instead of summation as follows:

$$T(E) \approx T_o \exp\left(-2\int\limits_{l_0}^{l} \rho(x)\,dx\right) \tag{4.2}$$

where the integration limits (l and l_o) correspond to the turning points, as shown in Fig. 4.1, and $\rho(x)$ is the wavevector inside the barrier. We encounter the form of this transmission coefficient when we introduced the WKB approximation method discussed in Sec. 2.10.

4.1 Transmission Through Potential Barriers

In this section we will discuss several examples that demonstrate the transmission (tunneling) of a particle, such as an electron, through different types of potential barriers. Let us first illustrate the transmission coefficient for an electron traveling through a δ-function barrier. This simple example is very important since the atoms in semiconductors or even small quantum dots can be seen by a traveling particle, such as an electron, as a δ-function potential. This approximation is very useful in understanding the basic idea of tunneling through potentials. Let us assume that the δ-function potential barrier is represented by Fig. 1.8 and is defined as $V(x) = \lambda\delta(x)$, where λ is the strength of the δ-function. The wave functions can be written as

$$\begin{aligned} \psi_{\mathrm{I}}(x) &= e^{ikx} + Ae^{-ikx} \qquad &\text{for } x < 0 \\ \psi_{\mathrm{II}}(x) &= Be^{ikx} \qquad &\text{for } x > 0 \end{aligned} \tag{4.3}$$

where $k = \sqrt{2mE/\hbar^2}$. The boundary conditions for δ-function potentials were discussed in Sec. 2.8 and are given by

$$\psi_{\mathrm{I}}(0) = \psi_{\mathrm{II}}(0) \qquad \text{and} \qquad \psi'_{\mathrm{II}}(0) - \psi'_{\mathrm{I}}(0) = \frac{2m\lambda}{\hbar^2}\psi_{\mathrm{II}}(0) \tag{4.4}$$

where m is the mass of the traveling particle. These boundary conditions yield a transmission coefficient of

$$T(k) = |B|^2 = \frac{k^2\hbar^4}{k^2\hbar^4 + m^2\lambda^2} \tag{4.5}$$

This simple result shows that the transmission probability is unity as $\lambda \to 0$ and decreases drastically as the strength of the δ-function, depicted in λ, is increased. The transmission coefficient behavior as a function of λ is shown in Fig. 4.2.

The step function potential discussed in Chap. 2 is a very simple case. A more complicated case is shown in Fig. 4.3 where the potential barrier is a smooth function. Let us assume that a particle is traveling

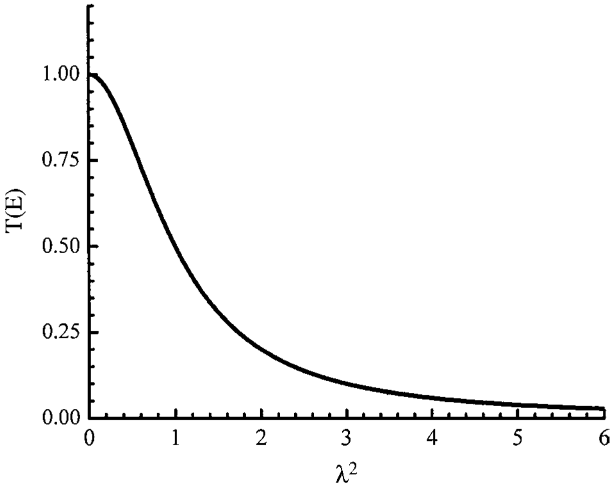


Figure 4.2 The transmission coefficient plotted as a function of λ for a particle traveling through a potential barrier with a strength λ.

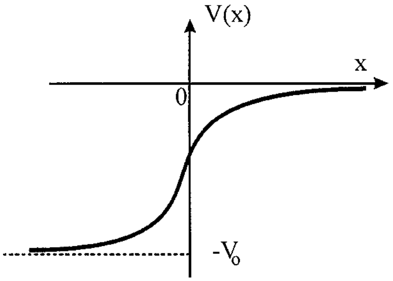


Figure 4.3 A step potential barrier changes continuously with x.

from left to right with an energy $E > 0$ and the potential barrier has the form $V(x) = -V_o/(e^{x/a} + 1)$ where V_o is shown in Fig. 4.3 and a is a positive number. This form of a function is common at solid surfaces as opposed to the step function shown in Fig. 2.1. The potential changes continuously over a distance of an interatomic separation such that $x > 0$ is inside the material and $x < 0$ is for outside the material. The Schrödinger equation for this potential can be written as

$$\left[-\frac{\hbar^2}{2m}\frac{d^2}{dx^2} - \left(E + \frac{V_o}{e^{x/a}+1}\right)\right]\psi(x) = 0 \tag{4.6}$$

This is a difficult equation to solve. However, by choosing the transformation

$$\zeta = -e^{-x/a} \qquad \psi(x) = \zeta^{-ika}u(\zeta) \qquad \text{where } k = \sqrt{\frac{2mE}{h^2}} \tag{4.7}$$

one can rewrite the equation of motion as

$$\zeta(1-\zeta)\frac{d^2u(x)}{dx^2} + (1-2ika)(1-\zeta)\frac{du(x)}{dx} - K_o^2a^2u(x) = 0 \tag{4.8}$$

where $K_o = \sqrt{2mV_o/h^2}$. The solution of this equation, which is for $x \to \infty(\zeta \to 0)$, is finite and behaves asymptotically as a traveling (propagating) wave. For the wave function inside the material ($x > 0$) the wave function is still a propagating wave. Without going through extensive algebra, we merely present the reflection coefficient R as

$$R = \left|\frac{\Gamma(2ika)\Gamma[-i(K+k)a]\Gamma[1-i(K+k)a]}{\Gamma(-2ika)\Gamma[i(K-k)a]\Gamma[1+i(K+k)a]}\right|^2 = \frac{\sinh^2[\pi a(K-k)]}{\sinh^2[\pi a(K+k)]} \tag{4.9}$$

where $K = \sqrt{2m(E+V_o)/h^2}$ and Γ is the well-know Γ-function. The transmission coefficient is then $T = 1 - R$. A plot of T and R is shown in Fig. 4.4 (solid lines). We also plotted T and R for the step function

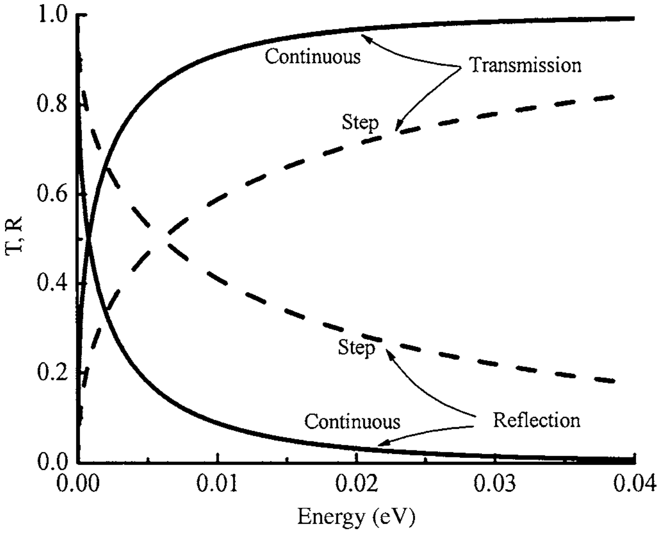


Figure 4.4 The transmission and reflection coefficients plotted for both the smooth step function (solid lines) and sharp step function (dashed lines).

potential (dashed lines) in this figure. It is clear that the transmission probability is higher in the case of a continuously changing potential step (Fig. 4.3) as compared to the sharp step (Fig. 2.1).

Another example of potential barriers is that when two semiconductor materials with different bandgaps are grown to form a heterojunction, a bandgap offset in both the conduction and valence bands is formed. In the absence of band bending, the conduction band discontinuity can be presented by Fig. 4.5. The potential form in this figure can be obtained if the material with a larger bandgap is graded. For example, the potential profile can be formed by growing $\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{As}$ on GaAs, where x is incrementally varied from 0.3 to 0 during growth. The band bending is usually formed when there is a separation of charges at

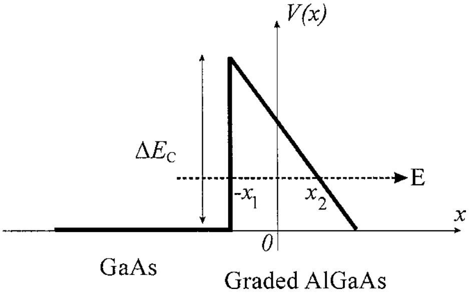


Figure 4.5 GaAs/AlGaAs heterojunction is plotted in the absence of band bending. The turning points are labeled $-x_1$ and x_2.

the modulation-doped heterojunction interfaces, as shown in Fig. 2.16. According to the WKB method (see Sec. 2.10), the transmission coefficient $T(E)$ is given by the following form:

$$T(E) \approx \exp\left[-2\int_{-x_1}^{x_2} |k(x)|\,dx\right] \tag{4.10}$$

where $-x_1$ and x_2 are the turning point, and

$$k(x) = \sqrt{\frac{2m^*}{\hbar^2}\left(\Delta E_c - e\mathcal{E}x - E\right)} \tag{4.11}$$

is the wavevector of an electron traveling from left to right with a kinetic energy E. Here ΔE_c is the conduction band offset and $V(x) = (\Delta E_c - e\mathcal{E}x)$ is the form of the graded potential. Notice that $e\mathcal{E}$ is the slope of the potential, which can be thought of as band bending due to an applied electric field $\mathcal{E}$. As a matter of fact, under the influence of an applied electric field all the band structures in the real space exhibit a band bending similar to the potential profile shown in Fig. 4.5. The kinetic energy E of the particle is determined according to the WKB approximation from the following relation: $V(x_2, -x_1) = E$. Substituting Eq. (4.11) into (4.10) and performing the integration we have

$$\begin{aligned} T(E) &\approx \exp\left[-2\int_{-x_1}^{x_2}\left|\sqrt{\frac{2m^*}{\hbar^2}(\Delta E_c - E - e\mathcal{E}x)}\right| dx\right] \\ &\approx \exp\left\{\left(\frac{2m^*}{\hbar^2}\right)^{1/2}\left[\frac{4}{3e\mathcal{E}}(\Delta E_c - E - e\mathcal{E}x)^{3/2}\right]\Bigg|_{-x_1}^{x_2}\right\} \\ &\approx \exp\left[-\frac{4}{3}\frac{\sqrt{2m^*}}{\hbar e\mathcal{E}}(\Delta E_c - E)^{3/2}\right] \end{aligned} \tag{4.12}$$

Notice that at $x = x_2$ we can approximate $(\Delta E_c - E - e\mathcal{E}x) = 0$, and at $x = -x_1$ we have $(\Delta E_c - E - e\mathcal{E}x) = (\Delta E_c - E)$. The transmission coefficient is plotted in Fig. 4.6 in units of $\frac{4}{3}\frac{\sqrt{2m^*}}{\hbar e\mathcal{E}}$, and the band offset is taken as $\Delta E_c = 0.3$ eV. The transmission coefficient is approaching unity when the energy of the electron reaches the value of the band offset as shown in the figure.

For a sharp step potential such as shown at $x = -x_1$ in Fig. 4.5, the WKB approximation may not provide a reasonable answer. A reasonable solution for the transmission coefficient would be if this step potential is smoothed as shown in Fig. 4.7. One may write the potential

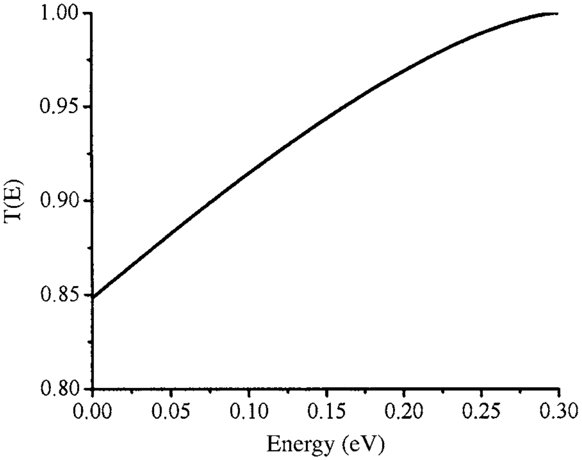


Figure 4.6 The transmission coefficient plotted as a function of the particle (electron) energy E.

profile in the following form: $V(x) = -Fx - e^2/4x$, where F is a positive constant. This potential represents the change of the potential near the surface of a solid with the second term known as the electrical image potential. The transmission coefficient can be written in the same form as Eq. (4.10) with

$$k(x) = \sqrt{\frac{2m^*}{\hbar^2}\left(|E| - Fx - \frac{e^2}{4x}\right)} \tag{4.13}$$

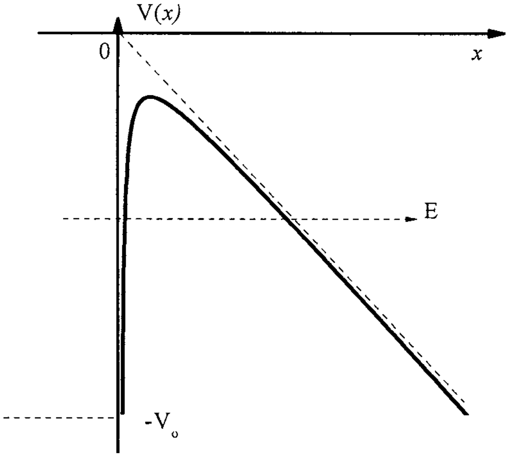


Figure 4.7 A potential profile near the surface of a solid is plotted as a function of the distance from the surface.

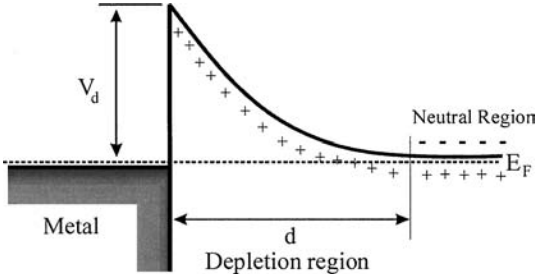


Figure 4.8 Schottky barrier formed between a metal and n-type semiconductor, such as GaAs.

The solution for the transmission coefficient is more complicated and will not be presented. However, the final expression of $T(E)$ is identical to Eq. (4.12) when the third term of Eq. (4.13) approaches zero.

Another well-known barrier is the Schottky barrier formed between a metal and a semiconductor when they are in contact, as shown in Fig. 4.8. The depletion region is formed near the interfaces where a built-in electric field exists due to separation of charges. The Fermi energy E_F is pinned at an energy $E_\varphi = eV_d$ below the conduction band of the semiconductor. The depletion potential V_d can be obtained as follows assuming that the semiconductor is uniformly doped:

$$V_d = \int_0^d \mathcal{E}\,dx = \int_0^d \frac{eN_d x}{\epsilon_o \epsilon_r}\,dx = \frac{eN_d d^2}{2\epsilon_o \epsilon_r} \tag{4.14}$$

where $\mathcal{E}$ = built-in electric field given by Gauss's law as $\mathcal{E} = eN_d/\epsilon_o\epsilon_r$
N_d = number of electrons transferred from ionized donor atoms
ϵ_o = permittivity of space
ϵ_r = dielectric constant of semiconductor

For an n-type GaAs, E_φ is about 0.7 eV with little variation for different metals.

If one assumes that the barrier has the form $V(x) = E_\varphi[1 - (x/d)^2]$, the transmission coefficient of an electron traveling with an energy equal to the Fermi energy can be written as

$$T \approx \exp\left\{-\frac{2}{\hbar}\int_0^d \sqrt{2m^* E_\varphi\left[1-\left(\frac{x}{d}\right)^2\right]}\,dx\right\} = \exp\left(-\frac{\pi d}{2\hbar}\sqrt{2m^* E_\varphi}\right)$$
$$= \exp\left(-\frac{d}{l}\right) \tag{4.15}$$

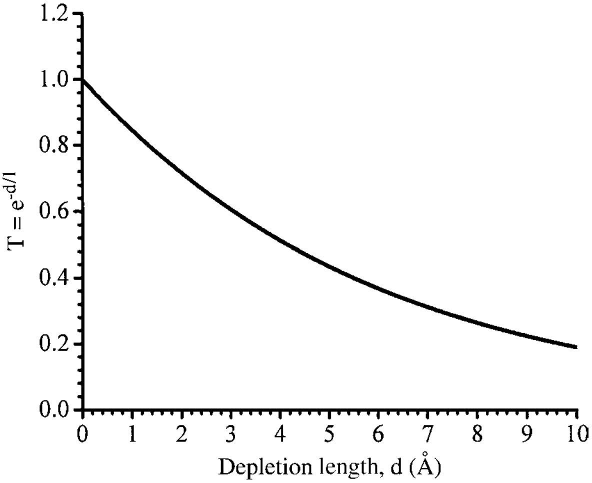


Figure 4.9 The transmission coefficient obtained for a Schottky potential barrier profile is plotted as a function of the depletion length inside the semiconductor, such as GaAs.

where l is defined as the decaying length and is given as $l = \sqrt{2\hbar^2}/\sqrt{(\pi^2 m^* E_\varphi)}$, which is approximately 6.0 Å for GaAs. The transmission coefficient in Eq. (4.15) is plotted as a function of the depletion length d for $l = 6.0$ Å as shown in Fig. 4.9. It is clear from this figure that the transmission coefficient approaches unity when the depletion length is very small. This requires that the semiconductor material be heavily doped.

4.2 Tunneling Through Pyramidal Potential Barriers

In Chap. 2 we discussed tunneling through a rectangular potential barrier, which resembles a quantum barrier. The two-barrier structure was also briefly discussed. In this section, we will discuss tunneling through barriers that resemble low-dimensional semiconductor structures such as quantum dots. A typical example of a quantum dot structure is shown in Fig. 4.10*a* where an $In_{0.3}Ga_{0.7}As$ quantum dot is grown on a GaAs buffer layer and then capped by undoped GaAs. Semiconductor quantum dots tend to grow in pyramidal-like shapes. To simplify the calculations of the transmission coefficient through the quantum dot, we assume a one-dimensional potential profile, as shown in Fig. 4.10*b*. The potential barrier that resembles a one-dimensional triangular shape

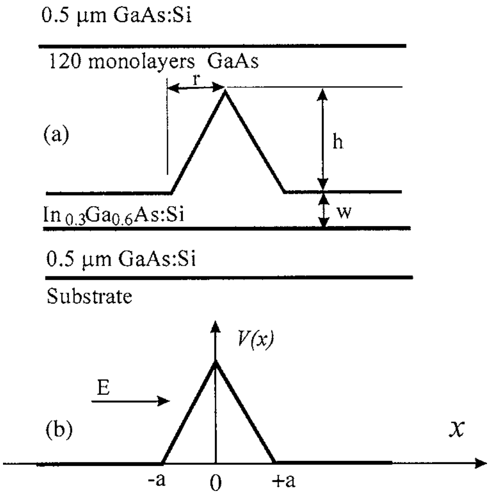


Figure 4.10 (*a*) A typical structure of a semiconductor quantum dot sandwiched between the buffer and barrier layers. (*b*) A potential energy profile resembles the shape of a one-dimensional quantum dot.

can be chosen as

$$V(x) = \begin{cases} V_o\left(1 + \dfrac{x}{a}\right) & \text{for } -a \leq x \leq 0 \\ V_o\left(1 - \dfrac{x}{a}\right) & \text{for } 0 \leq x \leq a \\ 0 & \text{for } |x| \geq a \end{cases} \tag{4.16}$$

Let us assume that the quantum dot is standing alone and the GaAs barrier does not exist. Moreover, assume that the potential profile takes the same shape as the pyramidal quantum dot. The wavevector of an electron tunneling through the dot can be characterized as

$$\begin{aligned} k &= \sqrt{\frac{2mE}{\hbar^2}} && \text{outside the dot} \\ K &= \sqrt{\frac{2m[V(x) - E]}{\hbar^2}} && \text{inside the dot} \end{aligned} \tag{4.17}$$

Schrödinger equations can now be written for the three regions as

$$\frac{d^2\psi}{dx^2} + k^2\psi = 0 \qquad \text{for } |x| \geq a \tag{4.18a}$$

$$\frac{d^2\psi}{dx^2} - \left(k_o^2 - k^2 + k_o^2\frac{x}{a}\right)\psi = 0 \qquad \text{for } -a \leq x \leq 0 \tag{4.18b}$$

$$\frac{d^2\psi}{dx^2} - \left(k_o^2 - k^2 - k_o^2\frac{x}{a}\right)\psi = 0 \qquad \text{for } 0 \leq x \leq a \tag{4.18c}$$

where $k_o = \sqrt{2mV_o/\hbar^2}$. By making the following transformation

$$\zeta = \left(\frac{a}{k_o^2}\right)^2 \left(k_o^2 - k^2 + k_o^2\frac{x}{a}\right) \qquad \text{and} \qquad \eta = \left(\frac{a}{k_o^2}\right)^2 \left(k_o^2 - k^2 - k_o^2\frac{x}{a}\right) \tag{4.19}$$

Schrödinger equations for a particle inside the potential barrier can be rewritten as

$$\begin{aligned} &\frac{d^2\psi}{d\zeta^2} - \zeta\psi = 0 \qquad \text{for} -a \le x \le 0 \\ &\frac{d^2\psi}{d\eta^2} - \eta\psi = 0 \qquad \text{for } 0 \le x \le a \end{aligned} \tag{4.20}$$

The form of this equation was encountered in the triangular well formed at the heterojunction interfaces in Chap. 2. The solutions of Eq. (4.20) can be written in terms of Airy functions Ai(x) and Bi(x) as

$$\psi = \begin{array}{ll} e^{ikx} + Ae^{-ikx} & \text{for } x \le -a \\ B\text{Ai}(\zeta) + C\text{Bi}(\zeta) & \text{for } -a \le x \le 0 \\ D\text{Ai}(\eta) + E\text{Bi}(\eta) & \text{for } 0 \le x \le a \\ Fe^{ikx} & \text{for } x \ge a \end{array} \tag{4.21}$$

By applying the boundary conditions at $x = -a$, 0, and a, the coefficients in Eq. (4.21) can be determined from the following relations:

$$\begin{aligned} e^{-ika} + Ae^{ika} &= B\text{Ai}(-\lambda) + C\text{Bi}(-\lambda) \\ i\sqrt{\lambda}(e^{-ika} - Ae^{ika}) &= B\text{Ai}'(-\lambda) + C\text{Bi}'(-\lambda) \\ B\text{Ai}(\mu) + C\,\text{Bi}(\mu) &= D\text{Ai}(\mu) + E\,\text{Bi}(\mu) \\ B\text{Ai}'(\mu) + C\,\text{Bi}'(\mu) &= -D\text{Ai}'(\mu) - E\,\text{Bi}'(\mu) \\ D\text{Ai}(-\lambda) + E\,\text{Bi}(-\lambda) &= Fe^{ika} \\ D\text{Ai}'(-\lambda) + E\,\text{Bi}'(-\lambda) &= -i\sqrt{\lambda}Fe^{ika} \end{aligned} \tag{4.22}$$

The integral forms of Airy functions are given as

$$\begin{aligned} \text{Ai}(x) &= \frac{1}{\pi}\int_0^\infty \cos\left(\frac{\zeta^3}{3} + \zeta x\right)d\zeta \\ \text{Bi}(x) &= \frac{1}{\pi}\int_0^\infty \exp\left(-\frac{\zeta^3}{3} + \zeta x\right)d\zeta + \frac{1}{\pi}\int_0^\infty \sin\left(\frac{\zeta^3}{3} + \zeta x\right)d\zeta \end{aligned} \tag{4.23}$$

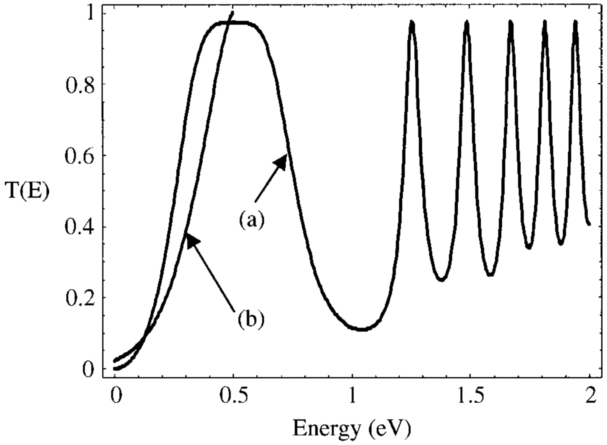


Figure 4.11 The transmission coefficient plotted as a function of energy for a pyramidal-shape potential barrier (curve *a*) using the exact solution described in Eq. (4.24). The WKB approximation yields similar results (curve *b*).

The prime in Eq. (4.22) indicates the first derivative of the Airy functions. The transmission coefficient is simply $|F|^2$, which can be obtained from Eq. (4.22) after performing a few algebraic steps and is given as

$$\begin{aligned} T(E) &= |F|^2 \\ &= \frac{\lambda}{[\mathrm{Bi}(\mu)\mathrm{Ai}'(-\lambda)-\mathrm{Ai}(\mu)\mathrm{Bi}'(-\lambda)]^2+\lambda[\mathrm{Bi}(\mu)\mathrm{Ai}(-\lambda)-\mathrm{Ai}(\mu)\mathrm{Bi}(-\lambda)]^2} \\ &\times \frac{1}{[\mathrm{Bi}'(\mu)\mathrm{Ai}'(-\lambda)-\mathrm{Ai}'(\mu)\mathrm{Bi}'(-\lambda)]^2+\lambda[\mathrm{Bi}'(\mu)\mathrm{Ai}(-\lambda)-\mathrm{Ai}'(\mu)\mathrm{Bi}(-\lambda)]^2} \end{aligned} \tag{4.24}$$

where $\lambda = (k_o a)^{2/3}(k^2/k_o^2)$ and $\mu = (k_o a)^{2/3}(1-k^2/k_o^2)$. The Airy functions were normalized such that $\mathrm{Ai}'(x)\mathrm{Bi}(x) - \mathrm{Ai}(x)\mathrm{Bi}'(x) = 1$. A plot of Eq. (4.24) is shown as curve *a* in Fig. 4.11. The transmission coefficient reaches unity as the energy of the particle reaches $V_o = 0.5$ eV and then starts to oscillate. These oscillations are called interference resonance, and they are discussed in Chap. 2 for potential wells.

The analytical derivation of the transmission coefficient is time-consuming. However, computer-assisted analysis is simple in this case. On the other hand, one can use the WKB approximation to obtain a similar result. In the WKB approximation, the transmission coefficient has the following form:

$$
\begin{aligned}
T(E) &\approx \exp\left(-2\int_{-a}^{a}\sqrt{\frac{2m}{\hbar^2}[V(x)-E]}\,dx\right) \\
&\approx \exp\left(-4\int_{0}^{a}\sqrt{\frac{2m}{\hbar^2}[V(x)-E]}\,dx\right) \\
&\approx \exp\left[-\frac{8}{3}\frac{a\sqrt{2m/\hbar^2}}{V_o}(V_o-E)^{3/2}\right]
\end{aligned}
\tag{4.25}
$$

This expression is plotted as curve *b* in Fig. 4.11. The upper limit value of the transmission coefficient is limited to the condition when $E = V_o$. In this case, V_o was chosen as 0.5 eV.

4.3 Double-Barrier Potential

Tunneling through a double-barrier structure is the basis for the resonant tunneling diode. This structure was briefly discussed in Chap. 2, but we will look at this problem in more detail in this section. For the simplest case, we have chosen the barriers and well widths to be identical, as shown in Fig. 4.12. By following the transfer matrix procedure described in Sec. 2.9, the transmission coefficient was presented in Eq. (2.129). To simplify the analysis for the transmission coefficient of an electron traveling with an energy $E < V_o$ from $x = -\infty$ to $x = +\infty$, one can assume that the widths of the well and the two barriers are the same. Typical materials for this structure are GaAs for the well and AlGaAs for the barriers. Since the electron effective mass does not change considerably in the well ($m^* = 0.067m_o$) and in the barrier ($m^* = 0.094m_o$), we assumed that it is the same for both materials.

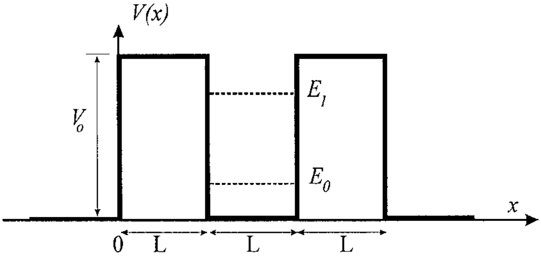


Figure 4.12 A double barrier-well structure commonly encountered in resonance tunneling diodes.

With these assumptions in mind, we derived an expression for the transmission coefficient as

$$T(E) = \frac{64[E(V_o - E)]^2}{D_1 + D_2} \tag{4.26}$$

where $D_1 = \{[(V_o - 2E)^2 - 4E(V_o - E)]\cosh(2\rho L_b)$
$\quad -V_o^2[2\sinh^2(\rho L_b)\cos(2kL_w) + 1]\}^2$
$D_2 = [4(V_o - 2E)\sqrt{E(V_o - E)}\sinh(2\rho L_b)$
$\quad +2V_o^2\sinh^2(\rho L_b)\sin(2kL_w)]^2$
L_w = well width
L_b = barrier width
$k = \sqrt{\frac{2m^*E}{\hbar^2}}$
$\rho = \sqrt{\frac{2m^*(V_o - E)}{\hbar^2}}$

The transmission coefficient is plotted as a function of the electron energy for a barrier of height $V_o = 0.10$ eV and three different widths ($L = 50$, 100, and 150 Å) as shown in Fig. 4.13. The peaks in the transmission coefficient correspond to the electron energy as being resonant with the confined energy levels in the well. As the well width increases, the number of confined energy levels is increased for a fixed potential barrier. When the electron energy is larger than the barrier, interference resonance peaks can be observed, which correspond to virtual states in the continuum. Notice that the number of resonance peaks and

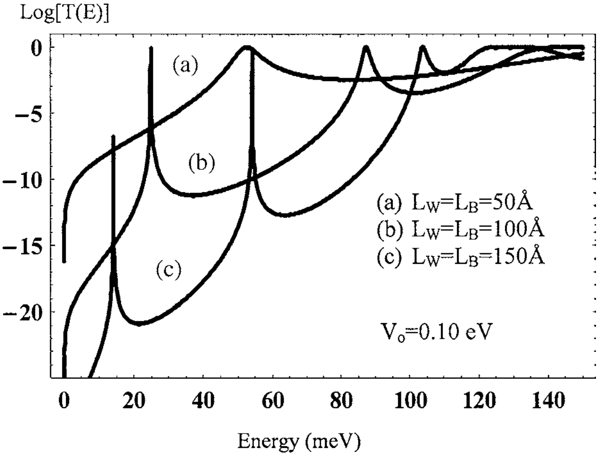


Figure 4.13 Transmission coefficient plotted as a function of electron energy in a double-barrier structure for three different well (L_W) and barrier (L_B) widths.

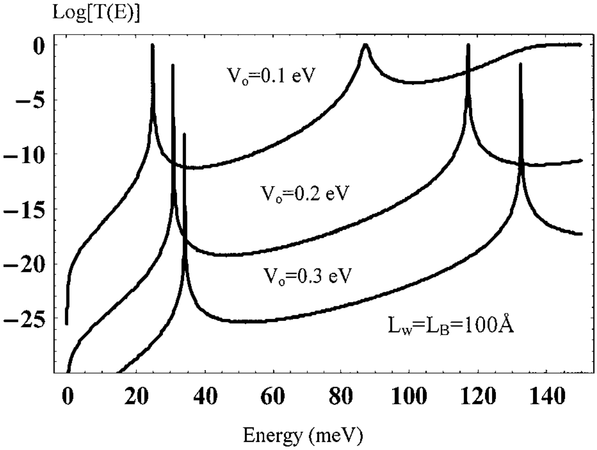


Figure 4.14 Transmission coefficient plotted as a function of electron energy for a fixed well width and three different barrier heights.

their energy positions change as the well width is varied. For example, there is only one state for a well thickness of 100 Å and a barrier height of 0.10 eV.

When the well width is fixed and the barrier height is increased, one would expect to observe additional bound states as illustrated in Fig. 4.14. In this figure, the transmission coefficient is plotted for three different barrier heights. The ground state is expected to slightly shift as the barrier height increases. Additionally, one can observe the resonance peaks to shift as a function of the barrier height.

The hand analysis of the double-barrier potential is very intensive, while computer-aided analyses of complex structures such as the double-barrier potential structure are easy to handle. The WKB method provides a useful approximation in tunneling problems. Let us assume that the rectangular double-barrier structure can be represented by a double-barrier potential with arbitrary shapes as shown in Fig. 4.15 (dashed curve). The turning points are labeled a_i ($i = 1, 2, 3$, and 4). By using the WKB boundary conditions at these turning points, one can write a 2×2 transfer matrix from which the transmission coefficient can be obtained (see, for example, Gildenblat et al. 1995):

$$T(E) = \frac{T_l T_r}{4\cos^2(I) + \frac{1}{4}(T_l + T_r)^2 \sin^2(I)} \tag{4.27}$$

where $T_l = \exp(-2\gamma_l)$ and $T_r = \exp(-2\gamma_r)$ are the transmission coefficients to the left and to the right, respectively. The parameters γ_l, γ_r,

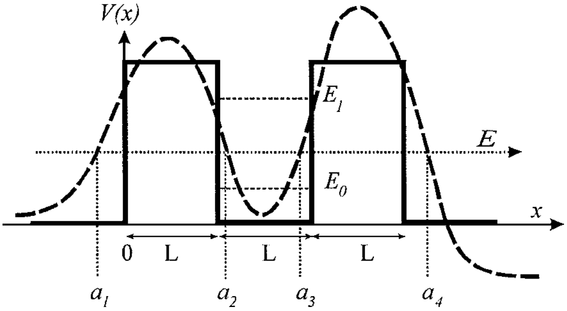


Figure 4.15 Arbitrary double well-barriers (dashed curve) superimposed on the identical double rectangular barriers. The turning points are labeled a_i.

and I are given by

$$\gamma_l = \int_{a_1}^{a_2} |k\,|\,dx, \qquad \gamma_r = \int_{a_3}^{a_4} |k\,|\,dx,$$
$$I = \int_{a_2}^{a_3} k\ dx \qquad \text{where } k = \sqrt{\frac{2m^*(E-V)}{\hbar^2}} \tag{4.28}$$

For symmetrical barriers, Eq. (4.27) is reduced to the standard form found in quantum mechanics textbooks (see, for example, Bohm 1953):

$$T(E) = \frac{1}{4\exp\left(4\int_{a_1}^{a_2} k\,dx\right)\cos^2\left(\int_{a_2}^{a_3} k\,dx\right) + \sin^2\left(\int_{a_2}^{a_3} k\,dx\right)} \tag{4.29}$$

A plot of Eq. (4.29) is shown in Fig. 4.16 for three different well sizes assuming that the electron effective mass is the same in the well and in the barriers. The height of the barriers is assumed to be $V_o = 0.3$ eV.

Again, the peaks in Fig. 4.16 are the electron energy that is coinciding with the confined energy levels in the well at which $T(E) = 1$. These energy levels can be derived from the following relation:

$$\int_{a_2}^{a_3} \sqrt{\frac{2m^*(E_n - V)}{\hbar^2}}\,dx = \left(n + \frac{1}{2}\right)\pi \qquad \text{for } E_n > V \tag{4.30}$$

For $\Delta E \ll E_n$, $T\,(\Delta E + E_n)$ can be expanded to give a Lorentzian line shape near the resonance

$$T(E_n + \Delta E) = \frac{\Gamma^2}{\Gamma^2 + (\Delta E)^2} \tag{4.31}$$

where Γ is the full width at half maximum.

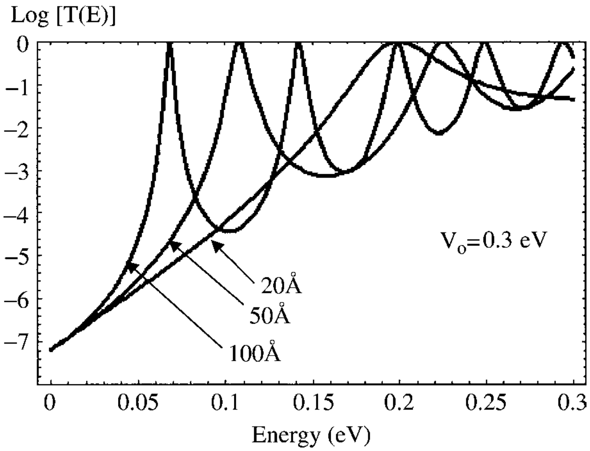


Figure 4.16 Transmission coefficient obtained from the WKB approximation is plotted as a function of electron energy in a double arbitrary barrier structure for three different well sizes.

4.4 The *pn*-Junction Tunneling Diode

Tunneling diodes are currently used in many applications including locking circuits, low-power microwave devices, and local oscillators. A typical structure of a *pn*-junction tunneling diode is shown in Fig. 4.17. One of the basic requirements for this homojunction is that both the *n*- and *p*-type junctions are degenerate, which means that they are heavily doped, such as the Fermi energy is pinned above (below) the conduction

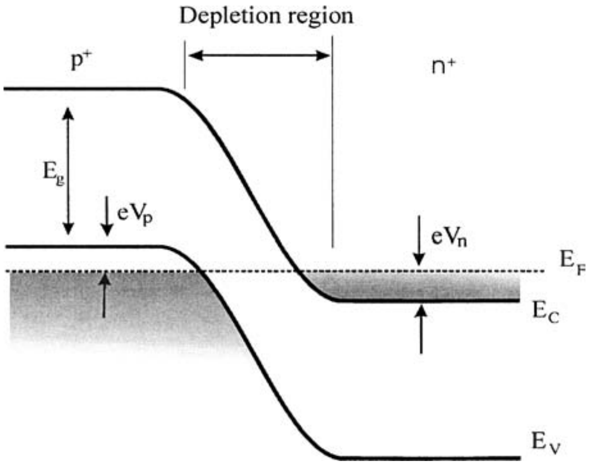


Figure 4.17 A sketch of a typical band diagram of a *pn*-junction tunneling diode.

(valence) band minimum (maximum) as shown in the figure. The depletion region acts as the potential barrier. The carrier tunneling requires that occupied energy states exist on the side from which the electron tunnels and that an empty state exists on the side to which the electron can tunnel. Both states should be at the same energy level, and tunneling can occur from the n-junction to the p-junction or vice versa.

The current of the tunneling diode is composed of three components (see, for example, Sze 2002): the tunneling, excess, and thermal currents. At the thermal equilibrium, the tunneling current from the valence band to the conduction band ($I_{v\rightarrow c}$) and the current from the valence conduction band to the valence band ($I_{c\rightarrow v}$) are balanced, and they can be expressed as

$$I_{v\rightarrow c} = A\int_{E_v}^{E_c} f_c(E)n_c(E)T_t[1 - f_v(E)]n_v(E)\,dE \tag{4.32a}$$

$$I_{c\rightarrow v} = A\int_{Ec}^{E_v} f_v(E)n_v(E)T_t[1 - f_c(E)]n_c(E)\,dE \tag{4.32b}$$

where $A =$ constant
$f_c(E),\ f_v(E) =$ Fermi-Dirac distribution functions for the conduction and valence bands, respectively
$n_c(E), n_v(E) =$ density of states in conduction and valence bands, respectively
$T_t =$ tunneling probability (transmission coefficient)

The tunneling probability is taken as

$$T_t = \exp\left(-\frac{\pi\sqrt{m^*}E_g^{3/2}}{2\sqrt{2}e\hbar\mathcal{E}}\right)\exp\left(-\frac{2E_\perp}{\overline{E}}\right) \tag{4.33}$$

where $E_g =$ bandgap of semiconductor
$\mathcal{E} =$ built-in electric field
$E_\perp =$ energy associated with the momentum perpendicular to direction of tunneling
$\overline{E} = 4\sqrt{2}e\hbar\mathcal{E}/(3\pi m^* E_g^{1/2}) =$ measure of the significant range of transverse momentum

The value of $\overline{E}$ is usually small, which means that only an electron with a small transverse momentum can tunnel. When a bias voltage is applied to the pn junction, the observed tunneling current I_t is

$$I_t = I_{c\rightarrow v} - I_{v\rightarrow c} = A\int_{E_c}^{E_v} [f_c(E) - f_v(E)]T_t n_c(E)n_v(E)\,dE \tag{4.34}$$

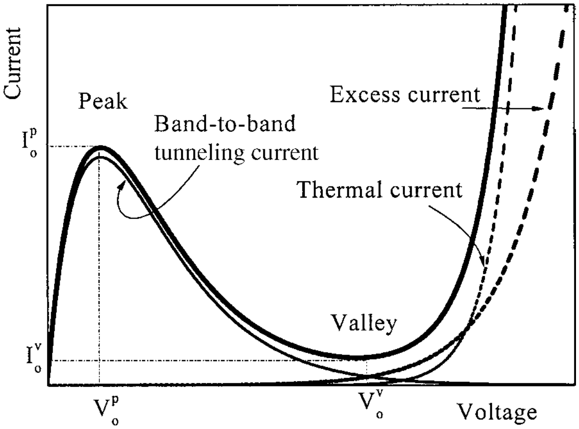


Figure 4.18 Static current-voltage characteristics of a typical tunnel diode. The current is composed of three components: band-to-band tunnel current, excess current, and thermal current. The sum of the three currents is shown as the thick curve with the well-known peak and valley.

The tunneling current in this equation is derived by Demassa and Knott (1970) and is simplified according the following expression:

$$I_t = I_o^p \frac{V}{V_o^p} \exp\left(1 - \frac{V}{V_o^p}\right) \tag{4.35}$$

where I_o^p and V_o^p are the peak current and peak voltage, respectively, defined in Fig. 4.18. The V_o^p was determined by Demassa and Knott to be $V_o^p = (V_n + V_p)/3$ where V_n and V_p are the degeneracy voltages defined in Fig. 4.17.

The excess current is mainly due to defect-assisted tunneling, where the carriers tunnel through defective states in the bandgap. This component of the total current is usually present when the bias voltage is higher than the normal operational conditions. The access current can be understood by inspecting Fig. 4.19, where the bias voltage V is high. Notice that V is multiplied by the electron charge e in order to project the voltage on the energy scale. The main process of the excess current is that an electron can drop from the conduction band (point A) to the defect state (point B) and then tunnel to the valence band (point C). Other routes are possible, but the presented process is the most common route for an electron to tunnel from the conduction band to the valence band. For a bias voltage V the energy E_x that the electron must have

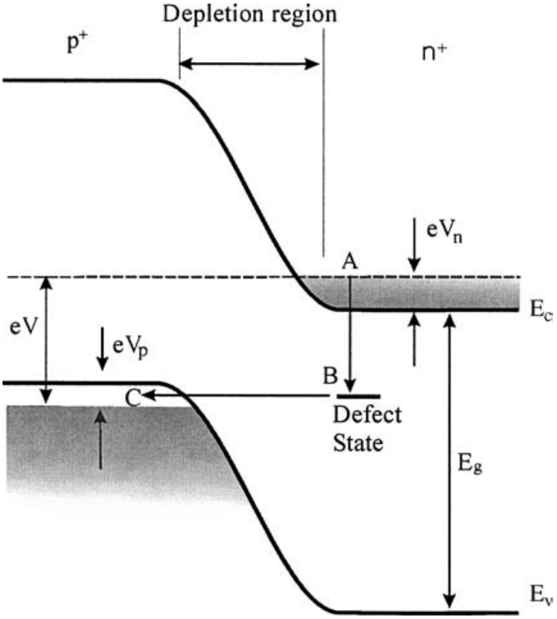


Figure 4.19 A sketch of the *pn*-junction bandgap showing defect-assisted tunneling, which causes the excess current under a high bias voltage.

to tunnel is given by

$$E_x \approx E_g - eV + e(V_n + V_p) \approx e(V_{\mathrm{bi}} - V) \tag{4.36}$$

where V_{bi} is the built-in potential. The tunneling probability is essentially identical to that given by Eq. (4.12) except $(\Delta E_c - E)$ is replaced by E_g:

$$T_x \approx \exp\left(-\frac{4}{3}\frac{\sqrt{2m^*}}{\hbar e\mathcal{E}}E_x^{3/2}\right) \approx \exp\left(-\frac{\alpha_x}{\mathcal{E}}E_x^{3/2}\right) \tag{4.37}$$

where $\alpha_x \approx 4\sqrt{2m^*}/(3\hbar e)$ and $\mathcal{E}$ is the electric field. The electric field across a step function can be written as $\mathcal{E} = 2(V_{\mathrm{bi}} - V)/W$, where W is the depletion region width given by

$$W = \left[\frac{2\epsilon_o}{e}\left(\frac{N_a + N_d}{N_a N_d}\right)(V_{\mathrm{bi}} - V)\right]^{1/2} \tag{4.38}$$

where ϵ_o = permittivity of space
N_a = concentration of acceptors
N_d = concentration of donors

The current density J_x associated with the excess current process can be written as

$$J_x \approx A D_x T_x \tag{4.39}$$

where D_x is the volume density of the occupied levels at energy E_x above the top of the valence band and A is a constant. Substituting Eqs. (4.36) through (4.38) into (4.39), we have (see Chynoweth et al. 1961)

$$J_x \approx A_1 D_x \exp\{-\alpha'_x[E_g - eV + 0.6(V_n - V_p)]\} \tag{4.40}$$

where A_1 is a constant. This relation shows that the excess current depends on the density of states and the applied voltage. Equation (4.40) can be rewritten as (see Roy 1971)

$$\begin{aligned} J_x &\approx J_V \exp\left[-\frac{4}{3}\sqrt{\frac{m^*\epsilon_o}{N^*}}\left(V - V_o^V\right)\right] \\ &\approx J_V \exp\left[A_2\left(V - V_o^V\right)\right] \end{aligned} \tag{4.41}$$

where J_V = valley current density
V_o^V = valley voltage
A_2 = constant
$N^* = N_a N_d / N_a + N_d$

Equation (4.41) is plotted as the long-dashed curve in Fig. 4.18.

Finally, the third component of the tunneling diode current density is the minority-carrier injection current given by

$$J_{\text{th}} = J_o\left[\exp\left(\frac{eV}{k_B T}\right) - 1\right] \tag{4.42}$$

where J_o = the reverse saturation current density
k_B = Boltzmann constant
T = temperature

The thermal current density is plotted as the short-dashed in Fig. 4.18. The sum of the three currents is plotted in this figure as the thick solid curve which shows the characteristic peak and valley encountered in tunneling diodes.

4.5 Resonant Tunneling Diodes

The bandgap alignment of the resonant band structure consists of two doped layers (bottom and top contact layers), two barriers, and one well with at least one bound energy state. When the structure is biased, the electrons tunnel from the bottom contact through the barriers, with a tunneling probability approaching unity, when the bound state in the

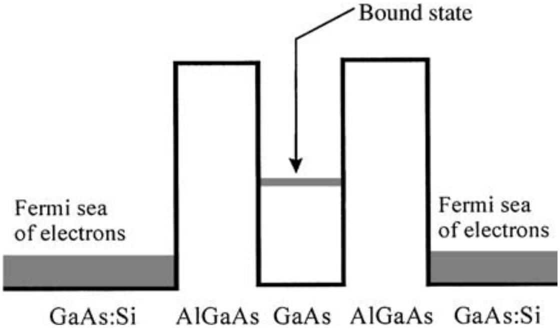


Figure 4.20 A sketch of a typical conduction band structure of a resonant tunneling diode is shown with the degenerate contact layers.

well is resonant with the Fermi energy level of the bottom layer. A typical structure consists of a thick degenerate GaAs:Si bottom layer, 50-Å $Al_{0.3}Ga_{0.7}As$ barrier, 50-Å GaAs well, 50-Å $Al_{0.3}Ga_{0.7}As$ barrier, and a thick degenerate GaAs:Si top contact layer. The structure is shown in Fig. 4.20 under zero bias. The *I*-*V* characteristic of the resonant tunneling diode can be understood by examining Fig. 4.21. For a small bias voltage, the bound state is assumed to be above the Fermi sea of electrons that is present in the bottom contact layer and the tunneling through the bound state is minimum, as shown in case (*a*). As the bias voltage is increased, the bound state becomes resonant with the Fermi energy level of the contact layer and the electrons start to tunnel giving rise to current. The current continues to rise and peaks when the bound state is aligned with the bottom of the electron band as shown in case (*c*). The current drops abruptly as the bound state moves further down from the Fermi electron sea by increasing the applied bias voltage. If the GaAs well contains more than one bound state, then the number of peaks in the *I*-*V* curve will increase accordingly. The negative differential conductivity exhibited in the *I*-*V* curve is very useful in amplifiers and oscillators.

Another example of resonant tunneling diodes is the InAs/AlSb double-barrier structure shown in Fig. 4.22*a*. The band offset is approximately 1.0 eV. This band offset can accommodate more energy levels as compared to the GaAs/AlGaAs system. In addition, the electron effective mass in InAs is about three times smaller than that of GaAs, which leads to higher mobility and better transport properties. Four confined energy levels are shown in this figure. The tunneling of electrons occurs mostly through the ground state E_1, but additional tunneling can occur through the excited states giving rise to additional peaks in the *I*-*V* characteristic curve as shown in Fig. 4.22*b*. The tunneling occurs when

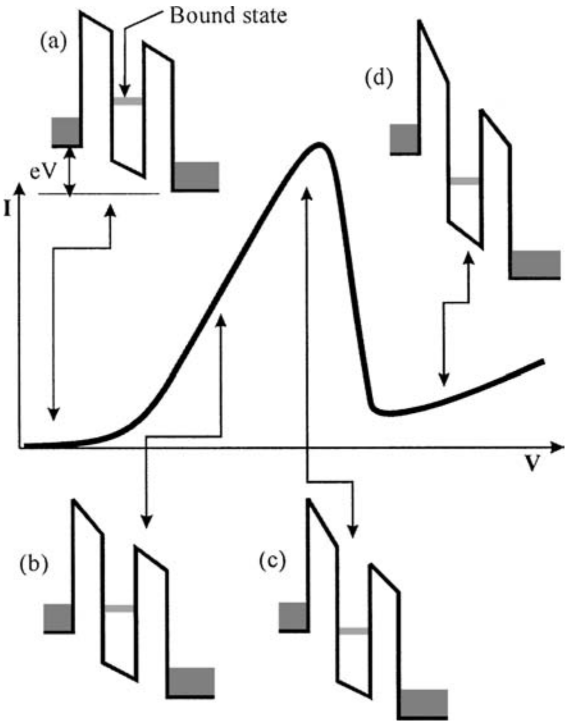


Figure 4.21 The I-V characteristic of the resonant tunneling diode with a single bound state is sketched as a function of bias voltage V.

$E_i (i = 1, 2, 3,$ and $4)$ is aligned, under bias voltage, with the Fermi sea of electrons in the InAs to the left of the structure. The typical I-V characteristic of the resonant tunneling diode is sketched in Fig. 4.20*b*, which exhibits four peaks corresponding to four bound energy levels in the well. In general, any peak voltage (V_o^p) should be larger than $2E_i/e$ due to voltage drops in accumulation and depletion regions.

For zero temperature and a δ-function line shape, the current density of the resonant tunneling diode is derived by Ferry (2001) as

$$J_z = \frac{e^2 V m^* \mathcal{E}_w}{2\pi^2 \hbar^3} \qquad \text{for } 2(\mathcal{E}_o - \mathcal{E}_F) \leq eV \leq 2\mathcal{E}_o \tag{4.43}$$

where $\mathcal{E}_w$ = width of transmission
V = applied bias voltage
$\mathcal{E}_o$ = energy of ground state in well
$\mathcal{E}_F$ = Fermi energy level

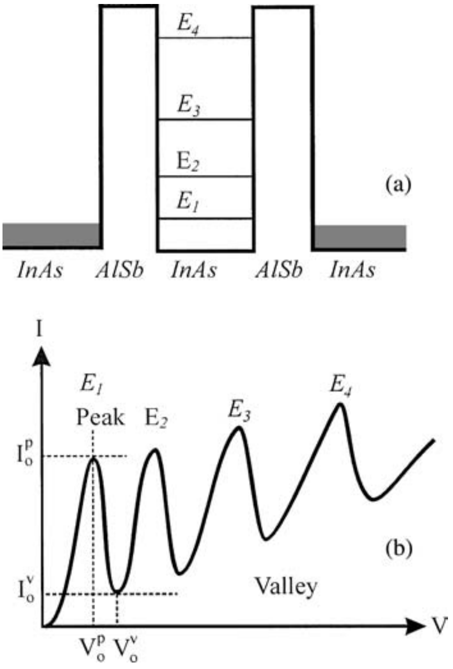


Figure 4.22 (*a*) A sketch of the InAs/AlSb double-barrier resonant tunneling diode showing four bound states. (*b*) A typical *I-V* characteristic of the tunneling resonant diode showing the peak and valley voltages.

Equation (4.43) is sketched in Fig. 4.23. For temperatures other than 0 K, one would expect to observe broadening in the *I-V* curve and the current density to be convoluted with a line shape such as a lorentzian. A sketch of the *I-V* curve at $T > 0$ K is shown as the dashed line in Fig. 4.23.

The limitation of the resonant tunneling diode is the "valley" current as shown in Fig. 4.18. For a device application, in particular digital circuits, it is desired to have a low valley current. In reality, it is difficult to achieve a zero valley current, but with creative designs one can reduce

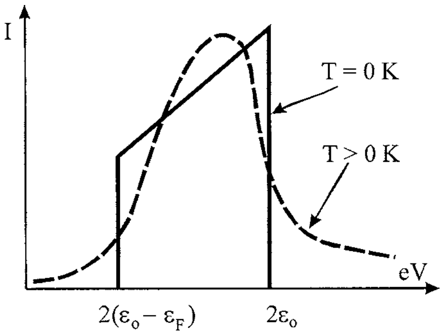


Figure 4.23 The *I-V* curves obtained from the simple model derived by Ferry (2001) is plotted for both δ-function ($T = 0$ K) and lorentzian ($T > 0$ K) line shapes.

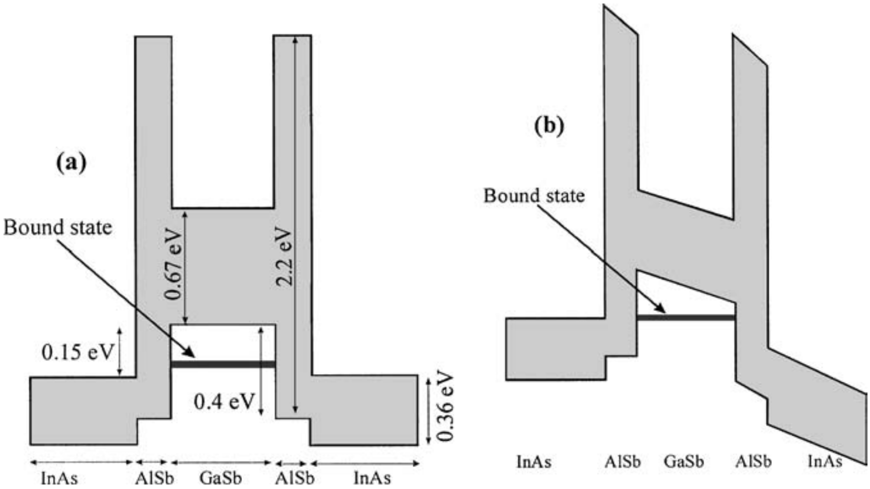


Figure 4.24 A schematic band-edge diagram of a InAs/AlSb/GaSb, AlSb/InAs double-barrier resonant interband tunneling diode under (*a*) zero bias voltage and (*b*) nonzero bias voltage.

the valley current to an acceptable value. One possible design is proposed by Kitabayashi et al. (1997) and is based on the InAs/AlSb/GaSb structure as shown in Fig. 4.24*a*. The structure in this figure is called the *resonant interband tunneling diode*.

The bound state in this case is in the valence band of the GaSb well. The tunneling current will flow when the bound state is lined up with a conduction band of the n-type InAs contact layer as illustrated in Fig. 4.24*b*. The electrons tunnel from the conduction band of the n-type InAs (left layer) through the bound state and then tunnel from the bound state to the conduction band of the n-type InAs on the right-hand side of the structure. With the large AlSb barriers, the peak current can be maintained at higher values by reducing the barrier thicknesses. The valley current in this structure is significantly reduced when the bound state becomes resonant with the bandgap of the InAs layer on the left of Fig. 4.24*b*.

4.6 Coulomb Blockade

For many electronic devices such as metal-oxide-semiconductor field-effect transistors and bipolar transistors, the number of electrons involved in the transport is vary large such that the energy quantization is irrelevant. However, the role of energy discreteness becomes increasingly very important in nanoscale devices where the capacitance in the structure is extremely small. The capacitance of parallel plates of an area A and separated by a distance d is given by

$$C = \frac{\epsilon \epsilon_o A}{d} \tag{4.44}$$

where ϵ is the dielectric constant of the material between the two plates and ϵ_o is the permittivity of space. Let us consider a *pn* junction made of GaAs and with a cross section of 0.1×0.1 mm^2 and a thickness of 10 nm. The energy associated with one electron process is (see Aleiner et al. 2002)

$$E_s = \frac{e^2}{2C} \tag{4.45}$$

Substituting Eq. (4.43) into (4.44), one can obtain a value of E_s to be much smaller than $\mathbf{k}_B T$ even for a temperature as low as that of liquid helium. On the other hand, if the junction cross section is on the order of 10×10 nm^2 and has a thickness of 20 Å, the single electron energy is on the order of 15 meV. This single electron energy is larger than $k_B T$ even for $T = 100$ K. This implies that the energy of a single-electron process is very important in nanostructures. One can think of the single-electron energy as the energy required to add one electron to the capacitor. The energy E_s is thus defined as the *charging energy*.

For any structure with a very small capacitance on the order of *atto*farads, the electrostatic potential caused by a single electron has a profound effect on the tunneling process. Generally speaking, this happens in quantum dot systems where the transport property is regulated by the quantization of the charge in units of the elementary charge e inside the nanostructure. This effect is called the *Coulomb blockade*. The conditions of observing the Coulomb blockade are such that the capacitance C and conductance G of the device satisfy the following inequalities:

$$C \ll \frac{e^2}{k_B T} \qquad \text{and} \qquad G \ll \frac{e^2}{2\pi\hbar} \tag{4.46}$$

where k_B is the Boltzmann constant and $e^2/(2\pi\hbar)$ is called the quantum conductance (inverse of the quantum resistance). The main feature of the Coulomb blockade is the total suppression of the current in a finite interval of external bias voltage such that

$$-\frac{e}{2C} < V_b < +\frac{e}{2C} \tag{4.47}$$

where V_b is the applied bias voltage. To illustrate this process, let us examine the current (I) versus the bias voltage (V_b) characteristic of a thin tunnel junction as shown in Fig. 4.25. For the Coulomb blockade effect to be observed the thermal energy should be smaller than

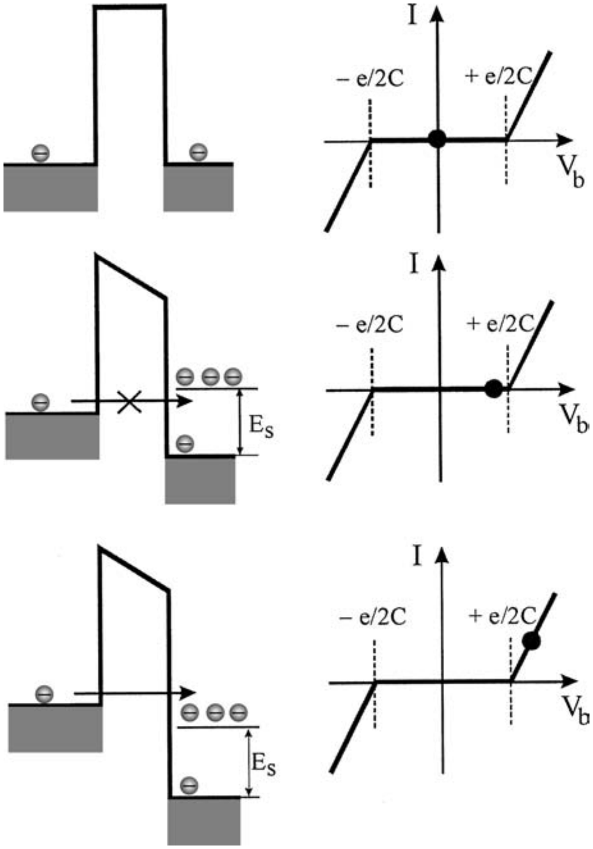


Figure 4.25 Illustration of Coulomb blockade in a thin junction with a small capacitance ($\sim aF$).

the charging energy. When the bias voltage is zero, there is no electron tunneling through the barrier. When the bias voltage is increased, electron flow through the barrier remains zero as long as the bias voltage energy eV_b is smaller than the charging energy E_s. Electron flow occurs when eV_b is larger than E_s as illustrated in the last panel. The current voltage profile is shown with a solid dot indicating the values of the bias voltage. The charge quantization and Coulomb blockade effect are the basis of electronic nanostructure devices such as single-electron transistors. More discussion on this effect will be presented in Chap. 9.

Summary

Tunneling of particles through potential barriers is a quantum effect. This effect was reviewed in this chapter and several examples were discussed. These examples cover simple structures, such as the δ-function and more complex structures such as a quantum well with a double barrier. Several devices based on the tunneling effects, such as the *pn*-junction tunneling diode and resonant tunneling diode, were presented. The *I-V* characteristic curve of the *pn*-junction tunneling diode is composed of three components that contribute to the overall tunneling diode current. These components are the tunneling current, the excess charge current, and the minority-carrier injection current. The sum of these three currents gives a peak-and-valley characteristic feature with a negative differential resistance. The negative differential resistance behavior is what makes these devices attractive for low-power microwave applications.

Coulomb blockage is an effect encountered in nanostructures where the capacitance of the structure is on the order of *atto*farads. The Coulomb blockade is a tunneling effect, and its main feature is the total suppression of the current in a finite interval of external bias voltage. For the Coulomb blockade effect to be observed, the thermal energy $k_B T$ should be smaller than a characteristic energy associated with one electron process known as the charging energy.

Problems

4.1 Consider Fig. 4.7 where the potential can be written as $V(x) = -Fx - e^2/(4x)$. Use the WKB approximation to determine the transmission coefficient.

4.2 Consider the identical double potential barriers shown in Fig. P4.2. Use the WKB approximation to derive an expression for the transmission coefficient. The turning points are labels x_1, x_2, x_3, and x_4.

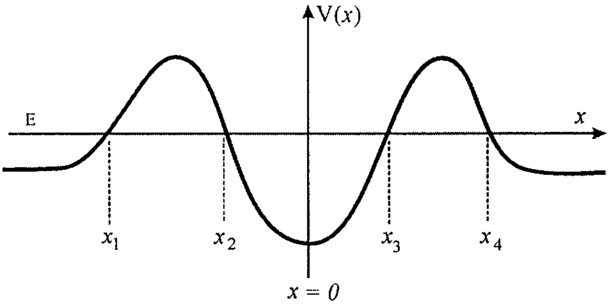


Figure P4.2

4.3 Show that the transmission coefficient of the double barrier-well structure shown in Fig. 4.12 is given by Eq. (4.26).

4.4 Use Eq. (4.26) to determine the number of bound states in a InAs/AlSb/InAs/AlSb/InAs double-barrier resonant tunneling diode. Assume that the well width is 80 Å and the barrier width is 25 Å. The conduction band offset in this system is ~1.0 eV and the electron effective mass in InAs is $0.023m_o$.

4.5 Assume that the peak voltage V_o^p in a resonant tunneling diode is 0.7 V. Calculate the ground-state energy level of a resonant tunneling diode made of GaAs/AlGaAs with a well width of 60 Å and barrier widths of 25 Å using Eq. (4.26). Assume the band offset is 0.3 eV. Compare your results with the value of $V_o^p = 0.7$ V. Give an explanation for the difference between the values obtained for E_1 and V_o^p.

4.6 The empirical tunneling current in the *pn* junction is given by Eq. (4.35). Derive an expression for the negative differential resistance. Find the largest negative differential resistance and the corresponding voltage. Assume that $V_o^p = 0.4$ V, and $I_o^p = 30$ mA.

4.7 Consider a thin GaAs tunnel junction with a thickness of 7 nm and area of 100 Å × 100 Å. What is the temperature needed to generate a thermal energy equivalent to the charging energy? Assume that the dielectric constant of GaAs is 11.56.

4.8 Derive expressions for the transmission and reflection coefficients of a particle of mass m traveling from right to left and tunneling through a δ-function potential barrier of form $V(x) = \lambda\delta(x)$.

4.9 Use a computer program, such as Mathematica to plot Eq. (4.24). Compare your results to Fig. 4.11.

This page intentionally left blank.

Chapter 5

Distribution Functions and Density of States

Semiconductor heterojunctions and nanostructures consist of large numbers of identical particles such as electrons, atoms, holes, and harmonic oscillators. In such cases, it is impossible to try to trace the motion of each individual particle. An alternative way of looking at these large numbers of particles is to settle for knowing averages of relevant dynamical quantities over the entire range of possible system configurations. This leads to the construction of the macroscopic properties of the system and to an understanding of how energy, velocity, and momentum are distributed among the particles that form the system. The branch of physics that addresses the distribution function of a system links the microscopic properties of the system to its macroscopic domains and is called *statistical mechanics*. For physical systems such as semiconductor materials, there are constraints associated with any distribution function. For example, the number of particles is finite, or the total energy of the system is constant. These constraints usually alter the probabilities associated with the possible system configurations.

The techniques of statistical mechanics have been applied to a variety of physical problems in many fields of study, including those involving gases, liquids, polymers, metals, semiconductors, transport theory, DNA, adsorption, spectroscopy, and optical and electrical properties of solids. Statistical thermodynamics is usually applied to a system in equilibrium. This branch of statistical mechanics links thermodynamics and molecular physics. Thermodynamics, on the other hand, provides connections between the properties of the system without supplying any information about the magnitude of any one of them, while statistical thermodynamics assumes the existence of atoms and

Copyright © 2005 by The McGraw-Hill Companies, Inc. Click here for terms of use.

molecules to calculate thermodynamic quantities from a molecular point of view. Statistical thermodynamics is further divided into two areas: (1) the study of systems of molecules in which molecular interaction is neglected, such as for dilute gases, and (2) the study of systems in which molecular interactions are of prime importance, such as for liquids.

To illustrate the terminology of statistical mechanics, let us consider the energy states of a particle in a three-dimensional infinite cubic potential well. These energy states are given by

$$E = E(k_x, k_y, k_z) = \frac{\pi^2\hbar^2}{2ma^2}\left(n_x^2 + n_y^2 + n_z^2\right) \tag{5.1}$$

where $a =$ length of one side of cube
$m =$ mass of the particle
$n_x, n_y, n_z =$ positive integers

The degeneracy $g(E)$ is given by the number of ways that an integer $M = 2ma^2E/\hbar^2$ can be written as the sum of the squares of three positive integers. The result could be an erratic and discontinuous function for small values of n_x, n_y, and n_z, but it becomes smooth for large values of n_x, n_y, and n_z. Consider a three-dimensional space spanned by $k_x = n_x/a, k_y = n_y/a$, and $k_z = n_z/a$ as shown in Fig. 5.1. Equation (5.1) is the equation for a sphere of radius **k**, where $k^2 = k_x^2 + k_y^2 + k_z^2$. Now, it is possible to calculate the number of states in the range dk. This is simply obtained by finding the volume of the shell between k and $k + dk$, which is given as

$$dV_k = 4\pi k^2 dk = 4\pi\left(\frac{2m}{\hbar^2}\right)^{3/2}\sqrt{E}\,dE \tag{5.2}$$

If n_x, n_y, and n_z are positive integers, then Eq. (5.2) should be divided by 8. Moreover, Eq. (5.2) should be divided by the volume of the unit cell in k-space, which is $(\pi/a)^3$, to give the density of state in the shell

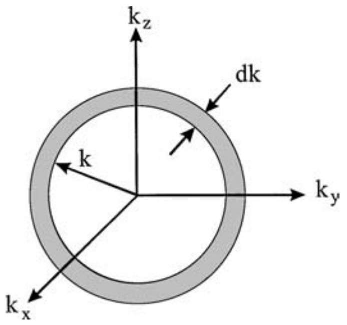


Figure 5.1 A spherical surface of a constant energy is plotted in the **k**-space. The shell of thickness $d\mathbf{k}$ is used to calculate the density of states of a particle in a box.

with thickness dE. The density of state can now be written as

$$g(E)\,dE = \frac{1}{8}\frac{1}{(\pi/a)^3}4\pi\left(\frac{2m}{\hbar^2}\right)^{3/2}\sqrt{E}dE$$

$$= \frac{a^3}{2\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2}\sqrt{E}dE \tag{5.3}$$

If one assumes that the energy $E = 3k_BT/2$, $T = 300$ K, $m = 9.11 \times 10^{-31}$ kg, and a =100 Å, the density of state can be easily obtained as $g(E)\,dE = 8.40 \times 10^{21} dE$. Thus, even for a system as simple as a particle in a box, the degeneracy can be very large at room temperature.

For a system consisting of N-noninteracting particles, the degeneracy is extremely high. The energy of the system is

$$E = \frac{\hbar^2}{2ma^2}\sum_j^N\left(n_{xj}^2 + n_{yj}^2 + n_{zj}^2\right) = \frac{\hbar^2}{2ma^2}\sum_j^N R_j^2 \tag{5.4}$$

where n_{xj}^2, n_{yj}^2, n_{zj}^2, and R_j^2 are positive integers. The degeneracy of the system can be calculated by generalizing this procedure for a one-particle system. The density of state can be written as

$$g(E)\,dE = \prod_j^N g_j(E)\,dE$$

$$= \frac{\sqrt{\pi}}{\Gamma(N+1)\Gamma(3N/2)}\left(\frac{a^3}{2\pi^2}\right)^N\left(\frac{2m}{\hbar^2}\right)^{3N/2} E^{(3N/2-1)}dE \tag{5.5}$$

where $\Gamma(x)$ is the gamma function given by

$$\Gamma(x) = \int_0^\infty e^{-t}t^{x-1}\,dt \tag{5.6}$$

The gamma function has the following properties:

$$\Gamma(x+1) = x\Gamma(x) \tag{5.7a}$$

$$\Gamma(n+1) = n! \qquad \text{for } n = \text{integer} \tag{5.7b}$$

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi} \tag{5.7c}$$

$$\Gamma\left(n+\frac{1}{2}\right) = \frac{(2n)!}{2^{2n}n!}\sqrt{\pi} \tag{5.7d}$$

The density of state $g(E)\,dE$ is calculated to be on the order of 10^N, where N is on the order of Avogadro's number. For $N = 10$, we obtain a density of state the order of 10^{1089}. Thus, the concept of density of state is very important for macroscopic systems.

5.1 Distribution Functions

Consider a liter of salt solution. From a macroscopic point of view, we can completely specify the system by a few parameters such as volume, concentration, and temperature. Regardless of the complexity of the system, it requires only a small number of parameters to describe it. From a microscopic point of view, there are enormous numbers of quantum states associated with the fixed macroscopic properties. Gibbs introduced the concept of *ensemble*, which is a virtual collection of a very large number of systems, denoted A, each constructed to be a replica on the macroscopic level of a particular system of interest. Suppose that an isolated system has a volume V, contains N molecules, and is known to have an energy E. Thus, the ensemble would have a volume AV, contain AN molecules, and have a total energy of AE. Each of the systems in this ensemble is a quantum mechanical system of N interacting molecules in a container of volume V. The values of N and V and the interaction between molecules are sufficient to determine the energy eigenvalues E_j of the Schrödinger equation along with their associated degeneracies $g(E_j)$. These energies are the only eigenvalues available to the system. The fixed energy of the system E is one of these E_j's, and there is a degeneracy $g(E)$. While the systems in the ensemble are identical at the macroscopic level, they may differ at the molecular level. Nothing has been said thus far about the distribution of the member of the ensemble with respect to the degeneracy of the possible quantum states.

The ensembles are required to obey the principle of *equal a priori probability*, which states that every $g(E)$ is represented an equal number of times in the ensembles. Thus, each $g(E)$ is treated equally, and the number of systems in the ensemble is an integer multiple of $g(E)$. An alternative interpretation of the principle of equal a priori probabilities is that an isolated system is equally likely to be in any of its $g(E)$ possible quantum states.

The most commonly used ensemble is called the *canonical ensemble*, in which the volume, number of particles, and the temperature are constant. The occupation number refers to the number of systems of the ensemble occupying a specific quantum state. The occupation numbers must satisfy the condition

$$\sum_j a_j = A \tag{5.8}$$

where a_j is the occupation number of the jth state and A is the total number of the systems in the ensemble, and the condition

$$\sum_j a_j E_j = E \tag{5.9}$$

where E_j is the energy of the jth state and E is the total energy of the ensemble. These two conditions mean that all the members of the ensemble are included in the calculations and the total energy of the system is fixed. The number of ways, $\Omega(\mathbf{a}) = \Omega(a_1, a_2, a_3, \ldots)$, that any particular distribution of the a_j's can be realized is the number of ways that A distinguishable objects (ones that we can label uniquely) can be arranged in groups, such as a_1 in the first group, a_2 in the second group, etc., is

$$\Omega(\mathbf{a}) = \frac{A!}{a_1!a_2!a_3!\ldots} = \frac{A!}{\prod_j a_j!} \tag{5.10}$$

The overall probability P_j that a system is in the jth quantum state is obtained by averaging the fraction a_j/A of the systems or members of the canonical ensemble in the jth state with an energy E_j. It can be written as

$$P_j = \frac{\overline{a}_j}{A} = \frac{1}{A}\frac{\sum_a \Omega(\mathbf{a})a_j(\mathbf{a})}{\sum_a \Omega(\mathbf{a})} \tag{5.11}$$

where the notation $a_j(\mathbf{a})$ indicates that the value a_j depends on the distribution and summations over all distributions that satisfy Eqs. (5.8) and (5.9). Given the probability that a system is in the jth state, one can calculate the canonical ensemble average of any property $\overline{M}$ from the following relation:

$$\overline{M} = \sum_j M_j P_j \tag{5.12}$$

where M_j is the value of M in the jth quantum state. For more discussion on this subject, see McQuarrie (1976).

5.2 Maxwell-Boltzmann Statistic

For a distinguishable number of particles, N, in a container with many compartments g_i, the number of ways that the particles can be distributed among the compartments can be written as

$$\Omega(N_1, N_2, \ldots, N_j) = \frac{N!}{\prod_{i=1}^{j} N_i!} \prod_{i=1}^{j} g_i^{N_i} \tag{5.13}$$

where $g_i^{N_i}$ is the degeneracy of finding the N_i particle in the g_i compartment. The problem of determining what set of values for the numbers $N_1, N_2, N_3, \ldots, \mathrm{N}_j$ will make Ω as large as possible, subject to the constraints of a constant number of particles, $\sum_{i=1}^{j} N_i = N =$ constant, and constant energy, $\sum_{i=1}^{j} \varepsilon_i N_i = E =$ constant (where ε_i is the total energy of the system consisting of N_i particles), is more or less a mathematical exercise. To maximize a function of many variables with a given constraint, one can apply the Lagrange method of undetermined multipliers as follows:

$$\frac{\partial f}{\partial x_i} + \alpha \frac{\partial g_1}{\partial x_i} + \beta \frac{\partial g_2}{\partial x_i} = 0 \tag{5.14}$$

where f = function to be maximized
α, β = undetermined multipliers
g_1, g_2 = constraints

To maximize Eq. (5.13) with two constraints, one can maximize the logarithm of Ω since the logarithm of products is converted to a sum which is much easier to handle mathematically. Hence, one can write

$$\frac{\partial}{\partial x_j}\left[\ln(\Omega) + \alpha \sum_{i=1}^{j} N_i + \beta \sum_{i=1}^{j} \varepsilon_i N_i\right] = 0 \tag{5.15}$$

The first derivative with respect to the N_j particle of the logarithm of Eq. (5.13) is

$$\begin{aligned}\frac{\partial \ln(\Omega)}{\partial N_j} &= \frac{\partial \ln N!}{\partial N_j} + \frac{\partial \sum_{i=1}^{j} N_i \ln g_i}{\partial N_j} - \frac{\partial \sum_{i=1}^{j} \ln N_i!}{\partial N_j} \\ &= 0 + \ln g_j - \ln N_j \\ &= \ln g_j - \ln N_j \end{aligned} \tag{5.16}$$

It follows that the derivatives of the two constraints are

$$\alpha \frac{\partial \sum_{i=1}^{j} N_i}{\partial N_j} = \alpha \qquad \text{and} \qquad \beta \frac{\partial \sum_{i=1}^{j} \varepsilon_i N_i}{\partial N_j} = \beta \varepsilon_j \tag{5.17}$$

Substituting Eqs. (5.16) and (5.17) into Eq. (5.15), we have

$$\ln g_j - \ln N_j + \alpha + \beta\varepsilon_j = 0 \tag{5.18}$$

The quantity N_j in this equation is the most probable number of particles to be found in the jth energy level, and g_j is the number of the quantum states associated with the jth energy level. Thus, N_j/g_j is the average number of particles per quantum state at that energy level, which is by definition the distribution function $f(\varepsilon_j)$. Equation (5.18) becomes

$$\frac{N_j}{g_j} = f(\varepsilon_j) = e^{\alpha}e^{\beta\varepsilon_j} \tag{5.19}$$

Without going through the thermodynamic derivation, it is found that

$$\beta = -\frac{1}{k_B T} \tag{5.20}$$

where k_B is the Boltzmann constant and T is temperature in kelvins. The task now is to determine the multiplier α. From Eq. (5.19) the number that occupies the jth quantum state can be written as

$$N_j = g_j e^{\alpha}e^{\beta\varepsilon_j} \tag{5.21}$$

It follows that the total number of particles in the system is

$$N = \sum_j N_j = \sum_j g_j e^{\alpha} e^{-\varepsilon_j/k_B T} = e^{\alpha}\sum_j g_j e^{-\varepsilon_j/k_B T} \tag{5.22}$$

which can be solved for the quantity e^{α} such as

$$e^{\alpha} = \frac{N}{\sum\limits_j g_j e^{-\varepsilon_j/k_B T}} \tag{5.23}$$

Substituting Eq. (5.23) into (5.21) we have

$$N_j = \frac{N g_j e^{-\varepsilon_j/k_B T}}{\sum\limits_j g_j e^{-\varepsilon_j/k_B T}} \tag{5.24}$$

When the system is composed of a quasi-continuum eigenvalues, the degeneracy g_j can be replaced by the density of state $g(E)\,dE$, the population N_j can be replaced by the function $N(E)\,dE$, and the summation can be replaced by integration over the region of allowed energies. These substitutions allow one to rewrite Eqs. (5.21) to (5.24) as

$$N(E)\,dE = f(E)g(E)\,dE = e^{\alpha}e^{-E/k_BT}g(E)\,dE \tag{5.25}$$

$$N = \int N(E)\,dE = e^{\alpha}\int e^{-E/k_BT}g(E)\,dE \tag{5.26}$$

$$e^{\alpha} = \frac{N}{\int e^{-E/k_BT}g(E)\,dE} \tag{5.27}$$

$$N(E)\,dE = \frac{Ne^{-E/k_BT}g(E)\,dE}{\int e^{-E/k_BT}g(E)\,dE} \tag{5.28}$$

For an ideal monatomic gas in a cubic container with sides of length a, the density of states can be presented by the expression shown in Eq. (5.3). By substituting this expression in to Eq. (5.27) we have

$$e^{\alpha} = \frac{N}{\frac{a^3}{2\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2}\int\limits_0^{\infty} e^{-E/k_BT}\sqrt{E}\,dE} \tag{5.29}$$

Let $x = E/k_BT$, which implies that $dE = k_BT\,dx$ and $\sqrt{E} = \sqrt{k_BT}\sqrt{x}$. Substituting these quantities back into Eq. (5.29) yields

$$e^{\alpha} = \frac{N}{\frac{a^3}{2\pi^2}\left(\frac{2mk_BT}{\hbar^2}\right)^{3/2}\int\limits_0^{\infty} e^{-x}\sqrt{x}\,dx} \tag{5.30}$$

The integral in this equation is a Γ-function [see Eq. (5.6)] with an argument of $\frac{3}{2}$. The function $\Gamma\left(\frac{3}{2}\right)$ can be evaluated from Eq. (5.7d) to be $\sqrt{\pi}/2$. Substitute this quantity back into Eq. (5.30) to obtain

$$e^{\alpha} = \frac{\sqrt{2}N}{a^3}\left(\frac{\pi\hbar^2}{mk_BT}\right)^{3/2} \tag{5.31}$$

Finally, substitute Eq. (5.31) into (5.19) to obtain the following expression for the distribution function:

$$f(E) = \frac{\sqrt{2}N}{a^3}\left(\frac{\pi\hbar^2}{mk_BT}\right)^{3/2} e^{-E/k_BT} \tag{5.32}$$

This expression is known as the Maxwell-Boltzmann distribution function, which is applicable to noninteracting particles in a system whose density of states is defined by Eq. (5.3). A plot of this function is shown

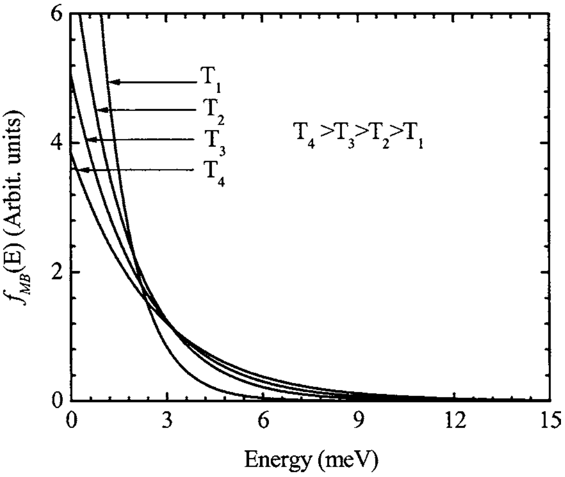


Figure 5.2 Maxwell-Boltzmann distribution function plotted as a function of energy for four different temperatures.

in Fig. 5.2 for four different temperatures. It is clear that this function becomes steeper as the temperature decreases.

5.3 Fermi-Dirac Statistics

The Maxwell-Boltzmann distribution function is applicable to classical systems where the particle can be identified and labeled. For quantum systems, there is no way one can distinguish between electrons or protons, for example. Quantum systems are composed of *inherently indistinguishable* particles, and therefore Maxwell-Boltzmann statistics cannot be applied. In addition to this point, the Pauli exclusion principle requires that the spin of particles be taken into consideration. These two points require a different distribution function known as the Fermi-Dirac distribution. The number of ways of realizing a distribution of N_j indistinguishable particles is determined as follows:

$$\Omega_{\mathrm{FD}}(N_1, N_2, N_3, \ldots, N_n) = \prod_j^n \frac{g_j!}{N_j!(g_j - N_j)!} \tag{5.33}$$

where g_j are the quantum states. The logarithm of this equation is

$$\ln \Omega_{\mathrm{FD}} = \sum_j \ln g_j! - \sum_j \ln N_j! - \sum_j \ln(g_j - N_j)! \tag{5.34}$$

Taking the first derivative of Eq. (5.34) with respect to N_i, we can write

$$\begin{aligned}\frac{\partial \ln \Omega_{\mathrm{FD}}}{\partial N_i} &= \frac{\partial \sum_j \ln g_j!}{\partial N_i} - \frac{\partial \sum_j \ln N_j!}{\partial N_i} - \frac{\partial \sum_j \ln(g_j - N_j)!}{\partial N_i} \\ &= 0 - \ln N_i + \frac{\partial \sum_j \ln(g_j - N_j)!}{\partial(g_i - N_i)} \\ &= -\ln N_i + \ln(g_i - N_i) \\ &= \ln\left(\frac{g_i}{N_i} - 1\right) \qquad (5.35)\end{aligned}$$

Lagrange's method of undetermined multipliers can now be applied to maximize Eq. (5.35) by using Eqs. (5.14) and (5.15), which yields

$$\ln\left(\frac{g_i}{N_i} - 1\right) = -\alpha - \beta E_i \qquad (5.36)$$

The quantity g_i/N_i is called the Fermi-Dirac distribution function $f_{\mathrm{FD}}(E_i)$ and can be written as

$$\frac{N_i}{g_i} = f_{\mathrm{FD}}(E_i) = \frac{1}{1 + e^{-(\alpha+\beta E_i)}} \qquad (5.37)$$

where β is as defined in Eq. (5.20) and α in the Fermi-Dirac distribution function is taken to be

$$\alpha = \frac{E_F}{k_B T} \qquad (5.38)$$

where E_F is known as the Fermi energy level. Substituting Eqs. (5.20) and (5.38) into Eq. (5.37) yields

$$f_{\mathrm{FD}}(E_i) = \frac{1}{1 + e^{(E_i - E_F)/k_B T}} \qquad (5.39)$$

This function, known as the Fermi-Dirac distribution function, is plotted as a function of energy for different temperatures as shown in Fig. 5.3. Notice that at $T = 0$ K, f_{FD} becomes a step function.

For quasi-continuous energy levels in which the degeneracy is represented by a density-of-state function, we can write

$$N(E)\,dE = f_{\mathrm{FD}}(E) g(E)\,dE = \frac{g(E)\,dE}{1 + e^{(E - E_F)/k_B T}} \qquad (5.40)$$

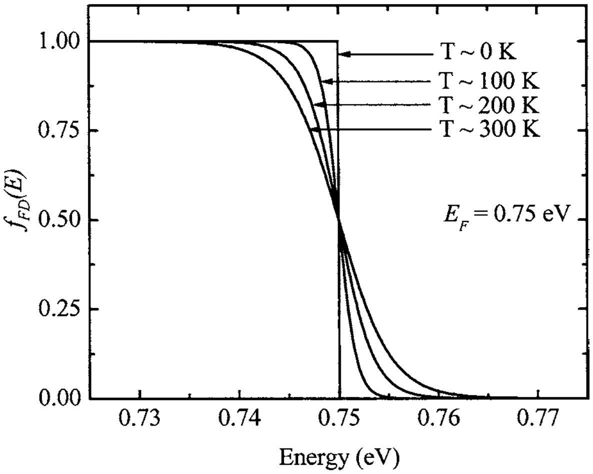


Figure 5.3 Fermi-Dirac function plotted as a function of energy for four different temperatures.

Substitute Eq. (5.39) into (5.40), and then integrate to give

$$N = \int f_{\mathrm{FD}}(E) g(E)\,dE = \int \frac{g(E)\,dE}{1 + e^{(E-E_F)/k_B T}}$$

$$= \frac{a^3}{2\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} \int_0^{\infty} \frac{\sqrt{E}\,dE}{1 + e^{(E-E_F)/k_B T}} \tag{5.41}$$

The integral in this equation is difficult to evaluate analytically. For $(E - E_F) \gg k_B T$, one can find a solution, for this integral, of the following form:

$$N = \frac{a^3}{2\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} \frac{e^{E_F/k_B T}\sqrt{\pi}(k_B T)^{3/2}}{2}$$

$$= \frac{a^3}{4}\left(\frac{2mk_B T}{\pi\hbar^2}\right)^{3/2} e^{E_F/k_B T} \tag{5.42}$$

or

$$E_F = k_B T \ln\left[\frac{4N}{a^3}\left(\frac{\pi\hbar^2}{2mk_B T}\right)^{3/2}\right] \tag{5.43}$$

The Fermi energy of the form of Eq. (5.43) is plotted for bulk GaAs materials as a function carrier concentration with respect to the conduction

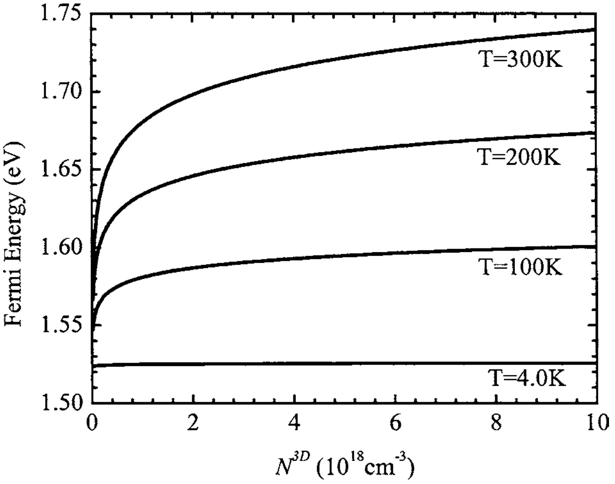


Figure 5.4 The Fermi energy plotted as a function of carrier concentration using Eq. (5.43) for different temperatures. The Fermi energy was taken with respect to the bottom of the conduction band minimum of GaAs.

band minimum as shown in Fig. 5.4. The curves obtained in this figure were plotted for different temperatures between 300 and 4.0 K. The conduction band minimum was included as a function of temperature as well. The sample size was chosen as a cubic specimen with a side of 1 cm. It is customary to divide the density of state by the volume in real space so that the volume of the sample would not show in the final expressions of either the Fermi energy or the carrier concentrations [Eq. (5.42)]. As the temperature decreases, the Fermi energy is reduced due to carrier freeze-out.

5.4 Bose-Einstein Statistics

Bose-Einstein statistics are used for particles that possess zero or integer spins ($S = 0, 1, 2, 3, \ldots$) which do not obey the Pauli exclusion principle. These particles are still indistinguishable. The most common particles that follow Bose-Einstein statistics are photons ($S = 1$). In this case, an arbitrary number of particles can occupy a single quantum state, and hence the number of ways of arranging N particles in the system can be written as

$$\Omega_{\mathrm{BE}}(N_1, N_2, N_3, \ldots, N_n) = \prod_{j}^{n} \frac{(N_j + g_j - 1)!}{N_j!(g_j - 1)!} \tag{5.44}$$

One now can proceed to maximize this function using Lagrange's method of undetermined multipliers as described in Secs. 5.2 and 5.3. The final results can be written as

$$f_{\mathrm{BE}}(E) = \frac{1}{e^{\alpha}e^{E/k_BT} - 1} \tag{5.45}$$

where $f_{\mathrm{BE}}(E)$ is the Bose-Einstein distribution function and α can be considered zero since the photons can be easily created and annihilated, and therefore the constraint of having a constant number of particles can be easily discarded. The function $f_{\mathrm{BE}}(E)$ can be reduced to

$$f_{\mathrm{BE}}(E) = \frac{1}{e^{E/k_BT} - 1} \tag{5.46}$$

The three distribution functions (Maxwell-Boltzmann, Fermi-Dirac, and Bose-Einstein) are plotted as a function of energy in Fig. 5.5 at $T = 300$ and 77 K with the Fermi energy chosen as $E_F = 0.2$ eV. It is obvious from this plot that the three distribution functions are in good agreement with each other when $(E - E_F) \gg k_BT$. The agreement is even improved as the temperature is reduced from 300 to 77 K. In many semiconductor cases, the Fermi-Dirac distribution function can

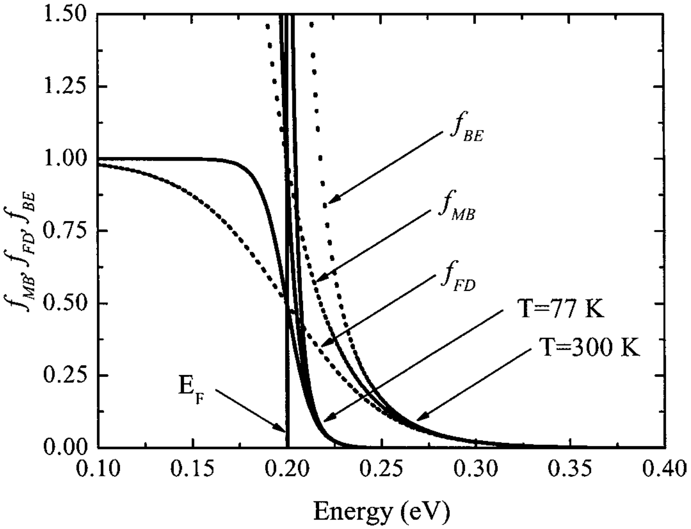


Figure 5.5 Fermi-Dirac (f_{FD}), Maxwell-Boltzmann (f_{MB}), and Bose-Einstein (f_{BE}) distribution functions plotted as a function of energy at $T = 300$ K (dashed lines) and at $T = 77$ K (solid lines).

be approximated as a Maxwell-Boltzmann distribution function, in particular when exp $[(E - E_F)/k_B T] \gg 1$.

5.5 Density of States

To understand the energy and momentum distribution among particles in a system, one needs to answer the question of how many states are available for these particles to occupy in a particular system. For a large number of particles in a three-dimensional system, such as electrons in a crystalline semiconductor, the answer to this question can be understood by applying Bloch's theorem (see Sec. 3.1) to the crystalline semiconductor where the wave function exhibits periodicity within the period structure (crystal) such as

$$\begin{aligned}\psi(x, y, z) &= \psi(x + L_x, y + L_y, z + L_z)\\ &= \exp\{i[k_x(x + L_x) + k_y(y + L_y) + k_z(z + L_z)]\}\\ &= \exp[i(k_x x + k_y y + k_z z)]\exp[i(k_x L_x + k_y L_y + k_z L_z)] \qquad (5.47)\end{aligned}$$

where $L_x = L_y = L_z = L$ is the period of a cubic crystal. For Bloch's theorem to be valid the second exponential of Eq. (5.47) must be unity, which implies that

$$k_x L = 2\pi n_x \qquad k_y L = 2\pi n_y, \qquad k_z L = 2\pi n_z \qquad (5.48)$$

where n_x, n_y, and n_z are integers. The volume of a unit cell $\left(V_k^o\right)$in the **k**-space occupied by one state is

$$V_k^o = k_x k_y k_z = \frac{(2\pi)^3}{L^3} \qquad (5.49)$$

Other states are obtained by assuming other values for n_x, n_y, and n_z such as (000), (100), (110), (200), and (210), which gradually fill a sphere of radius **k**. The Fermi energy is thus defined at zero temperature where the states within the sphere of radius $\mathbf{k}_F$ are all occupied and states for $\mathbf{k} > \mathbf{k}_F$ are all empty, where $\mathbf{k}_F$ is the Fermi wave vector.

We can define the density of states $g(E)$ as the number of states per unit energy per unit volume of real space (see, for example, Harrison 2000) such that

$$g(E) = \frac{\partial N}{\partial E} \qquad (5.50)$$

It follows that the total number of states, N, is equal to the degeneracy times the volume of the sphere in **k**-space divided by the volume occupied by one state (primitive unit cell) and divided again by the volume of real space such that

$$N = 2\frac{4\pi k^3}{3}\frac{1}{(2\pi/L)^3}\frac{1}{V} = 2\frac{4\pi k^3}{(2\pi)^3} \tag{5.51}$$

where we assume $V = L^3$. For electrons of spin $\frac{1}{2}$, the degeneracy is 2 for spin-up and spin-down. The density of states can be written as

$$g(E) = \frac{\partial N}{\partial \mathbf{k}}\frac{\partial \mathbf{k}}{\partial E} \tag{5.52}$$

where

$$\frac{\partial N}{\partial \mathbf{k}} = 2\frac{4\pi k^2}{(2\pi)^3} \tag{5.53}$$

From the effective mass approximation, the energy of the electrons is assumed to be parabolic in **k**-space as follows:

$$E = \frac{\hbar^2 k^2}{2m^*} \tag{5.54}$$

which yields

$$\frac{\partial \mathbf{k}}{\partial E} = \left(\frac{2m^*}{\hbar^2}\right)^{1/2}\frac{1}{2\sqrt{E}} \tag{5.55}$$

Substituting Eqs. (5.53) to (5.55) into Eq. (5.52), we obtain

$$g(E) = \frac{1}{2\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{3/2}\sqrt{E} \tag{5.56}$$

A plot of $g(E)$ as a function of energy is shown in Fig. 5.6 where the effective mass is assumed to be $m^* = 0.067m_o$. The inset is the three-dimensional sphere of radius $\mathbf{k}_F$ in the **k**-space. This is a typical example of electrons in bulk semiconductor material such as GaAs or silicon.

To understand the concept of Fermi energy and the distribution of electrons and holes in a semiconductor, let us first assume that the semiconductor is intrinsic, which means that the number of electrons in the conduction band is equal to the number of holes in the valence band. The density of states for both the conduction and valence bands

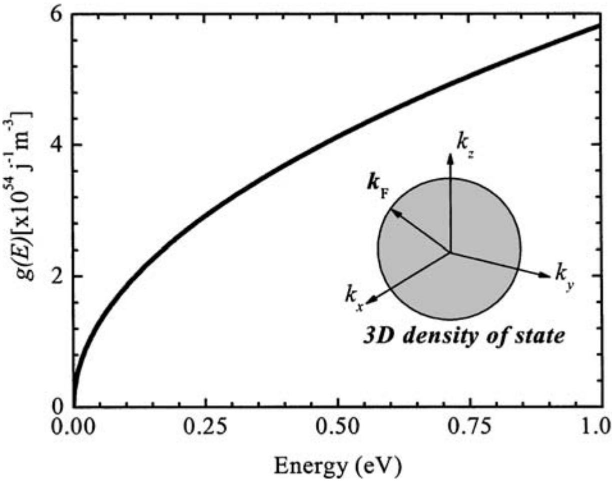


Figure 5.6 The density of states of a three-dimensional (3D) system, which is a typical bulk material such as semiconductor single crystals. The calculations were made for the electron effective mass in GaAs. The inset is a sphere in the **k**-space of radius $\mathbf{k}_F$.

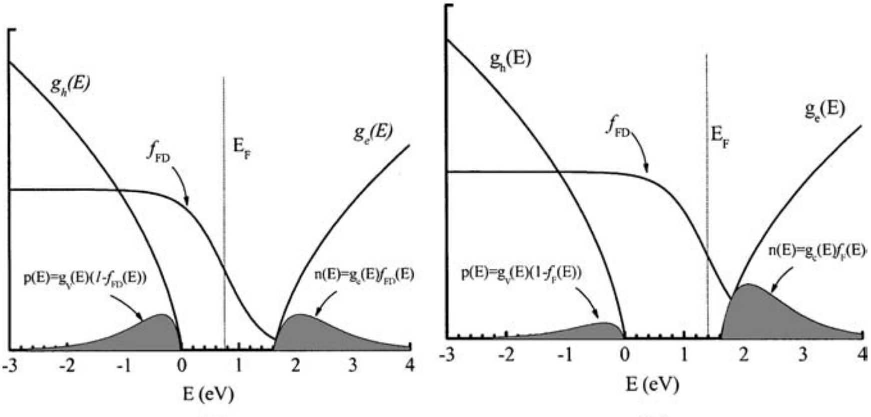
can be written as follows:

$$g_e(E) = \frac{1}{2\pi^2}\left(\frac{2m_e^*}{\hbar^2}\right)^{3/2}\sqrt{E - E_c} \qquad \text{and}$$
$$g_h(E) = \frac{1}{2\pi^2}\left(\frac{2m_h^*}{\hbar^2}\right)^{3/2}\sqrt{E_v - E} \tag{5.57}$$

where the subscripts e and h stand for electrons and holes, respectively, and $E_c$and E_v are the bottom and top of the conduction and valence bands, respectively. The result of plotting Eq. (5.57) is shown in Fig. 5.7. The electron distribution function is given by the Fermi-Dirac function Eq. (5.39), while the distribution function of holes can be expressed as the distribution function of unfilled states $(1 - f_{\text{FD}})$:

$$\begin{aligned} f_{\text{FD}}^h(E) &= 1 - f_{\text{FD}}(E) = 1 - \frac{1}{1 + e^{(E-E_F)/k_BT}} \\ &= \frac{1}{e^{-(E-E_F)/k_BT} + 1} \end{aligned} \tag{5.58}$$

The superscript h is introduced to refer to the hole distribution function. The electron and hole concentrations are shown as the shaded areas in Fig. 5.7. The carrier concentrations are plotted for intrinsic GaAs materials as shown in Fig. 5.7*a*. In this figure the number of electrons $[n(E)]$ is equal to the number of holes $[p(E)]$. For n-type GaAs,



(a) (b)

Figure 5.7 The distribution function, density of states, Fermi energy level, and carrier population for (*a*) intrinsic and (*b*) *n*-type GaAs are sketched as a function of energy. The *y*-axis unit is taken as an arbitrary unit since we have several overlaid parameters.

where the material is doped with donors, the Fermi energy is shifted toward the conduction band as shown in Fig. 5.7*b* and $n(E) > p(E)$ as indicated by the shaded area. The Fermi-Dirac distribution function is also shifted. For a degenerate *n*-type semiconductor (heavily doped semiconductor), the Fermi energy is pinned above the conduction band minimum. Similarly, for a *p*-type semiconductor, the Fermi energy will be shifted toward the valence band and for a degenerate *p*-type semiconductor, the Fermi energy is resonant in the valence band.

It was shown, as illustrated in Fig. 5.5, that the Fermi-Dirac distribution function can be approximated by the Maxwell-Boltzmann function for $(E_c - E_F) \gg k_B T$, where E_c is the bottom of the conduction band. This is a valid assumption for an intrinsic or lightly doped semiconductor where E_F is pinned near the midgap. This implies that $\exp[(E - E_F)/k_B T] \gg 1$, which yields

$$\begin{aligned} f_{\mathrm{FD}}(E) &= \frac{1}{1 + e^{(E-E_F)/k_B T}} \approx \frac{1}{e^{(E-E_F)/k_B T}} \\ &= e^{-(E-E_F)/k_B T} = e^{E_F/k_B T} e^{-E/k_B T} \end{aligned} \tag{5.59a}$$

and

$$\begin{aligned} f_{\mathrm{FD}}^{h}(E) &= 1 - \frac{1}{1 + e^{(E-E_F)/k_B T}} = \frac{1}{e^{-(E-E_F)/k_B T} + 1} \\ &\approx e^{+(E-E_F)/k_B T} = e^{-E_F/k_B T} e^{E/k_B T} \end{aligned} \tag{5.59b}$$

With this Maxwell-Boltzmann approximation, the electron and hole densities can be easily evaluated for a semiconductor at equilibrium as follows:

$$
\begin{aligned}
n_o &= \int f_{\mathrm{FD}}(E) g(E)\, dE = \int \frac{g(E)\, dE}{1 + e^{(E-E_F)/k_B T}} \\
&= \frac{1}{2\pi^2} \left(\frac{2m_e^*}{\hbar^2} \right)^{3/2} \int_{E_c}^{\infty} \frac{\sqrt{E - E_c}\, dE}{(1 + e^{(E-E_F)/k_B T})} \\
&= \frac{1}{2\pi^2} \left(\frac{2m_e^* k_B T}{\hbar^2} \right)^{3/2} e^{(E_F - Ec)/k_B T} \int_0^{\infty} e^{-x} \sqrt{x}\, dx \\
&= \frac{1}{2\pi^2} \left(\frac{2m_e^* k_B T}{\hbar^2} \right)^{3/2} e^{(E_F - E_c)/k_B T} \frac{\sqrt{\pi}}{2} \\
&= \frac{1}{4} \left(\frac{2m_e^* k_B T}{\pi \hbar^2} \right)^{3/2} e^{-(E_c - E_F)/k_B T} \\
&= N_c e^{-(E_c - E_F)/k_B T}
\end{aligned}
\tag{5.60}
$$

where n_o is the electron density and N_c is given by

$$
N_c = \frac{1}{4} \left(\frac{2m_e^* k_B T}{\pi \hbar^2} \right)^{3/2} \tag{5.61}
$$

The density of states used in this derivation is given by Eq. (5.57). Similarly, the hole concentration p_o can be obtained as

$$
p_o = N_v e^{-(E_F - E_v)/k_B T} \qquad \text{where } N_v = \frac{1}{4} \left(\frac{2m_h^* k_B T}{\pi \hbar^2} \right)^{3/2} \tag{5.62}
$$

The mass action law, $n_o p_o = n_i^2$, where n_i is the intrinsic carrier concentration, can now be written as

$$
\begin{aligned}
n_i = \sqrt{n_o p_o} &= \left[\frac{1}{4} \left(\frac{2m_h^* k_B T}{\pi \hbar^2} \right)^{3/2} \frac{1}{4} \left(\frac{2m_e^* k_B T}{\pi \hbar^2} \right)^{3/2} \right]^{1/2} \\
&\quad \times e^{-(E_F - E_v)/k_B T} e^{-(E_c - E_F)/k_B T} \\
&= \frac{1}{4} \left(\frac{2\sqrt{m_h^* m_e^*} k_B T}{\pi \hbar^2} \right)^{3/2} e^{-(E_c - E_v)/k_B T} \\
&= \frac{1}{4} \left(\frac{2\sqrt{m_h^* m_e^*} k_B T}{\pi \hbar^2} \right)^{3/2} e^{-E_g/k_B T}
\end{aligned}
\tag{5.63}
$$

where $E_g = E_c - E_v$ is the bandgap energy. The intrinsic carrier concentration is thus independent of the Fermi energy level. The Fermi energy for an intrinsic semiconductor can be evaluated by equating n_o and p_o, which yields

$$\frac{1}{4}\left(\frac{2m_e^* k_B T}{\pi\hbar^2}\right)^{3/2} e^{-(E_c-E_F)/k_BT} = \frac{1}{4}\left(\frac{2m_h^* k_B T}{\pi\hbar^2}\right)^{3/2} e^{-(E_F-E_v)/k_BT}$$

or (5.64)

$$e^{[2E_F-(E_c+E_v)]/k_BT} = \left(\frac{m_h^*}{m_e^*}\right)^{3/2}$$

Taking the natural log of both sides and rearranging, we have

$$E_F = \frac{1}{2}(E_c + E_v) + \frac{3}{4}k_BT\ \ln\left(\frac{m_h^*}{m_e^*}\right) \tag{5.65}$$

Usually the top of the valance band is taken as a reference point, which can be set as zero. The intrinsic Fermi energy at room temperature is ~0.78 eV for GaAs with $m_h^* = 0.45m_o$, $m_e^* = 0.067m_o$, and $E_c = 1.48$ eV.

5.6 Density of States of Quantum Wells, Wires, and Dots

The density of states in low-dimensional systems is derived in this section. To avoid confusion about how the low-dimensional systems are defined, we consider that the charge carriers have degree-of-freedom directions and confinement directions. For bulk materials, there are three degree-of-freedom directions and zero confined directions. Thus, bulk materials are called three-dimensional systems. Quantum wells are considered to be two-dimensional systems, which means that the charge carriers have two degree-of-freedom directions and one confined direction. In this case, the growth direction is the confined direction. Quantum wires on the other hand, have one degree-of-freedom direction and two confined directions. Thus, quantum wires are considered one-dimensional systems. When the charge carriers are confined in three directions, the structure is called a zero-dimensional system. We refer to this as a quantum dot system.

5.6.1 Quantum wells

The density of states in a quantum well system is restricted to the k_xk_y plane shown in the inset of Fig. 5.8 where the electrons or holes are now confined in this plane and their motion is restricted along the growth

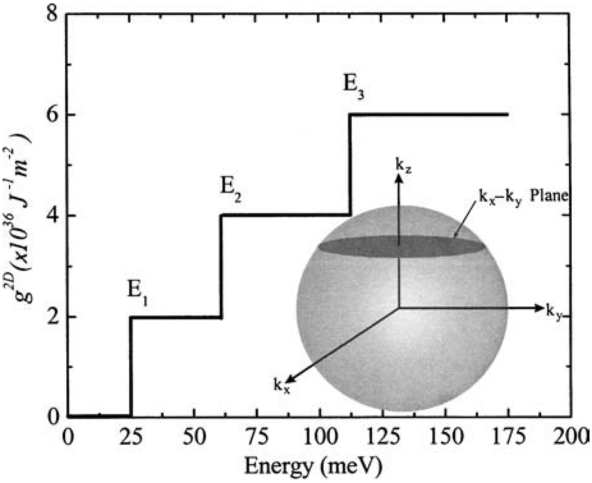


Figure 5.8 The density of states as a function of energy for a two-dimensional system such as GaAs/AlGaAs multiple quantum wells. The inset illustrates the two-dimensional confinement where the charge carriers are confined in the $k_x\ k_y$ plane.

axis (z direction in the real space, or k_z direction in the momentum space). The total number of states per unit cross-sectional area is given by the area in **k**-space divided by the area of the unit cell in **k**-space and divided by the area in real space:

$$N^{2\mathrm{D}} = 2\pi k^2 \frac{1}{(2\pi/L)^2}\frac{1}{L^2} = 2\frac{\pi k^2}{(2\pi)^2} \tag{5.66}$$

where Factor 2 = spin degeneracy of electrons
L^2 = real space square area
$2\pi/L^2$ = two-dimensional primitive unit cell in **k**-space

The density of state can be expressed as

$$\begin{aligned} g^{2\mathrm{D}}(E) &= \frac{\partial N^{2\mathrm{D}}}{\partial E} = \frac{\partial N^{2\mathrm{D}}}{\partial \mathbf{k}}\frac{\partial \mathbf{k}}{\partial E} \\ &= \left(\frac{k}{\pi}\right)\left(\frac{2m^*}{\hbar^2}\right)^{1/2}\frac{1}{2\sqrt{E}} = \frac{m^*}{\pi\hbar^2} \end{aligned} \tag{5.67}$$

where the energy E is defined in Eq. (5.54). Notice that the density of states is independent of energy. If there is more than one confined state in the quantum well system, the density of states at a given energy is the sum over all subbands below that particular energy. Hence Eq. (5.67) can be rewritten as

$$g^{\mathrm{2D}}(E) = \sum_{j=1}^{n} \frac{m^*}{\pi\hbar^2} Y(E - E_i) \tag{5.68}$$

where n is the total number of confined subbands below a particular energy and Y is a step function defined as

$$Y(E - E_i) = \begin{cases} 1 & \text{for } E > E_i \\ 0 & \text{for } E < E_i \end{cases} \tag{5.69}$$

The two-dimensional density of states is plotted in Fig. 5.8 for three confined subband energy levels. A typical system of two-dimensional structure is GaAs/AlGaAs multiple quantum wells, where three energy levels can be confined in a well of thickness 200 Å and a barrier height of ~0.3 eV. The barrier height is determined by the Al mole fraction of the AlGaAs layer.

For n-type GaAs/AlGaAs multiple quantum wells, the total number of electrons (n^{2D}) within a subband is given as

$$n^{\mathrm{2D}} = \int_{\text{Subband}} g^{\mathrm{2D}}(E) f_{\mathrm{FD}}(E)\, dE \tag{5.70}$$

where $f_{\mathrm{FD}}(E)$ is the Fermi-Dirac distribution function defined in Eq. (5.39). Substituting Eqs. (5.39) and (5.57) into Eq. (5.70) yields

$$n_j^{\mathrm{2D}} = \int_{E_j}^{\infty} \frac{m^*}{\pi\hbar^2} \frac{dE}{(e^{(E-E_F)/k_B T} + 1)} \tag{5.71}$$

To integrate this equation, we let $y = \exp[(E - E_F)/k_B T]$, which gives $dy = [y/(k_B T)]\, dE$ and $y_j = \exp[(E_j - E_F)/k_B T]$. Inserting this transformation into Eq. (5.71) we have

$$n_j^{\mathrm{2D}} = \frac{m^* k_B T}{\pi\hbar^2} \int_{y_j}^{\infty} \frac{dy}{(y+1)y} \tag{5.72}$$

The integral in this equation can be solved using integration by parts or by using the Mathematica software. The result of the integration gives

$$\int \frac{dy}{(y+1)y} = -\ln\left(1 + \frac{1}{y}\right) \tag{5.73}$$

Substituting Eq. (5.72) back into Eq. (5.71) and rearranging, the density of electrons in the jth subband is finally obtained as

$$n_j^{2\mathrm{D}} = \frac{m^* k_B T}{\pi\hbar^2}\left[-\ln\left(1+\frac{1}{y}\right)\right]\Bigg|_{y_j}^{\infty} = \frac{m^* k_B T}{\pi\hbar^2}\ln\left(1+\frac{1}{y_j}\right)$$

$$= \frac{m^* k_B T}{\pi\hbar^2}\ln\left(1+e^{(E_F-E_j)/k_B T}\right) \tag{5.74}$$

For n bound subbands in the quantum wells we have

$$n^{2\mathrm{D}} = \sum_{j=1}^{n} n_j^{2\mathrm{D}} = \frac{m^* k_B T}{\pi\hbar^2}\sum_{j=1}^{n}\ln\left(1+e^{(E_F-E_j)/k_B T}\right) \tag{5.75}$$

It appears from this equation that the Fermi energy E_F is explicitly independent of temperature when $(E_F - E_j) \gg k_B T$. However, for the limit $n^{2\mathrm{D}}\,\pi\hbar^2 \ll m^* k_B T$, the Fermi energy can be approximated as $E_F \approx E_j + k_B T\;\ln[n^{2\mathrm{D}}\pi\hbar^2/(m^* k_B T)]$. The Fermi energy in the latter limit is plotted as a function of the two-dimensional (2D) electron density as shown in Fig. 5.9. In this figure, the Fermi energy was plotted for different temperatures with respect to the first bound state (E_1) which was taken as 20 meV. The Fermi energy is below the bound state for population on the order of 5×10^{11} cm^{-2} and at $T > 100$ K. This can be understood by the fact that the Fermi energy in the preceding formalisms is merely a quasi-Fermi energy which describes the occupancy of the subband energy levels.

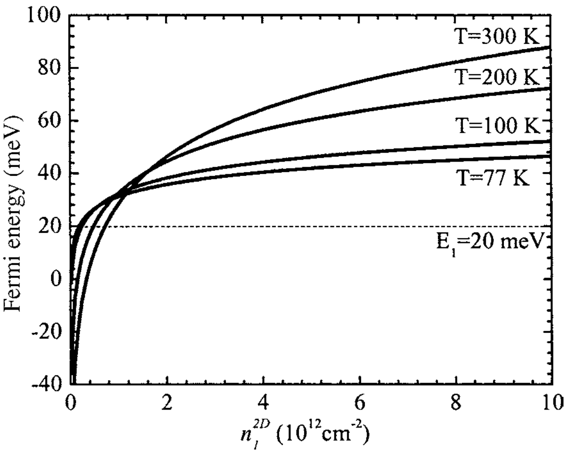


Figure 5.9 The Fermi energy plotted as a function of the 2D carrier concentration for the first subband energy level ($n_1^{2\mathrm{D}}$) at different temperatures. The first subband energy level (E_1) was taken as 20 meV.

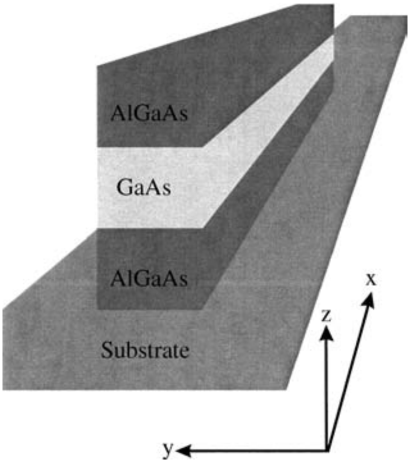


Figure 5.10 A schematic presentation of GaAs/AlGaAs quantum wire showing two directions of confinement (y and z directions) and one degree of freedom (x direction).

5.6.2 Quantum wires

The charge carrier confinement in semiconductors can be further decreased by reducing the number of degrees of freedom in the carrier momentum. This can be accomplished through photolithography or even self-assembled epitaxial growth of what is called quantum wires. A typical example of an n-type GaAs/AlGaAs quantum wire is sketched in Fig. 5.10 where the electrons in the GaAs layer are confined in both the growth direction (z direction) and the y direction, but they can move along the x direction. The z and y directions are called the directions of confinements, and the x direction is called the degree-of-freedom direction. The quantum wire usually refers to a one-dimensional (1D) system. Thus, for quantum wells, we have one direction of confinement and two degree-of-freedom directions. In contrast, the bulk materials have three degree-of-freedom directions and zero directions of confinement.

The density of states of the 1D system can be obtained by assuming that the electron momenta fill states along a line as shown in Fig. 5.11. The total number of states can be obtained by dividing the total length of the quantum wire ($2\mathbf{k}$) by the primitive unit cell and then dividing

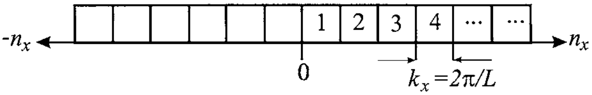


Figure 5.11 A quantum wire of length $2\mathbf{k}$ divided into one-dimensional primitive unit cells of length $k_x = 2\pi/L$.

by the length in the real space such as

$$n^{1\mathrm{D}} = 2\frac{2k}{(2\pi/L)}\frac{1}{L} = \frac{2\mathbf{k}}{\pi} \tag{5.76}$$

where $n^{1\mathrm{D}}$ is the total number of states, and the 2 is included to account for the electron spin degeneracy. The density of states can then be written as

$$g^{1\mathrm{D}}(E) = \frac{\partial n^{1\mathrm{D}}}{\partial E} = \frac{\partial n^{1\mathrm{D}}}{\partial \mathbf{k}}\frac{\partial \mathbf{k}}{\partial E} \tag{5.77}$$

where

$$\frac{\partial n^{1\mathrm{D}}}{\partial \mathbf{k}} = \frac{2}{\pi} \qquad \text{and} \qquad \frac{\partial \mathbf{k}}{\partial E} = \left(\frac{2m^*}{\hbar^2}\right)^{1/2}\frac{1}{2\sqrt{E}} \tag{5.78}$$

The second term of this equation is obtained from Eq. (5.54). Substituting Eq. (5.78) into (5.77) we have

$$g^{1\mathrm{D}}(E) = \left(\frac{2m^*}{\hbar^2}\right)^{1/2}\frac{1}{\pi\sqrt{E}} \tag{5.79}$$

Following the same discussion for the 2D system, the total density of states of the 1D system with an n number of confined energy levels is given as

$$g^{1\mathrm{D}}(E) = \left(\frac{2m^*}{\pi^2\hbar^2}\right)^{1/2}\sum_{j=1}^{n}\frac{1}{\sqrt{E-E_j}}Y(E-E_j) \tag{5.80}$$

where $Y(E - E_j)$ is a step function defined in Eq. (5.69). A plot of the density of states of a quantum wire is shown in Fig. 5.12 for four bound states. The inset is an illustration in the **k**-space for the two confinement directions (k_y and k_z directions) depicted as the two ellipses and the one degree-of-freedom direction depicted as the k_x line indicated by the arrow as the solid thick line. Notice that the units of the density of state are $\mathrm{J}^{-1}\cdot\mathrm{m}^{-1} = 1.602 \times 10^{-21}\ \mathrm{eV}^{-1}\cdot\mathrm{cm}^{-1}$.

The linear electron density in the quantum wire can be obtained in a fashion similar to that of the two-dimensional system where the population density for the jth subband can be expressed as

$$n_j^{1\mathrm{D}} = \int_0^\infty g^{1\mathrm{D}}(E)f_{\mathrm{FD}}(E)\,dE = \left(\frac{2m^*}{\pi^2\hbar^2}\right)^{1/2}\int_{E_j}^\infty \frac{dE}{\sqrt{E-E_j}(e^{(E-E_F)/k_BT}+1)} \tag{5.81}$$

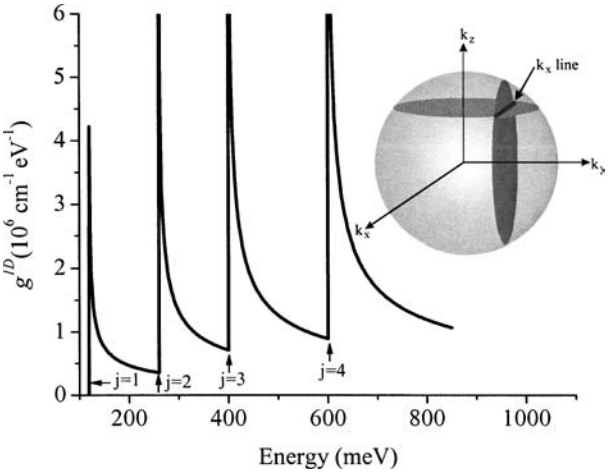


Figure 5.12 The density of states for a GaAs/AlGaAs quantum wire is plotted as a function of energy for four bound states labeled $j = 1, 2, 3$, and 4. The inset is the $\mathbf{k}$-space illustration of the two confinement directions (k_y and k_z) and the one degree-of-freedom direction (k_x).

The total linear density of states for a quantum wire with n bound states is thus

$$n^{1\mathrm{D}} = \sum_{j=1}^{n} n_j^{1\mathrm{D}} \tag{5.82}$$

The integral in Eq. (5.81) is difficult to solve but can be evaluated analytically in the limit of $(E - E_F)/k_BT \gg 1$ or in the limit of $(E - E_F)/k_BT \ll 1$ and numerically between these two limits. For example, when $(E - E_F)/k_BT \gg 1$, Eq. (5.81) becomes

$$\begin{aligned}
n_j^{1\mathrm{D}} &\approx \left(\frac{2m^*}{\pi^2\hbar^2}\right)^{1/2} \int_{E_j}^{\infty} \frac{dE}{\sqrt{E - E_j}\,(e^{(E-E_F)/k_BT})} \\
&\approx \left(\frac{2m^* k_BT}{\pi^2\hbar^2}\right)^{1/2} e^{(E_F - E_j)/k_BT} \int_0^{\infty} \frac{e^{-x}dx}{\sqrt{x}} \\
&\approx \left(\frac{2m^* k_BT}{\pi^2\hbar^2}\right)^{1/2} e^{(E_F - E_j)/k_BT}\sqrt{\pi} \\
&= \left(\frac{2m^* k_BT}{\pi\hbar^2}\right)^{1/2} e^{(E_F - E_j)/k_BT}
\end{aligned} \tag{5.83}$$

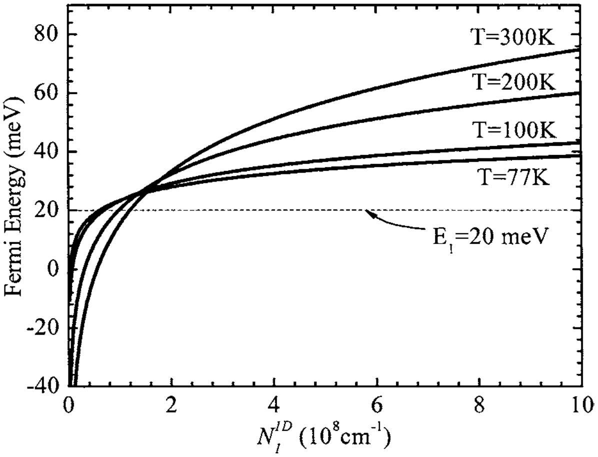


Figure 5.13 The Fermi energy plotted as a function of the 1D carrier concentration for the first subband energy level at different temperatures. The first subband energy level (E_1) was taken as 20 meV.

where we used the transformation $x = (E - E_j)/k_B T$. The general form of this equation is similar to that shown in Eq. (5.60). The Fermi energy is plotted with respect to the first bound state (E_1) as a function of the quantum wire carrier concentration for different temperatures as shown in Fig. 5.13. There is a similarity between the behavior of the Fermi energy in quantum wires and quantum wells as a function of the electron density as shown in Figs. 5.9 and 5.13.

5.6.3 Quantum dots

The quantum dot is characterized by having three confinement directions and zero degree-of-freedom directions as shown in the inset of Fig. 5.14 where we sketch the confinements in the **k**-space. The wavevector of the quantum dot is represented by the white dot where the three circles in the figures intercept as indicated by the arrow. Because of the lack of dispersion curves the wavevector selection rules are absent. The density of state is thus represented by the number of confined states divided by the energy interval. If the energy interval is approaching zero, then the density of states is simply a series of δ -functions centered on the confined energy levels ($E_1, E_2, E_3, \ldots$) as shown in Fig. 5. 14. The energy levels are entirely discrete and are given by

$$E_{n_x,n_y,n_z} = \frac{\pi^2\hbar^2}{2m^*}\left(\frac{n_x^2}{L_x^2} + \frac{n_y^2}{L_y^2} + \frac{n_z^2}{L_z^2}\right) \tag{5.84}$$

where L_x, L_y, and L_z are the dimensions of the quantum dot and n_x, n_y, and n_z are positive integers.

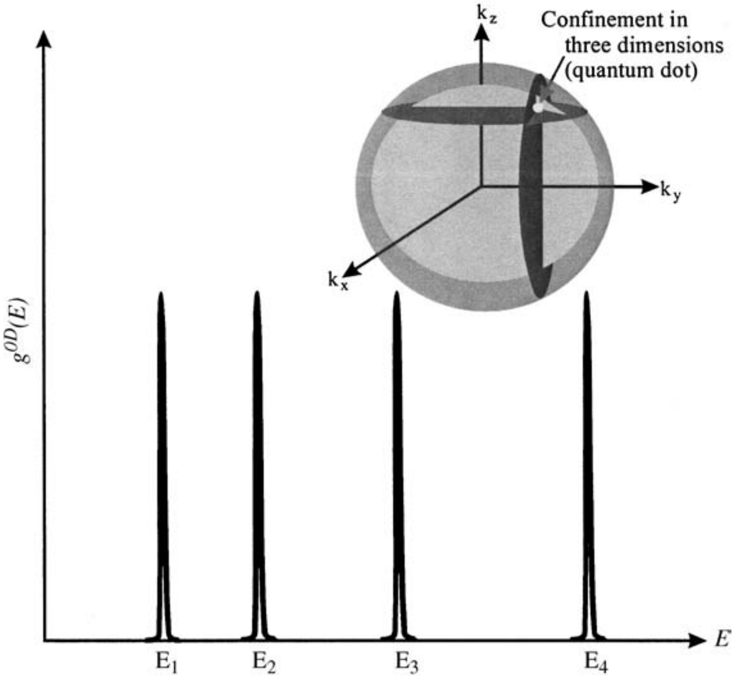


Figure 5.14 The density of states of a quantum dot is presented by a series of δ-functions centered on the confined energy levels. The inset is the $\mathbf{k}$-space presentation of the quantum dot showing confinement in three directions. The energy dispersion is absent and is represented by the white dot where the three circles intercept.

5.7 Density of States of Other Systems

This section focuses on deriving the density of states for systems that are occasionally encountered in semiconductor physics, in particular, superlattices and bulk materials under the influence of magnetic or electric fields. The density of states in quantum wells and wires under the influence of an external electric field will be briefly discussed.

5.7.1 Superlattices

Semiconductor superlattices were discussed in Chap. 3 where we approximated the confined energy levels by using minibands when the barriers were too thin. A typical example of superlattices is InAs/InGaSb type II superlattices. The quantized energy levels are given by Eq. (3.50). One approach to estimate the density of states in superlattices is to take the general form

$$g(E) = \sum_{j=1}^{n} \delta(E - E_j) \tag{5.85}$$

where the sum is over all the eigenvalues. This general form of the density of state is actually very convenient since the eigenvalues are quantized regardless of how small the separation is between them. The eigenvalues of the electrons in the superlattices can be expressed as (see Bastard 1988)

$$E(n, \mathbf{q}, \mathbf{k}_\perp) = \frac{\hbar^2 \mathbf{k}_\perp^2}{2m^*} + E_n(\mathbf{q}) \tag{5.86}$$

The density of state is thus

$$g(E) = 2 \sum_{n,\mathbf{q},\mathbf{k}_\perp} \delta \left[E - E_n(\mathbf{q}) - \frac{\hbar^2 \mathbf{k}_\perp^2}{2m^*} \right] \tag{5.87}$$

and the factor 2 is included for spin degeneracy. Converting the sum into an integral and using the δ-function definition, the density of state in the **k**-space becomes

$$\begin{aligned} g^s(E) &= 2 \frac{1}{(2\pi/L)^2} \frac{Nd}{L^2} \frac{m^*}{\hbar^2} 2 \int_0^{\pi/d} Y\left[E - E_n(\mathbf{q})\right] d\mathbf{q} \\ &= \frac{Nd}{\pi^2} \frac{m^*}{\hbar^2} \int_0^{\pi/d} Y\left[E - E_n(\mathbf{q})\right] d\mathbf{q} = \sum_n g_n^s(E) \end{aligned} \tag{5.88}$$

where $Y[E - E_n(\mathbf{q})]$ is a step function, Nd is the length of the superlattice, the superscript s was introduced to indicate that the density of states is for the superlattice, $g_n^s(E)$ is the density of states associated with the nth miniband, and the integral limits are the first Brillouin zone boundary. Since the miniband has a width such that $E_{\min} < E_n(\mathbf{q}) < E_{\max}$, $g_n^s(E) = 0$ for $E_n(\mathbf{q}) < E_{\min}$ and $g_n^s(E) = N\left[m^*/(\pi\hbar^2)\right]$ for $E_n(\mathbf{q}) > E_{\max}$. The jump from one miniband to the next is not abrupt as it is in the case for quantum wells or wires. This is due to the fact that $E_n(\mathbf{q}) = E_n + S_n + 2T_n \cos(\mathbf{q}d)$ is a function of the wavevector ($\mathbf{q}$) as described in Eq. (3.53). The final results of the density of states according to Bastard (1988) can be written as

$$g^s(E) = \begin{cases} 0 & \text{for } E < (E_n + S_n - 2|T_n|) \\ \dfrac{Nm^*}{\pi\hbar^2} \arccos\left(\dfrac{-E + E_n + S_n}{2\,|T_n|}\right) & \text{for } |E - E_n - S_n| < 2|T_n| \\ \dfrac{Nm^*}{\pi\hbar^2} & \text{for } E > (E_n + S_n + 2|T_n|) \end{cases} \tag{5.89}$$

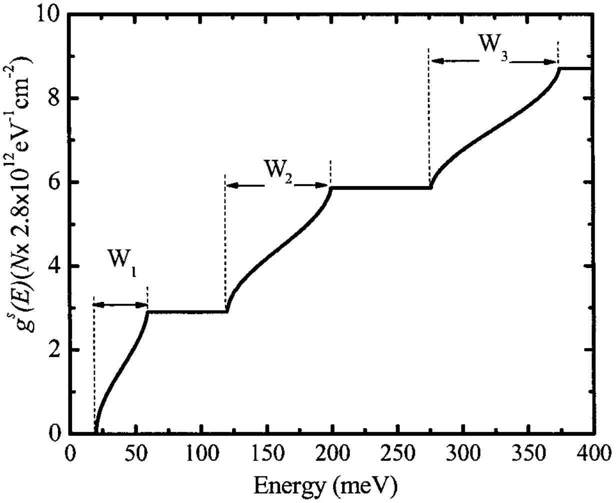


Figure 5.15 The density of states of a superlattice structure plotted as a function of energy. The width of the bands are indicated by W_i, where $i = 1, 2, 3$.

A plot of Eq. (5.89) is shown for three minibands in Fig. 5.15. Notice that the widths of the minibands increase as the subband quantum number n increases. The density of state is also multiplied by the number of superlattice periods, N.

5.7.2 Density of states of bulk electrons in the presence of a magnetic field

Bulk electrons here are assumed to be electrons in the conduction band of bulk semiconductor materials so that they have three degree-of-freedom directions. The allowed eigenvalues are quasi-continuum. In the presence of a magnetic field, each energy level splits into what is called Landau energy levels. The separation between the Landau energy levels is directly proportional to the strength of the magnetic field. The eigenvalues of an electron in a magnetic field parallel to the growth axis are given by

$$E_{n,k_z,\sigma_z} = \left(n + \frac{1}{2}\right)\hbar\omega_c + \frac{\hbar^2 k_z^2}{2m^*} + \sigma_z g^* \mu_B B \tag{5.90}$$

where n = Landau quantum number
ω_c = cyclotron frequency = $(eB)/(m^* c)$
μ_B = Bohr magneton
g^* = effective Lande g-factor
σ_z = spin eigenvalues $\pm\frac{1}{2}$

Using the general definition of the density of state [Eq. (5.85)], we have

$$g^B(E) = \sum_{n,k_z,\sigma_z} \delta(E - E_{n,k_z,\sigma_z}) \tag{5.91}$$

where the superscript B is introduced to indicate the presence of a magnetic field. Substituting Eq. (5.90) into (5.91), we obtain

$$\begin{aligned} g^B(E) &= \sum_{n,k_z,\sigma_z} \delta\left(E - \left(n + \frac{1}{2}\right)\hbar\omega_c - \frac{\hbar^2 k_z^2}{2m^*} - \sigma_z g^* \mu_B B\right) \\ &= \sum_{n,k_z,\sigma_z} \delta\left(x - \frac{\hbar^2 k_z^2}{2m^*}\right) = \sum_{n,k_z,\sigma_z} g_n^B(E) \end{aligned} \tag{5.92}$$

where

$$x = E - \left(n + \frac{1}{2}\right)\hbar\omega_c - \sigma_z g^* \mu_B B \tag{5.93}$$

The degeneracy of any Landau level in one-dimensional k_z-space becomes

$$g_n^B(E) = \frac{L_x L_y}{(2\pi)^2 l^2} \int_{-\infty}^{\infty} \delta\left(x - \frac{\hbar^2 k_z^2}{2m^*}\right) dk_z \tag{5.94}$$

where l is the magnetic length given by $(\hbar c/eB)^{1/2}$, and the quantity $L_x L_y/[(2\pi)^2 l^2]$ is included due to the degeneracy of any level in the $k_x k_y$ plane. Let $y = \hbar^2 k_z^2/(2m^*)$, which leads to $dy = (\hbar^2 k_z/m^*) dk_z$. Using this transformation, the density of states is further simplified such that

$$\begin{aligned} g_n^B(E) &= 2\frac{L_x L_y}{(2\pi)^2 l^2 \hbar} \int_0^{\infty} \sqrt{\frac{m}{2y}} \delta(x - y)\, dy \\ &= 2\frac{L_x L_y}{(2\pi)^2 l^2 \hbar} \sqrt{\frac{m}{2x}} = \frac{L_x L_y}{(2\pi)^2 l^2} \sqrt{\frac{2m}{\hbar^2}} \\ &\quad \times \left[E - \left(n + \frac{1}{2}\right)\hbar\omega_c - \sigma_z g^* \mu_B B\right]^{-1/2} \\ &= \frac{L_x L_y}{8\pi^2} \left(\frac{2m}{\hbar^2}\right)^{3/2} \hbar\omega_c \left[E - \left(n + \frac{1}{2}\right)\hbar\omega_c - \sigma_z g^* \mu_B B\right]^{-1/2} \end{aligned} \tag{5.95}$$

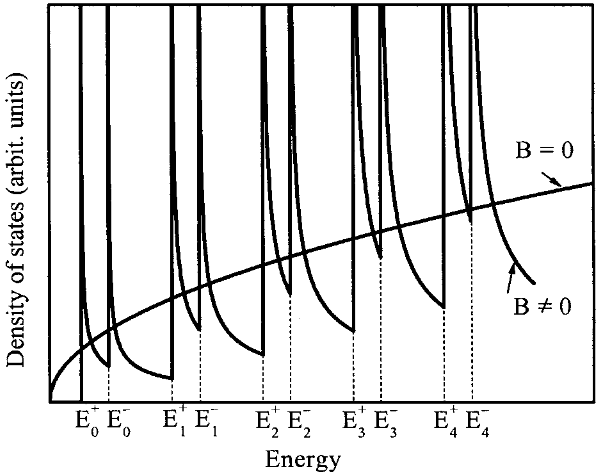


Figure 5.16 The density of states of electrons in the conduction band of a bulk semiconductor material plotted as a function of energy for both a zero magnetic field and in the presence of a magnetic field.

By including the spin degeneracy, the total density of states can be rewritten as

$$g^B(E) = \frac{1}{4\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2}\hbar\omega_c\sum_n\left[E-\left(n+\frac{1}{2}\right)\hbar\omega_c-\sigma_z g^*\mu_B B\right]^{-1/2} \tag{5.96}$$

Notice that we dropped L_xL_y from the last expression to obtain the density of states per unit area. A plot of the density of states described in Eq. (5.96) is shown in Fig. 5.16 for both $B = 0$ and $B \neq 0$. The density of states for $B \neq 0$ is zero only for $E < E_0^+$. The energies labeled E_n^+ correspond to $\sigma_z = +\frac{1}{2}$ and E_n^- correspond to $\sigma_z = -\frac{1}{2}$. This analysis indicates that only k_z is a good quantum number. The magnetic field produces energy quantization in the xy plane. One may imagine this situation by assuming that the electrons are trapped in the Landau circular orbits in the xy plane generated by the magnetic field, but the electrons can move along the z axis in a helical form. This form of motion is analogous to the confinement of electrons in a quantum wire as it is clear from the density of states in both cases that they have the same energy dependence as shown in Eqs. (5.80) and (5.96).

For two-dimensional systems such as multiple quantum wells, the confinement occurs along the growth axis (z direction), which is not a good quantum number. By applying a magnetic field parallel to the growth direction, the x and y directions are no longer good quantum numbers. This implies that the electrons are confined in the three directions without any degree-of-freedom directions. In this case, the density of states is similar to that of the quantum dots. In other words, the density of states is a series of δ-functions as shown in Fig. 5.14.

5.7.3 Density of states in the presence of an electric field

The density of states under the influence of an electric field is very complicated, and yet it is very important to understand the behavior of the density of states in devices that operate under applied bias voltage. In this section, we will follow the analyses of Davies (1998) and Davies and Wilkins (1988). When a semiconductor is experiencing an applied bias voltage, the conduction and valence bands bend or vary in a way that the properties of the device have to be solved using self-consistent calculations. An example of how the semiconductor bands are changed under bias voltage is shown in Fig. 5.17 for a type I quantum well. It is evident that the interfaces are modified into triangular shapes similar to the simple heterostructures discussed in Sec. 2.6 .

If one assumes a constant electric field is applied to a heterojunction, the electrostatic potential energy ($e\phi$) is

$$e\phi = e\mathcal{E}z \tag{5.97}$$

where e = charge of electron
$\mathcal{E}$ = electric field
z = distance from interface

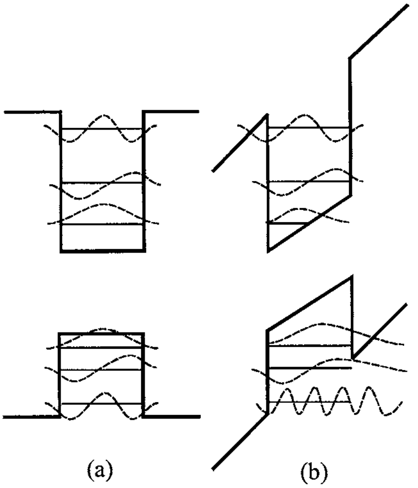


Figure 5.17 (*a*) A quantum well in the absence of an electric field. (*b*) The modified band structure in the presence of an electric field. As can be seen, the application of an electric field to a quantum well modifies the energy levels and wave functions.

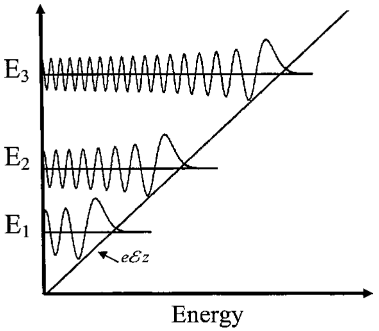


Figure 5.18 A triangular quantum well formed by applying a uniform electric field. Three energy levels are shown along with their wave functions. The wave functions have the form of Airy functions that satisfy the boundary conditions.

The stationary Schrödinger equation can be written as

$$-\frac{\hbar^2}{2m^*}\frac{d^2\psi(z)}{dz^2}+e\mathcal{E}z\psi(z)=E_n\psi(z) \tag{5.98}$$

The solution of this equation is expressed in terms of the Airy function

$$\psi(z)=\mathrm{Ai}\left(\frac{e\mathcal{E}z-E_n}{\mathcal{E}_o}\right) \tag{5.99}$$

where

$$\mathcal{E}_o=\left[\frac{(e\mathcal{E}\hbar)^2}{2m^*}\right]^{1/3}=e\mathcal{E}z_o \quad\text{and}\quad z_o=\left(\frac{\hbar^2}{2m^*e\mathcal{E}}\right)^{1/3} \tag{5.100}$$

The Airy function is plotted for three energy levels in a triangular quantum well as shown in Fig. 5.18. For any particular energy level, the wave function in Fig. 5.18 exhibits propagation behavior for $E < e\mathcal{E}z$ and tunneling behavior for $E > e\mathcal{E}z$. This feature has a very interesting effect on the local density of states at fixed values of z. To demonstrate this effect, consider the general definition of the density of states:

$$g(E,z)=\sum_{\mathbf{k}}|\psi_{\mathbf{k}}(z)|^2\delta(E-E_{\mathbf{k}}) \tag{5.101}$$

where the sum is over all eigenstates, labeled by **k**. The formalism of obtaining the density of states for quantum wells, quantum wires, and bulk materials under the influence of an electric field was reported by Davies (1998) and the final results are

$$g_{1\mathrm{D}}^{\mathcal{E}}(E,z)=\frac{2}{\hbar}\sqrt{\frac{2m^*}{\mathcal{E}_o}}\mathrm{Ai}^2\left(-\frac{E-e\mathcal{E}z}{\mathcal{E}_o}\right) \tag{5.102}$$

$$g_{2\mathrm{D}}^{\mathcal{E}}(E,z)=\frac{m^*}{2\pi}\mathrm{Ai}I(2^{2/3}S) \tag{5.103}$$

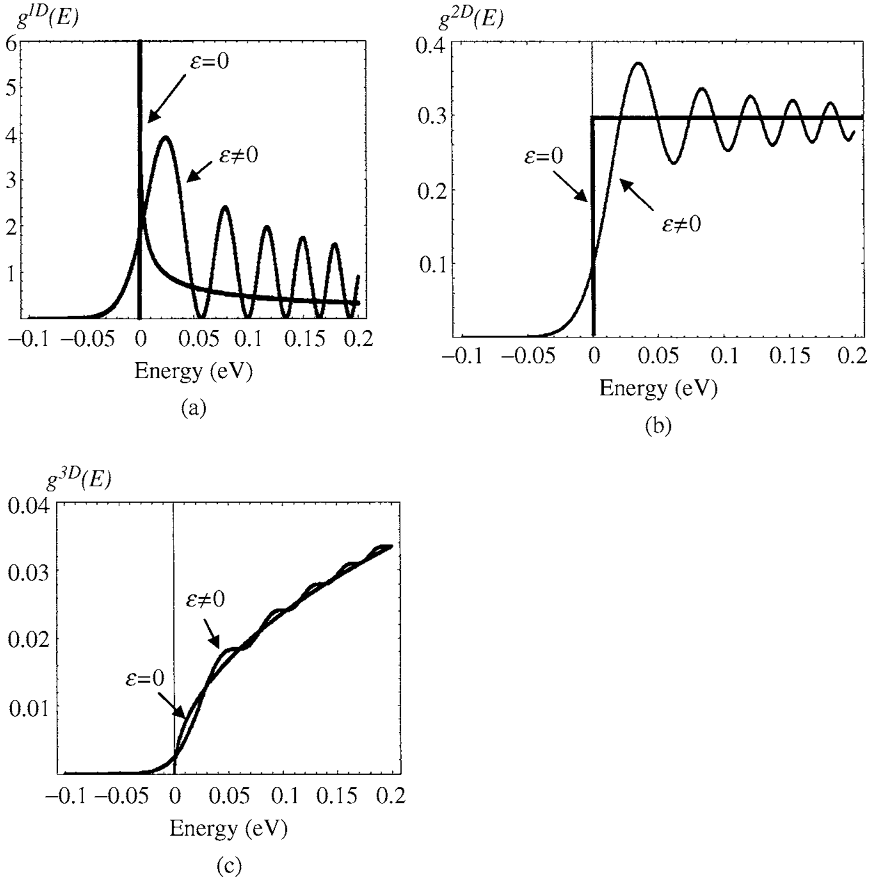


Figure 5.19 The density of states under the influence of a uniform electric field is plotted as a function of energy for (*a*) quantum wires, (*b*) quantum wells, and (*c*) bulk materials. (*After Davies 1998*).

where

$$S = \frac{E - e\mathcal{E}z}{\mathcal{E}_o} \tag{5.104}$$

and

$$\mathrm{Ai}I\,(x) = \int_x^\infty \mathrm{Ai}(y)\,dy \tag{5.105}$$

Finally, the density of states for a 3D system is

$$g_{3\mathrm{D}}^{\varepsilon}(E, z) = \frac{m^*}{\pi\hbar^3}\sqrt{2m^*\mathcal{E}_o}\{[\mathrm{Ai}'(S)]^2 - S\,[\mathrm{Ai}(S)]^2\} \tag{5.106}$$

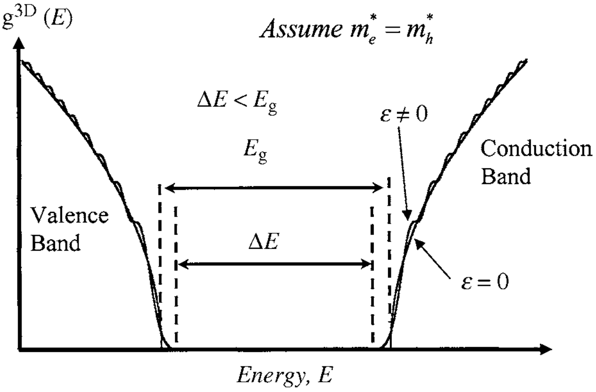


Figure 5.20 The density of states is plotted as a function of energy for both the conduction and valence bands with (rippled curves) and without (smooth curves) an applied electric field. Notice that the densities of states have leaked into the fundamental bandgap causing an apparent decrease in E_g. For simplicity, we assume the electrons and holes have the same effective mass.

where S is defined in Eq. (5.104). The superscript $\mathcal{E}$ is added to indicate the density of states under the influence of an applied electric field. Equations (5.102), (5.103), and (5.106) are plotted as a function of energy, as shown in Fig. 5.19. For simplicity, it is assumed in this figure that the ground-state energy values for the quantum wire and quantum well are zero and the conduction band minimum is also set to zero. It can be seen from Fig. 5.19 that the density of states tunneled below the energy levels in the three cases. An interesting feature in the bulk material is that the density of states has a tail that extends below the bottom of the conduction band minimum as shown in Fig. 5.19c. A similar result is obtained for the hole density of states as shown in Fig. 5.20 where we assumed that the effective mass of the electrons and holes are the same. The tunneling of states in the fundamental bandgap E_g, when a uniform electric field is applied, leads to the Franz-Keldysh effect. Photons with an energy of $\Delta E < E_g$ can be absorbed. The oscillations observed in the density of states when an electric field is applied are difficult to observe by using the optical absorption technique, due to the fact that most photons with energies above the fundamental bandgap energy are reflected or absorbed at the edge of the conduction band. However, these oscillations can be observed using the photoreflectance technique. The absorption tail due to the tunneling of states into the bandgap can be expressed as (see, for example, Mitin et al. 1999 and Fox 2001)

$$\alpha(E) \propto \exp\left[-\left(\frac{E_g - \hbar\omega}{\hbar\omega_F}\right)^{3/2}\right] \quad \text{for } \hbar\omega < E_g \qquad (5.107)$$

where

$$\omega_F = \frac{e^2}{2\hbar}\left(\frac{m_e^* + m_h^*}{m_e^* m_h^*}\right)\mathcal{E}^{2/3} \qquad (5.108)$$

where m_e^* and m_h^* are the electron and hole effective masses, respectively, and $\mathcal{E}$ is the constant applied electric field. The Franz-Keldysh effect has no significant applications in bulk semiconductors, but it can wash out the desired excitonic effect.

Summary

The distribution functions and density of states play a major role in the transport, electrical, and optical properties of semiconductor materials and devices. Thus, knowledge of these important parameters is necessary before proceeding. In this chapter we presented derivations for the Maxwell-Boltzmann, Fermi-Dirac, and Bose-Einstein distribution functions. The Fermi-Dirac distribution function is widely used in both bulk and low-dimensional semiconductor materials, since it describes the distribution of particles with one-half spin, such as electrons and holes. It should be pointed out that these distribution functions were derived for systems at equilibrium. For nonequilibrium cases, different analyses are applied. These analyses are presented in Chap. 7.

A fair amount of discussion in this chapter was devoted to the density of states in various systems. The density of states was derived for bulk semiconductors and then compared to the density of states in low-dimensional systems, such as quantum wells, wires, and dots. The density of states was also derived for semiconductor superlattices and bulk materials under the influence of a magnetic field. Electron motion in the presence of a magnetic field is confined to a two-dimensional plane. This condition is similar to confinement of electrons in quantum wires.

The density of states in bulk semiconductors, quantum wells, and quantum wires exhibits oscillatory behavior under the influence of an electric field. Additionally, the density of states leaks into the fundamental bandgap in the case of bulk materials causing an apparent decrease in the bandgap. It also exhibits a tail below the bound energy levels in the case of quantum wells and wires.

The distribution functions and density of states were used to obtain the Fermi energy level in bulk semiconductor, quantum well, and quantum wire systems. The expressions for the Fermi energy levels are always easy to handle. These expressions can yield an approximate behavior of the Fermi energy levels in certain regimes, such as high-or low-temperature regimes. A plot of the Fermi energy level as a function of temperature or as a function of carrier concentrations was shown for these systems.

Problems

5.1 The Γ-function is very useful in solving many statistical problems. Show that $\Gamma(n) = (n-1)\Gamma(n-1)$ and $\Gamma(n) = (n-1)!$.

5.2 The gaussian distribution $P(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e(x-\overline{x})^2/(2\sigma^2)$, is used occasionally to describe certain properties in semiconductors. For example, the diffusion of carriers can be described by a gaussian function. Show that $\int_{-\infty}^{\infty} P(x)dx = 1$. Plot $P(x)$ as a function of x for at least three different values of σ. What type of distribution do you obtain when $\sigma \to 0$?

5.3 Plot the Fermi energy as a function of carrier concentration for different temperatures (see Fig. 5.4) using Eq. (5.43) for a cubic GaAs sample of an edge of 10^{-6} cm.

5.4 Show that the hole concentration in an intrinsic semiconductor is given by Eq. (5.62).

5.5 Calculate the density of states for the following: (*a*) bulk GaAs, (*b*) the lowest state of the GaAs/AlGaAs quantum well, and (*c*) the lowest band GaAs/AlGaAs quantum wire. Express your answer in terms of energy, centimeters, and electronvolts.

5.6 Consider Fig. P5.6 where we plotted the energy levels in the GaAs/AlGaAs quantum well. The Fermi energy is shown to be above the bound state E_1.

(a) Calculate the Fermi energy position for a 2D electron density of $4 \times 10^{12}\text{cm}^{-2}$ at $T = 300$, 77, and 4.2 K.

(b) Calculate the Fermi energy levels at $T = 300$ and 77 K for the following 2D electron densities: $3 \times 10^{11}\ \text{cm}^{-2}$, $1 \times 10^{12}\ \text{cm}^{-2}$, and $5 \times 10^{13}\ \text{cm}^{-2}$.

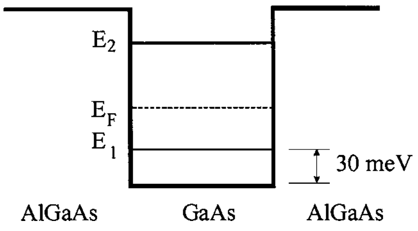


Figure P5.6

5.7 Use the general definition of the density of states as described by the summation of δ-functions [Eq. (5.85)] to derive the density of states for bulk semiconductors (3D system), quantum wells (2D system), and quantum wires (1D system).

5.8 For the Bose-Einstein distribution function (5.45), assume that the total number of the particles, N, with spin zero and mass m in a two-dimensional system is constant. Derive an expression for the parameter α.

5.9 The electron density in bulk GaAs can be written as

$$n_o = \frac{1}{2\pi^2}\left(\frac{2m_e^* k_B T}{\hbar^2}\right)^{3/2} \int_{E_c}^{\infty} \frac{\sqrt{E - E_c}\, dE}{(1 + e^{(E-E_F)/k_B T})} = N_c F_{1/2}(\eta),$$

where $F_{1/2}(\eta)$ is the Fermi integral and for $\eta > 1.25$ it is approximated as $F_{1/2}(\eta) = (4/3)\pi^{3/2}\ \eta^{3/2} + (\pi^{3/2}/6)\ \eta^{1/2}$, where $\eta = (E_F - E_c)/k_B T$. Plot the Fermi energy as a function of electron density for $T = 300$, 200, 100, 77, and 4.2 K.

5.10 The electron thermal energy in the conduction band of GaAs can be expressed as $k_B T$. Plot the magnetic field required to split the energy levels into Landau levels as a function of temperature. From the graph, find the magnetic field required to generate Landau levels at 4.2, 77, and 300 K.

5.11 Consider a quantum dot to be a cubic quantum box with a finite potential V_o outside the well. For bound states in the quantum well, the energy $E \leq 0$. Assume that the density of states is 3D-like inside the well. Calculate the number of states inside the quantum dot for $V_o = 0.6$ eV and for $L_x = L_y = L_z = 150$ Å.

5.12 Plot Eq. (5.107) for several values of the electric field. When do you start to see an effect on the band-edge absorption?

Chapter 6

Optical Properties

The optical properties of any material are the result of photon interactions with the constituents of the material. The aim of this chapter is to describe photon interactions with semiconductor materials, including low-dimensional systems, that lead to effects that are the basis for many technologies, such as detectors, emitters, optical communications, display panels, and optical oscillators. The interaction of photons with electrons in semiconductor materials is most important and gives rise to many phenomena. Electrons in semiconductor materials can absorb photons and be excited from the valence band to the conduction band. This is called the *interband* transition. The inverse of this process occurs when electrons decay from a higher energy level, such as a conduction band, to a lower energy level, such as a valence band, and photons are emitted. This is the basis for light-emitting diodes (LEDs) and laser diodes. Electrons can absorb photons and be excited from one state to another within a particular band, such as a conduction band. This transition is called an *intraband* transition. In low-dimensional systems, such as quantum wells, wires, and dots, electrons can be excited by photons and jump from one confined energy level to another. When the electrons are excited from a bound state to another bound state in the conduction band of a quantum well, for example, the transition is called an *intersubband* transition. These terminologies are also applied to heavy or light holes in semiconductors. These transitions are illustrated for a bulk material in Fig. 6.1*a* and for a quantum structure in Fig. 6.1*b*.

The band-to-band transition in a bulk material is usually referred to as the optical bandgap. In the case of a quantum structure, the conventional optical bandgap is no longer allowed, and the effective bandgap is referred to as the transition from the ground state in the valence band to the ground state in the conduction band, as illustrated in Fig. 6.1*b*.

Copyright © 2005 by The McGraw-Hill Companies, Inc. Click here for terms of use.

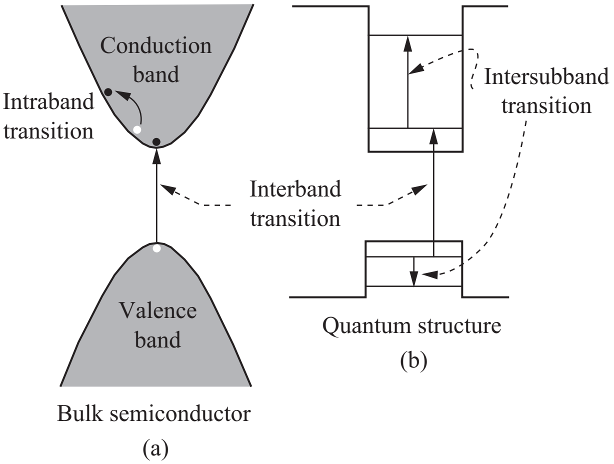


Figure 6.1 Illustrations of various electronic transitions are shown for (*a*) a bulk semiconductor material and (*b*) a quantum structure.

Thus, the effective bandgap in quantum structures is larger than the conventional optical bandgap in bulk materials. If an electron is excited from the valence band to the conduction band of a semiconductor, it leaves behind a positively charged hole. This process is called electron-hole pair generation. When the electron and hole interact with each other due to Coulomb interaction, the result is called an *exciton*. The excitonic energy levels are usually formed in the fundamental bandgap, as shown in Fig. 6.2*a*. The exciton may move about the crystal. In this case, the electron-hole pair is called a free exciton or a Wannier-Mott exciton. If the exciton is trapped by an impurity or an atom in the crystal, it is called a bound exciton or a Frenkel exciton, as shown in Fig. 6.2*b*. The binding energy of a free exciton is usually smaller than that of a bound exciton.

Many experimental techniques are used to probe electronic transitions in semiconductors. Essentially, the electron-photon interaction is the most dominant process in optoelectronic devices based on semiconductors and their nanostructures. In this chapter, we discuss various aspects of the optical properties of bulk semiconductors and low-dimensional systems.

6.1 Fundamentals

The interaction of photons with any material can be understood from Maxwell's classical electromagnetic theory. In MKS units, the four

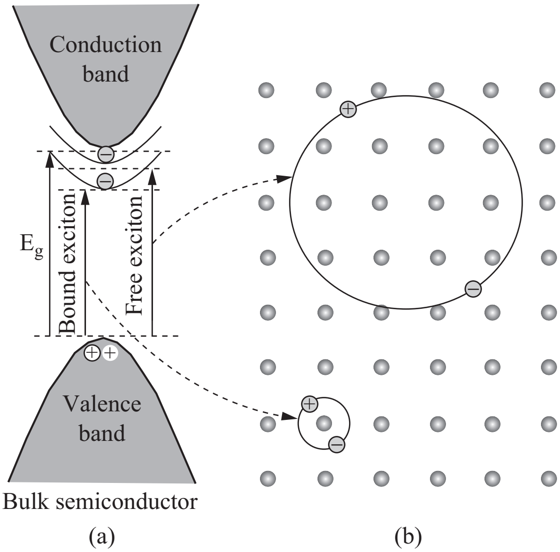


Figure 6.2 (*a*) A schematic presentation of excitons' energy levels with respect to the conduction band. (*b*) The bound electron-hole pairs for both free and bound excitons.

Maxwell equations that govern electromagnetic phenomena are

$$\nabla \cdot \boldsymbol{\mathcal{E}} = \frac{\rho}{\epsilon_o} \tag{6.1a}$$

$$\nabla \times \boldsymbol{\mathcal{E}} = -\mu_o \frac{\partial \mathbf{B}}{\partial t} \tag{6.1b}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{6.1c}$$

$$\nabla \times \mathbf{B} - \epsilon_o \frac{\partial \boldsymbol{\mathcal{E}}}{\partial t} = \mathbf{J} \tag{6.1d}$$

where $\boldsymbol{\mathcal{E}}$ = electric field
$\mathbf{B}$ = magnetic field
ρ = electric charge density
$\mathbf{J}$ = electric current density
ϵ_o = permittivity of free-space (8.854×10^{-12} F/m)
μ_o = permeability of free-space [$4\pi \times 10^{-7}$ W/(m · A)]

For the interaction of electromagnetic waves with electrically polarized material, we have

$$\mathbf{D} = \epsilon_o \boldsymbol{\mathcal{E}} + \mathbf{P} \tag{6.2}$$

where **D** is the electric displacement vector and **P** is the polarization vector. In the linear limit, the polarization vector can be written as

$$\mathbf{P} = \epsilon_o \overleftrightarrow{\chi} \cdot \boldsymbol{\mathcal{E}} \tag{6.3}$$

where $\overleftrightarrow{\chi}$ is the dielectric susceptibility tensor. Combining Eqs. (6.2) and (6.3) we have

$$\mathbf{D} = \epsilon_o (1 + \overleftrightarrow{\chi} \cdot \boldsymbol{\mathcal{E}} = \epsilon_o \overleftrightarrow{\epsilon} \cdot \boldsymbol{\mathcal{E}} \tag{6.4}$$

where $\overleftrightarrow{\epsilon} = 1 + \overleftrightarrow{\chi}$ is the dielectric tensor. These tensors can be written in terms of the scalar quantities ϵ and χ, such that $\overleftrightarrow{I}\epsilon = 1 + \overleftrightarrow{I}\chi$, where $\overleftrightarrow{I}$ is a unit tensor.

For a conductive medium, the current density is related to the electric field according to the following relation:

$$\mathbf{J}_T = \sigma \boldsymbol{\mathcal{E}} \tag{6.5}$$

where σ is the electrical conductivity, which may be a complex quantity, and $\mathbf{J}_T$ is the total current density composed of both the steady-state and the time-dependent current densities. For the optical properties of semiconductors, we are concerned with the time-dependent contribution to the current density. Hence, the steady-state contribution can be ignored. The time-dependent current density can now be written as

$$\mathbf{J}_T = \mathbf{J} = \frac{\partial \mathbf{P}}{\partial t} \tag{6.6}$$

Substituting Eq. (6.6) into (6.1*d*) and using Eq. (6.2) we have

$$\nabla \times \mathbf{B} = \frac{\partial \mathbf{D}}{\partial t} \tag{6.7}$$

By substituting Eq. (6.2) into (6.1*a*), we have

$$\begin{aligned} \nabla \cdot \boldsymbol{\mathcal{E}} &= \frac{\nabla \cdot (\mathbf{D} - \mathbf{P})}{\epsilon_o} = \frac{\rho}{\epsilon_o} \\ \therefore \nabla \cdot \mathbf{D} &= \nabla \cdot \mathbf{P} + \rho \end{aligned} \tag{6.8}$$

Substituting the continuity equation $\partial\rho/\partial t + \nabla \cdot \mathbf{J} = 0$ into Eq. (6.8), we obtain

$$\nabla \cdot \mathbf{D} = 0 \tag{6.9}$$

The wave equation for nonmagnetic materials can be derived by taking the curl of Eq. (6.1*b*):

$$\nabla \times \nabla \times \boldsymbol{\mathcal{E}} = -\mu_o \frac{\partial}{\partial t} \nabla \times \mathbf{B} \tag{6.10}$$

Substituting Eqs. (6.4) and (6.7) into Eq. (6.10), we obtain

$$\nabla \times \nabla \times \boldsymbol{\mathcal{E}} = -\mu_o \epsilon_o \epsilon \frac{\partial^2 \boldsymbol{\mathcal{E}}}{\partial t^2} = -\frac{\epsilon}{c^2} \frac{\partial^2 \boldsymbol{\mathcal{E}}}{\partial t^2} \tag{6.11}$$

where $c = 1/\sqrt{\mu_o \epsilon_o}$ is the speed of light. Recall that $\nabla \times \nabla \times \boldsymbol{\mathcal{E}} = \nabla(\nabla \cdot \boldsymbol{\mathcal{E}}) - \nabla^2 \boldsymbol{\mathcal{E}}$ and $\nabla \cdot \boldsymbol{\mathcal{E}} = 0$; hence Eq. (6.11) becomes

$$\nabla^2 \boldsymbol{\mathcal{E}} = \frac{\epsilon}{c^2} \frac{\partial^2 \boldsymbol{\mathcal{E}}}{\partial t^2} \tag{6.12}$$

A solution of this wave equation is a plane wave with the following form:

$$\boldsymbol{\mathcal{E}}(\mathbf{r}, t) = \boldsymbol{\mathcal{E}}_o \exp[i(\mathbf{k} \cdot \mathbf{r} - \omega t)] \tag{6.13}$$

where $\boldsymbol{\mathcal{E}}_o$ = amplitude of electric field
$\mathbf{k}$ = propagation vector
ω = angular frequency

Substituting this solution into the wave Eq. (6.12), one can obtain the following dispersion relation:

$$c^2 k^2 = \omega^2 \epsilon \tag{6.14}$$

The dielectric constant ϵ is frequency-dependent, and its explicit form is required to evaluate the dispersion relation. Substituting Eq. (6.14) back into (6.13) yields

$$\boldsymbol{\mathcal{E}}(\mathbf{r}, t) = \boldsymbol{\mathcal{E}}_o \exp\left[i\omega\left(\frac{\sqrt{\epsilon}}{c}\hat{\mathbf{k}} \cdot \mathbf{r} - t\right)\right] \tag{6.15}$$

The dielectric constant is related to the refractive index $n_r(\omega)$ according to the following relation:

$$n_r(\omega) = \sqrt{\epsilon(\omega)} \tag{6.16}$$

Both the refractive index and the dielectric constant are complex numbers and can be written as

$$n_r(\omega) = n_1(\omega) + i n_2(\omega) \tag{6.17a}$$

$$\epsilon(\omega) = \epsilon_1(\omega) + i\epsilon_2(\omega) \tag{6.17b}$$

where n_1 and ϵ_1 are the real parts and n_2 and ϵ_2 are the imaginary parts. Substitute Eqs. (6.16) and (6.17*a*) into Eq. (6.15) to obtain

$$\boldsymbol{\mathcal{E}}(\mathbf{r}, t) = \boldsymbol{\mathcal{E}}_o \exp\left[-\frac{\omega n_2(\omega)}{c}\hat{\mathbf{k}} \cdot \mathbf{r}\right] \exp\left[i\omega\left(\frac{n_1}{c}\hat{\mathbf{k}} \cdot \mathbf{r} - t\right)\right] \tag{6.18}$$

The intensity I of the electromagnetic wave is related to the electric field according to the following relation:

$$\begin{aligned} I &\propto |\mathcal{E}(\mathbf{r},t)|^2 \propto |\mathcal{E}_o|^2 \exp\left[-\frac{2\omega n_2(\omega)}{c}\widehat{\mathbf{k}}\cdot\mathbf{r}\right] \\ &\propto |\mathcal{E}_o|^2 \exp\left[-\alpha(\omega)\widehat{\mathbf{k}}\cdot\mathbf{r}\right] \end{aligned} \tag{6.19}$$

where $\alpha(\omega)$ is the optical absorption coefficient and is defined according to Eq. (6.19) as

$$\alpha(\omega) = \frac{2\omega n_2(\omega)}{c} = \frac{\omega\epsilon_2(\omega)}{cn_1(\omega)} \tag{6.20}$$

The optical absorption coefficient can also be obtained by using Beer's law:

$$I(z) = I_o \exp[-\alpha(\omega)z] \tag{6.21}$$

where $I(z)$ is the electromagnetic radiation intensity at a distant z inside the media and I_o is the intensity at $z = 0$.

6.2 Lorentz and Drude Models

The classical Lorentz model is applicable to solids with bandgaps. This model is analogous to the quantum-mechanically treated interband transitions. The Lorentz model assumes that the electron is bound to the nucleus like a mass attached to a spring. The motion of the jth electron in a solid can be described according to the following equation of motion:

$$m^*\frac{d^2\mathbf{x}_j}{dt^2} + m^*\Gamma\frac{d\mathbf{x}_j}{dt} + m^*\omega_o^2\mathbf{x}_j = -e\mathcal{E} \tag{6.22}$$

where m^* = effective mass of electron
Γ = damping constant
$\mathcal{E}$ = electric field

The second term on the right-hand side of this equation represents various dampings, such as collisions, and the third term is the Hooke's law restoring force. The time-dependent $\mathbf{x}$ and $\mathcal{E}$ can be taken as

$$\begin{aligned} \mathbf{x}_j &= \mathbf{x}_{oj} \exp[-i\omega t] \\ \mathcal{E} &= \mathcal{E}_o \exp[-i\omega t] \end{aligned} \tag{6.23}$$

The solution of Eq. (6.22) is thus given as

$$\mathbf{x}_j = \frac{e\mathcal{E}}{m^*\left[\left(\omega^2 - \omega_o^2\right) + i\Gamma\omega\right]} \tag{6.24}$$

The induced dipole moment $\mathbf{P}$ per unit volume V is given by

$$\mathbf{P} = -e\sum_j \frac{\mathbf{x}_j}{V} = -\frac{Ne^2\mathcal{E}}{m^*\left[\left(\omega^2 - \omega_o^2\right) + i\Gamma\omega\right]} \tag{6.25}$$

where N is the electron concentration. Using Eqs. (6.2) and (6.4), one can obtain the dielectric constant as

$$\epsilon(\omega) = 1 - \frac{Ne^2}{\epsilon_o m^*\left[\left(\omega^2 - \omega_o^2\right) + i\Gamma\omega\right]} \tag{6.26}$$

For electrically neutral solids with free electrons, there exists a plasma with equal concentrations of positive and negative charges. If the damping and Hooke's force are ignored in Eq. (6.22), the plasma frequency can be obtained as $\omega_p^2 = Ne^2/(\epsilon_o m^*)$. Interband electronic transitions in semiconductors contribute to the dielectric constants, and this contribution, labeled ϵ_∞, should be included in Eq. (6.26). The final expression for the complex dielectric constant is

$$\epsilon(\omega) = \epsilon_\infty\left[1 - \frac{\omega_p^2}{\left(\omega^2 - \omega_o^2\right) + i\Gamma\omega}\right] \tag{6.27}$$

where ω_p is redefined as $\omega_p^2 = Ne^2/(\epsilon_o\epsilon_\infty m^*)$. The real and imaginary parts of $\epsilon(\omega)$ can now be evaluated and given by the following expressions:

$$\epsilon_1(\omega) = \epsilon_\infty\left[1 - \frac{\omega_p^2\left(\omega^2 - \omega_o^2\right)}{\left(\omega^2 - \omega_o^2\right)^2 + \Gamma^2\omega^2}\right] \tag{6.28a}$$

$$\epsilon_2(\omega) = \epsilon_\infty\frac{\omega_p^2\Gamma\omega}{\left(\omega^2 - \omega_o^2\right)^2 + \Gamma^2\omega^2} \tag{6.28b}$$

Notice that $\sqrt{\epsilon_\infty} \approx n_r$ for $\omega \gg \omega_p$. The absorption coefficient defined in Eq. (6.20) can be rewritten as

$$\alpha(\omega) = \frac{\epsilon_\infty\omega^2\omega_p^2\Gamma}{cn_1(\omega)\left[\left(\omega^2 - \omega_o^2\right)^2 + \Gamma^2\omega^2\right]} \tag{6.29}$$

The absorption coefficient has a lorentzian lineshape. In the Lorentz model, the optical absorption is derived for band-to-band transitions. This is a very simplistic form of the optical absorption of interband transitions. In real semiconductor material, there are many effects that have to be included when deriving this coefficient. For example, momentum matrix elements or the oscillator strengths need to be considered.

For $\Gamma \ll \omega$ and $\omega_o = 0$, the absorption coefficient is reduced to

$$\alpha(\omega) \approx \frac{\epsilon_\infty \omega_p^2 \Gamma}{c n_1(\omega)\omega^2} \tag{6.30}$$

This absorption coefficient expression is actually the result of the Drude model. Notice that $\alpha(\omega) \propto \omega^{-2}$, which is the characteristic behavior of free-electron absorption. The electrical conductivity can be obtained according to the Drude model by setting the restoring force (Hooke's law) in Eq. (6.22) to zero. The solution of the equation of motion becomes

$$\mathbf{x}_j = \frac{e\mathcal{E}}{m^*\omega(\omega + i\Gamma)} \tag{6.31}$$

By taking the first derivative of $\mathbf{x}_j$ with respect to time, we have $\boldsymbol{\upsilon}_j = \partial \mathbf{x}_j/\partial t = -i\omega\mathbf{x}_j$, where $\boldsymbol{\upsilon}_j$ is the electron velocity. Substituting this derivative into Eq. (6.31) one obtains

$$\boldsymbol{\upsilon}_j = -\frac{ie\mathcal{E}}{m^*(\omega + i\Gamma)} \tag{6.32}$$

On the other hand, the conduction current density is $\mathbf{J} = -Ne\boldsymbol{\upsilon}_j$. Multiply Eq. (6.32) by the electron change and by the electron concentration to obtain

$$\mathbf{J} = \frac{iNe^2\mathcal{E}}{m^*(\omega + i\Gamma)} \tag{6.33}$$

One can see that the conductivity $\sigma(\omega)$ is

$$\sigma(\omega) = \frac{iNe^2}{m^*(\omega + i\Gamma)} \tag{6.34}$$

The real and imaginary parts of the dielectric constant in the Drude model can be obtained by setting $\omega_o = 0$ in Eq. (6.28). For $\omega = 0$, the conductivity is reduced to its direct current (dc) value of

$$\sigma(\omega) = \frac{Ne^2}{m^*\Gamma} = \frac{Ne^2\tau}{m^*} \tag{6.35}$$

where $\tau = \Gamma^{-1}$ and is designated as the scattering time.

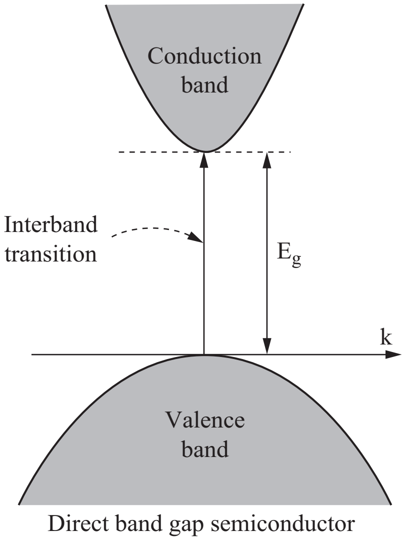


Figure 6.3 A sketch of the direct bandgap energy showing the vertical interband transition.

6.3 The Optical Absorption Coefficient of the Interband Transition in Direct Bandgap Semiconductors

A direct bandgap semiconductor is characterized by having its valence band maximum and conduction band minimum at the same k-value in reciprocal space, or momentum space, as shown in Fig. 6.3. It is customary to assume that this k-value is zero, which is designated as the center of the first Brillouin zone and in group theory is labeled as Γ-point symmetry. Many authors have taken different approaches to the calculation of the optical absorption of interband transitions in semiconductors. In our case, we follow the steps taken by Balkanski and Wallis (2000). The absorption coefficient is defined according to Beer's law as shown in Eq. (6.21). By taking the first derivative of the light intensity with respect to z, we have

$$\frac{dI(z)}{dz} = -I(z)\alpha(\omega) \tag{6.36}$$

For a sample with a cross-sectional area of A, the rate of energy absorption is

$$\frac{dE}{dt} = -A\,dI = I\alpha(\omega)A\,dz \tag{6.37}$$

where dI is the change in light intensity after the light passes through the sample. The rate of energy absorption can also be written as

$$\frac{dE}{dt} = \hbar\omega W_{vc}^{t} \tag{6.38}$$

where $\hbar\omega$ is the photon energy and W_{vc}^{t} is the total transition probability of an electron transition from a valence band state to a conduction band state. The absorption coefficient can be rewritten by combining Eqs. (6.37) and (6.38) as

$$\alpha(\omega) = \frac{\hbar\omega W_{vc}^{t}}{IA\,dz} \tag{6.39}$$

The two major items that need to be determined in this equation are the $I(z)$ and W_{vc}. The intensity can be obtained by assuming that it is the mean value of the Poynting vector $\mathbf{S} = \mathcal{E} \times \mathbf{B}$. The electric and magnetic fields can be written in terms of the vector potential $\mathbf{A}$ as $\mathcal{E} = -\partial\mathbf{A}/\partial t$ and $\mu_o\mathbf{B} = \boldsymbol{\nabla} \times \mathbf{A}$. Let $\mathbf{A} = \mathbf{A}_o\cos(\mathbf{k}\cdot\mathbf{r} - \omega t)$. The electric and magnetic fields can now be written as

$$\mathcal{E} = -\omega\mathbf{A}_o\sin(\mathbf{k}\cdot\mathbf{r} - \omega t) \tag{6.40a}$$

$$\mu_o\mathbf{B} = -\mathbf{k}\times\mathbf{A}_o\sin(\mathbf{k}\cdot\mathbf{r} - \omega t)] \tag{6.40b}$$

The Poynting vector takes the following form

$$\mathbf{S} = \frac{\omega}{\mu_o}\mathbf{A}_o \times [\mathbf{k}\times\mathbf{A}_o\sin^2(\mathbf{k}\cdot\mathbf{r} - \omega t)] \tag{6.41}$$

Thus, the intensity I can be written as

$$I = \langle\mathbf{S}\,\rangle = \frac{\omega^2 n_r(\omega)}{2\mu_o c}|\mathbf{A}_o|^2 \tag{6.42}$$

where $\langle\mathbf{S}\rangle$ is the time average of $\mathbf{S}$ over one period. In reaching the form shown in Eq. (6.42), the dispersion relation (6.14) and the vector analysis identity $\mathbf{A}\times(\mathbf{B}\times\mathbf{C}) = (\mathbf{A}\cdot\mathbf{C})\mathbf{B} - (\mathbf{A}\cdot\mathbf{B})\mathbf{C}$ were used assuming that $\mathbf{k}$ and $\mathbf{A}_o$ are orthogonal. Substituting Eq. (6.42) into (6.39), we have

$$\alpha(\omega) = \frac{2\mu_o c\hbar W_{vc}^{t}}{\omega n_r|\mathbf{A}_o|^2 V} = \frac{2\hbar^2}{\epsilon_o c n_r|\mathbf{A}_o|^2 V}\frac{1}{\hbar\omega}W_{vc}^{t} \tag{6.43}$$

The next step is to evaluate W_{vc}^{t}. The approach here is to evaluate the transition probability (W_{vc}) from one Bloch state in the valence band to another Bloch state in the conduction band. In obtaining the transition probability, the photon-electron interaction can be treated as a perturbation in the following Hamiltonian:

$$\begin{aligned}\mathbf{H} &= \frac{(\mathbf{p}+e\mathbf{A})^2}{2m_o} + V(\mathbf{r}) \\ &= \frac{\mathbf{p}^2}{2m_o} + \frac{e}{2m_o}(\mathbf{p}\cdot\mathbf{A} + \mathbf{A}\cdot\mathbf{p}) + \frac{e}{2m_o}\mathbf{A}^2 + V(\mathbf{r}) \qquad (6.44)\end{aligned}$$

where $\mathbf{p}$ = momentum
$\mathbf{A}$ = vector potential
$V(\mathbf{r})$ = electron potential energy

For low light intensities, $\mathbf{A}^2$ can be neglected. Because the photon momentum is negligible, the term rising from $\mathbf{p}$ acting on $\mathbf{A}$ can also be neglected. Notice that we use the free-electron mass instead of the electron effective mass. The photon-electron interaction Hamiltonian can be written as

$$\mathbf{H}' = \frac{e}{2m_o}\mathbf{A}\cdot\mathbf{p} \qquad (6.45)$$

The transition probability of an electron from the valence band to the conduction band is given by the Fermi golden rule:

$$W_{vc} = \frac{2\pi}{\hbar}|\langle\mathbf{k}_c'|\mathbf{H}'|\mathbf{k}_v\rangle|^2\delta(E_{\mathbf{k}_c'} - E_{\mathbf{k}_v} - \hbar\omega) \qquad (6.46)$$

where $\langle\mathbf{k}_c'|$ and $|\mathbf{k}_v\rangle$ are Bloch states in the conduction and valence bands, respectively. The δ-function is included to conserve the energy. If $\mathbf{A}$ has the form $\mathbf{A} = \mathbf{A}_o\cos(\mathbf{k}\cdot\mathbf{r} - \omega t)$, the interband matrix element of $\mathbf{H}'$ can be written as

$$\langle\mathbf{k}_c'|\mathbf{H}'|\mathbf{k}_v\rangle = \frac{e}{2m_o}\langle\mathbf{k}_c'|\mathbf{A}_o\cdot\mathbf{p}|\mathbf{k}_v\rangle \qquad (6.47a)$$

Substitute Eq. (6.47*a*) into (6.46) to obtain

$$W_{vc} = \frac{\pi e^2}{2\hbar m_o^2}|\mathbf{A}_o|^2|\langle\mathbf{k}_c'|\mathbf{p}_A|\mathbf{k}_v\rangle|^2\delta(E_{\mathbf{k}_c'} - E_{\mathbf{k}_v} - \hbar\omega) \qquad (6.47b)$$

where $\mathbf{p}_A$ is the momentum component along the $\mathbf{A}$ direction. In order for Eq. (6.47*b*) to be evaluated, the Bloch form of the valence and conduction bands are used such as

$$\langle\mathbf{k}_c'|\mathbf{p}_A|\mathbf{k}_v\rangle = \int_{\text{crystal}} e^{-i\mathbf{k}'\cdot\mathbf{r}}\varphi^*_{\mathbf{k}_c'}(\mathbf{r})\mathbf{p}_A e^{i\mathbf{k}\cdot\mathbf{r}}\varphi_{\mathbf{k}_v}(\mathbf{r})d^3\mathbf{r} \qquad (6.48)$$

Taking the integral over one primitive unit cell and summing over all the unit cells, Eq. (6.48) can be rewritten as

$$
\begin{aligned}
\langle \mathbf{k}_c' | \mathbf{p}_A | \mathbf{k}_v \rangle &= \sum_l \int_l e^{-i\mathbf{k}'\cdot\mathbf{r}} \varphi^*_{\mathbf{k}_c'}(\mathbf{r}) \mathbf{p}_A e^{i\mathbf{k}\cdot\mathbf{r}} \varphi_{\mathbf{k}_v}(\mathbf{r}) d^3\mathbf{r} \\
&= \sum_l \int_l e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{r}} \varphi^*_{\mathbf{k}_c'}(\mathbf{r}) (\hbar \mathbf{k}_A + \mathbf{p}_A) \varphi_{\mathbf{k}_v}(\mathbf{r}) d^3\mathbf{r} \qquad (6.49)
\end{aligned}
$$

where the summation is over all unit cells and the integration is over one unit cell labeled l. The term $\hbar\mathbf{k}_A$ is the result of the $\mathbf{p}_A$ operation on the exponential part of the wave function, and $\mathbf{k}_A$ is the component of the wavevector in the **A** direction. Utilizing the periodicity of the crystal where $\mathbf{r} = \mathbf{R}(l) + \mathbf{r}'$ and $\varphi^*_{n\mathbf{k}}(\mathbf{r}) = \varphi^*_{n\mathbf{k}}[\mathbf{R}(l) + \mathbf{r}'] = \varphi^*_{n\mathbf{k}}(\mathbf{r}')$, Eq. (6.49) can be written as

$$
\langle \mathbf{k}_c' | \mathbf{p}_A | \mathbf{k}_v \rangle = \sum_l e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{R}(l)} \int_{\text{cell } 0} e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{r}'} \varphi^*_{\mathbf{k}_c'}(\mathbf{r}') (\hbar \mathbf{k}_A + \mathbf{p}_A) \varphi_{\mathbf{k}_v}(\mathbf{r}') d^3\mathbf{r}' \qquad (6.50)
$$

where the sum is over all unit cells in the crystal and the integral is over the unit cell labeled "0." The sum over all unit cells obviously gives the number of cells in the crystal. The wave functions $\varphi_{\mathbf{k}_c'}(\mathbf{r})$ and $\varphi_{\mathbf{k}_v}(\mathbf{r})$ are orthogonal, which means that the first term of the integral is zero. The momentum matrix element can now be written as

$$
\langle \mathbf{k}_c' | \mathbf{p}_A | \mathbf{k}_v \rangle = N \delta_{\mathbf{k},\mathbf{k}'} \int_{\text{cell } 0} \varphi^*_{\mathbf{k}_c}(\mathbf{r}') \mathbf{p}_A \varphi_{\mathbf{k}_v}(\mathbf{r}') d^3\mathbf{r}' \qquad (6.51)
$$

where N is the total number of unit cells in the crystal. The momentum matrix element can now be written as

$$
W_{vc} = \frac{\pi e^2}{2\hbar m_o^2} |\mathbf{A}_o|^2 P^2 \delta(E_{\mathbf{k}_c'} - E_{\mathbf{k}_v} - \hbar\omega) \delta_{\mathbf{k},\mathbf{k}'} \qquad (6.52)
$$

where

$$
P = N \int_{\text{cell } 0} \varphi^*_{\mathbf{k}_c}(\mathbf{r}') \mathbf{p}_A \varphi_{\mathbf{k}_v}(\mathbf{r}') d^3\mathbf{r}' \qquad (6.53)
$$

The quantity P was briefly discussed in Sec. 3.7.2. It is a number known for many semiconductors. For example, $2P^2/m_o \approx 25.7$ (eV) for GaAs, ≈ 20.9 (eV) for InP, and ≈ 22.2 (eV) for InAs (see Singh 2003). The δ-function, $\delta_{\mathbf{k},\mathbf{k}'}$, in Eq. (6.52) gives the selection rules for the direct transition from the valence band to the conduction band.

The total transition probability from the valence band to the conduction band over all $\mathbf{k}$ and $\mathbf{k}'$ is obtained by summing over all $\mathbf{k}$ and $\mathbf{k}'$,

$$W_{vc}^{t} = 2\sum_{\mathbf{k}}\sum_{\mathbf{k}'} W_{vc} f_{\mathrm{FD}}^{\mathbf{k}v}(1 - f_{\mathrm{FD}}^{\mathbf{k}'c}) \tag{6.54}$$

where $f_{\mathrm{FD}}^{\mathbf{k}v}$ and $(1 - f_{\mathrm{FD}}^{\mathbf{k}'c})$ are Fermi-Dirac distribution functions for a full valence band and an empty conduction band, respectively. The factor 2 is added to account for the electron spin degeneracy. For a two-band model at $T = 0$ K, we have $f_{\mathrm{FD}}^{\mathbf{k}v} = 1$, $f_{\mathrm{FD}}^{\mathbf{k}'c} = 0$ and

$$\begin{aligned} E_{\mathbf{k}c} - E_{\mathbf{k}v} &= E_g + \frac{\hbar^2 k^2}{2m_c^*} + \frac{\hbar^2 k^2}{2m_v^*} \\ &= E_g + \frac{\hbar^2 k^2}{2m_r^*} \end{aligned} \tag{6.55}$$

where m_r^* is the reduced mass of the electron and hole system given by

$$\frac{1}{m_r^*} = \frac{1}{m_c^*} + \frac{1}{m_v^*} \tag{6.56}$$

Putting all these together, Eq. (6.54) can be rewritten as (see Balkanski and Wallis 2000)

$$W_{vc}^{t} = \frac{e^2 V}{8\pi^2 \hbar m_o^2} |\mathbf{A}_o|^2 P^2 \int \delta(E_{\mathbf{k}c} - E_{\mathbf{k}v} - \hbar\omega) d^3\mathbf{k} \tag{6.57}$$

where V is the volume of the semiconductor sample. The integral in Eq. (6.57) can be evaluated using spherical coordinates as

$$\begin{aligned} \int \delta(E_{\mathbf{k}c} - E_{\mathbf{k}v} - \hbar\omega) d^3\mathbf{k} &= 4\pi \int_0^\infty k^2 \delta\left(E_g - \hbar\omega + \frac{\hbar^2 k^2}{2m_r^*}\right) d\mathbf{k} \\ &= 4\pi \sqrt{\frac{2m_r^*}{\hbar^2}} \int_0^\infty \sqrt{E}\,\delta\left(E_g - \hbar\omega + \frac{\hbar^2 k^2}{2m_r^*}\right) \mathbf{k}\, d\mathbf{k} \\ &= 4\pi \sqrt{\frac{2m_r^*}{\hbar^2}} \frac{m_r^*}{\hbar^2} \int_0^\infty \sqrt{E}\,\delta(E_g - \hbar\omega + E) dE \\ &= 4\pi \sqrt{\frac{2m_r^*}{\hbar^2}} \frac{m_r^*}{\hbar^2} \sqrt{\hbar\omega - E_g} \\ &= 2\pi \left(\frac{2m_r^*}{\hbar^2}\right)^{3/2} \sqrt{\hbar\omega - E_g} \end{aligned} \tag{6.58}$$

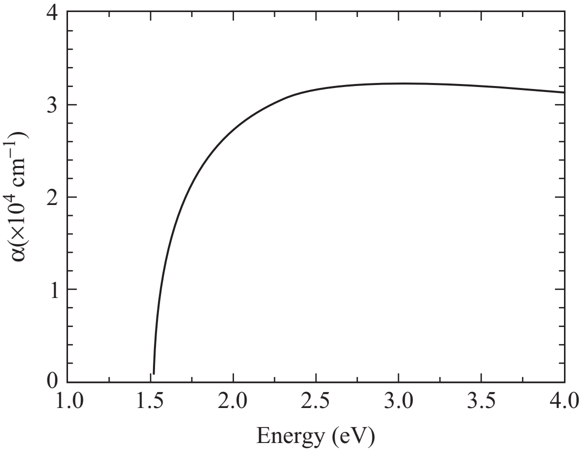


Figure 6.4 The optical absorption coefficients as a function of the photon energy for GaAs bulk material. The bandgap was chosen at 1.52 eV.

This expression is valid for $\hbar\omega \geq E_g$, and it is zero for $\hbar\omega \leq E_g$. Substituting Eqs. (6.58) and (6.57) into Eq. (6.43), we obtain the following expression for the optical absorption coefficient of direct interband transition in a bulk semiconductor:

$$\alpha(\omega) = \frac{2\mu_o c\hbar W_{vc}^t}{\omega n_r |\mathbf{A}_o|^2 V} = \frac{e^2\hbar}{2\pi\epsilon_o c n_r m_o^2} P^2 \left(\frac{2m_r^*}{\hbar^2}\right)^{3/2} \frac{1}{\hbar\omega}\sqrt{\hbar\omega - E_g} \qquad (6.59)$$

A plot of the absorption coefficient given by Eq. (6.59), using GaAs parameters, is shown in Fig. 6.4. The energy is plotted for $\hbar\omega \geq E_g$. The optical absorption coefficient can be expressed in terms of the oscillator strength f_{vc}, which is defined as $f_{vc} = 2P^2/(m_o\hbar\omega)$. The maximum value of f_{vc} can be obtained from the sum rule (see, for example, Wooten 1972) as

$$f_{vc} \approx \begin{cases} \left|1 - \dfrac{m_o}{m_e^*}\right| & \text{for electron} \\ 1 + \dfrac{m_o}{m_h^*} & \text{for hole} \end{cases} \qquad (6.60)$$

When the direct interband transition is forbidden at $\mathbf{k} = 0$, but allowed at $\mathbf{k} \neq 0$, the optical absorption coefficient depends on the photon energy as $\alpha(\omega) \propto (\hbar\omega - E_g)^{3/2}/(\hbar\omega)$ (see Pankove 1971). It should be noted that Fig. 6.4 does not include the absorption from either excitons or other valleys.

6.4 The Optical Absorption Coefficient of the Interband Transition in Indirect Bandgap Semiconductors

The interband transition in an indirect bandgap semiconductor such as Si occurs between the valence band maximum and conduction band minimum that are located at different k-values, as shown in Fig. 6.5. In a direct bandgap semiconductor, the interband transition is excited by only electron-photon interactions. The interband transition of an indirect bandgap semiconductor requires electron-photon and electron-phonon interactions. A phonon is the quanta of lattice vibrations. Thus, momentum and energy conservation require that

$$\mathbf{k}_c = \mathbf{k}_v \pm \mathbf{q} \tag{6.61a}$$

$$\hbar\omega = E_c - E_v \pm \hbar\omega_{ph} \tag{6.61b}$$

where $\mathbf{q}$ is the phonon wavevector and $\hbar\omega_{ph}$ is the phonon energy. The plus and minus signs are for emission or absorption of phonons, respectively. The optical absorption coefficient can be derived in a manner similar to that of the direct bandgap semiconductor. However, due to electron-phonon interactions, several steps have to be modified. For example, the electron-phonon matrix element must be included in the analysis. Furthermore, the argument of the δ-function must include the phonon energy. The number of phonons (from Bose-Einstein statistics)

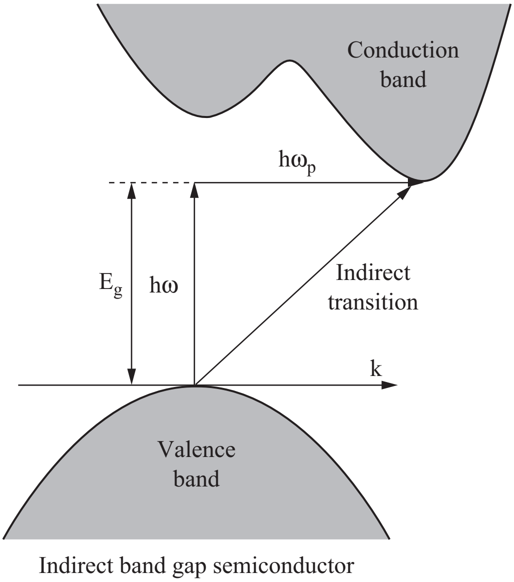


Figure 6.5 A sketch of an indirect bandgap semiconductor showing the indirect interband transition.

must also be included. The final result is

$$\alpha(\omega)=\frac{A(\hbar\omega+\hbar\omega_{ph}-E_g)^n}{\exp(\hbar\omega_{ph}/k_BT)-1}+\frac{B\exp(\hbar\omega_{ph}/k_BT)(\hbar\omega-\hbar\omega_{ph}-E_g)^n}{\exp(\hbar\omega_{ph}/k_BT)-1} \tag{6.62}$$

where A and B are constants, $n=2$ when a vertical transition is allowed, and $n=3$ when vertical transitions are not allowed (see Wooten 1972).

6.5 The Optical Absorption Coefficient of the Interband Transition in Quantum Wells

A typical example of an interband transition in a type I quantum well is shown in Fig. 6.1*b*, where the electrons are excited from the bound ground state in the valence band to the bound ground state in the conduction band. The steps used to calculate the interband transition in quantum wells are similar to those discussed previously for optical absorption of the interband transition in direct bandgap bulk semiconductor materials. There are, however, a few modifications that must be included:

1. The density of states appears explicitly in Eq. (6.59). Its form is given by Eq. (6.57) except that the effective mass is replaced by the reduced mass. Hence, the notation "reduced density of states" is introduced. For a two-dimensional system, such as a multiple quantum well, the reduced density of states

$$\frac{1}{2\pi^2}\left(\frac{2m_r^*}{\hbar^2}\right)^{3/2}\sqrt{\hbar\omega-E_g}$$

needs to be replaced by the reduced two-dimensional density of states

$$\frac{g_{cv}^{2D}}{L}=\frac{m_r^*}{\pi\hbar^2L}\sum_{m\cdot n}\langle g_v^m|g_c^n\rangle\Theta(E_{nm}-\hbar\omega) \tag{6.63}$$

where $$E_{nm}=E_g+E_c^n+E_v^m \tag{6.64}$$

and $\langle g_v^m|g_c^n\rangle$ = overlap integral between z-dependent envelop functions of conduction band and valence band
L = width of quantum well
Θ = Heaviside step function

Notice that E_{nm} is the photon energy required to excite the interband transition in the quantum well. The energies E_c^n and E_v^m correspond to the energies of the ground bound states in the conduction and valence bands, respectively. The width of the well is introduced in Eq. (6.64) to account for the transformation of the momentum matrix element [Eq. (6.53)] as it goes from the three-dimensional system to the two-dimensional system.

2. The number of wells, N_w, should be included in the final expression of the optical absorption coefficient.
3. The absorption coefficient is calculated for the wells only.
4. The overlap integral defined in (1) provides the selection rules for the transition. Let us assume that the envelope function has the form

$$F_{n\mathbf{k}_\perp}(r) = e^{i\mathbf{k}_\perp\cdot\mathbf{r}_\perp}\chi_n(z) \tag{6.65}$$

The overlap integral can now be written as

$$\langle g_v^m \mid g_c^n\rangle = \langle m\mathbf{k}_\perp \,|\, n\mathbf{k}'_\perp\rangle = \delta_{\mathbf{k}_\perp,k'_\perp}\int_{-L/2}^{L/2}\chi_m^h(z)\chi_n^e(z)\,dz \tag{6.66}$$

Thus, the overlap integral is nonzero if and only if $\chi_m^h(z)$ and $\chi_n^e(z)$ are both odd parity or both even parity.

By considering all these modifications, the optical absorption coefficient of the interband transition in type I multiple quantum wells can be written as

$$\alpha(\omega) = \frac{e^2 N_w m_r^*}{2\epsilon_o c n_r \hbar L m_o^2}\frac{P^2}{\hbar\omega}\sum_{n,m}\Theta(E_{nm} - \hbar\omega) \tag{6.67}$$

We assumed that the square of the overlap integral [Eq. (6.66)] is unity, and, therefore, it was not included in the absorption coefficient expression. Equation (6.67) is valid for both heavy holes and light holes. The only difference is that the reduced mass m_r^* is different. Notice that $\hbar\omega$ in the denominator is the minimum photon energy needed to cause an electronic transition from the valence band to the conduction band within the quantum well. If the definition of the oscillator strength is $f_{vc} = 2P^2/(m_o\hbar\omega)$, where its maximum value is given by Eq. (6.60), then the optical absorption coefficient can be rewritten as

$$\alpha(\omega) = \frac{e^2 N_w m_r^*}{4\epsilon_o c n_r \hbar L m_o} f_{vc}\sum_{n,m}\Theta(E_{nm} - \hbar\omega) \tag{6.68}$$

For example, the oscillator strength for an electronic transition in a GaAs/AlGaAs quantum well is $f_{vc} \approx |1 - 1/0.067| = 13.925$. This quantity is comparable to the value obtained from $f_{vc} = 2P^2/(m_o\hbar\omega) = 25.7/1.75 = 14.68$ for $\hbar\omega = 1.75$ eV. A plot of the absorption coefficient depicted in Eq. (6.68) is show in Fig. 6.6. The ladder-like behavior of the optical absorption is due to the step function, which is the characteristic signature of the reduced density of states in the quantum well.

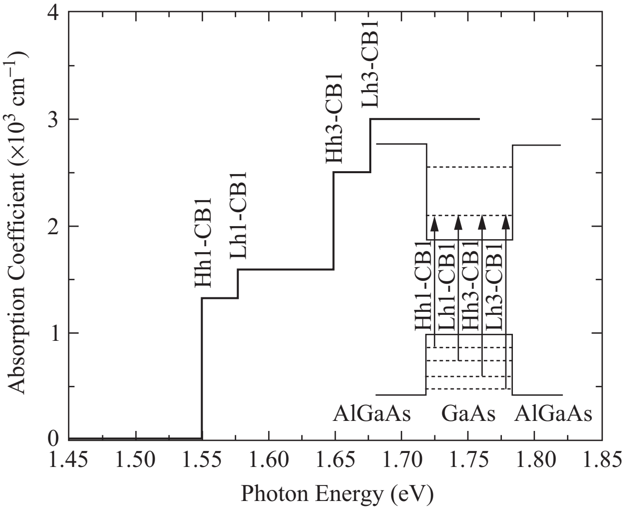


Figure 6.6 The optical absorption coefficient in a 30 Å GaAs/AlGaAs quantum well plotted as a function of photon energy.

6.6 The Optical Absorption Coefficient of the Interband Transition in Type II Superlattices

A typical example of a type II superlattice is an InAs/InGaSb structure as shown in Fig. 6.7. In this figure, we plotted the conduction band as a thick line and the valence band as a thin line. The bound states are shown as the dotted lines, and the wave functions for the ground

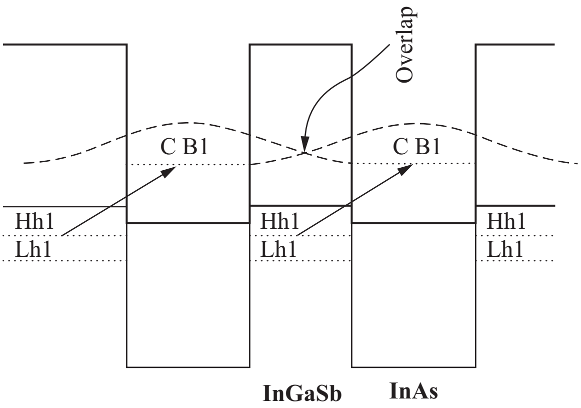


Figure 6.7 A sketch of the band alignment of the InAs/InGaSb superlattice is shown. Minibands are formed due to the overlap of the wave functions as indicated by the dashed line.

bound states in the conduction band are plotted as the dashed lines. The barrier material is grown thin enough to allow the wave functions to overlap, forming what are called minibands for both the holes and the electrons. The wave functions for the holes are not shown. The intriguing property of this system is that the interband transitions are indirect in real space as indicated by the arrows in the figure, but the system exhibits a direct bandgap in **k**-space. The energy dependence of the optical absorption of the band-to-band transitions can thus be described by the Pankove expression, $\alpha(\omega) \propto (\hbar\omega - E_g)^{3/2}/\hbar\omega$. This energy-dependence is valid at least near the band edge.

The analysis of the optical absorption coefficient for the type II superlattice is more complicated than interband transitions in type I superlattices. This is due to the fact that the overlap integral is no longer unity. In this section we simply report (for full derivation, see Bastard 1988) the absorption coefficient of the interband transition between the ground state of the heavy hole in an InGaSb layer and the ground state of the electron in the conduction band of an InAs layer as follows:

$$\alpha(\omega) = \frac{e^2 m_r^* P^2 P_b(E_1)}{\pi \epsilon_o n_r c m_o^2 \omega \hbar^2 L}\left[\frac{-x}{1+x^2} + \text{arc}\tan(x)\right] \tag{6.69}$$

where $P_b(E_1)$ is the probability of finding the electron in an InAs layer while in the CB1 state (see Fig. 6.7) and is given by

$$P_b(E_1) = \frac{B_c^2}{k_c} \tag{6.70}$$

The parameters B_c and k_c are the amplitude of the envelope wave function of the electron in the barrier and the corresponding propagation vector, respectively. In other words, the electron envelope wave function in the barrier (InGaSb) is given by $\chi_1^e(z) = B_c \exp[-k_c(z - L/2)]$. The subscript "1" indicates the wave function for the ground state (CB1). The parameter x in Eq. (6.69) is given by the following expression:

$$x = \sqrt{\frac{2m_h^*}{\hbar^2}}\,\frac{\sqrt{\hbar\omega - E_g^{\text{InAs}} + \Delta E_v - E_{\text{CB1}}}}{k_c} \tag{6.71}$$

where ΔE_v = valence band offset
E_g^{InAs} = InAs bandgap
E_{CB1} = electron ground state in the conduction band

The rest of the parameters in Eq. (6.69) were defined previously. The lineshape of the optical absorption coefficient is defined by the behavior

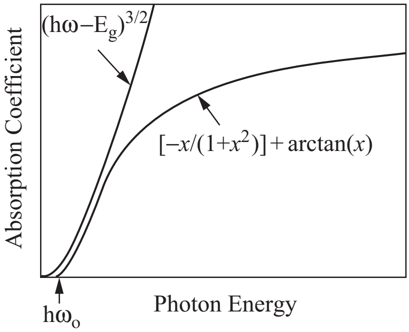


Figure 6.8 The lineshape of the optical absorption coefficient defined by Eq. (6.69) is plotted as a function of photon energy ($\hbar_\omega$). The lineshape defined by ($\hbar_\omega - E_g)^{3/2}$ is also shown.

of the quantity in the square brackets, which is plotted in Fig. 6.8. The onset of the optical absorption profile occurs at $\hbar\omega_o = E_g^{\text{InAs}} - \Delta E_v + E_{\text{CB1}}$. Above this onset, the lineshape appears to depend on the photon energy according to the relation $\alpha(\omega) \approx (\hbar\omega - E_g)^{3/2}$. This energy dependence implies that the direct electronic transition is forbidden. The latter relation is plotted in Fig. 6.8 for comparison purposes.

6.7 The Optical Absorption Coefficient of the Intersubband Transition in Multiple Quantum Wells

Intersubband transitions in low-dimensional quantum structures have been investigated for their infrared application as detectors and lasers. The intersubband transitions are generated in *n*- or *p*-type quantum well structures with at least one bound state as shown in Fig. 6.9. In this figure, we have shown (*a*) a bound-to-bound transition, where both the ground state and the first excited state are bound; (*b*) a bound-to-continuum transition, where the ground state is bound while the first excited state is resonant in the conduction band; (*c*) the transitions between the states depicted in **k**-space; (*d*) the optical absorption lineshape for the bound-to-bound transition; and (*e*) the optical absorption profile for the bound-to-continuum transition.

As indicated in Fig. 6.9*d* and *e*, the optical absorption profile of the electrons that undergo the intersubband transition from bound to bound is different from that of a bound-to-continuum transition. Let us first obtain the optical absorption coefficient for the bound-to-bound transition, which has a lorentzian lineshape. The envelope wave function for the two bound states in Fig. 6.9*a* can be written as

$$\psi_{n\mathbf{k}_\perp}(\mathbf{r}) = e^{i\mathbf{k}_\perp\cdot\mathbf{r}_\perp}\chi_n(z) \tag{6.72}$$

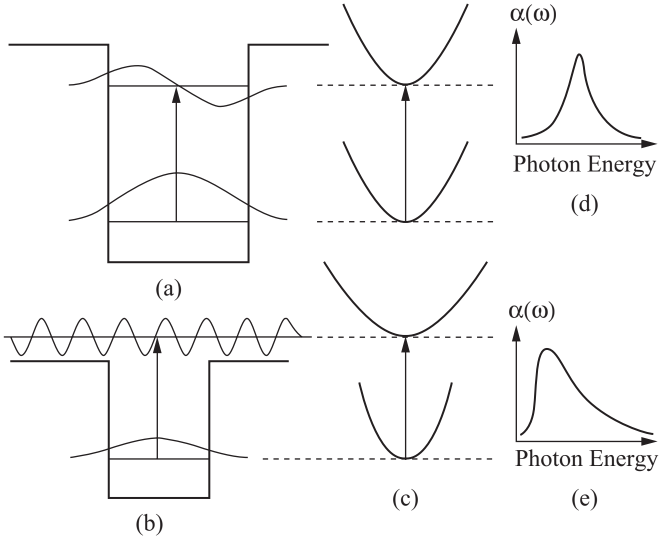


Figure 6.9 Bound-to-bound and bound-to-continuum configurations for intersubband transitions in n-type quantum well structures. The transitions are shown in **k**-space along with their optical absorption coefficient profiles.

where

$$\chi_n(z) = \begin{cases} \sqrt{\dfrac{2}{L}}\cos\left(\dfrac{n\pi z}{L}\right) & \text{for } n = \text{odd integer} \quad (6.73a) \\ \sqrt{\dfrac{2}{L}}\sin\left(\dfrac{n\pi z}{L}\right) & \text{for } n = \text{even integer} \quad (6.73b) \end{cases}$$

and $L =$ width of thickness of quantum well

$$\mathbf{k}_\perp = \sqrt{k_x^2 + k_y^2}$$
$$\mathbf{r}_\perp = x\widehat{\mathbf{x}} + y\widehat{\mathbf{y}}$$

Following the procedure discussed in Sec. 6.3, the optical absorption coefficient can be written as

$$\alpha(\omega) = \frac{2\pi e^2 N_w \hbar}{\epsilon_o c n_r m^{*2} V}\frac{1}{\hbar\omega}\sum_{i,j}|\langle i|p_z|j\rangle|^2\delta(E_j - E_i - \hbar\omega)\left(f_{\text{FD}}^i - f_{\text{FD}}^j\right) \tag{6.74}$$

where f_{FD} is the Fermi-Dirac occupation function and i and j are the initial and final states, respectively. The factor 2 is added for the electron spin degeneracy. There are a few approximations in this equation. First, the free-electron mass was replaced by the electron effective mass. Second, the number of quantum wells is included to account for the absorption from all quantum wells. Third, the momentum component was taken along the z direction, which is the growth direction. Fourth, i and j were used, but they stand for (n, $\mathbf{k}_\perp$) and $(n', \mathbf{k}'_\perp)$, respectively. The momentum matrix element can be expressed as

$$\langle n\mathbf{k}_\perp | p_z | n'\mathbf{k}'_\perp\rangle = \delta_{\mathbf{k}_\perp, \mathbf{k}'_\perp} \langle n | p_z | n'\rangle \tag{6.75}$$

The wavevectors $\mathbf{k}_\perp$ and $\mathbf{k}'_\perp$ were removed from the bracket since they are not good quantum numbers for p_z to operate on them. The δ-function was introduced to conserve the momentum. By using the wave functions described in Eq. (6.73), one can find that the nonvanishing matrix elements of p_z are those associated with the following selection rule:

$$n - n' = \text{odd integer} \tag{6.76}$$

which means that only transitions between subbands with opposite parity are allowed. For n is odd and n' is even, the matrix element can be written as

$$\begin{aligned}\langle n\mathbf{k}_\perp | p_z | n'\mathbf{k}'_\perp\rangle &= \int\limits_{-L/2}^{L/2} \sqrt{\frac{2}{L}} \cos\left(\frac{n\pi z}{L}\right)\left(-i\hbar\frac{d}{dz}\right)\sqrt{\frac{2}{L}}\sin\left(\frac{n'\pi z}{L}\right) dz \\ &= -\frac{2i\hbar n'}{L}\left\{\frac{\sin[(n'+n)\pi/2]}{n'+n} + \frac{\sin[(n'-n)\pi/2]}{n'-n}\right\}\end{aligned} \tag{6.77}$$

For $n = 1$ and $n' = 2$, the momentum matrix element is $-i8\hbar/3L$ and the absorption coefficient is

$$\alpha(\omega) = \frac{2\pi e^2 N_w \hbar}{\epsilon_o c n_r m^{*2} V}\frac{1}{\hbar\omega}\left(\frac{8\hbar}{3L}\right)^2 \sum_{\mathbf{k}_\perp, n, n'} \delta(E_{n'} - E_n - \hbar\omega)\left(f_{\text{FD}}^{n,\mathbf{k}_\perp} - f_{\text{FD}}^{n',\mathbf{k}'_\perp}\right) \tag{6.78}$$

If the quantum wells are doped and the Fermi energy is above the ground state ($n = 1$, $\mathbf{k}_\perp$), and if the excited state ($n' = 2$, $\mathbf{k}'_\perp$) is completely empty, then one can write

$$2\sum_{\mathbf{k}_\perp} f_{\text{FD}}^{n,\mathbf{k}_\perp} = N_1 \qquad \text{and} \qquad \sum_{\mathbf{k}^i_\perp} f_{\text{FD}}^{n',\mathbf{k}'_\perp} = 0 \tag{6.79}$$

The factor 2 is for electron spin degeneracy. Substituting Eq. (6.79) into (6.78), we have

$$\alpha(\omega) = \frac{n_1 \pi e^2 N_w \hbar}{\epsilon_o c n_r m^{*2} l}\frac{1}{\hbar\omega}\left(\frac{8\hbar}{3L}\right)^2 \delta(\Delta E - \hbar\omega) \tag{6.80}$$

where l is the total thickness of the quantum wells and $\Delta E = E_2 - E_1$. In this expression we substituted $N_1 = n_1$/area, where n_1 is the electron sheet density, or the two-dimensional electron gas density. The oscillator strength can now be defined as

$$f_{01} = \frac{2P^2}{m^*\hbar\omega} = \frac{2}{m^*\hbar\omega}\left(\frac{8\hbar}{3L}\right)^2 \tag{6.81}$$

where the subscript "01" stands for the electronic transition from the ground state to the first excited state and P is the value of the momentum matrix element along the z direction. An equivalent definition of the oscillator strength is $f_{01} = (2m^*\omega/\hbar)|\langle n\mathbf{k}_\perp|z|n'\mathbf{k}'_\perp\rangle|^2$. Notice that $\hbar\omega$ is the photon energy required to excite a transition from the ground to the excited state. For a GaAs/AlGaAs quantum well of thickness 100 Å and $\hbar\omega = 150$ meV, $f_{01} \approx 1.08$. A typical example of the intersubband transition in GaAs/AlGaAs multiple quantum wells (MQWs) is shown in Fig. 6.10, where the solid lines represent experimental measurements at 300 and 77 K and the dashed lines represent lorentzian lineshape fits for both spectra. The peak position shift of the intersubband transition is explained in terms of the many-body effect (see Manasreh 1993). Since the experimental measurements show broadening due to several effects, the δ-function in Eq. (6.80) can now be replaced by a Lorentzian lineshape. By inserting Eq. (6.81) into (6.80), we obtain

$$\alpha(\omega) = \frac{n_1\pi e^2 N_w\hbar}{2\epsilon_o c n_r m^* l} f_{01} \frac{\Gamma}{\pi[(\hbar\omega - \Delta E)^2 + \Gamma^2]} \tag{6.82}$$

where Γ is the half-width at half of the maximum.

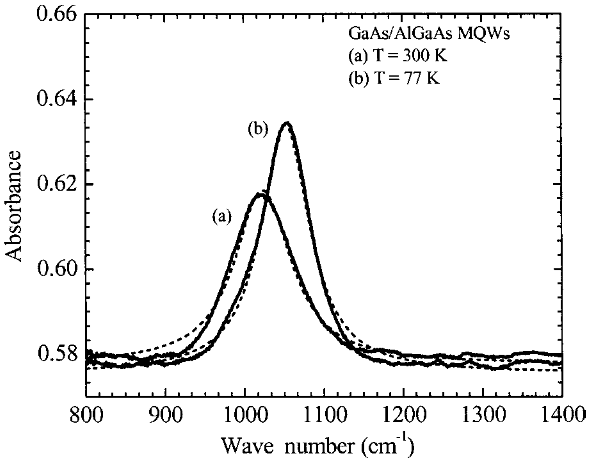


Figure 6.10 Absorbance of the intersubband transition in 75-Å GaAs/AlGaAs MQWs measured at 300 and 77 K (solid lines). The dashed lines are lorentzian fits of the data.

Intersubband transitions in n-type semiconductor quantum wells were found experimentally to be excited by a photon with an electric component parallel to the growth axis (z axis). Thus, the momentum matrix element should contain the following factor:

$$\widehat{\boldsymbol{\mathcal{E}}} \cdot \widehat{\mathbf{z}} \approx \sin\theta \tag{6.83}$$

where $\widehat{\boldsymbol{\mathcal{E}}}$ = unit vector of light polarization
θ = internal incident angle
$\widehat{\mathbf{z}}$ = unit vector along z direction.

If the light is polarized in the xy plane, the electron-photon coupling is zero. But if the light has a component parallel to the growth axis (z axis), then the electron-photon coupling is nonzero and the intersubband transition can be observed. If the light reaches the sample at an angle, then the light intensity has to be scaled by a factor of $\cos\theta$. The absorption coefficient for the intersubband transition in n-type quantum wells can now be rewritten as

$$\alpha(\omega) = \frac{n_1 \pi e^2 N_w \hbar}{2\epsilon_o c n_r m^* l} f_{01} \frac{\sin^2\theta}{\cos\theta} \frac{\Gamma}{\pi\,[(\hbar\omega - \Delta E)^2 + \Gamma^2]} \tag{6.84}$$

If θ is 45°, then the factor $(\sin^2\theta/\cos\theta)$ is ~0.71. One can now state the polarization selection rule for n-type quantum wells: *The electron-photon coupling for a spherically symmetric band in a quantum well is nonzero for photons polarized along the growth direction of the quantum well.*

For a bound-to-continuum intersubband transition, the calculation of the momentum matrix element is more complicated since the excited state is a propagating plane wave as shown in Fig. 6.9*b*. This problem has been discussed by Choi (1993) who derived the optical absorption lineshape as

$$\alpha(\omega) \propto \frac{\sqrt{\hbar\omega + E_1 - \Delta E_c}}{1 + C^2(\hbar\omega + E_1 - \Delta E_c)[\hbar\omega - (E_2 - E_1)]^2} \tag{6.85}$$

where ΔE_c = conduction band offset
E_1, E_2 = ground and first excited states, respectively
C = constant

For this expression to be valid, the excited state should be above ΔE_c. A similar expression was derived by Liu (1996).

Because of the small thickness of the quantum wells, the measured optical absorption of the intersubband transition is usually very small. One way to increase the absorption intensity is to fabricate a waveguide where the light will make multiple passes. Figure 6.11 shows the measurements of the absorbance of the intersubband transition for both the Brewster's angle and the waveguide configurations. Notice that the absorption coefficient can be obtained by dividing the absorbance by

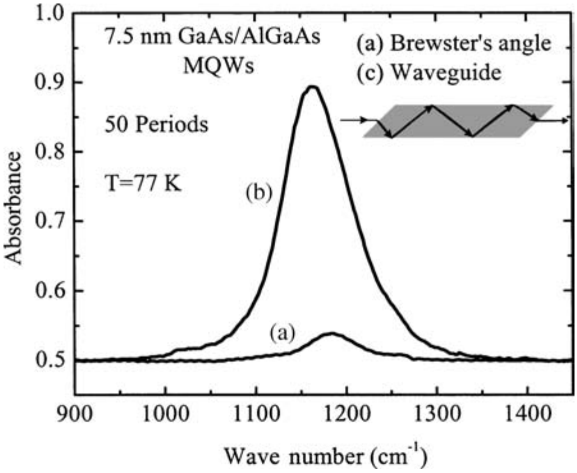


Figure 6.11 Intersubband transition in 75-Å GaAs/AlGaAs MQWs measured at $T = 77$ K using the (*a*) Brewster's angle and (*b*) waveguide configurations. The inset is a sketch of a waveguide with multiple passes.

the total thickness of the active region in multiple quantum wells. For a 75-Å well and 50 periods, the total thickness of the active region is 0.375 μm for a single pass. The effective optical active region in the waveguide configuration is 0.374 times the number of passes the photons will make before exiting the sample. It is clear from Fig. 6.11 that the signal obtained from using the waveguide configuration is much larger than that obtained using the Brewster's angle configuration due to multiple passes that the photons make inside the waveguide.

The preceding discussion is directed toward n-type multiple quantum wells, such as GaAs/AlGaAs, where the well is doped with a donor such as Si. It is quite possible, however, that intersubband transitions can be observed in p-type multiple quantum wells where the dopant is an accepter such as Be. In p-type multiple quantum wells, quantum confinement is in the valence band, as shown in Fig. 6.12. The momentum matrix elements show that normal incident photon-electron coupling is possible due to bands mixing as detailed by Brown and Szmulowicz

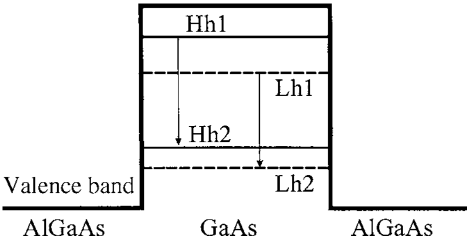


Figure 6.12 Intersubband transitions in p-type GaAs/AlGaAs multiple quantum wells shown for transitions for heavy hole (HH) bound states (solid lines) and for light hole (LH) bound state (dashed lines).

(1996) and Szmulowicz (1995). Transitions between heavy holes and light holes energy levels are also possible.

6.8 The Optical Absorption Coefficient of the Intersubband Transition in GaN/AlGaN Multiple Quantum Wells

III-nitride semiconductors have been studied for their applications in the visible and ultraviolet spectral regions. They can also be used for infrared applications by investigating the intersubband transitions in quantum structures such as GaN/AlGaN multiple quantum wells and quantum dots. The most common crystallographic structure of III-nitride materials is the wurtzite (hexagonal) structure. A large spontaneous polarization oriented along the c axis occurs due to the lack of inversion symmetry and the large ionicity associated with the covalent nitrogen bond (Bernardini et al. 1997). The electrostatic charge densities associated with the piezoelectric polarization field influence the carrier distributions, electric field, and consequently, a wide range of optical and electronic properties of nitride materials and devices. The total polarization at the GaN/AlGaN interface is the sum of the effective piezoelectric polarization and the difference spontaneous polarization (Morkoç et al. 1999 and Yu 2003). A typical value for the total polarization is on the order of $-0.096x$ C/m^2, where x is the Al mole fraction in the AlGaN (Morkoç et al. 1999). For $x = 0.3$, the total polarization is -0.0288 C/m^2. Notice that 1 C/m^2 $= 6.24 \times 10^{14}$ electrons/cm^2. Thus, one AlGaN/GaN interface can produce a total polarization charge on the order of $\sim 1.80 \times 10^{13}$ electrons/cm^2. A test of the polarization-induced charges is to measure the intersubband transition in certain GaN/AlGaN multiple quantum well structures, as shown in Fig. 6.13*a* and *b*. In Fig. 6.13*a*, we designed a sample where the well is ~35 Å GaN and the barrier is 100 Å bulk $Al_{0.35}Ga_{0.65}N$. Thus, this sample has one AlGaN/GaN interface, which contributes a sheet carrier density of $\sim 1.80 \times 10^{13}$ cm^{-2}. On the other hand, the structure in Fig. 6.13*b* is made of a similar well, but the barrier is composed of four 10 Å GaN/15 Å $Al_{0.65}Ga_{0.35}N$. Hence the total number of interfaces that contribute polarized induced charges is five (one from the well and four from the superlattice barrier). Indeed, when the intersubband transitions were measured for the two samples in the waveguide configuration, we obtained the absorbance spectra shown in Fig. 6.13*c*. By examining Fig. 6.13*c*, we noted that the spectrum intensity obtained for the sample with a superlattice barrier [spectrum (*a*)] is approximately five times larger than the spectrum intensity obtained for the sample with a bulk burrier [spectrum (2)]. Hence, the total polarization-induced sheet carrier density is $\sim 9.00 \times 10^{13}$ cm^{-2} for a sample with five interfaces. This is a straightforward test of measuring

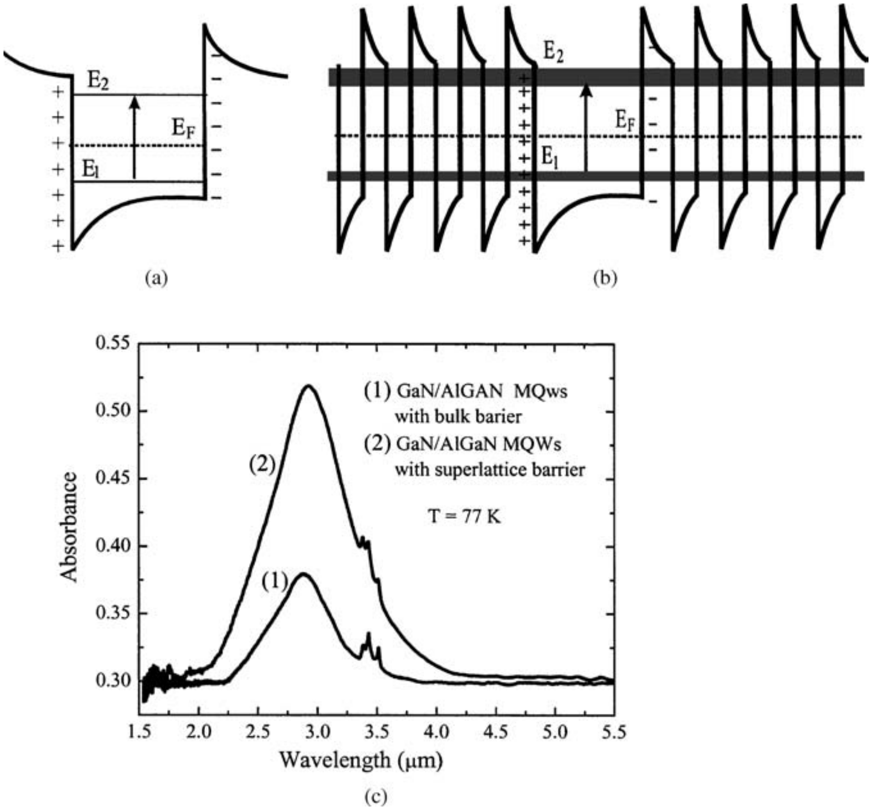


Figure 6.13 Structures of the GaN/AlGaN multiple quantum well designed such that (*a*) the barrier is bulk and (*b*) the barrier is composed of GaN/AlGaN short period superlattice. (*c*) The intersubband transition spectra for the two samples were measured at 77 K.

the polarization-induced sheet carrier density. The optical absorption coefficient of the intersubband transition in GaN/AlGaN multiple quantum wells is similar to that obtained for the intersubband transition in GaAs/AlGaAs multiple quantum wells.

III-nitride semiconductor materials have a wide range of applications, and their use as we have seen in Fig. 6.13*c* extends from the ultraviolet and visible spectrum to near the infrared spectral region. In addition to their optoelectronic applications, the III-nitride materials have been used for high-power modulation-doped field-effect transistors.

6.9 Electronic Transitions in Multiple Quantum Dots

The intersubband transitions in multiple quantum wells discussed in Sec. 6.8 show that there is a selection rule that permits electron-photon coupling at a certain angle of incident light. Maximum electron-photon coupling occurs at the Brewster's angle, while the waveguide configuration allows one to obtain a stronger absorption intensity due to the multiple passes that the light makes before exiting the sample. In the multiple quantum well case, the wave functions of the ground and excited states are highly symmetrical (s-type) and the transition is determined by the overlap integral or by the selection rules determined from the momentum matrix elements. In the case of quantum dots, there is a strong p-type mixture in the conduction band states (see Singh 2003). For an eight-band $\mathbf{k}\cdot\mathbf{p}$ model, the wave functions of the electronic energy levels in quantum dots can be written according to Singh as

$$\psi_n(x) = \sum_{j=1}^{8} \phi_{nj}(\mathbf{r}) u_j(x) \tag{6.86}$$

where ϕ_{nj} is the envelope part and u_j is the central cell part. The momentum matrix element can now be written as

$$p_{fi} = \sum_{jj'} \{\langle \phi_{fj'}|\mathbf{p}|\phi_{ij}\rangle\langle u_{j'}|u_j\rangle + \langle u_{j'}|\mathbf{p}|u_j\rangle\langle \phi_{fj'}|\phi_{ij}\rangle\} \tag{6.87}$$

The absorption coefficient can now be written for bound-to-bound transitions with a Lorentzian lineshape as

$$\alpha(\omega) = \frac{n_1 \pi e^2 N_q \hbar}{\epsilon_o c n_r m^{*2} l_{\text{av}}} \frac{1}{\hbar\omega} (\widehat{\mathcal{E}} \cdot \mathbf{p}_{f_i})^2 \frac{\Gamma}{\pi\left[(\hbar\omega - \Delta E)^2 + \Gamma^2\right]} \tag{6.88}$$

where n_1 = number of electrons per unit area in each quantum dot layer
N_q = number of quantum dot layers
l_{av} = average quantum dot layer thickness
$\widehat{\mathcal{E}}$ = direction of polarization unit vector

A typical example of a multiple quantum dot structure is depicted in Fig. 6.14*a*. In this figure we sketched $\text{In}_{0.3}\text{Ga}_{0.7}\text{As}$ triangular-shaped quantum dots grown by the molecular beam epitaxy technique using *Stranski-Krastanov* mode with GaAs being the barrier. The wetting layer and the average quantum dot height are shown. Two samples were designed such that the intersubband transition is bound-to-bound (Fig. 6.14*b*) in one sample and bound-to-continuum in the other (Fig. 6.14*c*). Waveguides were made from these two samples to allow the light to make multiple passes to increase the optical length. The average thicknesses (l_{av}) of the bound-to-bound and bound-to-continuum samples were 25 and 15 monolayers (ML), respectively. A monolayer of $\text{In}_{0.3}\text{Ga}_{0.7}\text{As}$ is approximately 2.588 Å.

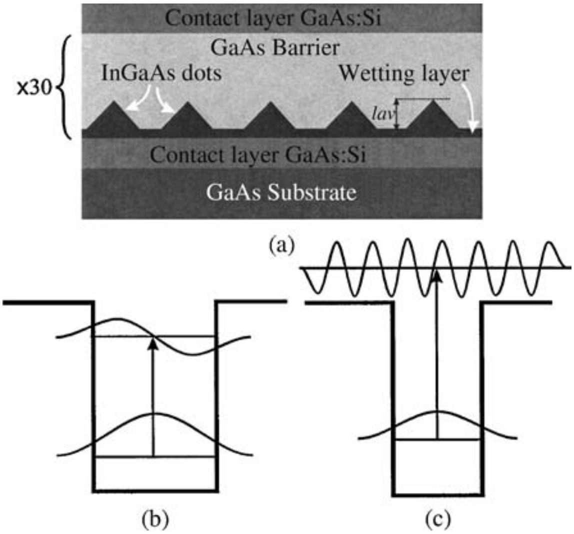


Figure 6.14 (*a*) A sketch of an InGaAs/GaAs multiple quantum dot structure used for optical absorption coefficient measurements showing the wetting layers, triangular-shaped quantum dots, and contact layers. (*b*) Bound-to-bound transition. (*c*) Bound-to-continuum transition.

Typical optical absorbance spectra for bound-to-bound and bound-to-continuum intersubband transitions obtained for multiple quantum dots are shown in Fig. 6.15. The solid lines are the experimental spectra and the dashed lines are theoretical spectra. For the bound-to-bound spectrum (*a*), a Lorentzian lineshape was used as given by Eq. (6.82). The bound-to-continuum spectrum (*b*) was fitted with a lineshape described by Eq. (6.85). The bound-to-continuum transition exhibits an asymmetrical lineshape due to the fact that the transition occurs between the bound ground state and all available states in the continuum, including the resonant state, which has the propagation property shown in Fig. 6.14*c*. Notice that the optical absorption coefficient of the intersubband transition contains the electron effective mass. This is due to the higher-order terms arising from the canonical transformations of the effective mass theory (Wallis 1958).

The sketch shown in Fig. 6.14*a* indicates that the dot size is the same for all dots in the structure. This picture, however, is very simplistic since the self-assembled quantum dot shows that there is a variation in size, shape, and strain causing a variation in the energy levels, which leads to an inhomogeneous broadening in the quantum dot ensemble properties. An example of this effect is the interband transitions in quantum dot ensembles, where the distribution of the quantum dot

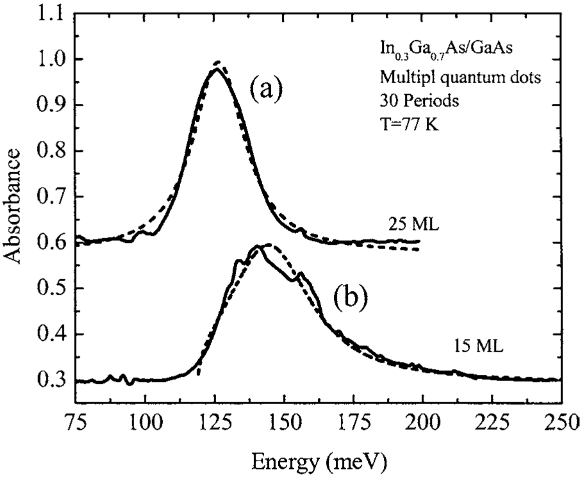


Figure 6.15 Optical absorbance spectra obtained for two multiple quantum dot structures with average dot height of (*a*) 25 monolayer (ML) and (*b*) 15 ML. The solid lines represent the experimental spectra, and the dashed lines represent the theoretical spectra.

size is assumed to be a gaussian lineshape of the form

$$G(a) = \frac{1}{\sqrt{2\pi}\,\sigma_a} \exp\left[-\frac{(a-a_o)^2}{2\sigma_a^2}\right] \tag{6.89}$$

where σ_a is the standard deviation and is given by $\sigma_a = \sqrt{\langle a - a_o\rangle^2}$. The quantum dots are assumed to be cubic in shape with an average side of length a_o (see Wu et al. 1987). The same analysis can be applied to spherical quantum dots with an average radius r_o. The optical absorption coefficient for the interband transition in a quantum dot can be written as

$$\alpha(\omega) = \frac{2\pi e^2\hbar}{\epsilon_o c n_r \hbar m_o^2 a^3}\frac{|P_n|^2}{\hbar\omega}\sum_{n,l}(2l+1)\delta\left(\hbar\omega - E_g - \frac{\pi^2\hbar^2 n^2}{2m_r^* a^2}\right) \tag{6.90}$$

where $\sum_l (2l+1)$ = degeneracy of energy level
P_n = momentum matrix element
E_g = bandgap

The selection rules dictate that $\Delta n = 0$, which means the allowed transitions are those between $\mathrm{HH_1}$ and $\mathrm{E_1}$, $\mathrm{HH_2}$ and $\mathrm{E_2}$, and so on. The reduced effective mass m_r^* was defined previously [see Eq. (6.56)]. The convolution of Eqs. (6.89) and (6.90) gives

$$\alpha(\omega) = \frac{2\pi e^2\hbar}{\epsilon_o c n_r \hbar m_o^2}\frac{|P_n|^2}{\hbar\omega}\frac{1}{\sqrt{2\pi}\sigma_a}\sum_l (2l+1)$$

$$\times \int_0^\infty \delta\left(\hbar\omega - E_g - \frac{\pi^2\hbar^2 n^2}{2m_r^* a^2}\right)\frac{1}{a^3}\exp\left[-\frac{(a-a_o)^2}{2\sigma_a^2}\right]da \qquad (6.91)$$

Letting

$$x^2 = \frac{2m_r^* a_o^2}{\pi^2\hbar^2}(\hbar\omega - E_g)$$

$$\xi = \frac{\sigma_a}{a_o},$$

$$A = \frac{2\pi e^2\hbar}{\epsilon_o c n_r \hbar m_o^2 a^3}\frac{|P_n|^2}{\hbar\omega}\frac{m_r^*}{\sqrt{2\pi}\pi^2\hbar^2}$$

we have

$$\alpha(\omega) = \frac{A}{a_o}\sum_{n,l}\frac{2l+1}{\xi n^2}\exp\left[-\frac{(n/x-1)^2}{2\xi^2}\right] \qquad (6.92)$$

A plot of Eq. (6.92) is shown in Fig. 6.16 for different values of ξ, assuming that the parameter A is the same for all transitions. The absorption peak position energies can be expressed as

$$\hbar\omega = E_g + x^2\frac{\pi^2\hbar^2}{2m_r^* a_o^2} \qquad (6.93)$$

where E_g is the bandgap of the bulk material and x can be read directly from Fig. 6.16. If one ignores the small red shift of the peak position

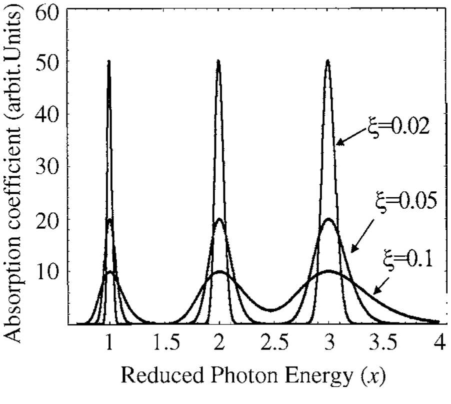


Figure 6.16 Absorption coefficient of interband transitions in quantum dot ensembles having a gaussian distribution plotted as a function of reduced photon energy. The spectra shown are for three different standard deviations.

energy due to the broadening effect, x is taken as an integer as shown in the figure. It is clear from Eq. (6.93) that the peak positions are determined solely by the size of the quantum dots for known electron and hole effective masses.

6.10 Selection Rules

6.10.1 Electron-photon coupling of intersubband transitions in multiple quantum wells

It was found experimentally that the intersubband transitions in n-type multiple quantum wells can be observed when the incident light has a polarization component in the z direction or growth direction. The light has electrical ($\mathcal{E}_x$) and magnetic (B_y) components which are orthogonal to each other and to the propagation direction, as illustrated in Fig. 6.17. For normal incident light, as illustrated in Fig. 6.17a, the electrical component is perpendicular to the z axis, and therefore, the electrical component along the z axis is zero. However, when the incident light reaches the surface of the sample at an angle φ from the

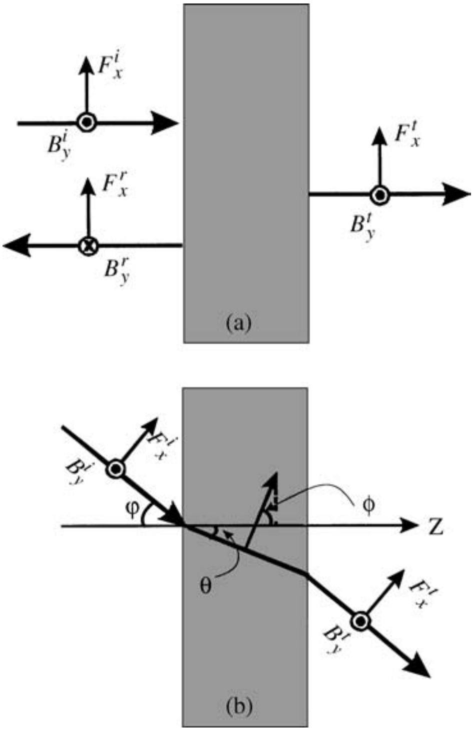


Figure 6.17 Reflection and transmission of an electromagnetic wave showing the electric ($\mathcal{E}_x$) and magnetic ($\mathbf{B}_y$) fields with respect to the direction of propagation. (a) Normal incidence gives a zero component of the electric field along the propagation direction. (b) Incidence at an angle φ from the normal yields a nonzero component of the electric field along the z direction.

normal, as illustrated in Fig. 6.17*b*, the electric field has a component in the z axis such that $\mathcal{E}_x \cdot \hat{z} = F_x \cos\phi = F_x \sin\theta$. The maximum electron-photon coupling occurs when φ is the Brewster's angle. For GaAs, with a refractive index of 3.27, φ is 73°. Thus, $\theta = \sin^{-1}[(\sin 73)/3.27] = 17°$. The electron-photon coupling selection rule for a spherically symmetric band in a quantum well is that this coupling is nonzero for photons polarized along the growth direction of the quantum well. For intersubband transitions in p-type multiple quantum wells, this selection rule is no longer valid and the photon can be absorbed at normal incident due to heavy hole and light hole wave functions mixing.

6.10.2 Intersubband transition in multiple quantum wells

The envelope functions of the bound states in the conduction quantum well are given by Eq. (6.73) for the even and odd states. The momentum matrix element is given by Eq. (6.75). Since the momentum operator, $p_z = -i\hbar d/dz$, changes the parity of the wave function, the wave functions in the integral $\langle n \,|\, p_z \,|\, n' \rangle$ must have opposite parities for a nonzero momentum matrix element. In other words, $|n\rangle$ and $|n'\rangle$ must have different parities, which leads to the selection rule $(n' - n) =$ odd integer. The same conclusion can be reached if the dipole matrix element is used instead of the momentum matrix element such that $\langle n \,|\, z \,|\, n' \rangle$. In this case, z is an odd function and, therefore, $|n\rangle$ and $|n'\rangle$ must have different parities for the integral to have a nonzero value.

6.10.3 Interband transition

The selection rules of interband transitions in multiple quantum wells can be understood by examining the wave functions of the valence and conduction bands, which can be written as

$$|i\rangle = \frac{1}{\sqrt{V}} u_v(\mathbf{r}) \varphi_{nh}(z) \exp(i\mathbf{k}_\perp \cdot \mathbf{r}_\perp) \tag{6.94a}$$

$$|f\rangle = \frac{1}{\sqrt{V}} u_c(\mathbf{r}) \varphi_{n'e}(z) \exp(i\mathbf{k}'_\perp \cdot \mathbf{r}'_\perp) \tag{6.94b}$$

where $|\mathbf{k}_\perp| = \sqrt{k_x^2 + k_y^2}$

$|\mathbf{r}_\perp| = \sqrt{x^2 + y^2}$

$u_v(\mathbf{r}), u_c(\mathbf{r}) =$ envelope functions for valence and conduction bands, respectively

$\varphi_{nh}(z), \varphi_{n'e}(z) =$ wave functions for bound states in valence and conduction bands, respectively

The exponentials are the plane waves for free motion in the xy plane. Either the momentum or the dipole matrix element can be used to determine the allowed interband transitions. The conservation of momentum allows one to set $\mathbf{k}_\perp = \mathbf{k}'_\perp$ since the photon momentum is very small compared to the electron momentum. The matrix element M can now be written as

$$M = \langle f|\mathbf{r}|i\rangle \tag{6.95}$$

For quantum wells, we have

$$\langle f|x|i\rangle = \langle f|y|i\rangle \neq \langle f|z|i\rangle \tag{6.96}$$

Since the interband transition is in a plane perpendicular to the growth axis or z direction, we are concerned about evaluating the matrix element along the x or y direction, which yields

$$\begin{aligned} M = \langle f|x|i\rangle &= \frac{1}{V}\int\int u_c^*(\mathbf{r})\varphi_{n'e}^*(z)xu_v(\mathbf{r})\varphi_{nh}(z)d^3\mathbf{r}\,dz \\ &= \frac{1}{V}\int u_c^*(\mathbf{r})xu_v(\mathbf{r})d^3\mathbf{r}\int \varphi_{n'e}^*(z)\varphi_{nh}(z)\,dz \\ &= \frac{1}{V}\langle u_c|x|u_v\rangle\langle n'e|nh\rangle = M_{cv}M_{nn'} \end{aligned} \tag{6.97}$$

where

$$\frac{1}{V}\langle u_c|x|u_v\rangle = M_{cv} \tag{6.98}$$

and

$$\langle n'e|nh\rangle = M_{nn'} \tag{6.99}$$

where $M_{nn'}$ is known as the electron-hole overlap. If one assumes that the wave functions of the bound states in the valance and the conduction bands have the forms

$$\begin{aligned} \varphi_{nh}(z) &= \sqrt{\frac{2}{L}}\cos\left(\frac{n\pi z}{L} + \frac{n\pi}{2}\right) \\ \varphi_{n'e}(z) &= \sqrt{\frac{2}{L}}\cos\left(\frac{n'\pi z}{L} + \frac{n'\pi}{2}\right) \end{aligned} \tag{6.100}$$

then the electron-hole overlap integral is

$$M_{nn'} = \frac{2}{L}\int_{-L/2}^{L/2} \cos\left(\frac{n\pi z}{L} + \frac{n\pi}{2}\right)\cos\left(\frac{n'\pi z}{L} + \frac{n'\pi}{2}\right)dz = \delta_{nn'} \tag{6.101}$$

From this equation, we have the following selection rule: $n = n'$ or $\Delta n = 0$.

6.11 Excitons

A brief discussion of excitons in bulk semiconductors and low-dimensional systems is presented in this section. Excitons in GaN thin films are discussed as an example for bulk materials. Then, excitons in quantum wells and quantum dots are discussed.

6.11.1 Excitons in bulk semiconductors

Excitons are quasi-particles used to describe electron-hole pairs coupled by Coulomb interaction in a manner similar to the hydrogen atom. As mentioned in the introduction of this chapter, there are two types of excitons; free and bound excitons as illustrated in Fig. 6.2. Excitons in semiconductors are stable so long as their binding energy is smaller than the thermal energy ($k_B T$). The optical absorption and photoluminescence emission of excitons affect the optical properties of the band edge of semiconductors and their heterojunctions. Exciton absorption is profound at low temperatures in most direct bandgap semiconductor materials, and it can even be observed at room temperature in semiconductors, such as GaN, where the binding energy of the exciton is slightly larger than the room temperature thermal energy. Figure 6.18

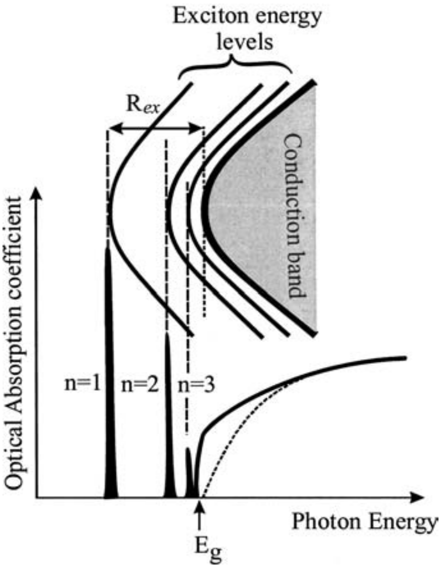


Figure 6.18 Illustration of the band edge absorption of a direct semiconductor in the absence (dashed curve) and in the presence of excitons (solid lines). The exciton energy levels ($n = 1$, 2, and 3) are shown. R_{ex} is the exciton binding energy in bulk semiconductors.

illustrates how the free exciton affects the band edge absorption of a pure semiconductor material at low temperature. The dashed line in this figure depicts the band edge absorption of a direct bandgap semiconductor without the exciton effect. The solid curve is the band edge absorption with the exciton effect included. The lines labeled $n = 1, 2,$ and 3 are the excitonic energy levels.

To obtain the exciton energy levels, one needs to solve the Schrödinger equation for a two-body problem. By considering the relative motion of the electron-hole system and ignoring the motion of the center of mass (the kinetic energy of the center of mass which is translation invariant), the Schrödinger equation can be written as

$$\left[-\frac{\hbar^2\mathbf{k}^2}{2\mu^*} - \frac{e^2}{4\pi\epsilon\epsilon_o r}\right]\psi_{\mathrm{ex}} = E_n\psi_{\mathrm{ex}} \tag{6.102}$$

where the first term is the relative motion of the electron-hole system (kinetic energy), the second term is the Coulomb interaction energy between the electron and hole, ϵ is the dielectric constant of the material, r is the distance between the electron and hole, and m^* is the exciton reduced effective mass $[1/\mu^* = (1/m_e^*) + (1/m_h^*)]$. The exciton wave function can be written as

$$\psi_{\mathrm{ex}} \propto \chi(r)\phi_c(r_e)\phi_v(r_h) \tag{6.103}$$

where $\chi(r)$ is the envelope function and $\phi_c(r_e)$ and $\phi_v(r_h)$ are Wannier functions that represent the electron and hole band edge states, respectively. Equation (6.102) can be solved in a manner similar to the hydrogen atom, where the energy levels can be written as

$$E_n = -\frac{\mu^* e^4}{2(4\pi\epsilon\epsilon_o)^2\hbar^2 n^2} = -\frac{\mu^*}{m_o\epsilon^2}\frac{R_H}{n^2} \tag{6.104}$$

where m_o is the free electron mass and R_H is the hydrogen atom Rydberg constant given by $R_H = m_o e^4/[2(4\pi\epsilon_o)^2\hbar^2] = 13.60$ eV. The quantity $R_{\mathrm{ex}} = \mu^* R_H/(m_o\epsilon^2)$ can now be called the exciton Rydberg constant. The radius of the electron-hole orbit can be written as

$$r_n = \frac{4\pi\epsilon\epsilon_o\hbar^2 n^2}{\mu^* e^2} = \frac{m_o\epsilon n^2}{\mu^*}a_H = n^2 a_{\mathrm{ex}} \tag{6.105}$$

where a_H is the Bohr radius of the hydrogen atom given by $a_H = 4\pi\epsilon_o\hbar^2/(m_o e^2) = 0.5293$ Å, and a_{ex} is the exciton Bohr radius. The exciton binding energy can be taken as R_{ex} or the energy of the ground

TABLE 6.1 Several Well-known Direct Bandgap Semiconductor Materials and Their Properties[a]

Material	E_g, eV	m_e^*	m_h^*	ϵ	$E_{\text{ex}}^{\text{bulk}}$, meV	$a_{\text{ex}}^{\text{bulk}}$, Å
InSb	0.23	0.013	0.40	16.8	0.61	706.2
InAs	0.35	0.027	0.40	15.15	1.50	317.0
GaSb	0.75	0.042	0.40	15.69	2.10	218.5
GaAs	1.52	0.067	0.54	13.18	4.67	117.0
InP	1.35	0.073	0.64	12.56	5.65	101.4
CdTe	1.48	0.086	0.60	10.6	10.02	67.8
ZnTe	2.39	0.12	1.30	8.7	19.90	41.9
GaN	3.44	0.20	0.60	9.50	22.60	33.5
ZnO	3.28	0.24	0.78	8.1	33.00	23.4

[a] E_g = bandgap; m_e^*, m_h^* = electron and heavy hole effective mass, respectively, in units of electron mass (9.11×10^{-31} kg); ϵ = dielectric constant; $E_{\text{ex}}^{\text{bulk}}$ = exciton binding energy; $a_{\text{ex}}^{\text{bulk}}$ = exciton radius.

state. For example, the binding energy of a free exciton in GaAs is found to be 4.35 meV, assuming that the electron and heavy hole effective masses are $0.067m_o$ and $0.54m_o$, respectively, and the dielectric constant is taken as 13.6. The Bohr radius of the free exciton in GaAs is calculated to be 12.07 nm. The equivalent temperature to the free exciton binding energy is ~50 K. Thus, the free exciton in GaAs is stable at temperatures below 50 K. In a highly pure GaAs sample with high mobility, the free exciton is observed at temperatures as high as 180 K. As a comparison, we calculated the binding energy of the free exciton in GaN material to be ~23.4 meV for electron and heavy hole effective masses of $0.20m_o$ and $0.60m_o$, respectively, and for a dielectric constant of 9.2. The Bohr radius is found to be ~3.24 nm. The equivalent temperature to the exciton in GaAs is ~271 K, which means that free excitons can be observed at room temperature in relatively pure GaN samples. The exciton binding energy and radius were calculated for a few semiconductor materials and are given in Table 6.1. The exciton binding energy increases as the bandgap increases. On the other hand, the exciton radius decreases with increasing bandgap.

Excitons are mostly observed at the high symmetry points in the Brillouin zone such as the Γ-point (the center of the Brillouin zone). At these points the slopes of the energy bands are zero and the group velocities of electrons and holes are the same, which is a necessary condition to observe excitons. The excitonic energy levels in a direct bandgap semiconductor can be written as

$$E_n = E_g - \frac{\mu^* e^4}{2(4\pi\epsilon\epsilon_o)^2\hbar^2 n^2} = E_g - \frac{R_{\text{ex}}}{n^2} \qquad (6.106)$$

If the motion of the center of mass of the exciton is included in the Schrödinger equation, the excitonic energy levels become

$$E_n = E_g + \frac{\hbar^2 k_{\text{ex}}^2}{2M} - \frac{R_{\text{ex}}}{n^2} \tag{6.107}$$

where $\mathbf{k}_{\text{ex}}$ is the exciton wavevector and $M = m_e^* + m_h^*$. The exciton center of mass in this equation behaves like a particle with mass M and a wavevector $\mathbf{k}_{\text{ex}}$. The translational energy of the exciton center of mass, which is usually very small, can be dropped from Eq. (6.107). The driving force of exciton generation is the Coulomb interaction between the electron-hole pair. If this interaction is zero, the exciton energy levels will vanish. The exciton functions are hydrogen atom-like functions. For example, the ground-state wave function can be written as

$$\phi^{100}(\mathbf{r}) = \frac{1}{\sqrt{\pi a_{\text{ex}}^3}} e^{-\mathbf{r}/a_{\text{ex}}} \tag{6.108}$$

The free-exciton radius in many semiconductors that have a bandgap in the range of 1 to 2 eV is on the order of 100 Å, which means that the exciton is spread over many unit cells as shown in Fig. 6.2. In wide-bandgap materials, such as GaN and ZnO, the exciton radius is smaller, but the binding energy is larger and free excitons are observable even at room temperature. The stability of the exciton at room temperature is very important for exciton-based device applications.

The optical absorption spectra of excitons have been reported for many direct bandgap semiconductor materials. For example, optical absorption measurements on wurtzite GaN thin films, grown on sapphire, exhibit three free excitons, as shown in Fig. 6.19. A room temperature spectrum shows excitonic behavior near the band edge absorption. When the sample is cooled to 10 K, the spectrum shows the three excitons, lines A, B, and C. These excitons are usually observed in epitaxially grown thin films with thicknesses ranging from 0.1 to 1.0 μm. Within this thickness range, absorption above the bandgap is possible since thick layers tend to absorb and/or reflect light just above the band edge. An alternative technique used to measure the absorption coefficient above the bandgap is ellipsometry. The absorption coefficient can then be obtained from the imaginary part of the dielectric constant.

The origin of the A, B, and C excitons can be understood by examining Fig. 6.20. In this figure, the band structure is sketched at the center of the Brillouin zone where the wurtzite structure has a nondegenerate energy level for the conduction band (CB) and a degenerate energy level for the valence band (VB). The valence band energy level splits into two energy levels (Γ_1, Γ_5) under the action of the axial crystal field (Δ_{cr}).

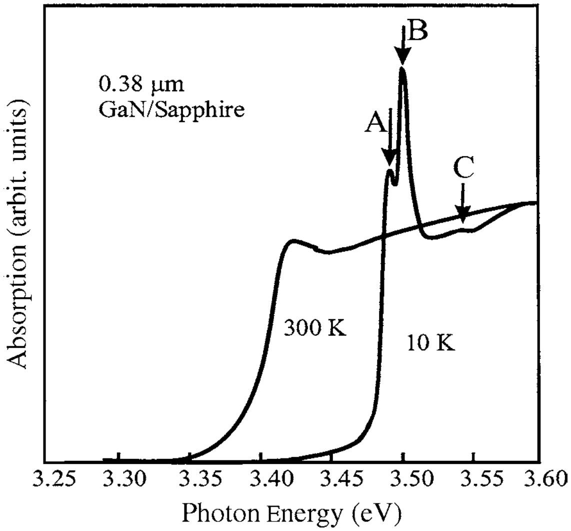


Figure 6.19 The optical absorption coefficient spectra of GaN thin film measured at 300 and 10 K. The three exciton lines, *A*, *B*, and *C*, are clearly visible in the spectrum measured at 10 K.

The spin-orbit interaction (Δ_{so}) causes a similar effect on the valence band. The combined actions of Δ_{cr} and Δ_{so} result in splitting of the valence band into three energy levels, labeled Γ_9, Γ_7, and Γ_7. The transitions from the three valence energy levels to a single conduction band energy level dominate the optical absorption near the band edge of GaN.

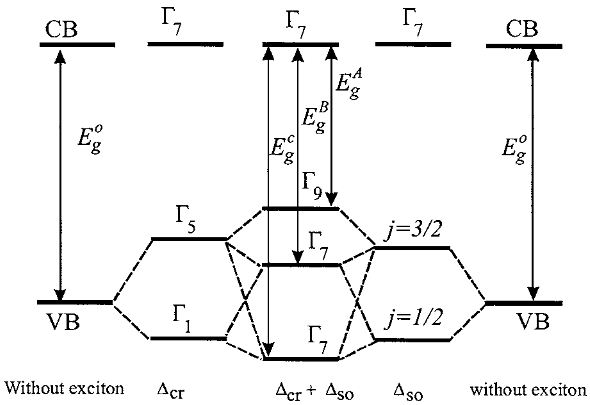


Figure 6.20 A sketch of the band structure near the fundamental band edge in wurtzite GaN showing the effect of the crystal field and spin-orbit interactions on the valence band.

The three exciton transitions are labeled E_g^A, E_g^B, and E_g^C, while E_g^o is the bandgap transition in the absence of the exciton effect.

The excitonic energy levels in GaN have been investigated by many techniques. GaN thin films are epitaxially grown on lattice mismatched substrates, such as sapphire or SiC. The interfaces are usually plagued by dislocations and extended defects. The structural property of the thin film usually improves as the thickness of the layer increases. Conversely, optical absorption above the band edge becomes difficult as the layer thickness increases. However, the photoreflectance technique is useful in this case, where the excitonic bound states are probed. Fig. 6.21 is a typical photoreflectance spectrum measured at 10 K for a 7.2-μm-thick GaN thin film grown on sapphire. The exciton transition for $n = 1, 2$ and ∞ are shown for excitons A and B. The reflectance from exciton C is very weak due to the fact that its energy is way above the bottom of the conduction band. In addition to the splitting of the valence band under the crystal field and spin-orbit interactions, a fine structure splitting of the exciton lines (on the order of 1 meV) due to electron-hole exchange interaction may occur. This fine structure, however, has not been observed yet.

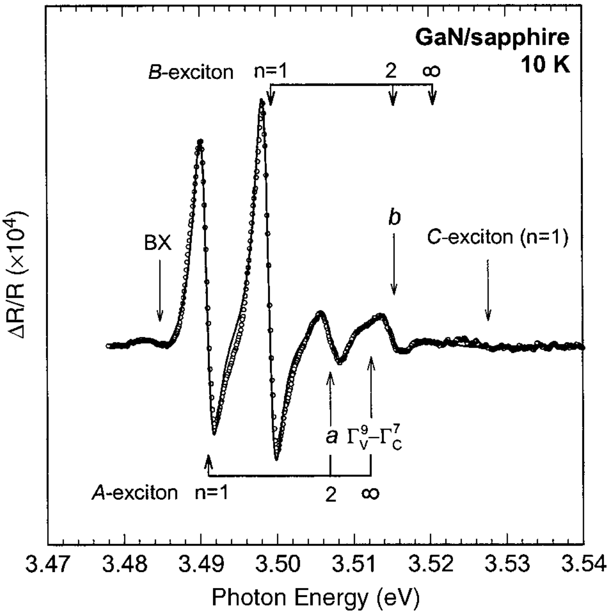


Figure 6.21 Photoreflectance spectrum from 7.2-μm-thick GaN grown by metal-organic chemical vapor deposition (MOCVD) on (0001) sapphire substrate is shown as a function of photon energy. The excitonic energy levels ($n = 1$ and 2) are shown for excitons A and B. (*After Schmidt and Song 2002.*)

The optical absorption of the exciton in bulk semiconductors was derived by Elliot (1957) and is given by

$$\alpha_{\text{ex}} = \begin{cases} \alpha_o \dfrac{2\pi\sqrt{a_{\text{ex}}}}{\sqrt{\hbar\omega - E_g}} & \text{for } \hbar\omega \approx E_g \\ \alpha_o \dfrac{\pi\zeta \exp(\pi\zeta)}{\sinh(\pi\zeta)} & \text{for } \hbar\omega > E_g \end{cases} \tag{6.109}$$

where α_o is the optical absorption in the absence of Coulomb interaction [see Eq. (6.68)] and $\zeta = a_{\text{ex}}/\sqrt{\hbar\omega - E_g}$.

6.11.2 Excitons in quantum wells

The Hamiltonian of the exciton in a quantum well does not have a simple analytical solution, but the problem can be solved using the variational method described in Chap. 1. If the variational wave function is assumed to have the form

$$\phi_n^{11}(\mathbf{r}) = \sqrt{\frac{2}{\pi a_{\text{ex}}^2}}\, e^{-\mathbf{r}/a_{\text{ex}}} \tag{6.110}$$

where $\mathbf{r}$ is the relative coordinate of the electron and hole in the xy plane, then the Hamiltonian in polar coordinates can be written as

$$\mathbf{H} = -\frac{\hbar^2}{2\mu^*}\left[\frac{1}{\mathbf{r}}\frac{\partial}{\partial \mathbf{r}}\left(\mathbf{r}\frac{\partial}{\partial \mathbf{r}}\right) + \frac{1}{\mathbf{r}^2}\frac{\partial^2}{\partial\theta^2}\right] - \frac{e^2}{4\pi\epsilon\epsilon_o\mathbf{r}} \tag{6.111}$$

Using the variational method, one can write the energy expectation value as

$$\langle E\rangle = \frac{\int \phi_n^{11^*}(\mathbf{r},\theta) H \phi_n^{11}(\mathbf{r},\theta)\mathbf{r}\, d\mathbf{r}\, d\theta}{\int \phi_n^{11^*}(\mathbf{r},\theta)\phi_n^{11}(\mathbf{r},\theta)\mathbf{r}\, d\mathbf{r}\, d\theta} \tag{6.112}$$

To simplify the solution of Eq. (6.112), let us set $\xi = 1/a_{\text{ex}}$, where ξ is the variational parameter. Equation (6.112) can be rewritten as

$$\begin{aligned}\langle E\rangle &= \int \exp(-\mathbf{r}\xi)\left\{-\frac{\hbar^2}{2\mu^*}\left[\frac{1}{\mathbf{r}}\frac{\partial}{\partial \mathbf{r}}\left(\mathbf{r}\frac{\partial}{\partial \mathbf{r}}\right) + \frac{1}{\mathbf{r}^2}\frac{\partial^2}{\partial\theta^2}\right] - \frac{e^2}{4\pi\epsilon\epsilon_o\mathbf{r}}\right\} \\ &\qquad \times \exp(-\mathbf{r}\xi)\mathbf{r}\, d\mathbf{r}\, d\theta \\ &= \left[\frac{\hbar^2\xi^2}{2\mu^*} - \frac{2e^2\xi}{4\pi\epsilon\epsilon_o}\right]\end{aligned} \tag{6.113}$$

The energy can now be maximized as follows. Take the first derivative of $\langle E\rangle$with respect to ξ and equate the results to zero to obtain

$$\zeta = \frac{2\mu^* e^2}{\hbar^2 4\pi\epsilon\epsilon_o} \tag{6.114}$$

Substituting Eq. (6.114) back into (6.113), the exciton ground-state energy E_o^{2D} is

$$E_o^{2D} = -\frac{2\mu^* e^4}{\hbar^2 (4\pi\epsilon\epsilon_o)^2} = -\frac{4\mu^* e^4}{2\hbar^2 (4\pi\epsilon\epsilon_o)^2} = -4E_o^{3D} \tag{6.115}$$

where E_o^{3D} is the exciton ground-state energy of the bulk material. One can generalize this result according to the following expression:

$$E_n^{2D} = -\frac{R_{\text{ex}}}{\left(n - \frac{1}{2}\right)^2} \qquad n = 1, 2, 3 \ldots \tag{6.116}$$

The exciton binding energy in quantum wells is equal to $4R_{\text{ex}}$. The Bohr radius of the exciton in a quantum well (a_{ex}^{2D}) is the inverse of the quantity given by Eq. (6.114), which is one-half the exciton radius (a_{ex}^{3D}) in bulk material ($a_{\text{ex}}^{2D} = a_{\text{ex}}^{3D}/2$). The theoretical value of the exciton binding energy in a quantum well presented in Eq. (6.115) is the upper limit, which is a difficult limit to reach experimentally. Experimentally, the exciton binding energy in quantum wells is $\sim 2.5E_o^{3D}$. This is still a substantial enhancement of the exciton binding energy, which is very important for many optoelectronic device applications. An example of exciton binding energy in quantum wells is shown in Fig. 6.22, where the binding energy is calculated as a function of the well width of an infinite-depth CdTe quantum well (see Harrison 2000). The binding

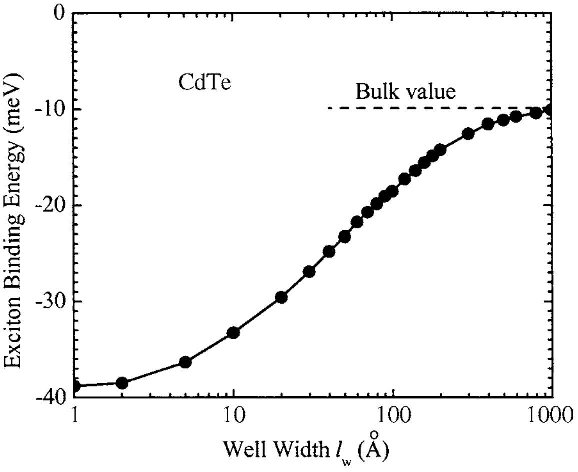


Figure 6.22 Exciton binding energy in an infinite-depth CdTe deep quantum well. (*After Harrison 2000.*)

energy limits in this figure satisfy the bulk limit (−10.1 meV) when the well thickness is too large (1000 Å) and the quantum well theoretical limit [~4 × (−10.1) meV] when the well thickness approaches zero. Notice that the energy is negative, which implies that the exciton ground state is a bound state.

6.11.3 Excitons in quantum dots

The calculation of exciton binding energy in quantum dots is very complicated. The most common approximation used is the variational method, which requires knowledge of a trial function and a Hamiltonian as discussed in Sec. 6.11.2. The analysis for quantum dots is more complicated due to the variation of the size and shape of the quantum dots. Generally speaking, the exciton binding energy is much higher in the case of quantum wells and dots as compared to bulk materials due to electron and hole confinement in quantum structures. Accompanying the increase in exciton binding energy is a reduction of the exciton Bohr radius. While reports in the open literature indicate a variety of results for various quantum dot shapes and sizes, a general consensus is that the binding energy in quantum dots increases as the size of the dot decreases. An example is reported by Grundmann et al. (1995) and shown in Fig. 6.23, where the exciton binding energy is plotted as a function of the base length of an InAs/GaAs pyramidal quantum dot size. The exciton binding energies (R_{ex}) in bulk GaAs and InAs are indicated in the figure. The behavior of the exciton binding energy in quantum dots shown in

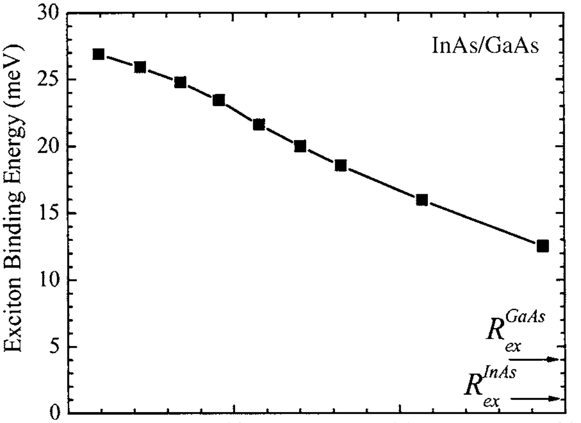


Figure 6.23 Exciton binding energy as a function of the InAs/GaAs pyramid base length. (*After Grundmann et al. 1995*).

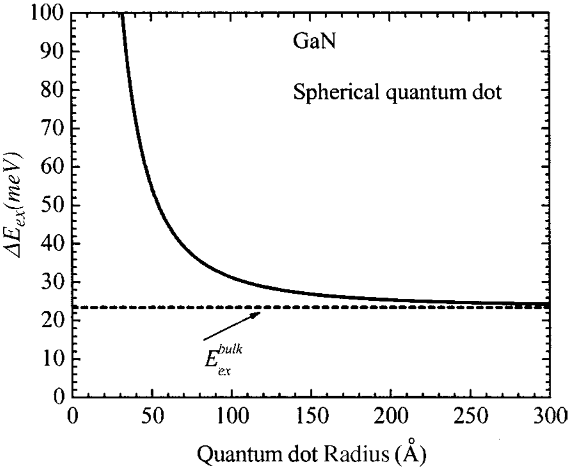


Figure 6.24 Variation of change in exciton binding energy as a function of dot size for GaN. (*After Strenger and Bajaj 2003.*)

Fig. 6.23 is the trend that most theoretical calculations exhibit. For example, the change in the binding energy in an ionic semiconductor spherical quantum dot was shown recently to have the following form (see Stenger and Bajaj 2003)

$$\Delta E_{\text{ex}} \propto \frac{\hbar^2 \pi^2}{2\mu^* R^2} + E_{\text{ex}}^{\text{bulk}} \tag{6.117}$$

where R is the radius of the sphere and $E_{\text{ex}}^{\text{bulk}}$ is the exciton binding energy in bulk materials. A plot of this equation is shown in Fig. 6.24 for a GaN spherical quantum dot. The dashed line in Fig. 6.24 represents the exciton binding energy in a bulk GaN material, and the solid line represents the change in the exciton binding energy as a function of the quantum dot radius. Again, the behavior of the binding energy in Fig. 6.24 seems to be universal for most quantum dot materials. While optical absorption from excitons in quantum dots has not been reported, perhaps due to the significantly small optical length, theoretical results indicate that excitons in quantum dots could have large oscillator strengths (see, for example, Bimberg et al. and references therein).

6.12 Cyclotron Resonance

The cyclotron resonance technique has been used to determine the effective masses of charge carriers in high-purity bulk semiconductors

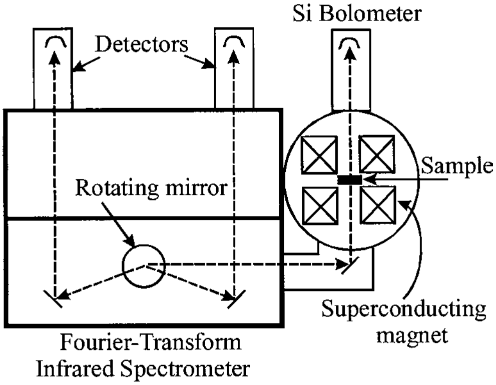


Figure 6.25 Cyclotron resonance experimental set up for the measurement of charge carrier effective masses.

as well as heterojunctions and quantum wells. This technique requires both electric and magnetic fields. Early experiments used microwave radiation in conjunction with the magnetic field. Then, the technique was developed to incorporate an infrared laser light instead of microwave radiation. With this configuration, the magnetic field is swept over a specific range to obtain the cyclotron resonance spectrum. Most recently, infrared light was used instead of laser light. The latter configuration provided a quick determination of the effective masses since the magnetic field is fixed and the infrared radiation is scanned using Fourier-transform infrared spectroscopy, as shown in Fig. 6.25.

An electron with a charge e and velocity $\mathbf{v}$, under the influence of a magnetic field $\mathbf{B}$, will experience a force $\mathcal{F}$ (Lorentz force) given by

$$\mathcal{F} = \mathbf{e}\mathbf{v} \times \mathbf{B} \tag{6.118}$$

When $\mathbf{v}$ is perpendicular to $\mathbf{B}$, the magnitude of this force is $|\mathcal{F}| = |\mathbf{e}|\mathbf{v}B$. On the other hand, the centripetal force $\mathcal{F}_c$, due to a uniform circular motion of a particle with mass m and acceleration v^2/r, is given by

$$\mathcal{F}_c = m\frac{v^2}{r} \tag{6.119}$$

By equating the two forces in Eqs. (6.118) and (6.119), one can obtain the radius r and the period T of the circular orbit as

$$r = \frac{mv}{eB} \quad \text{and} \quad T = \frac{2\pi r}{v} \tag{6.120}$$

The angular frequency of the particle is called the cyclotron frequency, which is given by

$$\omega_c = \frac{2\pi}{T} = \frac{v}{r} = \frac{|e|B}{m} \tag{6.121}$$

This equation can be modified for charge carriers in semiconductors as

$$\omega_c = \frac{|e|B}{m^*} \tag{6.122}$$

where m^* is the charged particle effective mass. For cyclotron resonance to work, the electric field $\mathcal{E}$ of the radiation should have a nonzero component in the plane of the cyclotron motion. The cyclotron resonance condition occurs when the radiation energy is equal to the energy needed for the charge carrier to make a transition between adjacent Landau energy levels. To determine the cyclotron resonance condition, the equation of motion of a free charged particle under the influence of electromagnetic and magnetic fields can be written as

$$m^* \left(\frac{d\boldsymbol{\upsilon}}{dt} + \frac{\boldsymbol{\upsilon}}{\tau} \right) = e(\boldsymbol{\mathcal{E}} + \boldsymbol{\upsilon} \times \mathbf{B}) \tag{6.123}$$

where the first term in parentheses on the left-hand side is due to the particle acceleration and the second term is due to collisions, which are characterized by the relaxation time τ of the carriers. The velocity $\boldsymbol{\upsilon}$ is the drift velocity under the influence of the electric field $\boldsymbol{\mathcal{E}}$. The magnetic field associated with the electromagnetic radiation is too small compared to the applied magnetic field, and hence, it is ignored.

Let us take the polarization of the electric field along the x direction and the magnetic field along the z direction. Since the particle motion is in the xy plane, the drift velocity has two components, one in each of the x and y directions. Thus, we can write the electric, magnetic, and velocity, fields, respectively, as

$$\begin{aligned} \boldsymbol{\mathcal{E}} &= \mathcal{E}_x \exp(i\omega t) \\ \mathbf{B} &= B_z \\ \boldsymbol{\upsilon} &= (\upsilon_x \hat{x} + \upsilon_y \hat{y}) \exp(i\omega t) \end{aligned} \tag{6.124}$$

Substituting Eq. (6.124) into (6.123) we have

$$m^* \left(-i\omega + \frac{1}{\tau} \right) \upsilon_x = e\mathcal{E}_x + e\upsilon_y B_z \tag{6.125a}$$

$$m^* \left(-i\omega + \frac{1}{\tau} \right) \upsilon_y = -e\upsilon_x B_z \tag{6.125b}$$

Solving these two equations for υ_x, we have

$$\upsilon_x = \frac{e\mathcal{E}_x}{m^*} \frac{-i\omega + \tau^{-1}}{(-i\omega + \tau^{-1})^2 + \omega_c^2} \tag{6.126}$$

where ω_c is defined by Eq. (6.122). The current density in the x direction can be written as

$$j_x = eN\upsilon_x = \sigma(\omega)\mathcal{E}_x \tag{6.127}$$

where N is the number of carriers. Combine Eqs. (6.127) and (6.126) to yield

$$\begin{aligned} j_x &= \frac{e^2 N \mathcal{E}_x}{m^*} \frac{-i\omega + \tau^{-1}}{(-i\omega + \tau^{-1})^2 + \omega_c^2} \\ &= \frac{e^2 N \tau \mathcal{E}_x}{m^*} \frac{1 - i\omega\tau}{\left(\omega_c^2 - \omega^2\right)\tau^2 + 1 - 2i\omega\tau} \end{aligned} \tag{6.128}$$

This leads to the following expression for the conductivity:

$$\begin{aligned} \sigma(\omega) &= \frac{e^2 N \tau}{m^*} \frac{1 - i\omega\tau}{\left(\omega_c^2 - \omega^2\right)\tau^2 + 1 - 2i\omega\tau} \\ &= \sigma_o \frac{1 - i\omega\tau}{\left(\omega_c^2 - \omega^2\right)\tau^2 + 1 - 2i\omega\tau} \end{aligned} \tag{6.129}$$

where σ_o is the dc conductivity given by $\sigma_o = e^2 N \tau / m^*$. The frequency-dependent conductivity is a complex quantity given by

$$\sigma(\omega) = \sigma_1(\omega) + i\sigma_2(\omega) \tag{6.130}$$

where $\sigma_1(\omega)$ and $\sigma_2(\omega)$ are the real and imaginary parts, respectively, of the conductivity and are given by

$$\sigma_1(\omega) = \sigma_o \frac{1 + \left(\omega_c^2 + \omega^2\right)\tau^2}{\left[1 + \left(\omega_c^2 - \omega^2\right)\tau^2\right]^2 + 4\omega^2\tau^2} \tag{6.131a}$$

$$\sigma_2(\omega) = \sigma_o \frac{2\omega\tau - \omega\tau\left[1 + \left(\omega_c^2 - \omega^2\right)\tau^2\right]}{\left[1 + \left(\omega_c^2 - \omega^2\right)\tau^2\right]^2 + 4\omega^2\tau^2} \tag{6.131b}$$

In the Faraday configuration, where the electric field is perpendicular to the magnetic field, the power absorbed by the carriers is given by

$$P(\omega) = \mathrm{Re}(j_x\mathcal{E}_x) = \sigma_o|\mathcal{E}_x|^2 \frac{1 + \left(\omega_c^2 + \omega^2\right)\tau^2}{\left[1 + \left(\omega_c^2 - \omega^2\right)\tau^2\right]^2 + 4\omega^2\tau^2} \tag{6.132}$$

For $\omega = \omega_c$ and $\omega_c\tau \gg 1$, the power is reduced to

$$P(\omega_c) = \frac{1}{2}\sigma_o|\mathcal{E}_x|^2 \tag{6.133}$$

A plot of the relative power absorbed $[P(\omega)/P_o]$, where $P_o = \sigma_o|\mathcal{E}_x|^2$, is shown in Fig. 6.26 for a GaAs/AlGaAs high electron mobility transistor

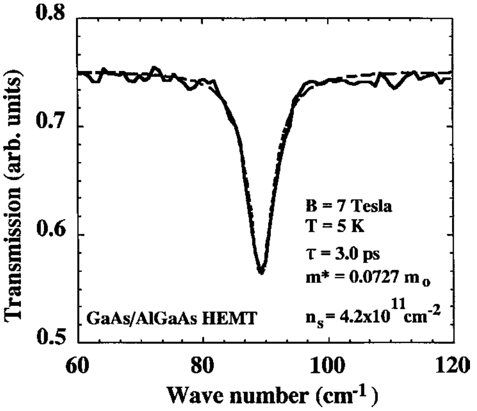


Figure 6.26 Cyclotron resonance spectrum (solid line) obtained for a GaAs/AlGaAs high electron mobility transistor (HEMT) structure. The dashed line is the theoretical fit using Eq. (6.132).

structure. This figure shows the transmission spectrum obtained from the Fourier-transform infrared spectroscopy setup shown in Fig. 6.25. The dashed line is a fit of the experimental spectrum using Eq. (6.132). The effective mass was determined from the cyclotron frequency, which was used as a fitting parameter. Notice that the electron effective mass of $0.0727m_o$ is slightly larger than the electron effective mass in bulk materials. The larger effective mass in heterostructures has been observed in many semiconductor quantum structures. Notice that Fig. 6.26 displays the transmission spectrum. Taking the negative of $[P(\omega)/P_o]$ is the proper form for fitting the transmission experimental results.

The cyclotron resonance is due to energy level quantization in the presence of electric and magnetic fields. These energy levels are known as Landau levels and can be obtained by solving the Schrödinger equation. Neglecting the crystal potential, the Hamiltonian can be written as

$$H = \frac{1}{2m^*}\left[P_x^2 + (P_y - e\mathbf{A})^2 + P_z^2\right] + g^*\mu_B\sigma_z B \qquad (6.134)$$

where $\mathbf{A}$ = vector potential given in Landau gauge as $\mathbf{A} = (0, Bx, 0)$
μ_B = Bohr magneton given by $e\hbar/(2m^*)$
g^* = effective g-factor
σ_z = electron spin quantum number given by $\pm 1/2$.

Notice that the magnetic induction ($\mathbf{H}$) is given by $\mathbf{H} = \mu_o\mathbf{B} = \nabla \times \mathbf{A}$. The wave function is an envelope function, which can be written as

$$\psi(x, y, z) = \exp(ik_y y + ik_z z)u(x) \qquad (6.135)$$

The Schrödinger equation can be written as

$$\left\{\frac{1}{2m^*}\left[P_x^2+(P_y-e\mathbf{A})^2+P_z^2\right]+g^*\mu_B\sigma_z B\right\}\psi(x,y,z)=E\psi(x,y,z) \tag{6.136}$$

Use momentum operators to give

$$\frac{\partial^2 u(x)}{\partial x^2}-\left(k_y-\frac{eBx}{\hbar}\right)^2 u(x)+\frac{2m^*}{\hbar^2}E'u(x)=0 \tag{6.137}$$

where

$$E'=E-\frac{\hbar^2 k_z^2}{2m^*}-g^*\mu_B\sigma_z B \tag{6.138}$$

Equation (6.137) can be rewritten as

$$-\frac{\hbar^2}{2m^*}\frac{\partial^2 u(x)}{\partial x^2}+\frac{m^*}{2}\left(\frac{eBx}{m^*\hbar}-\frac{\hbar k_y}{m^*}\right)^2 u(x)=E'u(x) \tag{6.139}$$

This equation is a one-dimensional harmonic oscillator equation with frequency ω_c and energy given by

$$E'=\left(n+\frac{1}{2}\right)\hbar\omega_c \tag{6.140}$$

Substituting Eqs. (6.138) into (6.140), we have

$$E_n=\left(n+\frac{1}{2}\right)\hbar\omega_c+\frac{\hbar^2 k_z^2}{2m^*}+g^*\mu_B\sigma_z B \tag{6.141}$$

As this equation indicates, the electronic energy level E will split under the influence of a magnetic field into Landau energy levels, with n being the Landau quantum number, separated by $\hbar\omega_c$. Furthermore, each Landau level will split into two levels due to the inclusion of electron spin. Thus, the electronic energy levels are quantized in the xy plane (the plane perpendicular to the magnetic field) and have translational energy $[\hbar^2 k_z^2/(2m^*)]$ along the z direction (the magnetic field direction).

6.13 Photoluminescence

With the increasing importance of nanostructures in optoelectronics, photoluminescence becomes a powerful technique that is used to characterize semiconductor micro- and nanostructures. This is because it provides information on many fundamental properties of semiconductors and nanostructures such as crystalline order, strain, composition, doping, surface carrier depletion depth, crystal damage, quality of interfaces, layer thickness, extended defects, microscopic defects, and surface

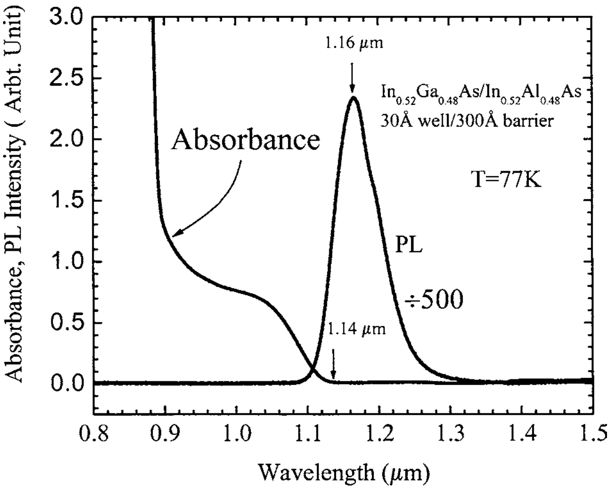


Figure 6.27 Absorption and photoluminescence (PL) spectra of $In_{0.52}Ga_{0.48}As/In_{0.52}Al_{0.48}As$ multiple quantum wells plotted as a function of the wavelength. The spectra were measured at 77 K.

quality. Thus, this technique is one of the most important and versatile for investigating compound semiconductors and their nanostructures. The interband optical absorption process in a semiconductor involves the excitation of an electron from the valence band to the conduction band after absorbing the photon. The reverse radiative process, where the photoexcited electron decays from the conduction band to the valence band, is called *photoluminescence*. In this process, the electron emits energy (photon) as it drops from the conduction band to the valence band. Luminescence can also be observed by injecting electrons into the semiconductor material, in which the injected electrons decay to the valence band by emitting photons. This process is called *electroluminescence*. Photon emission is more complicated than photon absorption in a semiconductor, but the emission results are easier to analyze. A comparison between absorption and emission is shown in Fig. 6.27, where we present the optical absorption and photoluminescence spectra for $In_{0.52}Ga_{0.48}As/In_{0.52}Al_{0.48}As$ multiple quantum wells. The optical absorption spectrum threshold occurs at ~1.14 μm (~1.088 eV), while the photoluminescence peak occurs at 1.16 μm (1.069 eV). The optical absorption threshold and the photoluminescence peak are expected to be identical since the bandgap is the same at a constant temperature. The reason for the difference between the absorption and emission is due to electron-phonon coupling. Electron-phonon coupling in a semiconductor involves extensive theoretical analysis. The simplest model used to explain electron-phonon coupling is the configuration coordinate model.

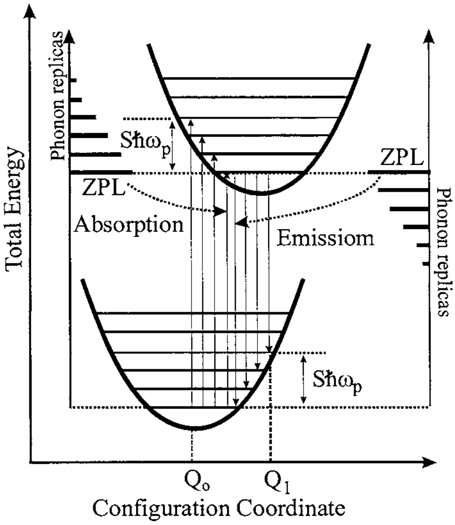


Figure 6.28 Schematic of the configuration coordinate interband transition in a direct bandgap semiconductor in the presence of electron-phonon coupling.

The configuration coordinate model is illustrated in Fig. 6.28. Let us consider the interband transition in a direct bandgap semiconductor material. In reality, the atoms vibrate in a solid, and the total energy of the electron is the sum of electronic and vibronic energies. The total energy of an electron in the valence band can be expanded in a Taylor series about a coordinate minimum Q_o such that

$$E(Q) = E(Q_o) + \frac{dE}{dQ}(Q - Q_o) + \frac{1}{2}\frac{d^2E}{dQ^2}(Q - Q_o)^2 + \cdots \qquad (6.142)$$

Since the expansion is made about an extrema, the first derivative in Eq. (6.142) is zero. Thus the valence band can be presented by a parabola around Q_o as shown in Fig. 6.28. The conduction band can be presented in a similar manner with a minimum at $Q_1 > Q_o$. The electron-phonon coupling is zero when $Q_1 = Q_o$. Each parabola represents a simple harmonic oscillator with a quantized energy

$$E_n^{Q_o,Q_1}(\omega_p) = \left(n + \frac{1}{2}\right)\hbar\omega_p \qquad n = 0, 1, 2, \ldots \qquad (6.143)$$

where ω_p is the phonon angular frequency and $E_n^{Q_o,Q_1}(\omega_p)$ is the phonon energies associated with the valence and conduction bands. The optical absorption of the interband transition is presented by a series of arrows pointing upward in the figure, where the electrons are excited from the phonon ground state $E_o^{Q_o}(\omega_p)$ in the valence band parabola to the phonon ground and excited states in the conduction band parabola $E_n^{Q_1}(\omega_p)$. On the other hand, the photoluminescence or emission transitions are presented by the arrows pointing downward, where the electrons decay from the phonon ground state, $E_o^{Q_1}(\omega_p)$, in the conduction band to the phonon levels, $E_n^{Q_o}(\omega_p)$, in the valence band. The transition (either absorption or emission) between $E_o^{Q_o}(\omega_p)$ and $E_o^{Q_1}(\omega_p)$ is called the zero phonon line (ZPL), which means it is a pure electronic transition with zero electron-phonon coupling. The electronic transitions between $E_n^{Q_1}(\omega_p)$ and $E_n^{Q_o}(\omega_p)$ for $n \neq 0$ are called phonon replicas. The difference between the absorption (E_a) and emission (E_e) energies

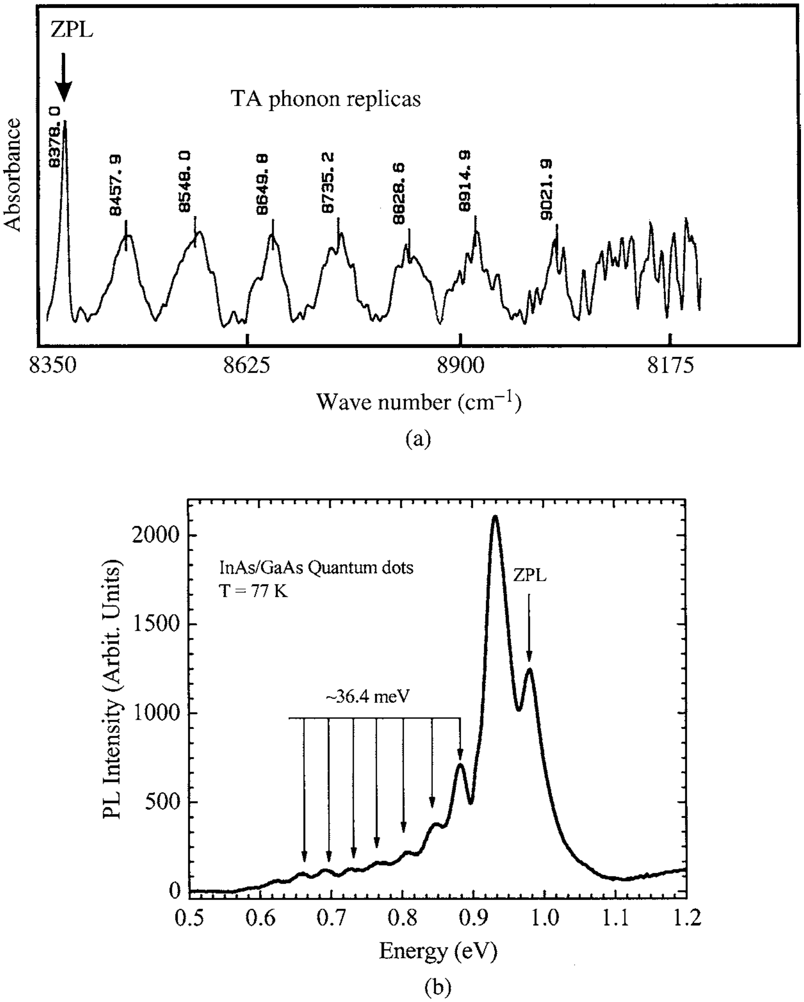


Figure 6.29 Zero phonon lines and their replicas in (*a*) EL2 defect in GaAs (Manasreh and Covington 1987) and (*b*) InAs/GaAs single-layer quantum dots grown by the molecular beam epitaxy technique.

is called the *Stokes shift* and can be written as

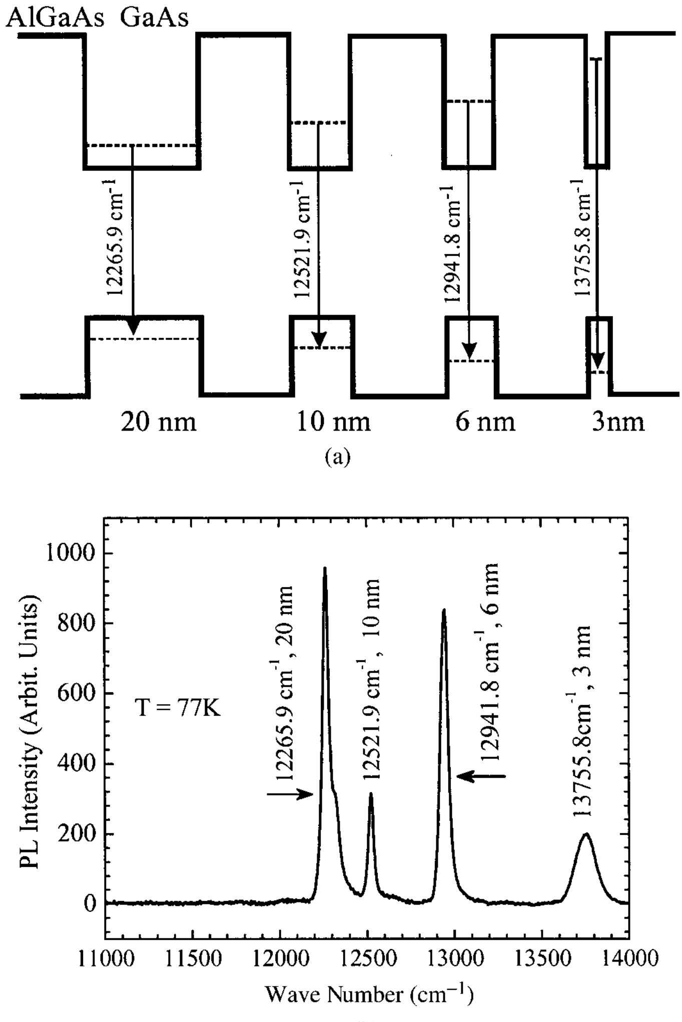
$$E_a - E_e = 2S\hbar\omega_p \tag{6.144}$$

where S is a dimensionless parameter called the Huang-Rhys factor. It is a measure of strong (large value of S) or weak (small value of S) electron-phonon coupling. Half of the Stokes shift is called the *Franck-Condon shift,* which is commonly referred to as $d_{\text{F}-\text{C}} = S\hbar\omega_p$.

Notice that phonon replicas are observed at the higher energy side of the zero phonon line in the case of the absorption spectrum, which indicates that phonons are absorbed by the electrons. In the case of emission (photoluminescence), the phonon replicas occur at the lower energy side of the zero phonon line, which means that the electrons emit phonons as they decay from the conduction band to the valence band. Two examples of zero phonon lines and their phonon replicas are shown in Fig. 6.29. The first example is the optical absorption of the zero phonon line associated with the EL2 defect in GaAs (Fig. 6.29*a*). The zero phonon lines occur at 8378 cm^{-1} (1.0387 eV), and the replicas are those of the TA phonon mode (~10 meV) in GaAs (recall that 1.0 eV = 8065.46 cm^{-1}). The second example is the photoluminescence zero phonon mode observed in InAs/GaAs single-layer quantum dots (Fig. 6.29*b*). The zero phonon line peak is indicated as ZPL. The stronger peak around 0.934 eV is also a zero phonon line due to the fact that the quantum dots have two dominant sizes. The ripples below 0.9 eV are due to phonon replicas separated by an average energy of ~36.4 meV. This phonon energy is most likely to be the optical phonon mode generated at the InAs/GaAs interfaces.

Photoluminescence (PL) can be observed in semiconductors and their nanostructures if electrons and holes are generated by optical excitation followed by radiation emission. If the electrons recombine with the holes without emitting radiation, the transition is called nonradiative. Photoluminescence technique is currently a standard technique in both industry and academia. It is used to calibrate epitaxial growth rate and growth quality, as illustrated in Fig. 6.30. A test of the epitaxial growth would be to grow a few quantum wells with different thicknesses, as shown in Fig. 6.30*b*, where GaAs/AlGaAs quantum wells grown on semi-insulating GaAs substrate were chosen as an example. The barrier thickness is usually chosen to be thick enough to prevent tunneling between wells. As the quantum well thickness is reduced from 20 to 3 nm, the bound states are squeezed outward and the interband transition energy is increased, as shown in Fig. 6.30*a*. The corresponding photoluminescence spectrum is shown in Fig. 6.30*b*, where the PL intensity is plotted as a function of photon energy. The PL energy (E_{PL}) can be written as

$$E_{\text{PL}}(L_z) = E_g + E_{\text{CB1}}(L_z) + E_{\text{HH1}}(L_z) - E_{\text{ex}}(L_z) \tag{6.145}$$



(a)

(b)

Figure 6.30 (*a*) A sketch of four GaAs/AlGaAs quantum wells grown by the molecular beam epitaxy technique on GaAs semi-insulating substrate. (*b*) Photoluminescence spectrum measured at $T = 77$ K for the structure described in (*a*).

where E_g = fundamental bandgap of bulk GaAs material, is taken as 1.50 eV (12,098 cm^{-1}) at $T = 77$ K
E_{CB1} = ground bound state in conduction band
E_{HH1} = bound ground state of heavy hole in valence band
E_{ex} = exciton binding energy
L_z = well width

The PL peak observed for the 20-nm quantum well exhibits a structure with a shoulder at the higher energy side. This is due to the presence of bound and free excitons in the quantum well, which is an indication of high structural interfaces.

The excess electron concentration N created by the laser excitation in the PL experiment is equal to the excess hole concentration, which is given by the rate equation as

$$\frac{dN}{dt} = -\frac{N}{\tau_r} \tag{6.146}$$

where τ_r is the lifetime for the carriers that undergo radiative recombination, which is the inverse of the Einstein coefficient for the spontaneous emission rate. Integration of this equation yields

$$N(t) = N_o \exp\left(-\frac{t}{\tau_r}\right) \tag{6.147}$$

where N_o is the excess electron concentration at $t = 0$. The radiative recombination rate R_r is defined as

$$R_r = \frac{dN}{dt} = -\frac{N}{\tau_r} \tag{6.148}$$

When the nonradiative recombination rate R_n is considered, the total spontaneous recombination rate R_s can be written as

$$R_s = R_r + R_n \tag{6.149}$$

For exponential decay, the internal quantum efficiency η_i is given by the carrier lifetime as

$$\eta_i = \frac{\tau_r^{-1}}{\tau_r^{-1} + \tau_n^{-1}} = \frac{1}{1 + \tau_n/\tau_r} \tag{6.150}$$

where τ_n is the lifetime of the carriers that undergo nonradiative recombination. The internal quantum efficiency of the interband transition is equal to unity when τ_n is zero.

It is possible to estimate the radiative recombination lifetime from the carrier concentration in direct bandgap semiconductors. The relationship between the lifetime and the carrier concentration is left as an exercise.

6.14 Lattice Vibrations and Phonons

In semiconductor crystals, the atoms are tightly coupled to one another and the binding energy is called cohesive energy, which is defined as the energy needed to separate a crystal into independent ions located large

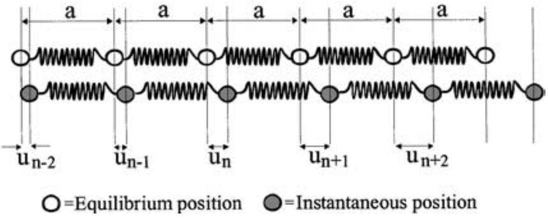


Figure 6.31 A one-dimensional illustration of a crystal with a lattice constant a showing the longitudinal displacement of a few atoms.

distances from each other. The thermal kinetic energy of the atoms in the crystal is simply the vibrational energy of motion, which propagates in the crystal as waves. These waves are called acoustical or sonic waves. The quanta of these waves are called phonons. Phonons in semiconductors can absorb or scatter light in the infrared spectral region. To understand how the acoustical waves propagate in a solid, let us first consider a one-dimensional monatomic lattice, as shown in Fig. 6.31. By including only the nearest-neighbor interaction and assuming that the vibrational amplitudes are smaller than the lattice spacing, one can write the force on the nth atom as

$$F_n = m\frac{\partial^2 u_n}{\partial t^2} = \gamma(u_{n+1} - u_n) - \gamma(u_n - u_{n-1}) = \gamma(u_{n+1} + u_{n-1} - 2u_n) \tag{6.151}$$

where m is the mass of the atom and γ is the force constant. For a solution having the character of a traveling wave, we have for the nth atom

$$u_n = Ae^{i(kx_n - \omega t)} = Ae^{i(kna - \omega t)} \tag{6.152}$$

where k = propagation constant
ω = angular frequency
$x_n = na$ (a = lattice constant)

Similarly, the solutions for the nearest-neighbor atoms are

$$\begin{aligned} u_{n+1} &= Ae^{i(kx_{n+1} - \omega t)} = Ae^{i[k(n+1)a - \omega t]} = e^{ika}u_n \\ u_{n-1} &= Ae^{i(kx_{n-1} - \omega t)} = Ae^{i[k(n-1)a - \omega t]} = e^{-ika}u_n \end{aligned} \tag{6.153}$$

The dispersion relation can be obtained by substituting Eqs. (6.152) and (6.153) into Eq. (6.151) and canceling u_n as follows:

$$\omega^2 = \frac{\gamma}{m}(2 - e^{ika} - e^{-ika}) = \frac{2\gamma}{m}(1 - \cos ka) = \frac{4\gamma}{m}\sin^2\left(\frac{ka}{2}\right) \tag{6.154}$$

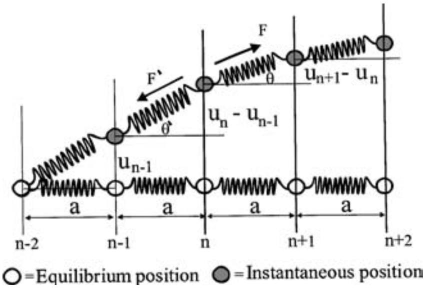


Figure 6.32 A one-dimensional model of a crystal with a lattice constant a showing the transverse displacement of a few atoms.

This equation can be rewritten as

$$\omega = \omega_m \left| \sin\left(\frac{ka}{2}\right) \right| \tag{6.155}$$

where $\omega_m = \sqrt{4\gamma/m}$ and the absolute value sign is given to indicate that ω is a positive quantity. The displacement of the atoms shown in Fig. 6.31 produces a longitudinal acoustic wave with a frequency described by Eq. (6.155).

For transverse acoustical waves, the atoms are displaced, as shown in Fig. 6.32. In addition to the atomic displacement shown in this figure, it is also possible to simultaneously displace the atoms perpendicular to the plane of the page. Thus, one may obtain two transverse modes. The equation of motion of the transverse modes of the nth atom can be obtained in a similar fashion as the longitudinal mode

$$F_n = m\frac{\partial^2 u_n}{\partial t^2} = \gamma_t(u_{n+1} + u_{n-1} - 2u_n) \tag{6.156}$$

where γ_t is the transverse force constant. The dispersion relation of the transverse mode is obtained as

$$\omega = \omega_m^t \left| \sin\left(\frac{ka}{2}\right) \right| \tag{6.157}$$

where $\omega_m^t = \sqrt{4\gamma_t/m}$. A plot of the dispersion relations for both the longitudinal (L) and transverse modes (T_1 and T_2) is shown in Fig. 6.33. When the force constants are the same for both the transverse modes, the dispersion relation becomes degenerate with $T_1 = T_2$.

For a more complicated case, let us consider a linear one-dimensional diatomic lattice model, as shown in Fig. 6.34, where the chain is composed of two alternating atoms of masses M and m (M is assumed to

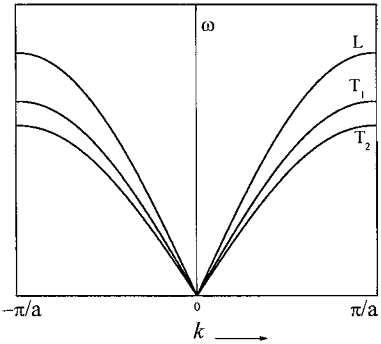


Figure 6.33 The angular frequency ω of the longitudinal (L) and transverse (T_1 and T_2) waves in a one-dimensional monatomic lattice is plotted as a function of the propagation constant k.

be larger than m). The equations of motion of atoms $2n$ (mass m) and $2n+1$ (mass M) can be written as

$$F_{2n} = m\frac{\partial^2 u_{2n}}{\partial t^2} = \gamma(u_{2n+1} + u_{2n-1} - 2u_{2n}) \tag{6.158a}$$

$$F_{2n+1} = M\frac{\partial^2 u_{2n+1}}{\partial t^2} = \gamma(u_{2n+2} + u_{2n} - 2u_{2n+1}) \tag{6.158b}$$

where γ is the force constant and the u's are the atomic displacements from the equilibrium. The solutions to Eqs. (6.158*a*) and (6.158*b*) can be expressed as

$$u_{2n} = Ae^{i(2kna-\omega t)} \qquad \text{and} \qquad u_{2n+1} = Be^{i[k(2n+1)a-\omega t]} \tag{6.159}$$

Similarly, the displacement of the atoms labeled $2n+2$ and $2n-1$ can be written as

$$\begin{aligned} u_{2n+2} &= Ae^{i[k(2n+2)a-\omega t]} = u_{2n}e^{i2ka} \\ u_{2n-1} &= Be^{i[k(2n-1)a-\omega t]} = u_{2n+1}e^{-i2ka} \end{aligned} \tag{6.160}$$

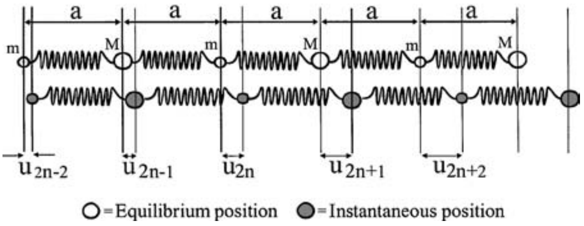


Figure 6.34 A one-dimensional chain of diatomic crystal with atomic masses M and m.

Substituting Eqs. (6.155) and (6.156) into Eq. (6.154) and evaluating the time derivative of u_{2n} and u_{2n+1}, the equations of motion can be rewritten as

$$\begin{aligned}(m\omega^2 - 2\gamma)u_{2n} + \gamma(1 + e^{-2ika})u_{2\mathrm{n}+1} &= 0\\ \gamma(1 + e^{2ika})u_{2n} + (M\omega^2 - 2\gamma)u_{2\mathrm{n}+1} &= 0\end{aligned} \tag{6.161}$$

These two homogeneous equations can be solved by equating the determinant to zero such that

$$\begin{vmatrix}(m\omega^2 - 2\gamma) & \gamma(1 + e^{-2ika})\\ \gamma(1 + e^{2ika}) & (M\omega^2 - 2\gamma)\end{vmatrix} = (m\omega^2 - 2\gamma)(M\omega^2 - 2\gamma) - 4\gamma^2\cos^2(ka) = 0 \tag{6.162}$$

where $1 + \cos(2ka) = 2\cos^2(ka)$ is used in this expression. The solution of Eq. (6.162) can be expressed as

$$\omega^2 = \frac{\gamma(m + M)}{mM}\left[1 \pm \sqrt{1 - \frac{4mM\sin^2(ka)}{(m + M)^2}}\right] \tag{6.163}$$

This dispersion relation has two roots (ω_+ and ω_-), which are plotted in Fig. 6.35. The two branches are called optical (ω_+) and acoustical (ω_-) modes. The maximum value of the optical branch is $\sqrt{\gamma(m + M)/(mM)}$

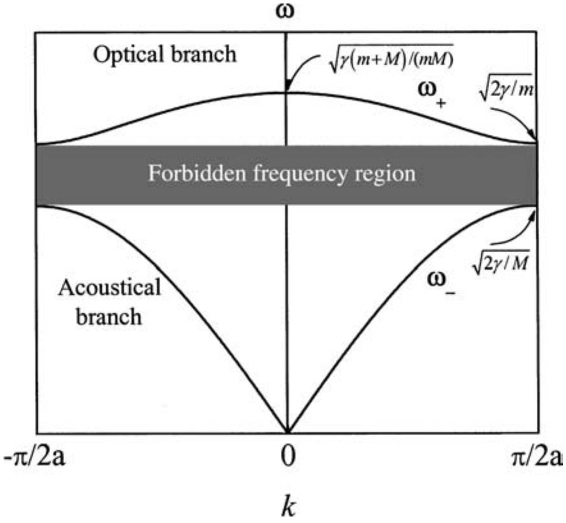


Figure 6.35 The dispersion curve for the diatomic one-dimensional lattice. The first Brillouin zone is extended between $-\pi/(2a)$ and $+\pi/(2a)$ since the unit cell is $2a$.

and there is a forbidden frequency gap at the Brillouin zone boundaries extended between $\sqrt{2\gamma/m}$ and $\sqrt{2\gamma/M}$. This gap is reduced to zero for $m = M$.

Phonon energy or frequency in semiconductors can be measured by infrared spectroscopy and Raman scattering techniques providing that the phonon density of states is large. Phonon modes can be infrared active and/or Raman active. Selection rules that govern the phonon absorption are outside the scope of this book, but detailed discussions and analyses are reported by many authors (see, for example, Birman 1984). To calculate the phonons' density of states, let us consider the simplest case of a monatomic one-dimensional crystal. To determine the number of phonon modes with different values of k that fall in the frequency interval from ω to $\omega + d\omega$, one can obtain from the dispersion relation, expressed in Eq. (6.157), the following:

$$\begin{aligned} d\omega &= \sqrt{\frac{4\gamma}{m}}d\left[\left|\sin\left(\frac{ka}{2}\right)\right|\right] = \sqrt{\frac{4\gamma}{m}}\frac{a}{2}\left|\cos\left(\frac{ka}{2}\right)\right| dk \\ &= a\sqrt{\frac{\gamma}{m}}\left|\cos\left(\frac{ka}{2}\right)\right| dk \end{aligned} \tag{6.164}$$

If one assumes that the one-dimensional crystal is a large circle that contains N atoms, where $N \gg 1$, then the atomic displacements satisfy the following conditions: $u_{N+n} = u_n$ and $e^{ikNa} = 1$. From the second condition, one can obtain the following relation:

$$k = \frac{2\pi p}{Na} \tag{6.165}$$

where p is an integer that satisfies the following relation: $-N/2 < p < N/2$. These relations indicate that k is discrete with N possible values corresponding to N different standing waves. The number of modes dn in the interval $d\omega$ is

$$dn = 2dp = \frac{N}{\pi}\sqrt{\frac{m}{\gamma}}\frac{d\omega}{|\cos(ka/2)|} \tag{6.166}$$

With the help of Eq. (6.157), the $\cos(ka/2)$ can be expressed as

$$\cos\left(\frac{ka}{2}\right) = \sqrt{1 - \sin^2\left(\frac{ka}{2}\right)} = \sqrt{1 - \frac{\omega^2 m}{4\gamma}} \tag{6.167}$$

The phonon density of states, g_{ph}, can now be defined as

$$g_{\text{ph}}(\omega) = \frac{1}{Na}\frac{dn}{d\omega} = \frac{2}{\pi a}\frac{1}{\sqrt{\omega_m^2 - \omega^2}} \tag{6.168}$$

where $\omega_m = \sqrt{4\gamma/m}$. The density of states expressed in Eq. (6.168) approaches infinity as ω approaches ω_m, and it has a constant value

TABLE 6.2 Energy of the Phonon Modes in Several Semiconductor Materials Reported in meV for Three Different Symmetry Points in the First Brillouin Zone

	$\Gamma(000)$		$X(100)$				$L(111)$			
Material	LO	TO	LO	TO	LA	TA	LO	TO	LA	TA
GaAs	35.8	33.0	35.4	30.9	27.8	9.7	29.2	32.1	25.6	7.6
InAs	29.9	26.9	20.0	26.2	18.0	13.5	23.7	26.6	18.0	9.0
InSb	24.5	22.0	16.0	21.7	14.7	4.9	19.6	20.9	12.3	4.1
AlSb	24.2	22.7	—	—	36.4	9.8	28.6	37.6	31.4	7.6
AlAs	49.5	44.2	—	—	12.7	12.7	—	—	—	9.9
GaSb	29.9	28.2								
Si	63.8	63.8	50.3	56.9	50.3	18.4	51.5	60.1	46.2	14.0
Ge	36.9	36.9	28.2	33.7	28.2	10.1	30.3	34.4	26.4	8.0

NOTE: LO= longitudinal optical, TO= transverse optical, LA = longitudinal acoustic, TA = transverse acoustic.

when ω approaches zero. The same analysis can be applied for a diatomic chain using the dispersion relation described in Eq. (6.163). The analysis, however, is a little bit more complicated.

For three-dimensional crystals, such as Si and GaAs, the calculations of the phonon modes and their density of states are more extensive and require computer analysis. Table 6.2 summarizes the energy of the phonon modes in several semiconductor materials. A typical example of phonon Raman scattering is shown in Fig. 6.36 for a semi-insulating GaAs sample. The Raman scattering spectrum in this figure

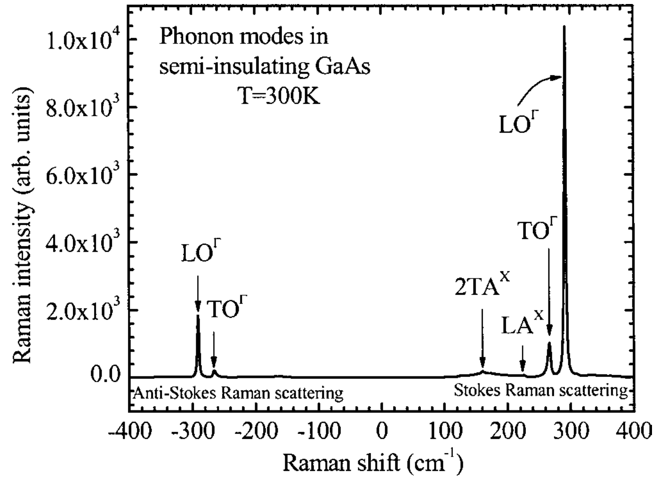


Figure 6.36 A Raman scattering spectrum obtained at 300 K for a semi-insulating GaAs sample. The spectrum shows Stokes and anti-Stokes phonon modes.

was obtained using a Fourier-transform infrared spectrometer in conjunction with a 1.06-μm YAG laser. The spectrum shows two sets of peaks called Stokes and anti-Stokes phonon scattering, which is known as the Raman effect. To understand this effect, consider an incident laser beam of an energy $\hbar\omega_i$, which is scattered by a semiconductor sample. The radiation consists of the laser beam ($\hbar\omega_i$) and weaker beams of energies $\hbar\omega_i \pm \hbar\omega$. The beam with the energy $\hbar\omega_i - \hbar\omega_i$ is called the Stokes Raman scattering line, and the beam with the energy $\hbar\omega_i + \hbar\omega$ is called the anti-Stokes Raman scattering line. These lines are shown in Fig. 6.36. The most important aspect of Raman scattering is that ω is independent of ω_i. The effect was predicted by Smekal (1923) and experimentally measured by Raman (1928). Raman scattering is considered as an inelastic scattering of light in which an internal form of motion (vibrational modes) of the scattering system is either excited or absorbed during the process.

Phonon modes in wurtzite structures such as GaN are more complicated than the phonon modes in diamond or zinc-blende structures. For more discussion on the subject see, for example, Manasreh and Jiang (2002) and Pattada et al. (2003). Raman spectroscopy is a very useful contactless technique in probing the charge carrier concentration in semiconductors. Charge carriers can be detected in Raman spectra through the coupling of the longitudinal optical phonon mode with plasma oscillations. The collective oscillation (plasmon) of an electron or hole gas in a solid is a longitudinal excitation, and its frequency ω_p can be written as

$$\omega_p = \sqrt{\frac{ne^2}{m^* \epsilon_o \epsilon_\infty}} \tag{6.169}$$

where n = charge carrier concentration
m^* = charge carrier effective mass
ϵ_o = permittivity of space
ϵ_∞ = high-frequency dielectric constant of material (related to refractive index n_r such that $\epsilon_\infty \approx n_r^2$)

A typical example of Raman scattering from the longitudinal phonon-plasmon coupled mode in a doped semiconductor quantum well is shown in Fig. 6.37 for an InGaAsN/GaAs single quantum well grown by metal-organic chemical vapor deposition technique on semi-insulating GaAs substrate. The macroscopic electric field of the plasma wave interacts with the polarization field associated with the longitudinal optical (LO) phonons in polar semiconductors, such as GaAs (zinc blende) and GaN (wurtzite) materials. This coupling splits the LO phonons into two LO-plasmon coupled (LOPC) modes, known as L_+ and L_-. The low-frequency mode L_- shifts from 0 cm^{-1} to the TO frequency, while the

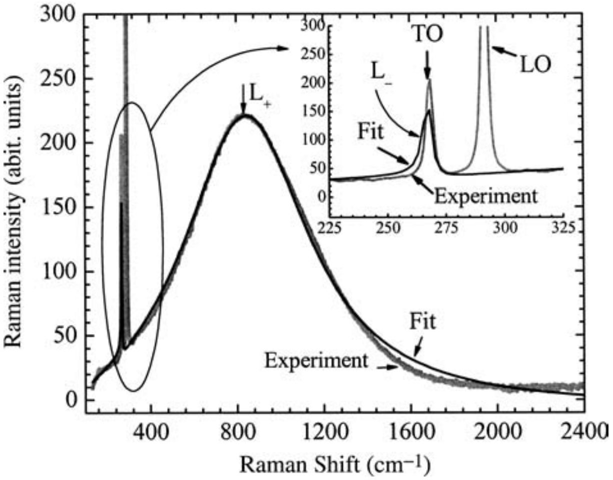


Figure 6.37 A Raman scattering spectrum obtained for an InGaAsN/GaAs single quantum well sample (gray line). The spectrum shows the LO and TO phonon modes and the L_+ branch of the LOPC mode. The solid black line is the result of the fitting analysis using Eqs. (6.170) and (6.171), which shows both the L_+ and L_- branches of the LOPC mode. The inset is the expansion of the spectral region in the vicinity of LO and TO phonon modes.

high frequency mode L_+ shifts from LO frequency to the plasma frequency ω_p for increasing the carrier concentration (see, for example, Mooradian and Wright 1966 and Absteiter et al. 1978).

The Raman intensity I_s is proportional to the imaginary part of the inverse of the total dielectric function (see, for example, Manasreh and Jiang 2002 and references therein):

$$I_s \propto \mathrm{Im}\left(-\frac{1}{\epsilon(\omega)}\right) \tag{6.170}$$

where the dielectric function contains the contribution from lattice vibration and the conduction electrons, which is given by

$$\epsilon(\omega) = \epsilon_\infty \left[1 + \frac{\omega_{\mathrm{LO}}^2 - \omega_{\mathrm{TO}}^2}{\omega_{\mathrm{TO}}^2 - \omega^2 - i\omega\Gamma} - \frac{\omega_p^2}{\omega(\omega - i\gamma)}\right] \tag{6.171}$$

The parameters Γ and γ are the damping constants of the phonon and plasmon, respectively. The plasmon frequency ω_p is obtained by fitting the LOPC Raman spectrum using Eqs. (6.170) and (6.171), with ω_p, Γ, and γ as fitting parameters. An example is shown in Fig. 6.37, where the gray line spectrum is the experimental result and the thin black line spectrum is the theoretical fit. The plasmon frequency in this case

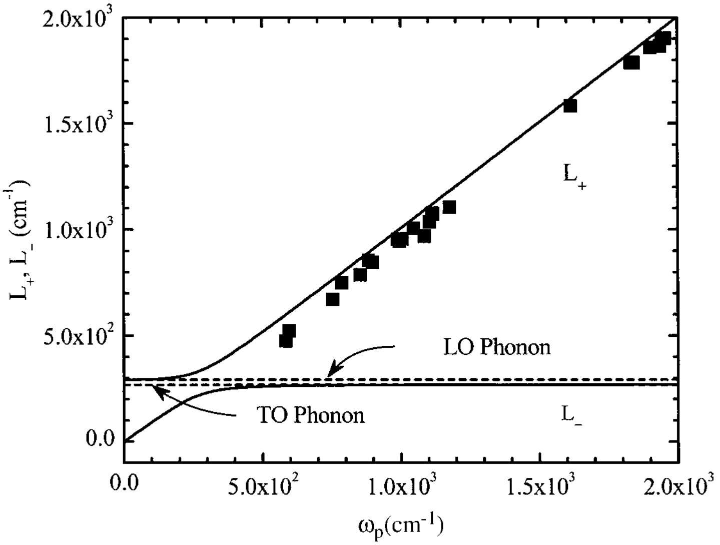


Figure 6.38 A plot of the L_+ mode as a function of the plasmon frequency for a series of InGaAsN/GaAs single quantum well samples (solid squares). The solid lines are plots of L_+ and L_- given by Eq. (6.172). The dashed lines represent the LO and TO phonon frequencies in an InGaAsN quantum well.

is obtained as 909 cm^{-1}. Using this value in Eq. (6.169), the carrier concentration is obtained as 7.56×10^{18} cm^{-3} for $\epsilon_\infty = 12.25$ and $m^* = 0.067m_o$.

As mentioned previously, the LOPC mode splits into two modes known as the L_+ and L_- branches. These two branches are approximately obtained by setting $\Gamma = \gamma = 0$ and solving Eq. (6.171) for $\epsilon(\omega) = 0$, which yields

$$L_\pm = \frac{1}{2}\left[\left(\omega_L^2 + \omega_p^2\right) \pm \sqrt{\left(\omega_L^2 + \omega_p^2\right)^2 - 4\omega_T^2\omega_p^2}\right]^{1/2} \tag{6.172}$$

The fitting analysis of the experimental spectrum in Fig. 6.37 reveals the presence of both L_+ and L_-. The L_- region along with the LO and TO phonon modes are replotted in the figure inset for clarity. The same fitting procedure was repeated for several InGaAsN/GaAs single quantum well samples with different nitrogen contents. The plasmon frequency ω_p was obtained for each sample by fitting the experimental spectra, as described previously. Additionally, the frequency maximum of the L_+ branch was obtained directly from the experimental LOPC mode spectra. To compare the experimental results to the theoretical predictions, the L_+ and L_- modes are plotted as a function of ω_p using

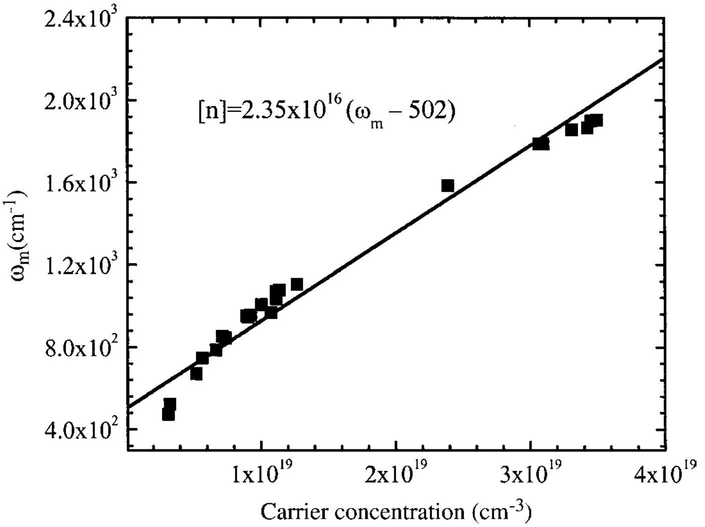


Figure 6.39 The frequency maximum ω_m of the L_+ branch as a function of the carrier concentration obtained from the data in Fig. 6.37. The solid line is a first-order linear fit of the data.

Eq. (6.172), as shown in Fig. 6.38. The experimental data were plotted in this figure as solid squares. The dashed lines represent the LO and TO phonon modes in the quantum well.

The plasmon frequency is used to calculate the carrier concentration in a series of samples with different nitrogen content. The results are shown in Fig. 6.39 where the frequency maximum of the L_+ branch is plotted as a function of the calculated carrier concentration. The solid line in this figure is the result of the linear fit of the data from which the following empirical expression is obtained: $[n] = 2.35 \times 10^{16}(\omega_m - 502)$ cm^{-3}, where $[n]$ is the carrier concentration. This expression can be used to obtain the carrier concentration directly from the peak of the L_+ mode, which is measured directly by Raman scattering in the unit of cm^{-1}.

In addition to the determination of the carrier concentration using Raman scattering, the plasmon damping rate γ, which is used in the fitting analysis, can be used to calculate the carrier drift mobility μ through the following relation: $\mu = e/(m^*\gamma)$. The drift mobility values estimated from the plasmon damping rate are on the order of 100 to 200 $\mathrm{cm}^2 \cdot \mathrm{V}^{-1} \cdot \mathrm{s}^{-1}$ which is in good agreement with those reported by Young et al. (2003). Even though the carrier concentration and drift mobilities are estimated from fitting a simple model based on Drude theory to Raman scattering spectra, the results provide a good indication of the material quality and its feasibility for device application.

Summary

The optical properties of bulk semiconductors and their low-dimensional quantum structures were discussed in this chapter. We started the chapter by defining the difference between the bulk and quantum well materials with the emphasis on interband and intersubband transitions. Bound and free excitons were illustrated in crystalline structures. The basic electromagnetic formalism was introduced with a fundamental discussion regarding the refractive index, dielectric constant, and linear optical absorption coefficient. The optical absorption coefficients of interband transitions in direct and indirect semiconductor materials were derived using the Fermi golden rule. The formalisms were extended to derive the optical absorption coefficients of interband transitions in type I and type II quantum wells. Detailed discussions on the optical absorption coefficient of the intersubband transitions in quantum wells and quantum dots were presented for both bound-to-bound and bound-to-continuum cases. An example of intersubband transition was presented for GaN/AlGaN multiple quantum wells, where the piezoelectric doping is significant. A complete section on the selection rules of both interband and intersubband transitions was presented.

Excitons in both bulk and quantum structures play a major role in the optoelectronic devices. Detailed analysis of the exciton binding energy and radius was presented for bulk semiconductors, quantum wells, and quantum dots. An attractive feature of semiconductor low-dimensional quantum structures is that the binding energies of the excitons are much higher than those of the excitons in bulk materials.

Selected techniques used to optically characterize the semiconductor quantum structures, such as cyclotron resonance, photoluminescence, and Raman scattering were discussed. Finally, lattice vibrations and phonons were briefly discussed at the end of the chapter.

Problems

6.1 The electric and magnetic fields can be written in terms of vector ($\mathbf{A}$) and scalar (ϕ) potentials such as $\mathcal{E} = -\frac{\partial \mathbf{A}}{\partial t} - \nabla\phi$ and $\mathbf{B} = \frac{1}{\mu_o}\nabla\times\mathbf{A}$. Rewrite the four Maxwell's equations in terms of these two potentials.

6.2 Show that Eq. (6.9) is valid.

6.3 Use the complex definition of the refractive index and the dielectric constant to show that the optical absorption coefficient can be expressed as $\alpha(\omega) = \omega\epsilon_2(\omega)/[cn_1(\omega)]$.

6.4 Derive expressions for the real (n_1) and imaginary (n_2) parts of the refractive index in terms of the real (ϵ_1) and imaginary (ϵ_2) parts of the dielectric constants. Plot n_1, n_2, ϵ_1, and ϵ_2 as a function of ω. Assume $N = 10^{17}\text{m}^{-3}$ and $m^* = 0.067m_o$.

6.5 The oscillator strength of the interband transition can be defined as $f_{vc} \approx 2P^2/[m_o(E_{kc} - E_{kv})]$. Additionally, the sum rule for a solid can be written as $\sum_{m \neq n} f_{mn} = |1 - m_o/m_e^*|$, where we sum the oscillator strength of electronic transitions from all m states to n states with the same **k**-value. Use these expressions to calculate the absorption coefficient of the interband transition in GaAs for $(h\omega - E_g) = 0.1$ eV.

6.6 Calculate the optical absorption coefficient of the interband transition in 20 Å GaAs/AlGaAs quantum well. Assume that the transition occurs from the ground state of the heavy hole in the valence band to the ground state of the electron in the conduction band. The photon energy required to excite this transition is 1.75 eV.

6.7 Calculate the oscillator strength and the optical absorption coefficient of a bound-to-bound intersubband transition in GaAs/AlGaAs multiple quantum wells. Assume that the number of wells is 50, the well width is 75 Å, the half-width at half of the maximum is 7 meV, the electron density is 5×10^{11} cm^{-2}, and the photon energy required to excite the transition is 180 meV.

6.8 Show that the Brewster's angle of GaAs is 73°. What does this angle mean?

6.9 Figure P6.9 is a waveguide made of GaAs with a thickness of 0.4 mm. The photons enter the sample at 45° from the formal as shown in the figure. Finish the design of this waveguide such that three passes will be made by the photons before they exit the sample. The GaAs refractive index is 3.4. What would be the length of the waveguide?

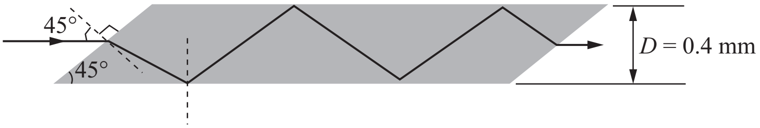


Figure P6.9

6.10 Use the definition of the oscillator strength of intersubband transition in GaAs/AlGaAs multiple quantum wells as $f_{01} = (2m^*\omega/\hbar)|\langle n\mathbf{k}_\perp|z|n'\mathbf{k}'_\perp\rangle|^2$ where $\langle n\mathbf{k}_\perp|z|n'\mathbf{k}'_\perp\rangle$ is known as the overlap integral. Calculate the oscillator strength for a 100-Å-thick well where the photon energy needed to excite the transition is $\hbar\omega = 0.15$ eV.

6.11 A time-dependent quantum operator can be written as $d\mathbf{M}(t)/dt = (i/\hbar)(H_o\mathbf{M} - \mathbf{M}H_o)$. Use the dipole matrix element to show that the oscillator strength can be written as $f_{01} = (2m^*\omega/\hbar)|\langle n\mathbf{k}_\perp|z|n'\mathbf{k}'_\perp\rangle|^2$.

6.12 Calculate the optical absorption band edge associated with an HH$_1$-E$_1$ transition of a InAs/GaAs quantum dot at room temperature. Assume that the quantum dot has a cubic shape with a side length of 6.5 nm. Compare your results to the bandgap of bulk InAs material.

6.13 Calculate the dipole matrix element of 1 → 2, 1 → 4, and 2 → 3 transitions in an infinite GaAs quantum well. The well width is of 10 nm, the effective mass is $0.067m_o$, and the wave functions of the bound states can be expressed as $\varphi_n(z) = \sqrt{2/d}\ \sin(n\pi z/L)$. Calculate the corresponding wavelengths of these three transitions.

6.14 Search the literature for the electron and heavy hole effect masses, bandgaps, and dielectric constants of five direct bandgap semiconductor bulk materials other than those listed in Table 6.1. Calculate and plot the exciton binding energy and the exciton radius in these materials as a function of their bandgaps.

6.15 Start from the Schrödinger equation and a trial function of the form $\psi = \exp(-r_\perp/a_{\mathrm{2D}})$, where $r_\perp = \sqrt{x^2 + y^2}$ and a_{2D} is the exciton radius in quantum wells. Show that the exciton binding energy in quantum wells can be written as $E_{\mathrm{ex}}^{\mathrm{2D}} = -4E_{\mathrm{ex}}^{\mathrm{3D}}$ and $a_{\mathrm{2D}} = 0.5a_{\mathrm{3D}}$, where 2D and 3D indicate two-dimensional (quantum wells) and three-dimensional (bulk material) systems, respectively.

6.16 A GaAs/AlGaAs quantum well has an effective electron mass of $0.072m_o$. A peak in the cyclotron resonance spectrum was observed at 20 meV. Calculate the magnetic field used to generate this peak. Estimate the splitting of Landau levels due to the electron spin. Assume $g^* = 1.75$ and $\mu_B = 9.27 \times 10^{-24}$ J/T.

6.17 In the photoluminescence experiment, the carriers spontaneous recombination rate can be written as $R_s = Bnp$, where B is a constant, n is the electron concentration, and p is the hole concentration. On the other hand, this rate can be written as the sum of the spontaneous rates at thermal equilibrium and in the presence of the excess carriers. Derive an expression for the lifetime of the excess carriers and express your answer for the two cases of high and low injection rates.

6.18 A band-to-band photoluminescence transition was observed for InGaAs thin film at 1825.68 nm. Calculate the In composition needed to produce this peak. Repeat the same process to obtain the Al fraction in AlGaAs thin film which has a PL peak at 689.62 nm. Search the literature to obtain the bandgap of $\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{As}$ and $\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{As}$ as a function of the mole fraction x.

6.19 Consider the diatomic one-dimensional crystal model shown in Fig. P6.19 where the force constants are indicated as γ_1 and γ_2 and the masses of the atoms are given as M_1 and M_2.

(a) Show that the dispersion relation can be written as

$$\omega^2 = \frac{\omega_o^2}{2}\left(1 \pm \sqrt{1 - \Gamma^2 \sin^2\left(\frac{ka}{2}\right)}\right)$$

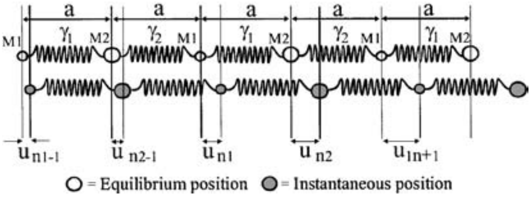


Figure P6.19

where

$$\omega_o = \sqrt{\frac{(\gamma_1 + \gamma_2)(M_1 + M_2)}{M_1 M_2}} \quad \text{and} \quad \Gamma^2 = \frac{4\gamma_1\gamma_2}{(\gamma_1 + \gamma_2)^2} \frac{4M_1M_2}{(M_1 + M_2)^2}$$

(b) Show that when $M_1 = M_2 = M$ and $\gamma_1 = \gamma_2 = \gamma$, the optical and acoustical branches of the dispersion relation can be written as

$$\omega_{\text{acoustical}} = \sqrt{\frac{4\gamma}{M}} \left| \sin\left(\frac{ka}{4}\right) \right| \quad \text{and} \quad \omega_{\text{optical}} = \sqrt{\frac{4\gamma}{M}} \left| \cos\left(\frac{ka}{4}\right) \right|$$

6.20 Show that the frequency gap width ($\Delta\omega$) between the optical and acoustical branches in Fig. 6.35 can be written as

$$\Delta\omega^2 = \omega_+^2(0) - \frac{4\gamma}{\sqrt{mM}}$$

6.21 Derive the imaginary and real parts of the inverse of the dielectric constant shown in Eq. (6.171). Plot the inverse of the imaginary part as a function of angular frequency ω using the following values for n-type GaN parameters: $\Gamma = 50$ cm^{-1}, $\gamma = 360$ cm^{-1}, $\omega_p = 550$ cm^{-1}, $\omega_{\text{LO}} = 734$ cm^{-1}, $\omega_{\text{TO}} = 531$ cm^{-1}, $m^* = 0.22m_o$, and $\epsilon_\infty = 4.84$. Calculate the electron concentration.

6.22 A plasmon damping rate of $\gamma = 950$ cm^{-1} was obtained by fitting the Raman spectrum of the LOPC mode with Eq. (6.170). Calculate the carrier drift mobility from this damping rate value. What is the carrier relaxation time that corresponds to this damping rate? Assume that the carrier effective mass is $0.067m_o$.

This page intentionally left blank.

Chapter

7

Electrical and Transport Properties

7.1 Introduction

Electric currents in semiconductors are due to the net flow of electrons and holes under bias voltages, and transport is the process that describes the motion of the charged particles. The two major transport processes are the drift and diffusion mechanisms. The drift mechanism is basically the movement of charged carriers under the influence of applied electric fields, and the diffusion mechanism is the flow of charged particles due to the density variation. Transport properties in semiconductors can be very complicated, depending on the actual size of the samples. Thus, it is worth discussing the classical and quantum limitations and regimes.

In order to define the limits of various transport regimes, one may scale the size of the sample against the de Broglie wavelength. This de Broglie wavelength λ can be expressed, as an example, for an electron traveling with a thermal kinetic energy in a semiconductor as

$$\lambda = \frac{h}{p} = \frac{h}{\sqrt{2m^* E}} = \lambda_o \sqrt{\frac{m_o}{m^*}} \tag{7.1}$$

where h = Plank's constant
p = momentum
E = energy
λ_o = de Broglie wavelength of a free electron
m^* = electron effective mass in semiconductor

The room temperature de Broglie wavelength of a free electron is ~76 Å, and that of an electron in GaAs is 295 Å. The de Broglie wavelength

Copyright © 2005 by The McGraw-Hill Companies, Inc. Click here for terms of use.

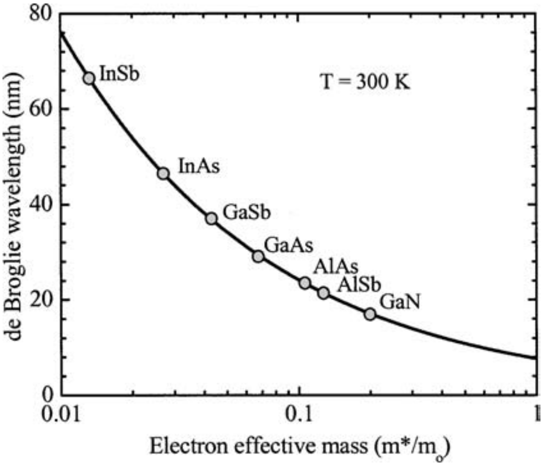


Figure 7.1 de Broglie wavelength plotted as a function of electron effective mass in several semiconductor materials. The electron energy is assumed to be that of room temperature thermal energy.

for a selection of semiconductor materials is plotted as a function of the electron effective mass, as shown in Fig. 7.1. The range of de Broglie wavelengths in this figure spans 660 to 167 Å. For temperatures as low as 4.2 K, the de Broglie wavelength upper limit increases to a fraction of a micron. This implies that the wavelength is comparable with the size of semiconductor structures and devices in the nanostructure limit. Hence, a quantum mechanical treatment of the transport properties in nanostructures must be considered.

When electrons in semiconductors lose their wavelike behavior, they can be treated classically. This could happen when the electron scattering from impurities and imperfections of the host crystal is dominant. Another reason why electrons lose their wavelike behavior is related to finite temperature and electron statistics. The electron scattering process in semiconductors and heterojunctions is dominated by scattering from impurities (including dopants), native defects, phonons, and interfaces. The scattering processes can further be divided into elastic scattering, where the particle energy is conserved while the momentum is changed, and inelastic scattering, where both the momentum and energy of the particle are changed. In elastic scattering, the motion of the electron remains coherent. The time τ_e between two successive elastic collisions is called the mean free time and can be used to define the mean free path l_e between scattering events, such that $l_e = \tau_e v$, where v is the electron group velocity ($v = p/m$). The wavelike properties of the electrons are coherent when they travel a distance l_e. Additionally, for elastic scattering, the electron wavelike properties remain coherent even for distances larger than l_e. For inelastic scattering, the electron wave functions have different energies, and the probability of finding the electron in any state is time-dependent. The distance between inelastic collisions, l_i, in which the electrons preserve their coherent properties, is called the inelastic scattering length. Generally speaking, l_i is larger than l_e, which means that the electrons undergo several collisions before losing their energy. The inelastic scattering length can be written as $l_i = \sqrt{D\tau_i}$, where τ_i is the time between inelastic collisions and D is the diffusion coefficient given by $D = \upsilon^2 \tau_e/\alpha$, with $\alpha = 3$ for bulk, $\alpha = 2$ for quantum wells, and $\alpha = 1$ for quantum wires.

The temperature effect can cause the destruction of quantum coherence of electrons in semiconductors. As discussed in Chap. 5, the Fermi-Dirac distribution function is broadened as the temperature increases. If the thermal energy $k_B T$ is much smaller than the Fermi energy, wave functions of the electrons maintain their amplitudes, but the phase varies slightly. If the variation of the phase is sufficiently small, the temperature broadening does not break the quantum coherence properties of the electrons. For temperatures high enough that electrons with different energies participate in the transport process, the wave function phase is spread, which leads to the destruction of the quantum coherence. The phase spreading time due to the temperature effect τ_T can be estimated from the uncertainty principle as $\tau_T = \hbar/(k_B T)$. The corresponding thermal diffusion length l_T is obtained as

$$l_T = \sqrt{D\tau_T} = \sqrt{\frac{D\hbar}{k_B T}} \tag{7.2}$$

This length is the distance that electrons travel before their quantum coherence is destroyed. The thermal dephasing of the electrons occurs for both elastic and inelastic scattering.

The dephasing effects caused by inelastic collisions and thermal spreading can occur simultaneously. The coherence length l_ϕ is thus determined by the smaller value of either the inelastic scattering length or the thermal diffusion length. The superposition of the electrons' wave functions determines the transport properties in heterojunctions and nanostructures. The coherence length defines the limit below which the electrons have wavelike characteristics. This leads to the definition of *mesoscopic* systems, which are characterized by physical dimensions smaller than the coherence length. Mesoscopic devices cannot be characterized by macroscopic transport material parameters such as conductivity and drift velocity. Mesoscopic device and system behavior is determined by wavelike phenomena and is strongly dependent on the geometry of the sample, contacts, and position of the scatterers.

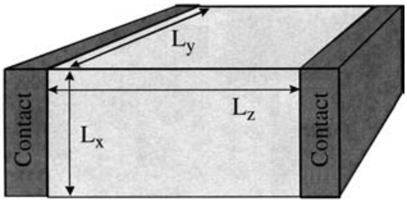


Figure 7.2 Device geometry with contacts used to define various transport regimes.

Let us assume that the dimensions of the device with the contacts described in Fig. 7.2 are $L_x < L_y < L_z$. The various regimes can be defined as shown in Table 7.1. Based on a comparison between device dimensions and the de Broglie wavelength λ, one can define the bulk device such that L_x, L_y, and L_z are all much larger than λ. For quantum well devices, L_z is on the order of λ, but L_x and L_y are much larger than λ. For quantum wire devices, L_x and L_z are on the order of λ, while L_y is much larger than λ. Finally, for a quantum box (dot), all the dimensions are on the order of λ.

The other aspects of transport properties are time and frequency. The time between successive collisions is defined as the lifetime, or free-flight time, which was previously labeled as τ_e. This time is usually much greater than the scattering duration time τ_s. In the classical regime, the relationship between lifetime and the size of the device is very important. For example, the transit time $t_{\text{tr}} = L_z/v$ determines the speed at which the signal propagates through the device, where v is the electron drift velocity. The inverse of the transit time determines the ultimate frequency at which the device can operate. For further discussion see Mitin et al. (1999).

The analysis of transport properties in quantum structures, such as quantum wells and dots, is more complicated than that for bulk

TABLE 7.1 Transport Regimes in Semiconductor Devices Given in Terms of Device Dimensions

Quantum regime	$L_z \approx \lambda$ is comparable with the electron wavelength.
Mesoscopic regime	$L_z \leq l_\phi$, where l_ϕ is the coherence length, also known as the dephasing length.
Classical regime	$L_z > l_\phi$
Classical ballistic regime	The mean free path of elastic collisions is larger than $L_z (l_e \geq L_z)$.
Classical transverse size effect	• Effects related to the mean free path: Both L_x and L_y are on the order of l_e. • Effects related to diffusion: Both L_x and L_y are on the order of l_i, where l_i is the inelastic scattering length.

NOTE: The wavelength λ is taken as the de Broglie wavelength, and the interconnect distance is L_z.

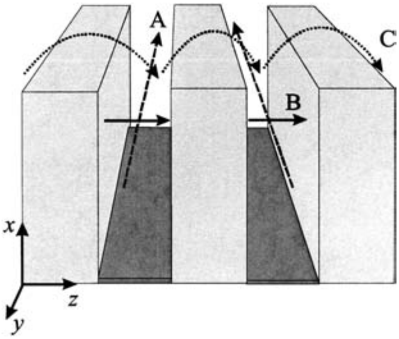


Figure 7.3 Transport mechanisms in quantum wells are shown by the arrows. (*A*) The dashed arrows represent the parallel transport, (*B*) the solid arrows represent the vertical transport by tunneling, and (*C*) the dotted arrows represent the vertical transport after photoexcitation or thermionic process.

materials. For example, three transport mechanisms can be distinguished in a multiple quantum well structure, as shown in Fig. 7.3. Mechanism *A*, depicted by the dashed arrows, is the parallel transport, where the electron's motion is along the y axis. Mechanism *B*, represented by the solid arrows, is the tunneling through barriers, where the electron transport is along the growth axis, or z direction. This transport is called vertical or perpendicular transport. Mechanism *C*, represented by the dotted arrows, is a vertical transport resulting from the excitation of carriers to higher energy levels that are close to the top of the barriers or resonant in the continuum. The excitation of carriers can be accomplished by electron-photon coupling, as is the case for intersubband transitions or by thermionic emission of the electrons over the barriers. Processes *A* and *C* can be analyzed classically or quantum mechanically, while process *B* is purely a quantum process.

7.2 The Hall Effect

Historically, Hall-effect measurements have been used extensively in determining majority carrier concentrations and their mobilities in bulk and thin-film materials. Two-dimensional electron gases formed in quantum wells and at heterojunction interfaces have been investigated by this technique. Electric and magnetic fields are essential to observe this effect. A sketch of the sample configuration is shown in Fig. 7.4. The motions of the electrons and holes under the influence of the electric and magnetic forces are shown. The configuration in this figure is constructed such that the electric current follows along the x axis, while the magnetic field is in the z direction. The force on both electrons and holes is in the $-y$ direction. In an n-type semiconductor, where the majority carriers are electrons, there is a buildup of negative charges at the $y = 0$ surface. For p-type material, positive charge buildup is also at the $y = 0$ surface. The net change produces an electric field in the $+y$ direction.

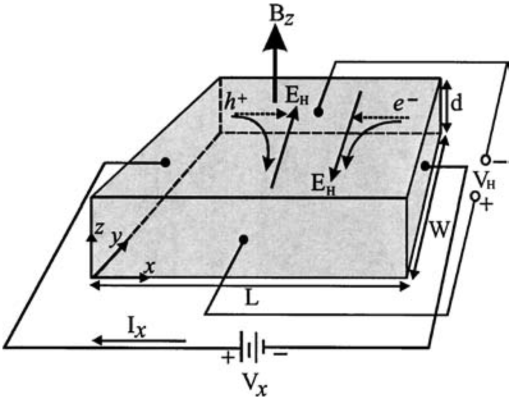


Figure 7.4 A sketch of a sample under the influence of electric and magnetic fields. This configuration is called the Hall bar.

In the steady-state case, the magnetic force is balanced by the electric force such that the net force is zero and can be expressed as

$$\begin{aligned}\mathbf{F} = 0 &= e(\boldsymbol{\mathcal{E}} + \mathbf{v} \times \mathbf{B}) \\ &= e(\mathcal{E}_x\hat{\mathbf{x}} + \mathcal{E}_y\hat{\mathbf{y}} + \mathcal{E}_z\hat{\mathbf{z}} - v_x\mathcal{E}_z\hat{\mathbf{y}}) \qquad (7.3)\end{aligned}$$

This equation yields

$$\mathcal{E}_y = v_x B_z \qquad (7.4)$$

where v_x is the drift velocity in the x direction. The electric field along the y direction expressed in Eq. (7.4) is called the Hall field, which produces the following voltage across the width W of the sample:

$$V_H = \mathcal{E}_y W = \mathcal{E}_H W \qquad (7.5)$$

where $\mathcal{E}_y = \mathcal{E}_H$ is called the Hall field. The voltage V_H is called the Hall voltage. It is negative for n-type semiconductors and positive for p-type semiconductors. Thus, the polarity of the voltage is used to determine whether the material is n-type or p-type. For n-type semiconductors, the Hall voltage can be obtained by substituting Eq. (7.4) into (7.5) to give

$$V_H = v_x B_z W \qquad (7.6)$$

Additionally, the drift velocity can be expressed as

$$v_x = -\frac{J_x}{en_H} = -\frac{I_x}{en_H A} = -\frac{I_x}{en_H Wd} \qquad (7.7)$$

where A is the area of the sample surface at $x = L$ in Fig. 7.4, which is given by the product of the sample's width W and thickness d, and n_H is the Hall electron concentration, which means the concentration obtained by Hall measurements. By substituting Eq. (7.7) into (7.6), the Hall voltage can be rewritten as

$$V_H = -\frac{I_x B_z}{edn_H} \tag{7.8}$$

The Hall voltage and the current can be measured experimentally. Hence, Eq. (7.8) can be used to determine the electron concentration

$$n_H = -\frac{I_x B_z}{eV_H d} \tag{7.9}$$

Similarly, the hole concentration in p-type semiconductors can be obtained as

$$p = \frac{I_x B_z}{eV_H d} \tag{7.10}$$

Hall mobility can now be obtained from the following relation:

$$I_x = J_x W d = en_H \mu_n \mathcal{E}_x W d = \frac{en_H \mu_n V_x W d}{L} \tag{7.11}$$

where L is the length of the sample and V_x is the applied voltage (see Fig. 7.4). From Eq. (7.11), one can obtain the electron Hall mobility as

$$\mu_n = \frac{I_x L}{eV_x n_H W d} = \frac{GL}{en_H W d} \tag{7.12}$$

where G is the sample conductance. The hole Hall mobility can be obtained in a similar manner as

$$\mu_p = \frac{I_x L}{eV_x p W d} = \frac{GL}{epWd} \tag{7.13}$$

Another parameter that is often discussed is the Hall coefficient R_H, which is defined as

$$R_H = \frac{r\mathcal{E}_y}{J_x B_z} = -\frac{r}{n_H e} \tag{7.14}$$

where r is the Hall factor, which is close to unity. For example, r is in the range of 1.0 to 1.3 for GaAs.

Generally speaking, the geometry of the sample plays a significant role in the concentration and mobility results obtained from Hall-effect measurements. The most common geometrical shape used for Hall-effect measurements is the van der Pauw geometry shown in Fig. 7.5. When using this geometry for the measurement of sheet resistance or sheet carrier concentration, one does not need to know the sample geometry. The thickness of the sample, however, should be known for volume resistivity and carrier concentrations. The validity of the van der Pauw

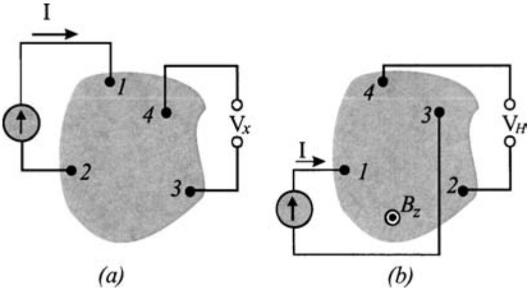


Figure 7.5 The van der Pauw configuration is shown for a sample with arbitrary shape. The configurations are for (*a*) resistivity and (*b*) Hall-effect measurements.

configuration requires that the sample has a flat, homogeneous, and isotropic surface.

The relationship between the current I and the voltage V_x in Fig. 7.5*a* is determined by mapping the arbitrarily shaped sample geometry onto a geometry that is more regular. The Laplace equation is then solved for the simpler geometry. The final results can be obtained as follows: The resistance between points i and j can be expressed as

$$R_{ij,kl} \equiv \frac{V_{kl}}{I_{ij}} \tag{7.15}$$

where the current enters contact i and leaves contact j, and V_{kl} is the voltage difference between contact k and contact l. For $B_z = 0$, the resistivity ρ is given by

$$\rho = \frac{\pi d}{\ln 2}\left(\frac{R_{21,34} + R_{32,41}}{2}\right) f \tag{7.16}$$

where d is the sample thickness and f is determined from the following equation:

$$\frac{Q-1}{Q+1} = \frac{f}{\ln 2}\operatorname{arccosh}\left[\frac{1}{2}\exp\left(\frac{\ln 2}{f}\right)\right] \tag{7.17}$$

where $Q = R_{21,34}/R_{32,41}$ if this ratio is greater than unity; otherwise $Q = R_{32,41}/R_{21,34}$ (see Look 1989). The factor f is usually close to unity for small values of Q and on the order of 0.3 for large values of Q.

Another useful approximation is to first obtain Q and then calculate α from

$$Q = \frac{\ln(0.5-\alpha)}{\ln(0.5+\alpha)} \tag{7.18}$$

and calculate f from

$$f = \frac{\ln(0.25)}{\ln(0.5 + \alpha) + \ln(0.5 - \alpha)} \tag{7.19}$$

The resistivity measurements can even be made more accurate when averaging ρ by including the two contact permutations and by reversing the current for all four permutations such that

$$\begin{aligned} \rho = \frac{\pi d}{8 \ln(2)} [&(R_{21,34} - R_{12,34} + R_{32,41} - R_{23,41}) f_A \\ &+ (R_{43,12} - R_{34,12} + R_{14,23} - R_{41,23}) f_B] \end{aligned} \tag{7.20}$$

where f_A and f_B are determined from Q_A and Q_B, respectively, by applying either Eq. (7.17) or (7.19). The quantities Q_A and Q_B are given by

$$Q_A = \frac{R_{21,34} - R_{12,34}}{R_{32,41} - R_{23,41}} \tag{7.21a}$$

$$Q_B = \frac{R_{43,12} - R_{34,12}}{R_{14,23} - R_{41,23}} \tag{7.21b}$$

The Hall voltage between contacts 4 and 2 can be written as

$$V_{H,42} = \frac{\rho \mu_n B_z I}{d} \tag{7.22}$$

and the Hall coefficient is obtained by averaging $V_{H,42}$ and $V_{H,31}$

$$R_H = \frac{d}{B_z} \left[\frac{R_{31,42} + R_{42,13}}{2} \right] \tag{7.23}$$

It is also useful to average the Hall coefficient over current and magnetic field polarities. Doing so minimizes the magnetoresistance and many other effects, such as contact resistance.

7.3 Quantum Hall and Shubnikov–de Haas Effects

Quantum transport in low-dimensional semiconductor systems is very interesting and offers the investigation of remarkable properties, such as the quantum Hall effect, the Shubnikov–de Haas effect, ballistic transport, and the fractional quantum Hall effect. For example, the Shubnikov–de Haas effect allows one to precisely measure the carrier concentrations formed at heterojunction interfaces. The investigation of two-dimensional systems in a perpendicular magnetic field provides quantization in Hall resistance (Klitzing et al. 1980), which results from the quantization of the energy in a series of Landau levels. The Landau

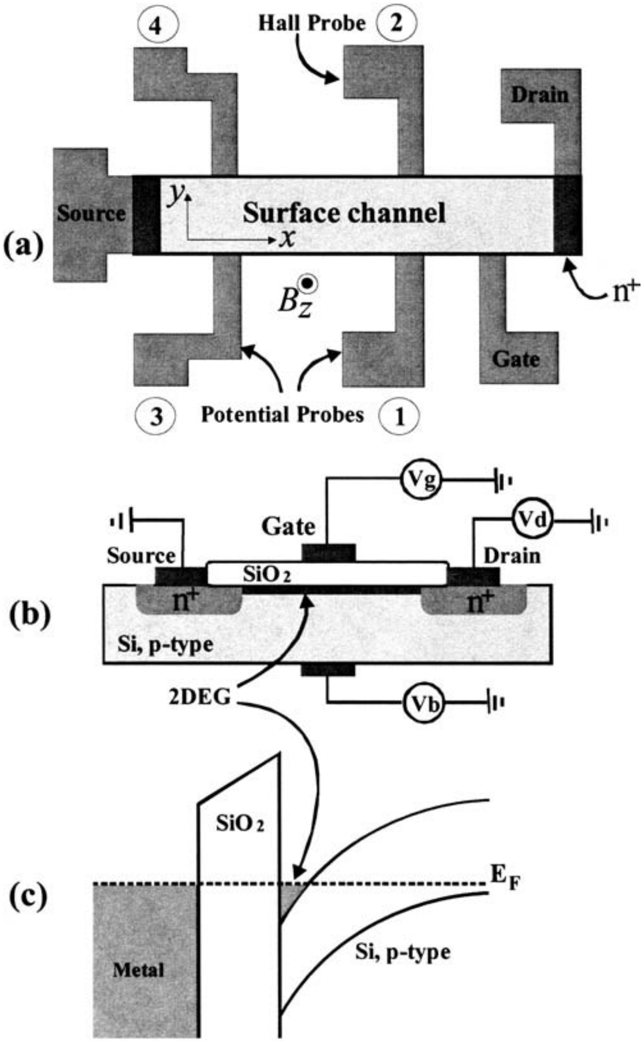


Figure 7.6 (*a*) A sketch of the device geometry used for both quantum Hall-effect and Shubnikov–de Haas measurements. (*b*) A cross section of the *n*-type MOSFET device showing the channel underneath the oxide (SiO2) layer. (*c*) The band bending near the oxide-Si interface showing the 2DEG.

magnetic length l_H (also known as the cyclotron radius of the lowest Landau energy level) assumes the role of wavelength in the quantum Hall effect, which is given by

$$l_H = \sqrt{\frac{\hbar}{eB}} \tag{7.24}$$

For $B = 10$ T, the magnetic length is $l_H \approx 8.12$ nm.

The original quantum Hall-effect device geometry used by Klitzing et al. (1980) is shown in Fig. 7.6*a*. The quantum Hall effect (QHE) measurements are made by probing the Hall voltage across points 1 and 2, while the Shubnikov–de Haas (SdH) measurements are made by probing the voltage across points 1 and 3. The device is symmetrical such that the QHE can be measured across points 3 and 4 and SDH measurements can be obtained across points 2 and 4. The initial QHE measurements were made on a Si metal-oxide semiconductor field-effect transistor as schematically shown in Fig. 7.6*b*. A two-dimensional electron gas (2DEG) is formed in the channel underneath the oxide layer as the gate voltage is applied. To create the channel, the gate voltage needs to be larger than the threshold voltage of approximately 0.7 V. The formation of the channel is very essential for the observation of both the QHE and SDH effects. The band bending at the oxide-Si interface is formed by applying a gate voltage larger than 0.7 V, as shown in Fig. 7.6*c*. The density of the 2DEG depends on the gate voltage, as well as on the drain source voltage.

Device geometry similar to that shown in Fig. 7.6 *a* has been applied to many semiconductor heterojunctions and quantum wells. The quantum Hall-effect and Shubnikov–de Haas measurements from a device with such a geometry are shown in Fig. 7.7 for an InAs/AlGaSb single quantum well. A gate, in this case, is not needed since the 2DEG is formed in

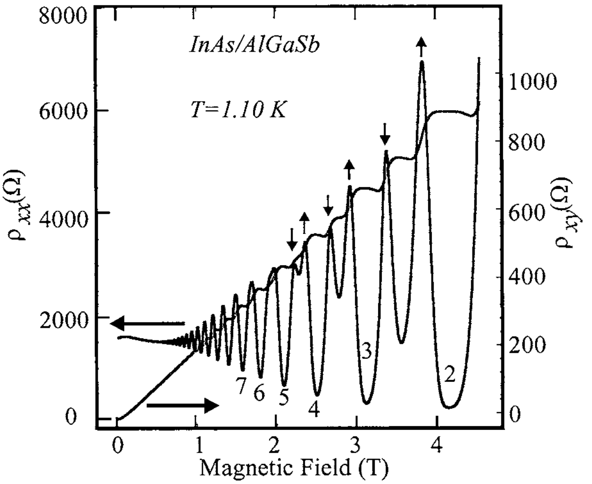


Figure 7.7 The quantum Hall-effect resistivity ρ_{xy} observed as a function of the magnetic field. The parallel resistivity ρ_{xx} represents the Shubnikov–de Haas effect. The vertical arrows indicate electron spin-up or spin-down, and the integer numbers represent the filling factor. Notice that ρ_{xy} and ρ_{xx} are sheet resistivities and their unit is ohm.

the quantum well due to the quantization of the energy levels in the two-dimensional nature of the quantum well. The InAs/AlGaSb single quantum well was chosen due to its high electron mobility and large band offset, which provide good carrier confinement. Furthermore, this system exhibits a large spin splitting due to the large effective g^* value (~ -7.6) in InAs, as compared to other systems, such as GaAs/AlGaAs quantum wells ($g^* \sim -0.2$). The g^* is obtained from the following expression, which is derived from the fourth-order effective mass theory (Palik et al. 1961):

$$g^* = 2\left(1 - \frac{(1-x)}{(2+x)}\frac{(1-y)}{y}\right) \tag{7.25}$$

where $x = 1/(1 + \Delta/E_g)$
$y = m^*/m_o$
and Δ = spin-orbit splitting energy in valence band

7.3.1 Shubnikov–de Haas effect

This effect manifests itself in the oscillations of the parallel resistivity ρ_{xx} obtained for the 2DEG in an InAs/AlGaSb single quantum well system in the presence of a high magnetic field, as shown in Fig. 7.7. The oscillations observed in ρ_{xx} are periodic as a function of $1/B_z$ in two-dimensional systems due to the constant density of states for Landau levels. The periodicity of ρ_{xx} can be used to extract the 2DEG carrier density. Since the resistivity is expected to be minimum when the Fermi level lies between two Landau levels, where the density of states is the smallest, one can define the Landau level filling factor ν as

$$\nu \equiv \frac{n_s h}{e B_z} \tag{7.26}$$

where n_s is the density of the 2DEG and the filling factor ν is an integer $(1, 2, 3, \ldots)$. This equation assumes degenerate spin and valley Landau levels. Thus, for adjacent Landau levels, we have

$$n_s = \frac{e}{h}\,\frac{1}{\Delta(1/B_z)} \tag{7.27}$$

An accurate measurement of n_s is obtained for larger ν (small values of B_z) where the spin-splitting is minimum, as shown in Fig. 7.7. The sheet carrier density obtained by this method is more accurate than that obtained by conventional Hall-effect measurements. This is mainly due to the fact that the conventional Hall effect does not distinguish between 2D and 3D carriers, but the results from both techniques are usually very close in value. The carrier concentration can also be obtained by

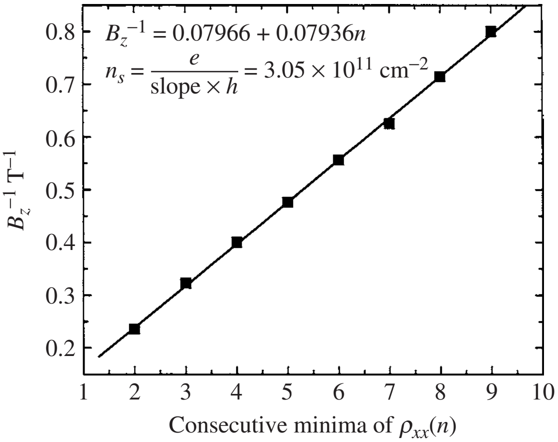


Figure 7.8 The inverse of the magnetic field plotted as a function of the consecutive minima obtained from ρ_{xx} in Fig. 7.7. The line is a linear fit to the data. The slope of the line is used to calculate the density of the 2DEG

plotting B_z^{-1} against the consecutive minima of ρ_{xx} (n), as shown in Fig. 7.8. The slope of the plot is related to the 2DEG density through the following relation, $n_s = e/(\text{slope} \times \text{h})$.

While the effective mass is not included in the SdH oscillations, it can be determined by investigating the oscillation amplitudes as a function of temperature and magnetic field of the low-field oscillatory conductivity expression derived by Ando et al.

$$\rho_{xx}^{-1} = \sigma_{xx} = \frac{n_s e^2 \tau_f}{m^*} \frac{1}{1 + (\omega_c \tau_f)^2} \left[1 - \frac{2(\omega_c \tau_f)^2}{1 + (\omega_c \tau_f)^2} \frac{2\pi^2 k_B T}{\hbar \omega_c} \right]$$
$$\times \cosh\left(\frac{2\pi^2 k_B T}{\hbar \omega_c} \right) \cos\left(\frac{2\pi E_F}{\hbar \omega_c} \right) \exp\left(-\frac{\pi}{\omega_c \tau_f} \right) \quad (7.28)$$

where E_F is the Fermi energy given by

$$E_F = \frac{\hbar^2 k_F^2}{2m^*} = \frac{2\pi \hbar^2 n_s}{m^*} \quad (7.29)$$

and τ_f = scattering time corresponding to dephasing of Landau state
ω_c = cyclotron angular frequency = $\frac{|e|B_z}{m^*}$
k_B = Boltzmann's constant
T = temperature

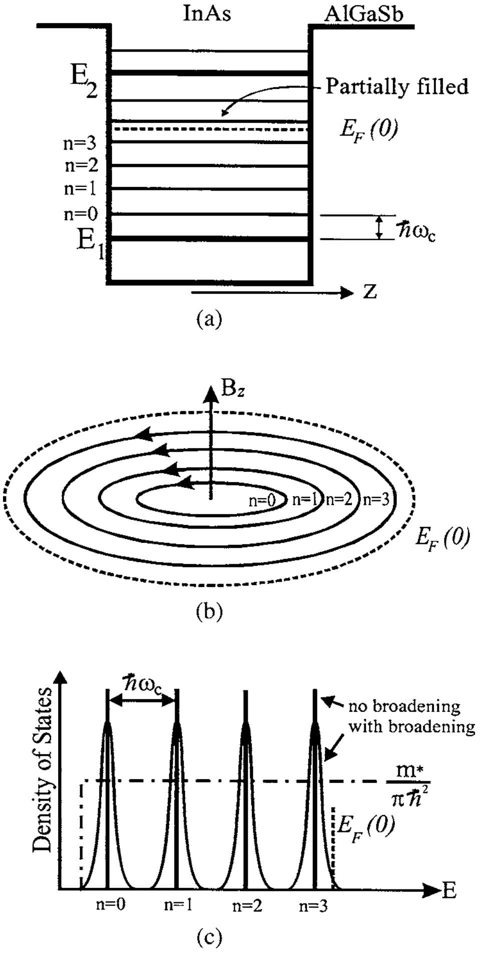


Figure 7.9 (*a*) A sketch of an InAs/AlGaSb single quantum well showing two bound states (E_1 and E_2), the Fermi energy level, $E_F(0)$, and the Landau levels. (*b*) Landau levels are filled up to the Fermi energy level, which contains all allowed states when the magnetic field is zero. (*c*) Energy representation of Landau levels and Fermi level. The Landau levels are broadened due to various scattering mechanisms.

The scattering time τ_f can also be extracted from Eq. (7.28). Both the effective mass and scattering time values can be quite different from the values obtained from the Hall-effect and cyclotron resonance measurements.

The origin of the oscillations in ρ_{xx} can be understood by examining Fig. 7.9. An n-type doped InAs/AlGaSb single quantum well is sketched in Fig. 7.9*a*, where we assume two bound states E_1 and E_2 exist with the Fermi energy $E_F(0)$ at a zero magnetic field assumed to be between E_1 and E_2. By applying a magnetic field along the growth axis (z direction), each electronic energy level splits into an n number of Landau levels with an energy described in Sec. 6.12, Eq. (6.140). The separation between Landau levels is $\hbar\omega_c$. The density of states per unit area of each Landau level is obtained from the following relation:

$$(\hbar\omega_c)\left(\frac{m^*}{2\pi\hbar^2}\right) = \frac{m^*\omega_c}{2\pi\hbar} = \frac{eB_z}{\hbar} \tag{7.30}$$

As discussed in Chap. 5, the density of states for bulk semiconductor material under the influence of a magnetic field resembles the quantum wire density of states due to the confinement of the electrons in Landau orbits in the xy plane. In quantum wells and heterojunctions, the electrons are confined in the z direction. By applying a static magnetic field along the z direction, the electrons are further confined in the xy plane as shown in Fig. 7.9*b*, leading to a zero degree of freedom (confinement in the three directions). Thus, the density of states of each Landau energy-level is simply a δ-function with a degeneracy of eB_z/h. This is shown in Fig. 7.9*c* as the solid vertical lines labeled "No broadening." In reality, however, the impurities, alloy fluctuations, interface roughness, and crystal imperfections will broaden the δ-function density of states. This broadening of Landau levels is depicted as a gaussian lineshape (see the curves labeled "With broadening" in Fig. 7.9*c*). The dashed-dotted line in Fig. 7.9*c* is the two-dimensional density of states of the electronic energy levels E_1 in the absence of the magnetic field. The states at the tails of the Landau levels are called localized states, and they play an important role in the quantum Hall effect. As the magnetic field is increased, Landau levels are swept across the Fermi levels giving rise to the observed oscillations in ρ_{xx}.

7.3.2 Quantum Hall effect

As a starting point, it is very beneficial to understand the Drude classical model of the magnetoresistance in semiconductors. The classical equation of motion of an electron in the presence of magnetic ($\mathbf{B}$) and electric ($\boldsymbol{\mathcal{E}}$) fields can be expressed as

$$m^*\frac{d\boldsymbol{\upsilon}}{dt} + m^*\frac{\boldsymbol{\upsilon}}{\tau} = -e(\mathcal{E} + \boldsymbol{\upsilon} \times \mathbf{B}) \tag{7.31}$$

where $\boldsymbol{\upsilon}$ is the drift velocity and τ is the scattering time. The magnetic field is applied along the z axis, and $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{\upsilon}$ are assumed to vary with time as exp($-\omega t$). This equation can be expressed in its three components as

$$\begin{aligned} m^*\frac{d\,\upsilon_x}{dt} + m^*\frac{\upsilon_x}{\tau} &= -e\mathcal{E}_x - e\upsilon_y B_z \\ m^*\frac{d\,\upsilon_y}{dt} + m^*\frac{\upsilon_y}{\tau} &= -e\mathcal{E}_y + e\upsilon_x B_z \\ m^*\frac{d\,\upsilon_z}{dt} + m^*\frac{\upsilon_z}{\tau} &= -e\mathcal{E}_z \end{aligned} \tag{7.32}$$

By multiplying Eq. (7.32) by the carrier concentration n_s and the electron charge $-e$, and comparing the results with the relation

$$\mathbf{j} = \overset{\leftrightarrow}{\sigma} \cdot \boldsymbol{\mathcal{E}} \tag{7.33}$$

where $\overset{\leftrightarrow}{\sigma}$ is the conductivity tensor, one can obtain the components of the conductivity tensor as

$$\begin{aligned} \sigma_{xx} &= \sigma_{yy} = \frac{\sigma_o(1 - i\omega\tau)}{1 - \left(\omega^2 - \omega_c^2\right)\tau^2 - 2i\omega\tau} \\ \sigma_{zz} &= \frac{\sigma_o}{1 - i\omega\tau} \\ \sigma_{xy} &= -\sigma_{yx} = \frac{\sigma_o \omega_c \tau}{1 - \left(\omega^2 - \omega_c^2\right)\tau^2 - 2i\omega\tau} \\ \sigma_{xz} &= \sigma_{zx} = \sigma_{yz} = \sigma_{zy} = 0 \end{aligned} \tag{7.34}$$

where $\sigma_o = n_s e^2 \tau / m^*$ is the conductivity in the absence of the magnetic field. For the steady-state case where $d\mathbf{\upsilon}/dt = 0$, the conductivity tensor can be written as

$$\overset{\leftrightarrow}{\sigma} = \frac{\sigma_o}{1 + (\omega_c\tau)^2} \begin{pmatrix} 1 & -\omega_c\tau & 0 \\ \omega_c\tau & 1 & 0 \\ 0 & 0 & 1 + (\omega_c\tau)^2 \end{pmatrix} \tag{7.35}$$

Thus, the conductivity in a two-dimensional system in the presence of a magnetic field applied along the z direction can be expressed as

$$\overset{\leftrightarrow}{\sigma} = \frac{\sigma_o}{1 + (\omega_c\tau)^2} \begin{pmatrix} 1 & -\omega_c\tau \\ \omega_c\tau & 1 \end{pmatrix} \tag{7.36}$$

and the resistivity tensor is related to the conductivity tensor as

$$\overset{\leftrightarrow}{\rho} = \overset{\leftrightarrow}{\sigma}^{-1} \tag{7.37}$$

The resistivity tensor can now be written as

$$\overset{\leftrightarrow}{\rho} = \frac{1}{\sigma_{xx}^2 + \sigma_{xy}^2} \begin{pmatrix} \sigma_{xx} & -\sigma_{xy} \\ \sigma_{xy} & \sigma_{xx} \end{pmatrix} \tag{7.38}$$

The condition $\omega_c\tau \gg 1$ implies that the carriers are collisionless. By applying this condition to Eq. (7.34), one can obtain $\sigma_{xx} \approx 0$ and $\sigma_{xy} \approx -n_s e/B$. In the presence of collisions, where $\omega_c\tau \geq 1$, we have

$$\begin{aligned} \sigma_{xx} &= \frac{n_s e}{B_z} \frac{\omega_c^2 \tau^2}{1 + \omega_c^2 \tau^2} \\ \sigma_{xy} &= -\frac{n_s e}{B_z} - \frac{\sigma_{xx}}{\omega_c \tau} \end{aligned} \tag{7.39}$$

where these conductivity components are simply the sum of the collision and collisionless parts. When the Fermi energy level is between Landau levels labeled n and $n+1$, no elastic scattering can occur at low temperatures ($T \le 4.2$ K), and the energy separation between consecutive Landau levels is $\hbar\omega_c$. This case is thus equivalent to the condition $\omega_c\tau \gg 1$, which gives $\sigma_{xx} \approx 0$, and σ_{xy} is given by its classical collisionless value. From the density of states per Landau level, eB/h, one can write the carrier density n_s as $n_s = neB/h$, where n is the nth Landau level. The Hall conductivity σ_{xy} can be expressed as

$$\sigma_{xy} = \frac{n_s e}{B_z} = \frac{e}{B_z}\frac{neBz}{h} = n\frac{e^2}{h} \qquad \text{and} \qquad \rho_{xy} = \frac{1}{n}\frac{h}{e^2} \tag{7.40}$$

This equation shows that the Hall resistivity takes quantized values of $25812.87/n$ whenever the Fermi energy level lies between filled-broadened Landau levels, as illustrated by the plateaus in Fig. 7.7. This is called the quantum Hall effect.

The quantum Hall effect is observed for integer filling factors as described in Eq. (7.26). However, at low temperatures ($T < 5.2$ K), a fractional value of the filling factor ν has been observed for the lowest Landau level in many heterojunction systems with high mobility. In this case, ν can take on values of p/q, where p and q are integers. This is called the fractional quantum Hall effect (see Tsui et al. 1983). Laughlin (1983) provided an explanation of the fractional quantum Hall effect based on the condensation of electrons or holes into a collective ground state due to electron-electron or hole-hole interactions. This ground state is separated from the nearest excited state by an energy of $0.03e^2/l_H$, where l_H is the Landau magnetic length. The possibility of a repulsive interaction between carriers of the same charge, leading to a condensation, is related to the two-dimensional character of the system. The condensed phase consists of quasi-particles called *anyons*, of fractional charge $2/l$, where $l = 3, 5, 7, \ldots$, that follow statistics intermediate between Fermi-Dirac and Bose-Einstein formalisms.

7.4 Charge Carrier Transport in Bulk Semiconductors

As discussed in the introduction of this chapter, there are several mechanisms that impact charge transport in bulk and low-dimensional systems. For example, tunneling, which is discussed in previous chapters, is a quantum effect that cannot be explained in terms of classical theory. In this section, we discuss various transport properties of bulk semiconductors.

7.4.1 Drift current density

The resultant movement of the electrons and holes in semiconductors under the influence of an applied electric field is called drift, which gives rise to the drift currents. The equation of motion of an electron with mass m^* under the influence of an electric field $\mathcal{E}$ is given by

$$m^* \frac{d\,\upsilon_d(t)}{dt} = -e\mathcal{E} \tag{7.41}$$

where $\upsilon_d(t)$ is the drift velocity, which, after integration, is given as

$$\upsilon_d(t) = -\frac{e\mathcal{E}}{m^*}t \tag{7.42}$$

The drift velocity increases linearly with time between collisions. The mean value of the drift velocity is

$$\langle \upsilon_d \rangle = \int_0^\infty \upsilon_d(t)\mathcal{P}(t)\,dt = -\frac{e\mathcal{E}}{m^*}\int_0^\infty t\mathcal{P}(t)\,dt = -\frac{e\tau}{m^*}\mathcal{E} \tag{7.43}$$

where τ is the time that it takes for a carrier to suffer two successive collisions and $\mathcal{P}(t)$ is the probability that a carrier has *not* made a collision at time t and is given by

$$\mathcal{P}(t) = \frac{1}{\tau}\exp\left(-\frac{t}{\tau}\right) \tag{7.44}$$

From Eq. (7.43), the electron mobility can be expressed as $\mu = |e|\tau/m^*$. The current density can now be written as

$$j_e = -ne\upsilon_e = ne\mu_e\mathcal{E} \tag{7.45}$$

where n is the electron density and the subscript e stands for electrons. For holes, the current density is

$$j_h = pe\upsilon_h = pe\mu_h\mathcal{E} \tag{7.46}$$

where p is the hole density. For the preceding current densities we have assumed that the drift velocity is linearly dependent on the electric field and that the mobility is independent of the electric field. This may not be the case for high electric fields ($\mathcal{E} > 10^4$ V/cm). In the case of the high electric field regime, the relaxation time, drift velocity, and mobility can all be dependent on the electric field. For additional discussion on the saturation of the drift velocity, see Look (1989) and Sze (2002). For mixed conduction, where both electrons and holes are present, the total current density is the sum of Eqs. (7.45) and (7.46), which gives a total conductivity of

$$\sigma = ne\mu_e + pe\mu_h = e^2\left(\frac{n\tau_e}{m_e^*} + \frac{p\tau_h}{m_h^*}\right) \tag{7.47}$$

The mobility in this equation is called conductivity mobility. The carrier mobility can be determined by different methods, such as the Hall-effect and magnetoresistance techniques, which may lead to different values for the mobility. The mobility determined from various techniques depends, however, on the scattering time or the relaxation time, which was defined previously as the time between two successive collisions or scattering events. The determination of the scattering time depends on several effects that take place as the charge carrier is drifting from one end of a material to the other under an applied electric field. For example, the scattering mechanisms in GaAs include defect scattering, such as intrinsic defects, charged and neutral impurities, and alloying; carrier-carrier scattering; and lattice scattering. Lattice scatterings may be due to intervalley scattering (acoustical and/or optical phonons) and intravalley scattering (phonons, deformation potential, piezoelectric, etc.). The scattering time τ can be written as

$$\frac{1}{\tau} = \frac{1}{\tau_1} + \frac{1}{\tau_2} + \frac{1}{\tau_3} + \cdots \tag{7.48}$$

where the subscripts indicate different types of scattering. Consequently, the mobility of electrons can be expressed as

$$\frac{1}{\mu} = \frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{\mu_3} + \cdots \tag{7.49}$$

For example, the mobility due to lattice scattering was shown to depend on $T^{-3/2}$, where T is the temperature (Smith 1978), and the mobility due to impurity scattering varies as $T^{3/2}/N_i$, where N_i is the total impurity concentration.

The mobility also depends on the effective mass of the charge carriers. When the effective mass is obtained from the conductivity measurements, it is called the *mobility effective mass*. The values of the mobility effective mass may differ from those obtained from the cyclotron resonance and Shubnikov–de Haas experiments.

For low values of an applied electric field, the drift velocity of charge carriers in semiconductor materials and devices exhibits a linear relationship as a function of the electric field. However, many devices operate at high electric fields ($\mathcal{E} \approx 1$ to 100 kV/cm) where the drift velocity is no longer linear with $\mathcal{E}$. An example of the drift velocity under

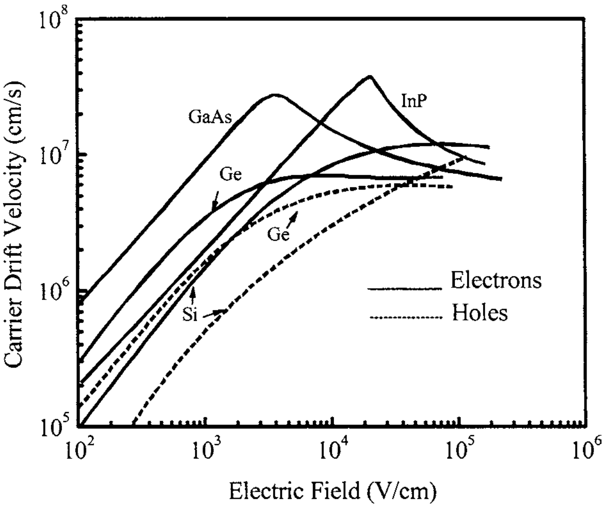


Figure 7.10 Carrier drift velocities as a function of the electric field for SiGe, GaAs, and InP.

the influence of a high electric field is shown in Fig. 7.10, where the drift velocity becomes almost independent of the electric field for Si and Ge. The drift velocity saturation, which is independent of the electric field, in this figure is due to the fact that electrons (holes) gain high energy (*hot electrons* or *holes*) from the electric field and their scattering rates are increased, leading to a reduction in the scattering time (scattering lifetime). The reduction in lifetime causes the mobility to decrease.

The curves related to GaAs and InP in Fig. 7.10 exhibit a negative differential mobility at high electric fields, which produces a negative differential resistance. This characteristic, however, is useful in the design of oscillators and low-power microwave devices. The drift velocity–electric field behavior in GaAs, as well as in many direct bandgap materials can be explained in terms of the conduction valley occupancy (see, for example, Singh 2003). As shown in Fig. 7.11, the electrons move in the high-mobility (μ_Γ) Γ valley at low electric fields, where the effective mass is $0.067m_o$. The velocity peaks at around 4 to 5 × 10^5 V/cm, where most of the electrons are still in the Γ valley. At higher electric fields, the electrons gain enough energy to transfer to the L valley where the electron effective mass is much heavier ($\sim 0.22m_o$) and the mobility (μ_L) is lower. The transfer of the electrons from the Γ valley to the L valley is the cause of the negative differential mobility, which leads to the negative differential resistance. To observe the negative differential resistance, the energy separation between the L and Γ valleys

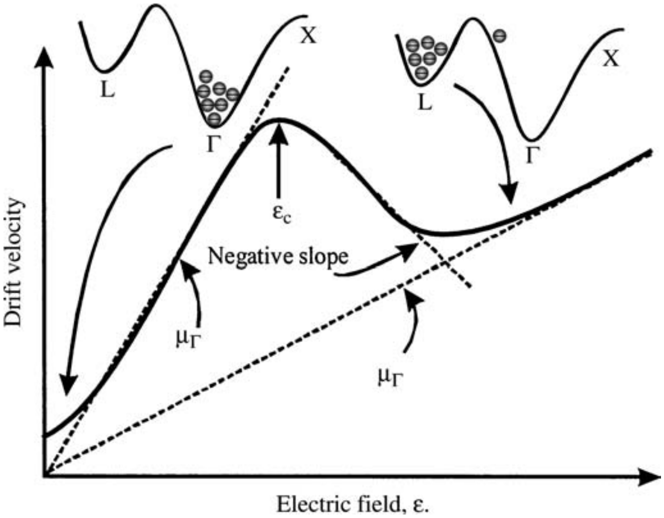


Figure 7.11 Illustration of the electron transfer from the Γ valley to the L valley in the conduction band of GaAs as the applied electric field is increased. The associated drift velocity behavior as a function of the electric field is shown with a negative slope on the right-hand side of the peak.

should be much larger than $k_B T$ so that the L-valley will not be thermally populated with electrons. In the case of GaAs, this separation energy is ~0.32 eV. An additional condition necessary to the observation of the negative differential resistance is that the separation between the L and Γ valleys should be less than the bandgap of the semiconductor to avoid populating the L valley by exciting carriers from the valence band to this valley through mechanisms such as impact ionizations.

When electrons are injected into a semiconductor by application of an electric field, they suffer several collisions in a certain period of time (several picoseconds) before they reach a steady-state distribution. If electrons are injected into the upper valley, where the effective mass is high, the injected electrons may have velocities lower than the steady-state velocity for a short period of time. This leads to what is called *velocity undershoot*. *Velocity overshoot* is when the electrons are injected ballistically into the sample and stay in the Γ valley with velocities higher than the steady-state velocity. Eventually the electrons suffer scattering, and their velocity decreases in time to the steady-state velocity.

The negative differential resistance is very useful in microwave devices and oscillators. The negative-slope region in the drift velocity versus electric field curve usually occurs when a high electric field

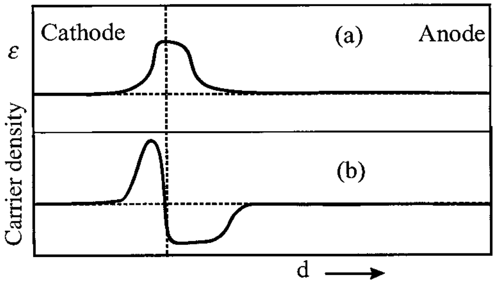


Figure 7.12 (*a*) Electric field and (*b*) carrier density as a function of distance for a domain moving from the cathode to the anode in a GaAs thin sample under the influence of high electric field ($> 10^5$ V/cm).

($> 5 \times 10^5$ V/cm) is applied to semiconductor materials, such as GaAs or InP. In this region, instability can arise and current oscillation can occur. These oscillations were first observed by Gunn in 1963 and are called Gunn oscillations. These oscillations are observed in thin samples (on the order of 10 μm) under the influence of an electric field higher than a critical field ($\mathcal{E}_c$), as shown in Fig. 7.11.

The frequency of the oscillations is found to be equal to the electron drift velocity divided by the length of the sample. The origin of Gunn oscillations is that there is a fluctuation called the electric field *domain* formed near the cathode, as shown in Fig. 7.12*a*, where the carriers pile up on the left-hand side of the domain, while the carriers on the right-hand side of the domain are depleted, as shown in Fig. 7.12*b*. Because of the negative differential resistance for $\mathcal{E} > \mathcal{E}_c$, the increase in the field inside the domain causes further slowing down of the electrons inside the domain, which leads to more charge pileup. The pileup process continues until most of the applied field is across the domain.

Only one domain can exist inside the sample at one time. The domain drifts across the sample from the cathode toward the anode at the saturation velocity under the influence of an applied bias voltage. The domain disappears once it reaches the anode, and a new domain is formed, giving rise to current oscillation. If the saturation velocity is 10^7 cm/s and the length of the sample is 10^{-4} cm, the oscillation frequency is 10 GHz. This frequency is in the microwave region. Thus, Gunn diodes are known as microwave generators and have applications in radar and communications.

7.4.2 Diffusion current density

When there is a spatial variation of carrier concentration in semiconductors, the carriers move from regions of high concentration to regions of low concentration. The movement of the carriers results in what is called *diffusion current*. The carrier diffusion is governed by Fick's law, which states that the carrier flux $\mathcal{F}_n$ is proportional to the concentration

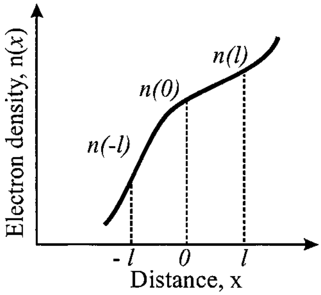


Figure 7.13 An example of electron concentration variation as a function of distance used to illustrate Fick's law.

gradient. For electrons, Fick's law has the following form:

$$\mathcal{F}_n = -D_e \frac{dn}{dx} \tag{7.50}$$

where D_e is the electron diffusion coefficient and n is the electron concentration. Fick's law can be verified by assuming that the electron concentration in a semiconductor at a constant temperature varies along the x axis such that $n(x)$ is described by the curve in Fig. 7.13. The average electron flux $\mathcal{F}_1$ crossing the concentration profile from the left can be expressed as follows:

$$\mathcal{F}_1 = \frac{n(-l) \cdot l}{2\tau} = \frac{n(-l)\upsilon_{\text{th}}}{2} \tag{7.51}$$

where τ = mean free time between collisions
l = mean free path
υ_{th} = electron thermal velocity ($\upsilon_{\text{th}} = l/\tau$)

Similarly, the average electron flux $\mathcal{F}_2$ crossing from right to left is

$$\mathcal{F}_2 = \frac{n(l)\upsilon_{\text{th}}}{2} \tag{7.52}$$

The net carrier flow from left to right is thus the difference between the two fluxes,

$$\mathcal{F}_n = \mathcal{F}_1 - \mathcal{F}_2 = \frac{\upsilon_{\text{th}}}{2}[n(-l) - n(l)] \tag{7.53}$$

One can now expand the carrier concentration at $x = \pm l$ by using the Taylor series to the first order to obtain

$$\begin{aligned} \mathcal{F}_n &= \frac{\upsilon_{\text{th}}}{2}\left[n(0) - l\frac{dn}{dx} - n(0) - l\frac{dn}{dx}\right] \\ &= -\upsilon_{\text{th}} l \frac{dn}{dx} = -D_e \frac{dn}{dx} \end{aligned} \tag{7.54}$$

where $D_e = \upsilon_{\text{th}} l$. The diffusion current density for conduction electrons can now be expressed as

$$J_e = eD_e \frac{dn}{dx} \tag{7.55}$$

Similarly, the hole diffusion current is

$$J_h = -eD_h \frac{dp}{dx} \tag{7.56}$$

where D_h is the hole diffusion constant and p is the hole concentration. When both the electric field and concentration gradient are present, the current densities for electrons and holes can be written as

$$J_e = ne\mu_e \mathcal{E} + eD_e \frac{dn}{dx} \tag{7.57a}$$

$$J_h = pe\mu_h \mathcal{E} - eD_h \frac{dp}{dx} \tag{7.57b}$$

For mixed conduction in three dimensions, the total current density, which consists of the drift and diffusion components for both electrons and holes, can be generalized as

$$J = ne\mu_e E + pe\mu_h \mathrm{E} + eD_e \nabla n(r) - eD_h \nabla p(r) \tag{7.58}$$

For a semiconductor at equilibrium, the current density of each type of carrier must be zero. For electrons, Eq. (7.57a) can be written as

$$ne\mu_e \mathcal{E} = -eD_e \frac{dn}{dx} \tag{7.59}$$

Furthermore, the electric field is related to the electric potential $V(r)$ according to the following relation:

$$\mathcal{E} = -\nabla V(r) \tag{7.60}$$

Substituting Eq. (7.60) into (7.59), we have

$$n\mu_e \nabla V(r) = D_e \nabla n(r) \tag{7.61}$$

The carrier concentration under nondegenerate conditions can be written as

$$n(r) = N_c \exp\left[\frac{E_c - eV(r) - E_F}{k_B T}\right] \tag{7.62}$$

where the conduction band edge is modified in the presence of the applied voltage $V(r)$. By taking the gradient of Eq. (7.62), one can obtain

$$\nabla n(r) = \frac{en(r)\nabla V(r)}{k_B T} \tag{7.63}$$

Substituting Eq. (7.63) into (7.61) gives

$$D_e = \frac{k_B T \mu_e}{e} \tag{7.64}$$

Similarly, the hole diffusion coefficient is

$$D_h = \frac{k_B T \mu_h}{e} \tag{7.65}$$

Equations (7.66) and (7.67) are known as the *Einstein relations*. Substituting Eqs. (7.64) and (7.65) into Eqs. (7.57*a*) and (*b*), respectively, we obtain

$$\mathbf{J}_e = \mu_e(ne\mathcal{E} + k_B T \nabla n) \tag{7.66a}$$

$$\mathbf{J}_h = \mu_h(pe\mathcal{E} - k_B T \nabla p) \tag{7.66b}$$

It is clear from these equations that the current density is proportional to the mobility even in the presence of carrier diffusion.

In high-frequency electronic components, an additional current density contribution, called the *displacement current density,* becomes important. Assume that the electrons in a semiconductor material are subject to an alternating current (ac) electric field of the form

$$\mathcal{E} = \mathcal{E}_o \exp(-i\omega t) \tag{7.67}$$

The displacement current density $\mathbf{J}_d$ is given by

$$\mathbf{J}_d = \frac{\partial \mathcal{D}}{\partial t} \tag{7.68}$$

where $\mathcal{D}$ is the electric displacement given by

$$\mathcal{D} = \epsilon\epsilon_o\mathcal{E} \tag{7.69}$$

where ϵ is the dielectric constant and ϵ_o is the permittivity of space (8.85×10^{-12} F/m). Combining Eqs. (7.67) to (7.69) yields

$$\mathbf{J}_d = -i\omega\epsilon\epsilon_o\mathcal{E} \tag{7.70}$$

By combining Eq. (7.70) with the static contribution of the current density derived above, we have

$$\mathbf{J} = (\sigma - i\omega\epsilon\epsilon_o)\mathcal{E} \tag{7.71}$$

Thus, the electrical conductivity is composed of dc and ac components. Again, the ac component is significant and cannot be neglected in the case of high-frequency devices.

7.4.3 Generation and recombination

For a semiconductor at thermal equilibrium and zero bias voltage, the product of electron (n) and hole (p) concentrations is given by $np = n_i^2$, where n_i is the intrinsic carrier concentration. Many electronic devices, such as bipolar transistors and *pn*-junction diodes, operate on the principle of carrier injection. For excess carriers, the semiconductor is no longer at equilibrium and $np > n_i^2$. The generation of excess carriers can be accomplished by several techniques, but the most common methods are either applying a bias voltage or illuminating the sample with photons. The thermionic process is when electrons gain enough thermal energy to allow them to make transitions to higher energy levels. The introduction of excess carriers is called *carrier generation*. When the system is at nonequilibrium, a process exists to restore the system back to equilibrium. This mechanism is called *recombination*. For example, when a semiconductor sample is illuminated with light, electrons absorb the photons to make the transition from the valence band to the conduction band, leaving behind holes with positive charges. The excited electrons recombine with holes in the valence band, releasing energy in the form of photons (luminescence) or phonons (thermal energy). If photons are emitted as a by-product of the recombination, the process is called *radiative recombination*. *Nonradiative recombination* occurs when the energy of the electron is absorbed by the lattice. When the excited electrons recombine directly with holes in the valence band, the process is called *direct recombination*. If the recombination process is made through centers with energy levels lying in the fundamental bandgap, the process is called *indirect recombination*.

Direct recombination is common in direct bandgap materials such as GaAs and GaN. Figure 7.14 illustrates the generation and recombination processes in a direct bandgap semiconductor. The quantity $\mathcal{g}_l$ is the light generation rate, $\mathcal{g}_t$ is the thermal generation rate, and $\mathcal{R}$ is the recombination rate. The units of these rates are number/(cm^3·s). For

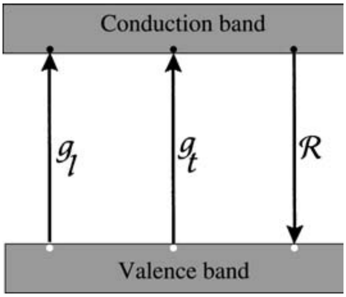


Figure 7.14 Direct generation and recombination of electron-hole pairs during illumination of the sample with photons.

thermal equilibrium, g_l is zero and $g_t = \mathcal{R}$. For direct recombination, where the bottom of the conduction band and the top of the valence band are lined up, the recombination rate is given by

$$\mathcal{R} = \alpha_r pn \tag{7.72}$$

where α_r is the recombination rate proportionality constant. For the nonequilibrium case when an n-type semiconductor specimen is subject to illumination by light, the recombination rate can be written as

$$\mathcal{R} = \alpha_r p_n n_n = \alpha_r\left(n_n^o + \Delta n\right)\left(p_n^o + \Delta p\right) \tag{7.73}$$

where $n_n =$ total majority carrier concentration
$p_n =$ total minority concentration
$n_n^o =$ equilibrium majority carrier concentration
$p_n^o =$ equilibrium minority carrier concentration

Δn and Δp are the excess carrier concentrations defined as

$$\Delta n = n_n - n_n^o \qquad \text{and} \qquad \Delta p = p_n - p_n^o \tag{7.74}$$

To maintain charge neutrality, Δn and Δp must be equal. The total generation rate $\mathcal{G}$ is the sum of the thermal and light generation rates. Thus, the net rate of change of the hole concentration can be expressed as

$$\frac{dp_n}{dt} = \mathcal{G} - \mathcal{R} = g_l + g_t - \mathcal{R} \tag{7.75}$$

For the steady-state case, the left-hand side of Eq. (7.75) is zero and

$$g_l = \mathcal{R} - g_t \tag{7.76}$$

Thus, g_l can be considered as the net recombination rate. For thermal equilibrium, we have

$$g_t = \mathcal{R} = \alpha_r p_n^o n_n^o \tag{7.77}$$

Substituting Eqs. (7.73) and (7.77) into Eq. (7.76) and taking $\Delta n = \Delta p$, we obtain for the net recombination rate, the following expression:

$$\begin{aligned} g_l &= \mathcal{R} - g_t = \mathcal{R} - \alpha_r p_n^o n_n^o \\ &= \alpha_r p_n n_n - \alpha_r p_n^o n_n^o = \alpha_r \left(n_n^o + \Delta n\right)\left(p_n^o + \Delta p\right) - \alpha_r p_n^o n_n^o \\ &= \alpha_r \Delta p \left(n_n^o + p_n^o + \Delta p\right) \end{aligned} \tag{7.78}$$

For $p_n^o \ll n_n^o$ and $\Delta p \ll n_n^o$, the net recombination rate becomes

$$g_l = \alpha_r \, \Delta p n_n^o = \frac{p_n - p_n^o}{\tau_p} \tag{7.79}$$

where τ_p is the excess minority lifetime $[\tau_p = 1/(\alpha_r n_n^o)]$. This equation describes the net recombination rate when the sample is subject to light illumination with photon energy larger than the fundamental bandgap energy. For indirect semiconductors such as Si, the derivation of the net rate is left as an exercise (see Prob. 7.8).

In addition to direct and indirect recombination, there are other recombination mechanisms. An example of these mechanisms is surface recombination. Since the lattice structure of any semiconductor material at the surface contains a large number of dangling bonds, called *surface states*, the recombination rate may be enhanced at the surface. Another example of recombination is *Auger recombination*, where the energy from electron-hole recombination is released to a third particle (either an electron or hole). The third particle loses its energy to the lattice, rendering Auger recombination as a nonradiative recombination process. There are several types of Auger recombinations, but they will not be discussed here.

7.4.4 Continuity equation

The continuity equation combines the drift, diffusion, generation, and recombination processes into a single equation. To construct the continuity equation in one dimension (along the x axis), consider an element of thickness dx in a sample of cross-sectional area A, as shown in Fig. 7.15. The overall rate of change in the number of electrons in this slice can be written as the sum of the electrons entering at x minus the number of electrons leaving at $dx + x$ plus the electron generation rate minus the electron recombination rate. This can be written as

$$\frac{\partial n}{\partial t} A \, dx = \left[\frac{J_n(x)A}{-e} - \frac{J_n(x+dx)A}{-e} \right] + (\mathcal{G}_n - \mathcal{R}_n) A \, dx \tag{7.80}$$

where $A \, dx$ = volume of slice
$\mathcal{G}_n$ = electron generation rate
$\mathcal{R}_n$ = electron recombination rate
$-e$ = charge of electron

Expanding $J_n(x + dx)$ in terms of a Taylor series and retaining the first-order terms, we obtain

$$J_n(x + dx) = J_n(x) + [\partial J_n(x)/\partial x] dx + \cdots \tag{7.81}$$

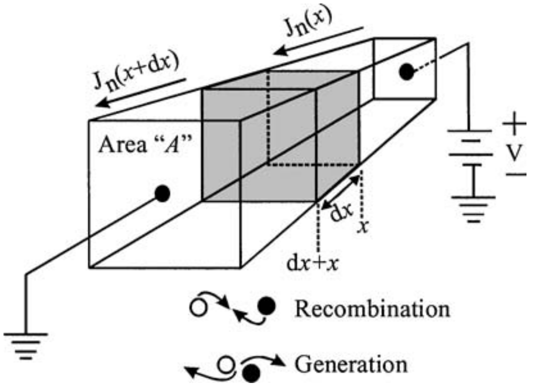


Figure 7.15 A sketch of a sample used to illustrate the derivation of the continuity equation. The four processes occurring in the segment with thickness dx are recombination, generation, flow-in current density $J_n(x)$, and flow-out current density $J_n(x+dx)$.

the continuity equation becomes

$$\frac{\partial n}{\partial t} = \frac{1}{e}\frac{\partial J_n(x)}{\partial x} + (\mathcal{G}_n - \mathcal{R}_n) \tag{7.82}$$

A similar expression can be obtained for the holes as

$$\frac{\partial p}{\partial t} = -\frac{1}{e}\frac{\partial J_p(x)}{\partial x} + (\mathcal{G}_p - \mathcal{R}_p) \tag{7.83}$$

where $\mathcal{G}_p$ is the hole generation rate, $\mathcal{R}_p$ is the hole recombination rate, and the minus sign of the first term on the right-hand side is due to the positive charge of the holes. Substituting Eqs. (7.57a) and (b) into Eqs. (7.82) and (7.83), we obtain

$$\frac{\partial n}{\partial t} = n\mu_n\frac{\partial \mathcal{E}}{\partial x} + \mu_n\mathcal{E}\frac{\partial n}{\partial x} + D_n\frac{\partial^2 n}{\partial x^2} + \mathcal{G}_n - \frac{n}{\tau_n} \tag{7.84}$$

$$\frac{\partial p}{\partial t} = -p\mu_n\frac{\partial \mathcal{E}}{\partial x} - \mu_n\mathcal{E}\frac{\partial p}{\partial x} + D_p\frac{\partial^2 p}{\partial x^2} + \mathcal{G}_p - \frac{p}{\tau_p} \tag{7.85}$$

where the regeneration rate is obtained from Eq. (7.72). For minority carriers, we have the following expressions for electrons (n_p) in p-type material and holes (p_n) in n-type material:

$$\frac{\partial n_p}{\partial t} = n\mu_n\frac{\partial \mathcal{E}}{\partial x} + \mu_n\mathcal{E}\frac{\partial n_p}{\partial x} + D_n\frac{\partial^2 n_p}{\partial x^2} + \mathcal{G}_n - \frac{n_p - n_p^o}{\tau_n} \tag{7.86}$$

$$\frac{\partial p_n}{\partial t} = -p\mu_p\frac{\partial \mathcal{E}}{\partial x} - \mu_p\mathcal{E}\frac{\partial p_n}{\partial x} + D_p\frac{\partial^2 p_n}{\partial x^2} + \mathcal{G}_p - \frac{p_n - p_n^o}{\tau_p} \tag{7.87}$$

Notice that $\Delta n = (n_p - n_p^o)$ and $\Delta p = (p_n - p_n^o)$ are the excess minority carriers. Also n_p^o and p_n^o are constants, and the derivatives of Δn and Δp are simply the derivatives of n_p and p_n. This is because n_p^o and p_n^o are constants. Moreover, the first derivative of the electric field with respect to x can be written in terms of the charge density inside the semiconductor, ρ_s, through Poisson's equation

$$\frac{\partial \mathcal{E}}{\partial x} = \frac{\rho_s}{\epsilon \epsilon_o} \tag{7.88}$$

where ϵ is the dielectric constant of the semiconductor and ϵ_o is the permittivity of space. Assuming that the acceptors and donors in the semiconductors are totally ionized, the charge density can be expressed as

$$\rho_s = e(p - n + N_d - N_a) \tag{7.89}$$

where N_d and N_a are the ionized donor and acceptor concentrations, respectively.

The continuity equations (7.86) to (7.88) can be solved with imposed boundary conditions and physical approximations. Let us assume the following:

1. If the charge neutrality condition is imposed, then we have

$$\Delta p = \left(p_n - p_n^o\right) = \Delta n = \left(n_p - n_p^o\right) \tag{7.90}$$

2. The generation rates of the electrons and holes are equal:

$$\mathcal{G}_n = \mathcal{G}_p = \mathcal{G} \tag{7.91}$$

3. The recombination rates of the electrons and holes are equal,

$$\mathcal{R}_n = \frac{\Delta n}{\tau_n} = \mathcal{R}_p = \frac{\Delta p}{\tau_p} = \mathcal{R} \tag{7.92}$$

4. The minority carriers are equal, $n_p \approx p_n$.

With these assumptions, Eqs. (7.86) and (7.87) become

$$\frac{\partial n_p}{\partial t} = n\mu_n \frac{\partial \mathcal{E}}{\partial x} + \mu_n \mathcal{E} \frac{\partial n_p}{\partial x} + D_n \frac{\partial^2 n_p}{\partial x^2} + \mathcal{G} - \mathcal{R} \tag{7.93}$$

$$\frac{\partial n_p}{\partial t} = -p\mu_p \frac{\partial \mathcal{E}}{\partial x} - \mu_p \mathcal{E} \frac{\partial n_p}{\partial x} + D_p \frac{\partial^2 n_p}{\partial x^2} + \mathcal{G} - \mathcal{R} \tag{7.94}$$

Multiply Eq. (7.93) by $\mu_p p$ and Eq. (7.94) by $\mu_n n$, add the two equations, and then divide by $(\mu_n n + \mu_p p)$ to obtain the following:

$$\frac{\partial n_p}{\partial t} = \mu \mathcal{E} \frac{\partial n_p}{\partial x} + D \frac{\partial^2 n_p}{\partial x^2} + \mathcal{G} - \mathcal{R} \tag{7.95}$$

where

$$\mu = \frac{\mu_n \mu_p (p - n)}{\mu_n n + \mu_p p} \tag{7.96}$$

and

$$D = \frac{\mu_n n D_p + \mu_p p D_n}{\mu_n n + \mu_p p} = \frac{D_p D_n (n + p)}{D_n n + D_p p} \tag{7.97}$$

Equations (7.96) and (7.97) are called the *ambipolar* mobility and diffusion coefficients, respectively. For low-level carrier injection, Eqs. (7.96) and (7.97) are reduced to the following expressions:

$$\mu = \begin{cases} \mu_n & \text{for } p\text{-type} \\ -\mu_p & \text{for } n\text{-type} \end{cases} \tag{7.98}$$

and

$$D = \begin{cases} D_n & \text{for } p\text{-type} \\ D_p & \text{for } n\text{-type} \end{cases} \tag{7.99}$$

It is clear from these relations that the ambipolar mobility and ambipolar diffusion coefficients are reduced respectively to the minority mobility and diffusion coefficients. Thus, the behavior of the excess majority carriers is determined by the minority carrier parameters. This behavior is the general characteristic of semiconductor devices, such as *pn*-junction diodes and bipolar transistors, based on the principle of carrier injection.

The continuity equation can be solved for a steady-state system where the time first derivative of the carrier concentration is zero. Another way of solving the continuity equation for a system in thermal equilibrium is by assuming both the concentration gradient and electric field are absent. A more complicated problem is encountered when a constant electric field is applied across the sample and carrier diffusion is present. Let us assume that a mechanism exists such that a finite number of electron-hole pairs are generated in an *n*-type semiconductor sample at $t = 0$ and $x = 0$, where t is the time and x is the spatial coordinate. Since the electric field $\mathcal{E}$ is constant, the gradient of $\mathcal{E}$ is zero. Assume that the generation rate $\mathcal{G}$ is zero at $t > 0$. The continuity equation of the minority carriers (holes) can now be expressed as

$$\frac{\partial p_n}{\partial t} = -\mu_p \mathcal{E} \frac{\partial p_n}{\partial x} + D_p \frac{\partial^2 p_n}{\partial x^2} - \frac{p_n - p_n^o}{\tau_p} \tag{7.100}$$

A possible solution for this equation is

$$\Delta p = p_n - p_n^o = p_n'(x,t)\exp\left(\frac{-t}{\tau_p}\right) \tag{7.101}$$

Substituting this solution into the continuity Eq. (7.100) yields

$$\frac{\partial p_n'(x,t)}{\partial t} = -\mu_p\mathcal{E}\frac{\partial p_n'(x,t)}{\partial x} + D_p\frac{\partial^2 p_n'(x,t)}{\partial x^2} - \frac{p_n'(x,t)}{\tau_p} \tag{7.102}$$

This equation can now be solved using the Laplace transformation technique which yields a solution of gaussian form:

$$p_n'(x,t) = \frac{1}{\sqrt{4\pi D_p t}}\exp\left[-\frac{(x-\mu_p\mathcal{E}t)^2}{4D_p t}\right] \tag{7.103}$$

Substituting Eq. (7.103) into (7.101), we obtain the final solution as

$$\Delta p(x,t) = p_n - p_n^o = \mathcal{N}\frac{\exp(-t/\tau_p)}{\sqrt{4\pi D_p t}}\exp\left[-\frac{(x-\mu_p\mathcal{E}t)^2}{4D_p t}\right] \tag{7.104}$$

where $\mathcal{N}$ is the number of electrons or holes generated per unit area. A plot of this solution is shown in Fig. 7.16 with and without an applied electric field. As the excess minority carriers (holes) are generated at $t = 0$ and $x = 0$ when the electric field is zero, as shown in Fig. 7.16*a*,

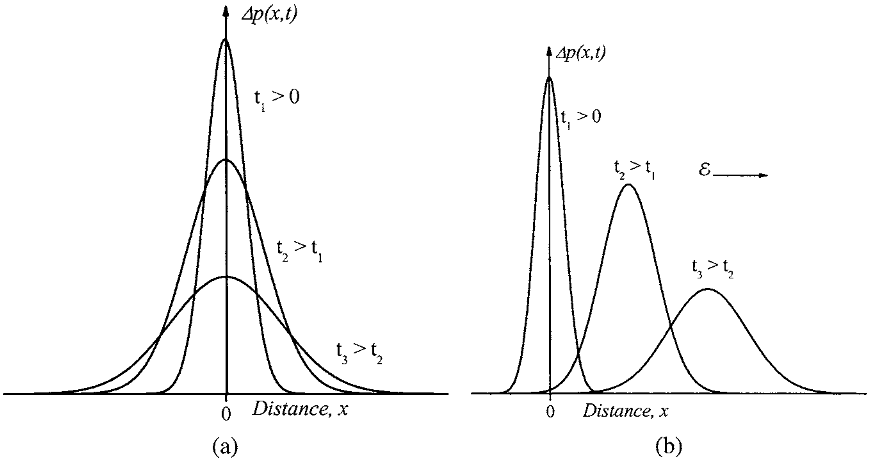


Figure 7.16 Equation (7.104) is plotted as a function of time and distance for (*a*) without applied electric field and (*b*) with applied electric field.

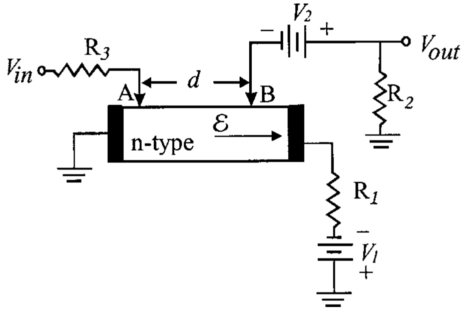


Figure 7.17 An illustration of the Haynes-Shockley experiment.

the holes start to diffuse in both the $+x$ and $-x$ directions. The excess majority carriers (electrons) generated during the process diffuse exactly at the same rate as the holes. As time passes, the diffused electrons and holes start to combine, leading to the disappearance of the holes as t approaches infinity. Thus, diffusion and recombination occur at the same time. When the electric field is not zero, the excess minority carriers drift in the same direction, as shown in Fig. 7.16*b*, since the holes have positive charge. Since the recombination is present as the holes drift in the direction of the electric field, the excess electrons seem to drift in the same direction even though their charge is negative.

The first experiment to measure excess carrier behavior was reported by Haynes and Shockley in 1951. The basic concept of the Haynes and Shockley experiment is illustrated in Fig. 7.17. A rectangular input pulse, shown in Fig. 7.18, is introduced at point A at time $t = 0$. The excess carriers drift along the semiconductor when an electric field $\mathcal{E}_1$, is applied to the sample, producing an output voltage signal at point B after time t_1 has lapsed, as shown in Fig. 7.18. If the electric field is reduced to $\mathcal{E}_2 < \mathcal{E}_1$, the signal received at point B will arrive at time

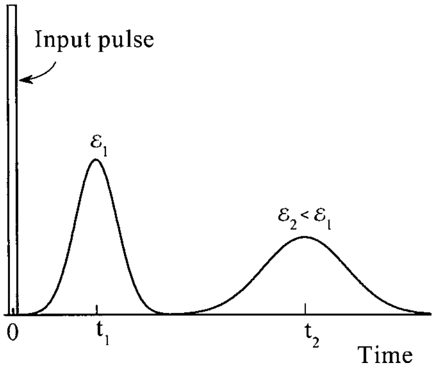


Figure 7.18 Carrier diffusion in the Haynes-Shockley experiment where the input signal is a rectangular narrow pulse applied at point A in Fig. 7.17. The minority carrier pulse is received at point B in Fig. 7.17 at two different electric field values.

$t_2 > t_1$, as shown in the figure. This is because the drift velocity is smaller for lower electric field values. During this longer time period, there is more diffusion and recombination and the excess carrier pulse is smaller, as shown in Fig. 7.18.

The Haynes-Shockley experiment allows one to measure the mobility, diffusion coefficient, and relaxation time of the minority carriers. However, the most accurate parameter that can be extracted from this experiment is the mobility of the excess minority carriers.

7.5 Boltzmann Transport Equation

When charge carriers, such as electrons in semiconductors, are at equilibrium (absence of external perturbations), their statistical distribution obeys the Fermi-Dirac distribution function f_k^o given by

$$f_k^o = \frac{1}{e^{(E_k - E_F)/k_B T} + 1} \tag{7.105}$$

where E_F is the Fermi energy level and k_B is the Boltzmann constant. When electrons are subjected to an external perturbation such as an applied electric field, diffusion, or scattering, their distribution function is no longer described by the Fermi-Dirac function, but by a function f_k that depends on time, space, and momentum. The Boltzmann approach is used to evaluate the behavior of the nonequilibrium distribution function f_k with time. The evolution of f_k as a function of time due to scattering, diffusions, and an external field can be written as

$$\frac{df_k}{dt} = \left.\frac{\partial f_k}{\partial t}\right|_{\text{scatterings}} \tag{7.106}$$

This equation is known as the Boltzmann equation. Since f_k is a function of time, $\mathbf{r}$, and $\mathbf{k}$, the total derivative can be expanded as follows:

$$\frac{df_k}{dt} = \frac{\partial f_k}{\partial t} + \frac{\partial f_k}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial f_k}{\partial y}\frac{\partial y}{\partial t} + \frac{\partial f_k}{\partial z}\frac{\partial z}{\partial t} + \frac{\partial f_k}{\partial k_x}\frac{\partial k_x}{\partial t} + \frac{\partial f_k}{\partial k_y}\frac{\partial k_y}{\partial t} + \frac{\partial f_k}{\partial k_z}\frac{\partial k_z}{\partial t} \tag{7.107}$$

Since

$$\upsilon = \frac{\partial x}{\partial t}\widehat{\mathbf{x}} + \frac{\partial y}{\partial t}\widehat{\mathbf{y}} + \frac{\partial z}{\partial t}\widehat{\mathbf{z}} \qquad \text{and} \qquad \frac{\partial k_x}{\partial t}\widehat{\mathbf{k}}_x + \frac{\partial k_y}{\partial t}\widehat{\mathbf{k}}_y + \frac{\partial k_z}{\partial t}\widehat{\mathbf{k}}_z = \frac{\mathcal{F}}{\hbar} \tag{7.108}$$

where $\boldsymbol{\upsilon}$ is the electron velocity and $\mathcal{F}$ is the external force acting on the system. Substituting Eqs. (7.108) into (7.106) and knowing that

$$\frac{\partial f_k}{\partial x}\widehat{\mathbf{x}} + \frac{\partial f_k}{\partial y}\widehat{\mathbf{y}} + \frac{\partial f_k}{\partial z}\widehat{\mathbf{z}} = \nabla_r f_k \quad \text{and} \quad \frac{\partial f_k}{\partial k_x}\widehat{\mathbf{k}}_x + \frac{\partial f_k}{\partial k_y}\widehat{\mathbf{k}}_y + \frac{\partial f_k}{\partial k_z}\widehat{\mathbf{k}}_z = \nabla_k f_k \tag{7.109}$$

we have

$$\frac{\partial f_k}{\partial t} + \boldsymbol{\upsilon}\cdot\boldsymbol{\nabla}_r f_k + \frac{1}{\hbar}\mathcal{F}\cdot\boldsymbol{\nabla}_k f_k - \left.\frac{\partial f_k}{\partial t}\right|_{\text{scattering}} = 0 \tag{7.110}$$

This equation is known as the Boltzmann transport equation. The term labeled “scattering” represents the distribution function due to scattering between electrons and their surroundings, which can be defined as

$$\left.\frac{\partial f_k}{\partial t}\right|_{\text{scattering}} = -\int [f_k(1-f_{k'})W_{k,k'} - f_{k'}(1-f_k)W_{k',k}]\,dk' \tag{7.111}$$

where the term $(1-f_k)$ represents the probability of having a vacancy in the state k, the term $(1-f_{k'})$ represents the probability of having a vacancy in the state k', and $W_{k,k'}$ and $W_{k',k}$ are the rates at which the electron makes a transition from state k to state k' and from state k' to state k, respectively. These rates are also called transition matrix elements. There is a whole subfield of transport theory devoted to the calculation of these matrix elements for various scattering mechanisms. While the Boltzmann transport equation provides a very useful description of many transport processes in semiconductors, it is still a strictly classical method of describing the transport properties. This is because the distribution function is specified in terms of position, momentum, and time. The simultaneous description of position and momentum is in contradiction to the Heisenberg uncertainty principle and, therefore, the Boltzmann transport equation is not a valid description of quantum effects.

Since the Boltzmann transport equation includes various nonequilibrium mechanisms, such as scattering, recombination, generation, drift, and diffusion, an exact solution for this equation is extremely difficult to obtain. Even approximate solutions require sophisticated numerical analyses, such as the Monte Carlo and drift-diffusion methods. One possible approximation is the relaxation-time method, which assumes that the scattering term in Eq. (7.111) can be replaced by a constant relaxation term. This approximation reduces Eq. (7.110) to a regular differential equation. Thus, the right-hand side of Eq. (7.106) can be replaced by

$$\left.\frac{\partial f_k}{\partial t}\right|_{\text{collisions}} = -\frac{f_k - f_k^o}{\tau} \tag{7.112}$$

This equation indicates that it will take the system a characteristic time τ, called the relaxation time, to relax from the nonequilibrium state to the equilibrium state. With this approximation, the Boltzmann transport equation can be rewritten as

$$\frac{\partial f_k}{\partial t} + \mathbf{\upsilon} \cdot \nabla_r f_k + \frac{1}{\hbar}\mathcal{F} \cdot \nabla_k f_k = -\frac{f_k - f_k^o}{\tau} \tag{7.113}$$

Let us assume that the system is in steady-state, where $\partial f_k/\partial t = 0$, and that the distribution function is spatially uniform such that the spatial gradient of f_k is zero, i.e., $\nabla_r f_k = 0$. Moreover, let us assume that the external force $\mathcal{F}$ is only due to a constant applied electric field $\mathcal{E}$ such as $\mathcal{F} = -e\mathcal{E}$. Finally, the term $(1/\hbar)\nabla_k$ is simply $(1/m^*)\nabla_\upsilon$, where m^* is the electron effective mass. With these approximations, the Boltzmann transport equation is reduced to

$$-\frac{e}{m^*}\mathcal{E} \cdot \nabla_\upsilon f_k = -\frac{f_k - f_k^o}{\tau} \tag{7.114}$$

If we assume that the electric field is along the x axis, Eq. (7.114) becomes

$$\frac{e\tau\mathcal{E}_x}{m^*}\frac{\partial f_k}{\partial \upsilon_x} = f_k - f_k^o \tag{7.115}$$

Another useful approximation is that if f_k is assumed to be not far from f_k^o, then the derivative of these two functions with respect to the velocity is approximately the same:

$$\frac{\partial f_k}{\partial \upsilon_x} \approx \frac{\partial f_k^o}{\partial \upsilon_x} \tag{7.116}$$

The equilibrium distribution function is assumed to be a Fermi-Dirac distribution function given by Eq. (7.105). For simplicity, let us assume our reference point is the Fermi energy, which can be set to zero in Eq. (7.105). Let us also assume that $e^{E_k/k_BT} \gg 1$. Thus, f_k^ocan be written as

$$f_k^o = e^{-E_k/k_BT} \tag{7.117}$$

which is the form of the Maxwell-Boltzmann distribution function. The energy E_k can be taken as $\frac{1}{2}m^*\upsilon_x^2$. The derivative of Eq. (7.117) with

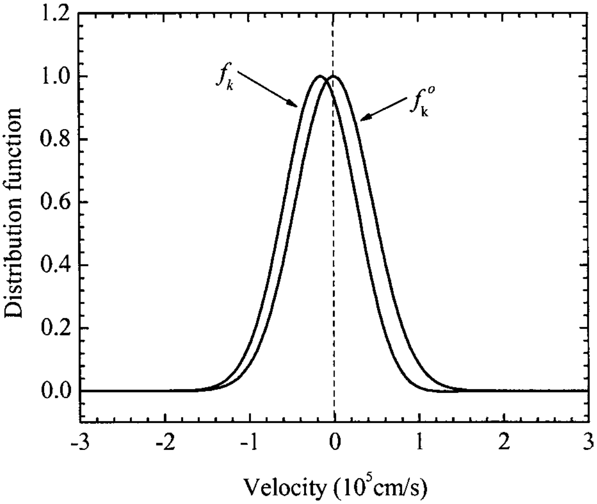


Figure 7.19 The distribution function plotted as a function of carrier velocity for equilibrium (f_k^o) and nonequilibrium (f_k) cases. Notice that the peak of f_k is shifted from $v_x = 0$.

respect to velocity is obtained as follows:

$$\frac{\partial f_k^o}{\partial v_x} = \frac{\partial}{\partial v_x}(e^{-m^* v_x^2/2k_BT}) = -\frac{m^* v_x}{k_BT} e^{-m^* v_x^2/2k_BT} \approx \frac{\partial f_k}{\partial v_x} \tag{7.118}$$

Substituting Eq. (7.118) into (7.115) and rearranging yields

$$f_k = f_k^o \left(1 - \frac{e\tau \mathcal{E}_x v_x}{k_BT}\right) \tag{7.119}$$

A plot of the nonequilibrium distribution function expressed in Eq. (7.119) as a function of the drift velocity is shown in Fig. 7.19 for a free electron. For this plot, the temperature is assumed to be 300 K, the applied electric field is 5×10^5 V/cm, and the relaxation time τ is 0.4 ps. The x component of the velocity can be taken as $v \cos\theta$. Notice that f_k^o is centered at $v_x = 0$ and f_k is slightly shifted toward the left. If the minus sign in the parenthesis in Eq. (7.119) is positive, then f_k will shift to the right.

The nonequilibrium distribution function can also be realized by shifting the wave vector **k** by $e\tau\mathcal{E}/\hbar$. This can be accomplished by considering the distribution function for electrons in a parabolic band at equilibrium, which is given by Eq. (7.105). By setting the Fermi energy to zero, $E_k = \hbar^2 k^2/(2m)$, and by replacing **k** with $\mathbf{k} - e\tau\mathcal{E}/\hbar$, one can obtain the distribution function for the nonequilibrium case as shown in Fig. 7.20. If the applied electric field is along the x direction, the distribution will shift only for k_x. In equilibrium, there is a net cancellation

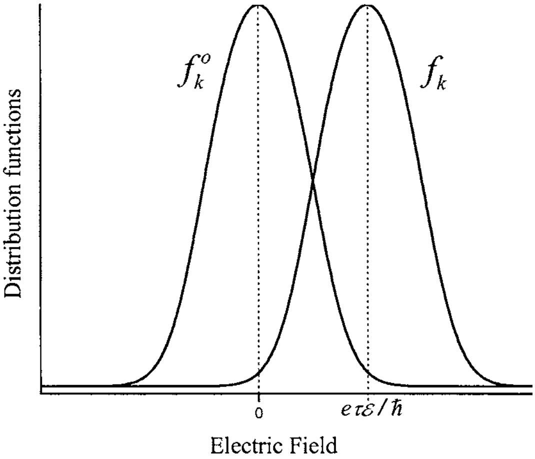


Figure 7.20 The displaced distribution function shows the effect of an applied electric field.

between positive and negative momenta, but when an electric field is applied, there is a nonzero net shift in the electron momenta given by $\delta\mathbf{p} = \hbar\delta\mathbf{k} = -e\tau\mathcal{E}$.

7.6 Derivation of Transport Coefficients Using the Boltzmann Transport Equation

Many of the transport coefficients can be derived from the Boltzmann transport equation in the framework of a relaxation time approximation as described in Eq. (7.112). The relaxation time depends on various scattering mechanisms. This relaxation time depends on the energy and mass of the scattered particles (electrons, for example) according to the following relation:

$$\tau = \tau_o(m^*)^{\alpha}\,(E)^{\beta} \tag{7.120}$$

where τ_o is a constant. Constants α and β characterize the scattering mechanism and depend on the type of scattering mechanism. For example, α is $\frac{1}{2}$ and β is $\frac{3}{2}$ for electron-ionized impurity scattering, while alloy scattering yields $\alpha = -\frac{1}{2}$ and $\beta = -\frac{3}{2}$.

For simplicity, let us assume that we have an n-type semiconductor in which an applied electric field, magnetic field, and temperature gradient are present. The Boltzmann transport equation for the steady-state case can now be written as

$$-\mathbf{v}\cdot\nabla_r f_k + \frac{e}{\hbar}(\mathcal{E} + \mathbf{v}\times\mathbf{B})\cdot\nabla_k f_k = \frac{f_k - f_k^o}{\tau} \tag{7.121}$$

where $\mathcal{E}$ is the applied electric field and **B** is the applied magnetic field. Using $m^* \upsilon = \hbar k$ in Eq. (7.121) we obtain the Boltzmann transport equation in the following form:

$$-\boldsymbol{\upsilon} \cdot \nabla_r f_k + \frac{e}{m^*}(\boldsymbol{\mathcal{E}} + \boldsymbol{\upsilon} \times \mathbf{B}) \cdot \nabla_\upsilon f_k = \frac{f_k - f_k^o}{\tau} \tag{7.122}$$

An analytical solution for this equation is difficult to obtain without additional approximations. One good approximation is to assume that the solution of the function f_k can be written in terms of f_k^o and a first-order correction term such that

$$f_k = f_k^o - \boldsymbol{\upsilon} \cdot \mathcal{Q}(E) \frac{\partial f_k^o}{\partial E} \tag{7.123}$$

where $\mathcal{Q}(E)$ is an unknown vector function that depends only on the energy of the electron, E. For the small perturbation case, where $(f_k - f_k^o) \ll 1$, the expressions in Eq. (7.122) can be approximated as

$$\boldsymbol{\upsilon} \cdot \nabla_r f_k \approx \boldsymbol{\upsilon} \cdot \nabla_r f_k^o = \boldsymbol{\upsilon} \cdot (\nabla_r T) \left(\frac{E_F - E}{T} \frac{\partial f_k^o}{\partial E} \right) \tag{7.124a}$$

$$\boldsymbol{\mathcal{E}} \cdot \nabla_\upsilon f_k \approx \boldsymbol{\mathcal{E}} \cdot \nabla_\upsilon f_k^o = \boldsymbol{\mathcal{E}} \cdot (\nabla_\upsilon E) \frac{\partial f_k^o}{\partial E} = \boldsymbol{\mathcal{E}} \cdot (m^* \boldsymbol{\upsilon}) \frac{\partial f_k^o}{\partial E} \tag{7.124b}$$

$$(\boldsymbol{\upsilon} \times \mathbf{B}) \cdot \nabla_\upsilon f_k \approx -\boldsymbol{\upsilon} \cdot (\mathbf{B} \times \mathcal{Q}(\mathbf{E})) \frac{\partial f_k^o}{\partial E} \tag{7.124c}$$

Substituting Eqs. (7.124*a*) to (*c*) and (7.123) into (7.122), we obtain

$$-e\tau(\boldsymbol{\mathcal{E}} \cdot \boldsymbol{\upsilon}) + \frac{e\tau}{m^*} \boldsymbol{\upsilon} \cdot [\boldsymbol{B} \times \mathcal{Q}(E)] + \tau \frac{E_F - E}{T} \boldsymbol{\upsilon} \cdot (\nabla_r T) - \boldsymbol{\upsilon} \cdot \mathcal{Q}(E) = 0 \tag{7.125}$$

The velocity in this equation can be factored out to obtain the Boltzmann transport equation for the steady-state case under applied electric and magnetic fields, and a temperature gradient. To obtain a solution for $\mathcal{Q}(E)$, let us assume that the applied electric field and temperature gradient lie in the *xy* plane, while the magnetic field is applied along the *z* direction. The *x* and *y* components of $\mathcal{Q}(E)$are

$$\mathcal{Q}_x(E) = \frac{\tau\left(-e\mathcal{E}_x + \frac{E_F - E}{T}\frac{\partial T}{\partial x}\right) - \omega_c \tau^2 \left(-e\mathcal{E}_y + \frac{E_F - E}{T}\frac{\partial T}{\partial y}\right)}{1 + \omega_c^2 \tau^2} \tag{7.126a}$$

$$\mathcal{Q}_y(E) = \frac{\tau\left(-e\mathcal{E}_y + \frac{E_F - E}{T}\frac{\partial T}{\partial y}\right) + \omega_c \tau^2 \left(-e\mathcal{E}_x + \frac{E_F - E}{T}\frac{\partial T}{\partial x}\right)}{1 + \omega_c^2 \tau^2} \tag{7.126b}$$

where ω_c is the cyclotron angular frequency (eB_z/m^*). By knowing $\mathcal{Q}(E)$, one can use Eq. (7.123) to obtain the transport parameters of a semiconductor in the nonequilibrium case.

7.6.1 Electrical conductivity and mobility in *n*-type semiconductors

Let us assume that the electric field is applied in the x and y directions in an n-type semiconductor. Let the magnetic field and temperature gradient be zero. With these assumptions, Eq. (7.126) is reduced to

$$\mathcal{Q}_x(E) = -\tau e\mathcal{E}_x \tag{7.127a}$$

$$\mathcal{Q}_y(E) = -\tau e\mathcal{E}_y \tag{7.127b}$$

The general expression for the electron current density is

$$J_x = -env_x = -e\int_0^\infty v_x f(E) g^{3\mathrm{D}}(E)\,dE \tag{7.128}$$

where $f(E)$ is given by

$$f(E) = \left(f_k - f_k^o\right) = -\boldsymbol{v}\cdot\boldsymbol{\mathcal{Q}}(E)\frac{\partial f_k^o}{\partial E} \tag{7.129}$$

and $g^{3\mathrm{D}}(E)$ is the density of states per unit volume in a bulk semiconductor and is given by

$$g^{3\mathrm{D}}(E) = \frac{1}{2\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2}\sqrt{E} \tag{7.130}$$

assuming that the bottom of the conduction band energy is the reference point, which can be set to zero. By combining Eqs. (7.127) through (7.130), the current density can be written as

$$\begin{aligned} J_x = -env_x &= e\int_0^\infty v_x^2 \mathcal{Q}(E)\frac{\partial f_k^o}{\partial E} g^{3\mathrm{D}}(E)\,dE \\ &= -e^2\mathcal{E}_x \int_0^\infty \tau v_x^2 \frac{\partial f_k^o}{\partial E} g^{3\mathrm{D}}(E)\,dE \\ &= \frac{2e^2\mathcal{E}_x}{3m^* k_B T}\int_0^\infty \tau E g^{3\mathrm{D}}(E) f_k^o\,dE \end{aligned} \tag{7.131}$$

where we assumed that $v_x^2 = v_y^2 = v_z^2 = 2E/(3m^*)$. The Fermi-Dirac distribution function, f_k^o, is approximated as a Maxwell-Boltzmann function given by Eq. (7.117), and its derivative is given by

$$\frac{\partial f_k^o}{\partial E} = -\frac{f_k^o}{k_B T} \tag{7.132}$$

The electrical conductivity can now be written as

$$\sigma = \frac{J_x}{\mathcal{E}_x} = \frac{2e^2}{3m^* k_B T} \int_0^\infty \tau E g^{3\mathrm{D}}(E) f_k^o \, dE$$

$$= \frac{2ne^2}{3m^* k_B T n} \int_0^\infty \tau E g^{3\mathrm{D}}(E) f_k^o \, dE$$

$$= \frac{ne^2}{m^*} \frac{\int_0^\infty \tau E g^{3\mathrm{D}}(E) f_k^o \, dE}{\int_0^\infty \left(\frac{3k_B T}{2}\right) g^{3\mathrm{D}}(E) f_k^o \, dE}$$

$$= \frac{ne^2}{m^*} \frac{\int_0^\infty \tau E^{3/2} f_k^o \, dE}{\int_0^\infty E^{3/2} f_k^o \, dE} = \frac{ne^2 \langle \tau \rangle}{m^*} \tag{7.133}$$

where the average total kinetic energy is given by $E = \frac{3}{2} k_B T$, the average relaxation time is given by

$$\langle \tau \rangle = \frac{\int_0^\infty \tau E^{3/2} f_k^o \, dE}{\int_0^\infty E^{3/2} f_k^o \, dE} \tag{7.134}$$

and the electron density is given by

$$n = \int_0^\infty g^{3\mathrm{D}}(E) f_k^o \, dE \tag{7.135}$$

In general, the relaxation time is a function of electron energy for most scattering processes in semiconductors as expressed in Eq. (7.120). If we assume that α is zero, i.e., the relaxation time is independent of the electron effective mass, Eq. (7.120) is reduced to

$$\tau(E) = \tau_o E^\beta \tag{7.136}$$

For a nondegenerate semiconductor and for Maxwell-Boltzmann statistics, the relaxation time is

$$\langle \tau \rangle = \tau_o \frac{\int_0^\infty E^{\beta+3/2} e^{-(E-E_F)/k_B T} \, dE}{\int_0^\infty E^{3/2} e^{-(E-E_F)/k_B T} \, dE} = \tau_o (k_B T)^\beta \frac{\Gamma\left(\frac{5}{2} + \beta\right)}{\Gamma\left(\frac{5}{2}\right)} \tag{7.137}$$

where Γ is the Γ-function described in Chap. 5. Substituting Eq. (7.137) into (7.133), we obtain the following expression for the electrical conductivity:

$$\sigma = \frac{ne^2\tau_o}{m^*}(k_B T)^\beta \frac{\Gamma\left(\frac{5}{2}+\beta\right)}{\Gamma\left(\frac{5}{2}\right)} \tag{7.138}$$

From this the electron mobility is obtained as

$$\mu_n = \frac{e\tau_o}{m^*}(k_B T)^\beta \frac{\Gamma\left(\frac{5}{2}+\beta\right)}{\Gamma\left(\frac{5}{2}\right)} \tag{7.139}$$

For $\beta = 0$, the expression for the mobility is reduced to that obtained when the system is in equilibrium. Various scattering mechanisms are discussed in Sec. 7.7.

The electrical conductivity and mobility previously discussed were derived for a single-valley model with a spherical constant energy surface for the conduction band. These requirements are usually encountered in III-V semiconductor materials such as GaAs, GaSb, and InP. The electron effective mass in Eq. (7.138) is usually isotropic. However, for semiconductors with multivalley conduction bands, such as Si, the effective mass is anisotropic even though the electrical conductivity remains isotropic due to the cubic crystal structure. For multivalley semiconductors, the effective mass is given by

$$m_\sigma^* = \left[\frac{1}{3}\left(\frac{1}{m_l} + \frac{1}{m_t}\right)\right]^{-1} \tag{7.140}$$

where m_l and m_t are the electron longitudinal and transverse effective masses, respectively, along the two main axes of the ellipsoidal energy surface near the conduction band edge. The subscript σ in introduced in Eq. (7.140) to indicate that the mass is the conductivity effective mass.

Additionally, the Hall mobility described in Eq. (7.139) is not required to be equal to the drift mobility where the magnetic field is zero. These mobilities are related according to the following relation:

$$\mu_H = \frac{\langle\tau^2\rangle}{\langle\tau\rangle^2}\mu \tag{7.141}$$

where μ_H is the Hall mobility ($B_z \neq 0$) and μ is the drift mobility in the absence of a magnetic field.

7.6.2 Hall coefficient R_H

The Hall coefficient for a nondegenerate n-type semiconductor with a single-valley spherical energy band can be derived by assuming that the magnetic field in the Hall measurement is weak, where $(\omega_c\tau^2) \ll 1$.

For the steady-state case in the absence of a temperature gradient, the vector function $\mathcal{Q}(E)$ given by Eq. (7.126), is reduced to the following:

$$\mathcal{Q}_x(E) = -\tau e\mathcal{E}_x + \omega_c \tau^2 e\mathcal{E}_y \tag{7.142a}$$

$$\mathcal{Q}_y(E) = -\tau e\mathcal{E}_y - \omega_c \tau^2 e\mathcal{E}_x \tag{7.142b}$$

By following the same procedure as in Sec. 7.6.1 and using the preceding expressions for $\mathcal{Q}(E)$, the current densities can be written as

$$\begin{aligned} J_x = -env_x &= e\int_0^\infty v_x^2 \mathcal{Q}(E) \frac{\partial f_k^o}{\partial E} g^{3\mathrm{D}}(E)\, dE \\ &= \frac{2e^2}{3m^* k_B T} \int_0^\infty \tau E(\mathcal{E}_x - \omega_c \tau \mathcal{E}_y) g^{3\mathrm{D}}(E) f_k^o\, dE \end{aligned} \tag{7.143}$$

and

$$\begin{aligned} J_y = -env_y &= e\int_0^\infty v_y^2 \mathcal{Q}(E) \frac{\partial f_k^o}{\partial E} g^{3\mathrm{D}}(E)\, dE \\ &= \frac{2e^2}{3m^* k_B T} \int_0^\infty \tau E(\mathcal{E}_y + \omega_c \tau \mathcal{E}_x) g^{3\mathrm{D}}(E) f_k^o\, dE \end{aligned} \tag{7.144}$$

From the definition of the electron Hall coefficient in Eqs. (7.14), (7.143), and (7.144), we can write

$$R_H = \left.\frac{E_y}{J_x B_z}\right|_{J_y=0} = -\left(\frac{3k_B T}{2e}\right) \frac{\int_0^\infty \tau^2 E g^{3\mathrm{D}} F(E) f_k^o\, dE}{\left[\int_0^\infty \tau E g^{3\mathrm{D}} F(E) f_k^o\, dE\right]^2} = -\frac{1}{en}\frac{\langle \tau^2 \rangle}{\langle \tau \rangle^2} \tag{7.145}$$

The Hall factor r shown in Eq. (7.14) can now be defined as

$$r = \frac{\langle \tau^2 \rangle}{\langle \tau \rangle^2} \tag{7.146}$$

The expression for the Hall coefficient for the holes is the same as that for the electron except that the minus sign in Eq. (7.145) is replaced by a plus sign.

Many other transport parameters can be analytically derived from the Boltzmann transport equation, depending on the conditions and the initial assumptions. For example, the Seebeck coefficient can be derived in the steady state in the absence of a magnetic field, but in the presence of a temperature gradient. The analysis can be accomplished as follows: First, set $\omega_c = 0$ in Eq. (7.126a), which gives

$$\mathcal{Q}_x(E) = \tau\left(-e\mathcal{E}_x + \frac{E_F - E}{T}\frac{\partial T}{\partial x}\right) \tag{7.147}$$

Substitute Eq. (7.147) into the current density J_x shown in Eq. (7.128), and then set J_x to zero to give

$$J_x = -env_x = -e\int_0^\infty v_x^2\tau\left(-e\mathcal{E}_x + \frac{E_F - E}{T}\frac{\partial T}{\partial x}\right)g^{3\mathrm{D}}(E)\,dE = 0 \tag{7.148}$$

Rearrange to obtain

$$\begin{aligned} S_n = \frac{\mathcal{E}_x}{\partial T/\partial x} &= -\frac{1}{eT}\left[\frac{\int_0^\infty \tau E^2 g^{3\mathrm{D}}(E)\frac{\partial f_k^o}{\partial E}\,dE}{\int_0^\infty \tau E g^{3\mathrm{D}}(E)\frac{\partial f_k^o}{\partial E}\,dE} - E_F\right] \\ &= -\frac{1}{eT}\left(\frac{\langle\tau E\rangle}{\langle\tau\rangle} - E_F\right) \end{aligned} \tag{7.149}$$

where S_n is the Seebeck coefficient. Notice that we used $v_x^2 = 2E/(3m^*)$. Other parameters that can be easily derived from the Boltzmann transport equation are the transverse magnetoresistance, Nernst coefficient, Ettingshausen coefficient, and Righi-Leduc coefficient.

7.7 Scattering Mechanisms in Bulk Semiconductors

There are several scattering mechanisms that play major roles in the determination of carrier mobilities and conductivities in semiconductors. These mechanisms are summarized in Fig. 7.21. This figure lists the scattering relaxation time as a function of the energy and the effective mass of the charge carrier. The mobility as a function of the sample temperature and the effective mass of the carrier is also shown in the figure. This figure does not include electron-electron or electron-hole scattering. The carrier-carrier scatterings become important in heavily doped semiconductors where impact ionization and Auger processes are significant. The theoretical analyses of scattering processes are usually treated using quantum mechanics. The Fermi golden rule is employed to calculate the scattering rate or the transition matrix element when the charge carrier is scattered from one state to another.

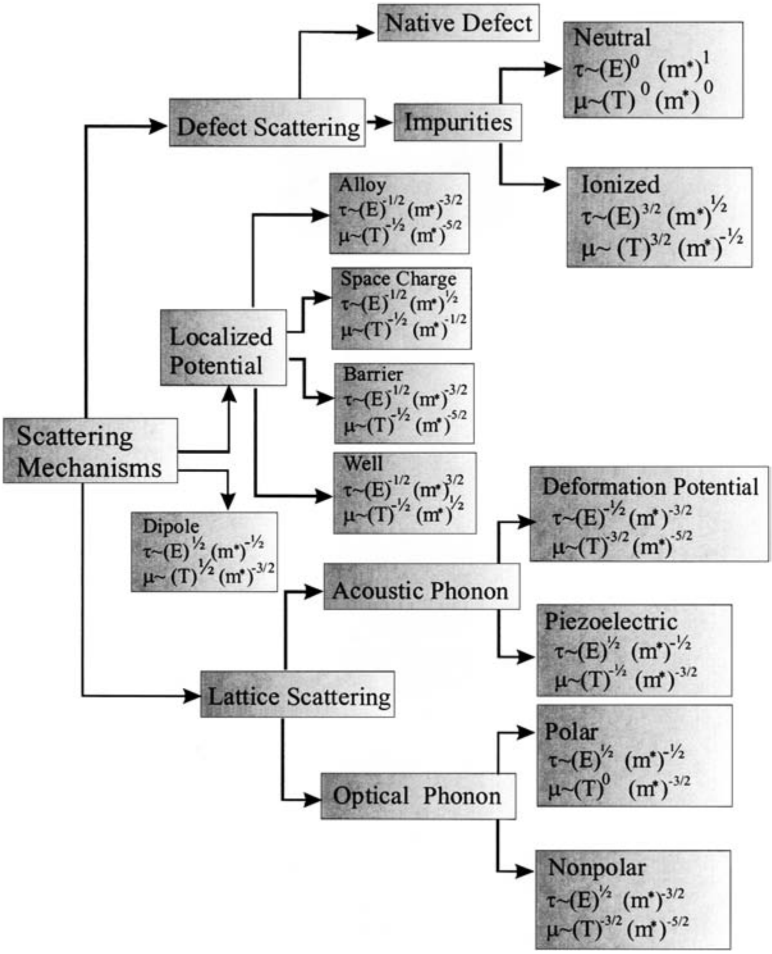

Native Defect
Defect Scattering
Impurities
Neutral
$\tau \sim (E)^{0}$ $(m^*)^{1}$
$\mu \sim (T)^{0}$ $(m^*)^{0}$
Ionized
$\tau \sim (E)^{3/2}$ $(m^*)^{1/2}$
$\mu \sim (T)^{3/2}$ $(m^*)^{-1/2}$
Alloy
$\tau \sim (E)^{-1/2}$ $(m^*)^{-3/2}$
$\mu \sim (T)^{-1/2}$ $(m^*)^{-5/2}$
Space Charge
$\tau \sim (E)^{-1/2}$ $(m^*)^{1/2}$
$\mu \sim (T)^{-1/2}$ $(m^*)^{-1/2}$
Localized Potential
Barrier
$\tau \sim (E)^{-1/2}$ $(m^*)^{-3/2}$
$\mu \sim (T)^{-1/2}$ $(m^*)^{-5/2}$
Scattering Mechanisms
Well
$\tau \sim (E)^{-1/2}$ $(m^*)^{3/2}$
$\mu \sim (T)^{-1/2}$ $(m^*)^{1/2}$
Dipole
$\tau \sim (E)^{1/2}$ $(m^*)^{-1/2}$
$\mu \sim (T)^{1/2}$ $(m^*)^{-3/2}$
Deformation Potential
$\tau \sim (E)^{-1/2}$ $(m^*)^{-3/2}$
$\mu \sim (T)^{-3/2}$ $(m^*)^{-5/2}$
Acoustic Phonon
Piezoelectric
$\tau \sim (E)^{1/2}$ $(m^*)^{-1/2}$
$\mu \sim (T)^{-1/2}$ $(m^*)^{-3/2}$
Lattice Scattering
Polar
$\tau \sim (E)^{1/2}$ $(m^*)^{-1/2}$
$\mu \sim (T)^{0}$ $(m^*)^{-3/2}$
Optical Phonon
Nonpolar
$\tau \sim (E)^{1/2}$ $(m^*)^{-3/2}$
$\mu \sim (T)^{-3/2}$ $(m^*)^{-5/2}$

Figure 7.21 Summary of the major scattering mechanisms that influence the mobility and relaxation time of electron transport in semiconductors. The mobility temperature dependence for the nonpolar scattering is for $k_B T \gg \hbar\omega_o$, where ω_o is the frequency of the nonpolar optical phonon.

The general procedure of the calculation is to identify the scattering potential, and then to use first-order perturbation theory to calculate the transition matrix element. The scattering relaxation time can be obtained from the following relation:

$$\tau^{-1} = N_t \sigma_t \upsilon_{\text{th}} \tag{7.150}$$

where N_t = density of total scattering centers
σ_t = total scattering cross section
υ_{th} = average thermal velocity = $(3k_B T/m^*)^{1/2}$

The total scattering cross section can be obtained from the differential cross section using

$$\sigma_t = 2\pi \int_0^{\pi} \sigma(\theta')(1 - \cos\theta') \sin\theta' d\theta' \tag{7.151}$$

where $\sigma(\theta')$ is the differential cross section defined as the total number of particles that make the transition from one state to another per unit solid angle per unit time divided by the incident flux density, which can be written as

$$\sigma(\theta') = \frac{V^2 k'^2 |W_{\mathbf{k},\mathbf{k}'}|^2}{(2\pi\hbar \upsilon_{\mathbf{k}'})^2} \tag{7.152}$$

where V = volume of crystal
$W_{\mathbf{k},\mathbf{k}'}$ = transition matrix element for the particle that is scattered from state $\mathbf{k}$ to state $\mathbf{k}'$
$\upsilon_{\mathbf{k}'}$ = particle velocity in state $\mathbf{k}'$

For elastic collisions, both momentum and energy are conserved, which means that $\upsilon_{\mathbf{k}} = \upsilon_{\mathbf{k}'}$ and $\mathbf{k} = \mathbf{k}'$. Derivations of the mobility and scattering relaxation time are thus straightforward even though extensive mathematical manipulation is unavoidable. In this section, we provide the final results of the mobility and relaxation time for several scattering centers without going through the derivation.

7.7.1 Scattering from an ionized impurity

A typical example of elastic scattering in semiconductors is the scattering of electrons by an ionized shallow donor impurity. This is due to the fact that the mass of the electron is much smaller than the mass of the ionized impurity, so the change in the electron energy is negligible. The relaxation time for ionized impurity scattering with an ionic charge of e is given by

$$\frac{1}{\tau_i} = \frac{e^4 N_i \ln\left[\frac{8m^* E \epsilon \epsilon_o k_B T}{\hbar^2 e^2 n'}\right]}{16\pi (2m^*)^{1/2} \epsilon^2 \epsilon_o^2 E^{3/2}} \tag{7.153}$$

and

$$\mu_i = \frac{e\langle\tau_i\rangle}{m^*} = \frac{64\sqrt{\pi}\epsilon^2 \epsilon_o^2 (2k_B T)^{3/2}}{N_i e^3 (m^*)^{1/2} \ln\left[\frac{12m^*(k_B T)^2 \epsilon\epsilon_o}{\hbar^2 e^2 n'}\right]} \tag{7.154}$$

where N is the density of the ionized impurity, n' is the density of the screening electrons surrounding the ionized donor impurity, and the subscript i is introduced to indicate ionized impurity scattering. The natural log of the quantity in the brackets is a slowly varying function with temperature and n'. A slight variation is obtained for the quantity in the bracket when the bare Coulomb potential is used as the perturbing Hamiltonian (see Conwell and Weisskopf 1950).

7.7.2 Scattering from a neutral impurity

The scattering of electrons by a neutral impurity was first derived by Erginsoy (1956), who assumed that the scattering is elastic scattering. The relaxation time and mobility for this type of scattering are given by

$$\frac{1}{\tau_{\text{ni}}} = \frac{80\pi\epsilon\epsilon_o N_n \hbar^3}{m^{*2}e^2} \tag{7.155}$$

and

$$\mu_{\text{ni}} = \frac{e\tau_{\text{ni}}}{m^*} = \frac{m^* e^3}{80\pi\epsilon\epsilon_o N_n \hbar^3} \tag{7.156}$$

where the subscript ni stands for neutral impurity and N_n is the density of the neutral impurity. The mobility and scattering time are independent of both temperature and the energy of the electron. The charge carrier scattering from neutral impurities is more significant at low temperatures since carrier freezeout may occur at shallow-level impurity centers in extrinsic semiconductors. However, the low-temperature mobility in many semiconductor materials does not always agree with the theoretical behavior of the mobility given by Eq. (7.156).

7.7.3 Scattering from acoustic phonons

7.7.3.1 Deformation potential. Electron scattering from longitudinal acoustic phonons is very significant in intrinsic semiconductors. It is usually considered elastic scattering since the change in the electron energy is proportional to the ratio of the sound velocity in a solid ($\sim 3 \times 10^5$ cm/s) and the average thermal velocity of the electron ($\sim 10^7$ cm/s). This ratio is smaller than unity for temperatures higher than 100 K. The acoustic-mode lattice vibrations induce changes in lattice spacing, which induces a local fluctuation in the bandgap. The potential resulting from this fluctuation is called the *deformation potential*. This potential may be estimated each time the bandgap is changed per unit strain. Electron scattering from the deformation potential is important in undoped silicon and germanium at room temperature. The relaxation time and mobility due to the deformation potential were derived by Bardeen and Shockley (1950) and given as

$$\tau_{\text{dp}} = \frac{\pi\hbar^4 C_1}{\sqrt{2}(m^*)^{3/2} k_B T\, E_1^2 E^{1/2}} \tag{7.157}$$

and

$$\mu_{\mathrm{dp}} = \frac{e\langle\tau_{\mathrm{dp}}\rangle}{m^*} = \frac{2\sqrt{2\pi}e\hbar^4 C_1}{3(m^*)^{5/2}(k_B T)^{3/2}E_1^2} \tag{7.158}$$

where C_1 = longitudinal elastic constant

$$= \begin{cases} \frac{C_{11}+C_{12}+C_{44}}{2} = \rho \upsilon_s^2 & \text{for wave propagation along (110) direction} \\ C_{11} & \text{for wave propagation along (100) direction} \end{cases}$$

ρ = crystal density
υ_s = the sound velocity in crystal
E_1 = deformation potential constant

and the subscript dp stands for deformation potential. Values reported for E_1 include $|E_1| = 16$ eV for Si, $|E_1| = 9.5$ eV for Ge, and $|E_1| = 9.3$ eV for GaAs.

The theoretical expression for mobility indicates that mobility is proportional to $T^{-3/2}$, which is in good agreement with the experimental measurements obtained for silicon and germanium at $T < 200$ K. At higher temperatures, intervalley optical phonon scattering contributes substantially to electron mobility. This leads to a different relation between mobility and temperature such as $\mu_{\mathrm{pd}} \approx T^n$, where n lies between 1.5 and 2.7.

7.7.3.2 Piezoelectric potential. For polar semiconductors, such as compound semiconductors with zinc-blende and wurtzite structures, the bonds are partially ionic and the unit cell does not possess inversion symmetry. A strain-induced electric field can be generated due to the piezoelectric effect. The piezoelectric potential is thus generated by the acoustic-mode lattice vibrations. A relaxation time due to carrier-piezoelectric potential scattering can be defined as

$$\tau_{\mathrm{pz}} = \frac{2\sqrt{2}\pi\hbar^2\epsilon\epsilon_o}{e^2(m^*)^{1/2}k_{\mathrm{B}}T\,P^2E^{1/2}} \tag{7.159}$$

and the corresponding mobility is

$$\mu_{\mathrm{pz}} = \frac{e\langle\tau_{\mathrm{pz}}\rangle}{m^*} = \frac{2\sqrt{2\pi}\,\hbar^2\epsilon\epsilon_o}{3(m^*)^{3/2}eP^2(k_{\mathrm{B}}T)^{1/2}} \tag{7.160}$$

where the subscript pz stands for piezoelectric and P is the piezoelectric coupling coefficient, which is on the order of $\sim 5 \times 10^{-2}$ for many zinc-blende semiconductors and about an order of magnitude higher for many wurtzite compound semiconductors. The piezoelectric scattering rate is several orders of magnitude smaller than that for deformation potential scattering. Thus, piezoelectric scattering is not that important in compound semiconductors, but it can be significant at very low temperatures.

7.7.4 Optical phonon scattering—polar and nonpolar

Scattering from dipole moments formed by the interaction of the ionic charges on atoms with optical-mode lattice vibrations is called the polar optical-mode scattering process. This scattering mechanism is the dominant process in semiconductors at a high temperature or a high electric field. Solving the Boltzmann equation (7.112) is very difficult in this case since the relaxation time becomes a function of the perturbation strength in addition to the charge particle energy. The relaxation time and mobility are obtained for this scattering process using specific temperature conditions. Further details are discussed by Look (1989). For example, the relaxation time and mobility for $0 \leq (T_{\mathrm{po}}/T) \leq 5$, where T_{po} is the energy of the longitudinal optical phonon divided by the Boltzmann constant, are given by

$$\tau_{\mathrm{po}} = \frac{2\sqrt{2}\pi\hbar^2(e^{T_{\mathrm{po}}/T} - 1)\chi(T_{\mathrm{po}}/T)E^{1/2}}{e^2(m^*)^{1/2}(k_B T_{\mathrm{po}})\left(\epsilon_\infty^{-1} - \epsilon^{-1}\right)\epsilon_o} \tag{7.161}$$

and

$$\mu_{\mathrm{pz}} = \frac{e\langle\tau_{\mathrm{po}}\rangle}{m^*} = \frac{2\sqrt{2}\pi\hbar^2(e^{T_{\mathrm{po}}/T} - 1)}{e(m^*)^{3/2}(k_{\mathrm{B}} T_{\mathrm{po}})^{1/2}\left(\epsilon_\infty^{-1} - \epsilon^{-1}\right)\epsilon_o} \tag{7.162}$$

where $\chi(T_{\mathrm{po}}/T)$ is a slowly varying function of temperature and ϵ_∞ is the high-frequency dielectric constant.

The optical phonon modes produce fluctuations in the bandgap similar to those produced by the acoustic phonon modes. Electrons are scattered by the deformation potential produced by the optical phonon modes. This type of scattering is called nonpolar scattering. The relaxation time for nonpolar scattering is given by

$$\tau_{\mathrm{npo}} = \frac{2\sqrt{2}\pi\rho\hbar^3\omega_o}{D_o(m^*)^{3/2}n_o\left[\sqrt{E + \hbar\omega_o} + \mathcal{H}(E - \hbar\omega_o)(n_o + 1)n_o^{-1}\sqrt{E - \hbar\omega_o}\right]} \tag{7.163}$$

where $\hbar\omega_o$ = optical phonon energy

ρ = density of crystal

D_o = deformation potential (energy per unit strain)

$\mathcal{H}$ = Heaviside step function

$$n_o = \frac{1}{\exp(\hbar\omega_o/k_B T) - 1}$$

The corresponding mobilities are

$$\mu_{\mathrm{npo}} = \begin{cases} \dfrac{2\sqrt{2\pi}\,\rho\hbar^4\omega_o^2 e}{3D_o^2(m^*)^{5/2}(k_BT)^{3/2}} & \text{for } k_BT \gg \hbar\omega_o \\ \dfrac{\pi\rho\hbar^4 e\sqrt{2\hbar\omega_o}}{D_o^2(m^*)^{5/2}n_o} & \text{for } k_BT \ll \hbar\omega_o \end{cases} \tag{7.164}$$

It is very difficult to estimate D_o, but this scattering process is believed to be much weaker than other lattice scattering mechanisms, at least for electron scattering.

7.7.5 Scattering from short-range potentials

Charge carriers can be scattered from a variety of potentials, which can be approximated as short-range potentials that have constant strength over a small volume and zero strength elsewhere. Brief discussions about several short-range potentials are presented next.

7.7.5.1 Scattering from dislocations. Charge carriers (both electrons and holes) can be scattered from dislocations in semiconductors. Dislocations may be considered as a line charge, and scattering can be viewed as scattering from ionized impurity centers. On the other hand, dislocations create strain fields, which produce deformation potential–like scattering. The scattering from dislocations is significant for dislocation densities larger than 10^8 cm^{-2}, and the relaxation time can be estimated by assuming that the dislocation line is cylindrical with radius R and length L. This leads to an electron scattering time as follows:

$$\tau_{\mathrm{dis}} = \frac{3}{8N_dRv} \tag{7.165}$$

where v is the electron velocity and N_d is the dislocation density. The electron mobility is obtained directly from Eq. (7.165) and is given by

$$\mu_{\mathrm{dis}} = \frac{e\langle\tau_{\mathrm{dis}}\rangle}{m^*} = \frac{3e}{8N_dR}\frac{1}{\sqrt{3m^*k_BT}}\frac{4\sqrt{2}}{3\sqrt{\pi}} \approx \frac{3e}{8N_dR}\frac{1}{\sqrt{3m^*k_BT}} \tag{7.166}$$

As shown in this equation, the mobility is proportional to $T^{-1/2}$. Dislocation scattering can be significant in materials where dislocation densities are high, such as GaN thin films grown on sapphire.

7.7.5.2 Scattering from δ-function and alloy potentials. A typical example of a localized potential is the δ-function potential of the form

$$V = V_\delta E_\delta \delta(\mathbf{r} - \mathbf{r}_o) \tag{7.167}$$

where V_δ is a small volume and E_δ is the strength of the potential. An expression for the relaxation time for a density of scattering centers,

N, is obtained as

$$\tau_\delta = \frac{\pi}{\sqrt{2}} \frac{\hbar^4}{N V_\delta^2 E_\delta^2 (m^*)^{3/2} E^{1/2}} \tag{7.168}$$

The corresponding mobility is derived (see Anselm and Askerov 1962) and given by

$$\mu_\delta = \frac{2\sqrt{2\pi}}{3} \frac{e\hbar^4}{N V_\delta^2 E_\delta^2 (m^*)^{5/2} (k_B T)^{1/2}} \tag{7.169}$$

The alloy scattering time can be obtained (Mott and Jones 1958) by setting $N V_\delta^2 E_\delta^2$ in Eq. (7.168) to

$$N V_\delta^2 E_\delta^2 = V_c x(1-x) E_{\mathrm{AB}}^2 \tag{7.170}$$

and assuming that the alloy is composed of two binary compounds A and B. The quantity E_{AB} is the difference between the bandgaps of compounds A and B, x is the fraction of compound A, and V_c is the volume of the unit cell. The mobility, thus, can be obtained as

$$\mu_{\mathrm{al}} = \frac{2\sqrt{2\pi}}{3} \frac{e\hbar^4}{V_c E_{\mathrm{AB}}^2 x(1-x)(m^*)^{5/2} (k_B T)^{1/2}} \tag{7.171}$$

Notice that E_{AB} defines the difference in the scattering potential between compounds A and B.

7.7.6 Scattering from dipoles

When acceptor and donor atoms in semiconductors are close together, they may scatter electrons as a dipole instead of as individual monopoles. The scattering relation time for the unscreened case was first derived by Dimitrov (1976) and is given as

$$\tau_{\mathrm{dipole}} = \frac{2\sqrt{2}\pi\, 3\hbar^2 \epsilon_o^2 \epsilon^2 E^{1/2}}{(m^*)^{1/2} e^2 N q_d^2} \tag{7.172}$$

where q_d is the dipole moment given by $e^* l$ (where l is the distance between the charges and e^* is the ionic charge) and N is the density of the dipoles. The derivation of the mobility for the nondegenerate electrons is straightforward from Eq. (7.172) and is given by

$$\mu_{\mathrm{dipole}} = \frac{2^{9/2}\sqrt{\pi}\hbar^2 \epsilon_o^2 \epsilon^2 (k_B T)^{1/2}}{(m^*)^{3/2} e N q_d^2} \tag{7.173}$$

If the charge of the ion is assumed to be e and the distance between the two ions is averaged to l, then the mobility can be written as

$$\mu_{\mathrm{dipole}} = 3.57 \times 10^7 \frac{T^{1/2}}{(l\,\mathrm{cm})^2 (N\,\mathrm{cm}^{-3})}\,\mathrm{cm}^2 \cdot \mathrm{V}^{-1} \cdot \mathrm{s}^{-1} \tag{7.174}$$

If l is 10 Å and the dipole density is 10^{17} cm^{-3}, then the electron mobility due to dipole scattering is 6.18×10^5 cm^2/(V · s) at 300 K. This high value for mobility indicates that dipole scattering is insignificant. However, if l is 100 Å, then the electron mobility drops to 6.18×10^3 cm^2/(V · s), which means that dipoles contribute strong scattering. In reality, dipole scattering does not play a major role in semiconductors.

7.8 Scattering in a Two-Dimensional Electron Gas

Electrons generated at the interfaces of heterojunctions, such as GaAs/AlGaAs structures, are confined along the growth directions and are free to move in the perpendicular plane. A typical example of this type of heterostructure is shown in Fig. 7.22, where the barrier (AlGaAs) is doped instead of the GaAs layer. The portion of the AlGaAs layer near the interface is not doped and is called the *spacer*. A typical doping level in the AlGaAs layer is on the order of 10^{18} cm^{-3}. Experimentally, it is found that the mobility increases as the spacer thickness increases while the two-dimensional electron gas density is decreased as the spacer thickness increases. Thus, there is a tradeoff between the electron concentration, mobility, and spacer thickness. The electron mobility for GaAs structures can reach values higher than 1×10^6 cm^2 · V^{-1} · s^{-1}. The reason for the high mobility is that electron scattering from ionized impurities is negligible since scattering from ionized impurities decreases as the spacer thickness increases.

Doping in the barrier is called modulation doping, and the structure shown in Fig. 7.22 is the basis for the modulation-doped field-effect transistor (MODFET). This is also known as a high electron mobility transistor (HEMT). A comparison between the mobility for bulk GaAs and a GaAs MODFET structure is shown in Fig. 7.23 as a function of temperature. The ionized impurity concentration in bulk GaAs material is typically on the order of 10^{17} to 10^{18} cm^{-3}. Scattering from ionized impurities dominate the mobility behavior at low temperatures for bulk material. Essentially impurity scattering is almost eliminated in the MODFET structure, as shown in Fig. 7.23, where the mobility continues

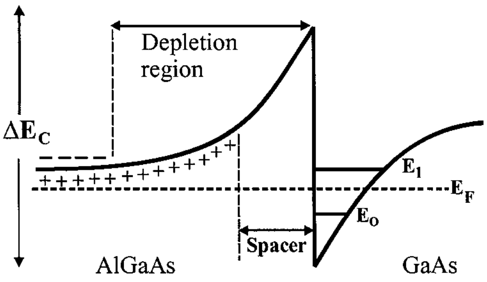


Figure 7.22 A sketch of the conduction band structure of a GaAs/AlGaAs heterojunction showing the confined states at the interfaces, the Fermi energy, the spacer, the depletion region, and the conduction band offset.

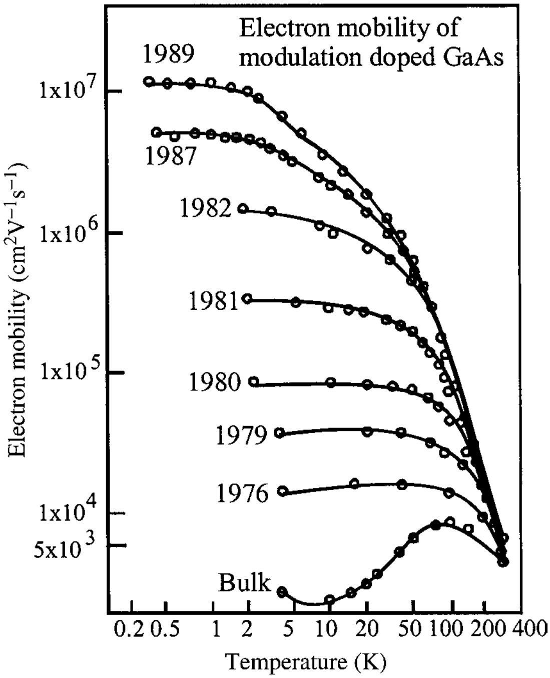


Figure 7.23 The electron mobility of *n*-type doped bulk GaAs and *α* modulation-doped GaAs/AlGaAs heterostructure plotted as a function of temperature. The results show how the electron mobility in a modulation-doped GaAs/AlGaAs heterojunction has risen over the years. (*After Pfeiffer et al. 1989.*)

to increase as the temperature is reduced and reaches a plateau at temperatures less than 10 K. This figure shows the improvement of mobility over the years, which accompanied the advancement of epitaxial growth techniques, in particular molecular beam epitaxy.

The theory of electron scattering in low-dimensional systems such as quantum wells, wires, and dots can be extensive and lengthy. In this section, we present the most important scattering mechanisms in quantum structures without going through the theoretical derivation, bearing in mind that the theory of charge carrier scattering in semiconductors and their heterojunctions contains many approximations and assumptions. As mentioned previously, the scattering mechanisms are described in terms of the transition probability $W_{\mathbf{k},\mathbf{k}'}$ when the charge carrier is scattering from the initial state $\mathbf{k}$ to the final state $\mathbf{k}'$. This transition probability is derived using Fermi's golden rule. Any scattering mechanisms can be described by the relaxation time, which is defined as the average time between scattering events. The inverse of the average relaxation time is called the scattering rate. Moreover, the relaxation time can further be classified as the population, momentum, or energy relaxation time. For example, a system at equilibrium can be described by a physical parameter, such as population, momentum, or energy. When a scattering event occurs, the system is no longer at equilibrium and its physical parameters have different values than at equilibrium. The time required for the physical parameters of the system to relax back to the equilibrium state is called the relaxation time of that particular physical parameter.

Electrons in low-dimensional quantum systems scatter from imperfections, interface roughness, alloys, surface charges, dislocations, phonons, and other charge particles. Sections 7.8.1 and 7.8.2 briefly describe the most important scattering processes in low-dimensional systems.

7.8.1 Scattering by remote ionized impurities

The density of background impurities in the conduction channels of quantum wells and wires is usually small. For modulation-doped structures such as the structure shown in Fig. 7.22, scattering from remote ionized impurities can be significant, especially if the spacer thickness is too small. The perturbation part of the Hamiltonian of the electrostatic potential of the remote ionized impurities is mainly due to the fluctuation of the potential that breaks the translational symmetry parallel to the heterojunctions of the quantum well (see, for example, Mitin et al. 1999). This scattering, which causes a change of the momentum and average relaxation time, is called the average momentum relaxation time, $\tau_{p,i}$, where subscripts p and i stand for momentum and ionized impurities. This relaxation time is derived (Mitin et al. 1999) as follows

$$\begin{aligned}
\langle \tau_{p,i} \rangle^{-1} &= \frac{\pi m^* N_D^{\mathrm{2D}}}{\hbar^3} \left(\frac{2e^2}{\epsilon_o \epsilon} \right)^2 \int_0^{\pi} \frac{(1-\cos\theta)}{\mathbf{k}_{\mathrm{sc}} + 2\mathbf{k}_F \sin(\theta/2)} \\
&\quad \times \exp\left[-4\mathbf{k}_F |Z_o| \sin\frac{\theta}{2} \right] d\theta \\
&\approx \frac{\pi m^* N_D^{\mathrm{2D}}}{\hbar^3} \left(\frac{2e^2}{\epsilon_o \epsilon} \right)^2 \int_0^{\pi} \frac{\theta^2/2}{\mathbf{k}_F^2 (1+\theta)} \exp[-2\mathbf{k}_F |Z_o| \theta]\, d\theta \\
&\approx \frac{\pi m^* N_D^{\mathrm{2D}}}{2\mathbf{k}_F^2 \hbar^3 (2\mathbf{k}_F |Z_o|)^3} \left(\frac{2e^2}{\epsilon_o \epsilon} \right)^2 \int_0^{\infty} x^2 \exp[-x]\, dx \\
&\approx \frac{2\pi m^* N_D^{\mathrm{2D}}}{2\mathbf{k}_F^2 \hbar^3 (2\mathbf{k}_F |Z_o|)^3} \left(\frac{2e^2}{\epsilon_o \epsilon} \right)^2
\end{aligned} \tag{7.175}$$

where the integration is made over the impurity coordinate and

where $N_D^{2\text{D}}$ = sheet concentration of ionized impurity = electron concentration in well = n_s

$\mathbf{k}_F$ = Fermi wavevector = $\sqrt{2\pi n_s}$ (where n_s = density of two-dimensional electron gas)

θ = angle defined as $\mathbf{k} = 2\mathbf{k}_F \sin\frac{\theta}{2}$ (where $\mathbf{k}$ is the electron wavevector before scattering)

$\mathbf{k}_{\text{sc}}$ = characteristic wavevector that determines the range of wavevectors in which the electrons effectively screen the electric field

Z_o = distance at which a thin layer of impurities is present

The doping is assumed to be δ doping. The integral in Eq. (7.175) is evaluated after assuming that $\theta \ll 1$, which leads to $\sin(\theta/2) \approx \theta/2$ and $(1-\cos\theta) \approx \theta^2/2$, $\mathbf{k}_F|Z_o| \gg 1$, and $\mathbf{k}_{\text{sc}} \approx \mathbf{k}_F$. Moreover, we assume that $x = 2\mathbf{k}_F|Z_o|\theta$ and the upper limit of the integral is approximated to ∞, which leads to a value of 2 for the integral. Numerical calculations of the integral may lead to values different than 2. By putting all these assumptions together, the electron mobility associated with scattering by remote ionized impurities takes the following form:

$$\begin{aligned}\mu_{p,i} = \frac{e\langle\tau_{p,i}\rangle}{m^*} &= 16\frac{(\epsilon_o\epsilon)^2\mathbf{k}_F^5\hbar^3|Z_o|^3}{4e^3 2\pi n_s(m^*)^2}\\ &= 16\frac{(\epsilon_o\epsilon)^2(2\pi n_s)^{5/2}\hbar^3|Z_o|^3}{4e^3 2\pi n_s(m^*)^2}\\ &= 16\frac{(\epsilon_o\epsilon)^2(\pi n_s)^{3/2}\hbar^3|Z_o|^3}{\sqrt{2}e^3(m^*)^2}\end{aligned}\tag{7.176}$$

Notice that the mobility increases as the third power of the distance between the two-dimensional electron gas in the well and the doped layer. For a GaAs/AlGaAs quantum well, where $Z_o = 150$ Å, $n_s = 10^{12}$ cm^{-2}, and the dielectric constant = 12.91, the mobility is estimated to be $\mu_{p,i} \approx 2.13 \times 10^5$ cm^2/(V·s). It is obvious that electron scattering by remote ionized impurities is negligible when the spacer in Fig. 7.22 is made thicker than 150 Å. Notice that the electron mobility described in Eq. (7.176) does not depend explicitly on the temperature.

7.8.2 Scattering by interface roughness

Since semiconductor heterojunctions and quantum wells are composed of materials of different bandgaps, interfaces between these materials possess various degrees of roughness. Ideally, the interfaces should be abrupt. In reality, there is a fluctuation in the material thicknesses

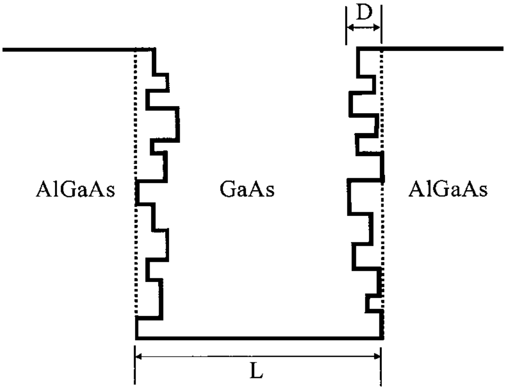


Figure 7.24 A sketch of a GaAs/AlGaAs quantum well showing the variation of the well thickness, which is the source of the interface roughness. The dotted lines represent prompt interfaces.

at the interfaces, which causes charge carriers to scatter. In Fig. 7.24, the interface roughness between GaAs and AlGaAs is presented as a variation of the well thickness. The ideal interfaces are shown as dotted lines in the figure. If the average fluctuation of the GaAs well is taken as D and the spatial correlation of the roughness is described by a correlation length λ, the electron momentum relaxation time for an infinite quantum well is derived (Mitin et al. 1999) as

$$(\tau_{p,n})^{-1} = \frac{4\pi m^* D^2 \lambda^2 E_n^2}{\hbar^3 L^2} \int_0^{2\pi} \frac{1}{2\pi} \frac{1-\cos\theta}{[1+(\mathbf{k}_{\mathrm{sc}}/2\mathbf{k}_F)\sin(\theta/2)]^2} \times \exp\left(-\lambda^2 \mathbf{k}_F^2 \sin^2\frac{\theta}{2}\right) d\theta \tag{7.177}$$

where L = width of quantum well
$\mathbf{k}_F$ = Fermi wavevector
$\mathbf{k}_{\mathrm{sc}}$ = (defined previously) screening length = $2/a^*$ (where a^* is the bulk effective Bohr radius, which is ~100 Å for GaAs)
E_n = quantized energy levels = $\hbar^2\pi^2 n^2/(2m^* L^2)$ for $n = 1, 2, 3, \ldots$

The relaxation time for the first subband ($n = 1$) can be written as

$$(\tau_{p,1})^{-1} = \frac{2\pi^3 D^2 \lambda^2 E_1}{\hbar L^4} \int_0^{2\pi} \frac{1}{2\pi} \frac{1-\cos\theta}{[1+(k_{\mathrm{sc}}/2k_F)\sin(\theta/2)]^2} \times \exp\left(-\lambda^2 k_F^2 \sin^2\frac{\theta}{2}\right) d\theta \tag{7.178}$$

For a GaAs/AlGaAs quantum well numerical calculation, let us set L = 125 Å, $m^* = 0.067m_o$, $D/L = 0.03$, and $\lambda/L = 0.2$. For a carrier concentration of 10^{12} cm^{-2}, the Fermi wavevector is $k_F = \sqrt{2\pi n_s} = 2.5 \times 10^6\,\text{cm}^{-1}$ and $\lambda k_F = 0.62$. For a screening length of $k_{\text{sc}} = 2/a^* = 2/100\,\text{Å}$, we obtain $k_{\text{sc}}/k_F = 0.798$. Assuming that $\theta < 1$, we obtain for the integral an absolute value of 0.205. Every level E_1 is obtained for the first subband as 35.99 meV. Substituting these values in Eq. (7.178), we obtain $\tau_{p,1} = 3.97 \times 10^{-11}$ s. The mobility associated with this scattering time is calculated as $\mu_{p,1} \approx 1.04 \times 10^6\,\text{cm}^2/(\text{V}\cdot\text{s})$. If the 2π in the integral is dropped, the mobility reduces to $\mu_{p,1} \approx 1.7 \times 10^5\,\text{cm}^2/(\text{V}\cdot\text{s})$. Scattering from the interface roughness in quantum structures can dominate many scattering mechanisms at low temperatures.

7.8.3 Electron-electron scattering

Electron-electron interaction is usually considered as a part of the many body effects. Electron-electron interaction is classified into two different interactions: long-range and short-range. The long-range nature of the Coulomb interaction that gives rise to the collective response of the electrons comes from the collective oscillation of the electron gas. The resulting oscillations are called plasma oscillations and have a range greater than the characteristic screening length of the system. Plasmon is the quanta of the plasma oscillations. On short-range scales, electron gas behaves more as a collection of individual charged particles. The short-range electron-electron interaction is considered to be elastic scattering, where the momentum and energy of the entire system are conserved. For short-range electron-electron scattering, the relation time is given (Brennan 1999) by

$$(\tau_{e-e})^{-1} = \frac{m^* e^4 k_{12}}{4\pi\hbar^3 \epsilon^2 k_{\text{sc}}^{-2}\left(k_{\text{sc}}^{-2} + k_{12}^2\right)} \tag{7.179}$$

where electron spin is neglected, k_{sc} is the screening length, and k_{12} is the magnitude of the difference between the wavevectors of two electrons in their initial states. For an electron gas in a GaAs quantum well, $m^* = 0.067 m_o$, $k_{\text{sc}} = 2/a^* = 2 \times 10^{-8}$ m (where a^* is the effective Bohr radius taken as 10^{-8} m), $\epsilon = 12.91$, and it is assumed that $k_{12} = 10^{-2}$ m. The assumed value for k_{12} is quite reasonable knowing that the Fermi wavevector is 2.5×10^8 m for a carrier concentration of 10^{12} cm^{-2}. This means that the two electrons have almost identical wavevectors in their initial states. These values give a relation time of $\tau_{e-e} \approx 1.54 \times 10^{-11}$ s. The electron mobility associated with the relaxation time is obtained as $\mu_{e-e} \approx 4.03 \times 10^5\,\text{cm}^2/(\text{V}\cdot\text{s})$. This value is the same order of magnitude as obtained for remote ionized impurity and interface roughness mobilities. Electron-electron scattering is usually negligible for carrier concentrations on the order of 10^{17} cm^{-3}. For detailed discussions regarding the relaxation time due to long-range electron-electron interactions, see Brennan (1999) and references therein.

7.9 Coherence and Mesoscopic Systems

Electron or hole transport in commercially available electronic devices is governed by various scattering mechanisms. On the other hand, devices based on coherent transport (transport without scattering) are still at the developmental stage. To understand the coherent length or dephasing length, let us first consider an electron that undergoes an elastic collision, where the initial, $\psi_i(\mathbf{r}, t)$, and final, $\psi_f(\mathbf{r}, t)$, wave functions (Mitin et al. 1999) are

$$\psi_i(\mathbf{r}, t) = e^{-i\omega t} e^{i\mathbf{k}\cdot\mathbf{r}} \quad \text{and} \quad \psi_f(\mathbf{r}, t) = e^{-i\omega t} \sum_{\mathbf{k}', \mathbf{k}'=\mathbf{k}} A_{\mathbf{k}'} e^{i\mathbf{k}'\cdot\mathbf{r}} = e^{-i\omega t}\psi(\mathbf{r}) \tag{7.180}$$

where $\mathbf{k}$ and $\mathbf{k}'$ are the wavevectors before and after the scattering event. For elastic scattering, we have $\mathbf{k} = \mathbf{k}'$, which means that the momentum is conserved, and $|A_{k'}|^2$ is the probability of finding the electron with a wavevector $\mathbf{k}'$ after scattering. From Eq. (7.180), one can obtain $|\psi_f(\mathbf{r}, t)|^2 = |\psi(\mathbf{r})|^2$, which means that the spatial distribution remains independent of time after scattering. The incident and scattered wave functions can produce complex wave patterns, but one of the most important properties of elastic scattering is that the phase of the electron is not destroyed, or elastic scattering does not destroy the coherence of the electron motion even for a distance larger than the elastic mean free path.

After inelastic scattering, the electron wave function has different energy and time dependences according to the following:

$$\psi_i(\mathbf{r}, t) = e^{-i\omega(\mathbf{k})t} e^{i\mathbf{k}\cdot\mathbf{r}} \qquad \text{and} \qquad \psi_f(\mathbf{r}, t) = \sum_{\mathbf{k}', \mathbf{k}'\neq\mathbf{k}} A_{\mathbf{k}'} e^{-i\omega(\mathbf{k}')t} e^{i\mathbf{k}'\cdot\mathbf{r}} \tag{7.181}$$

The time-dependent component of the wave function of the scattered electron cannot be factored out of the sum, and $|\psi_f(\mathbf{r}, t)|^2$ is now a function of time. For inelastic scattering, the electron preserves its quantum coherence for a distance equal to or less than the inelastic scattering length l_i. In general, l_i is larger than the de Broglie wavelength. A comparison between various characteristic lengths as compared to the de Broglie wavelength is shown in Fig. 7.25.

The electron dephasing length, or coherent length l_ϕ, is the distance that the electrons travel before losing their quantum mechanical coherence, which is a result of a large spreading of the wave function phases.

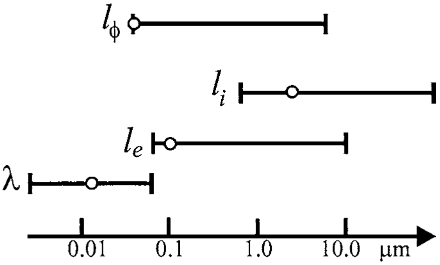


Figure 7.25 Intervals for the characteristic lengths: de Broglie wavelength λ, mean free path l_e, inelastic scattering length l_i, and coherence length in semiconductor materials l_ϕ. As an example, the lengths are marked by circles for Si at $T = 77$ K assuming that the electron mobility equals 10^4 cm/(V·s). (*After Mitin et al. 1999.*)

The dephasing effect is caused by either inelastic collisions, or temperature spreading of phases, or both, which leads to the assumption that the dephasing length is determined by the smaller value of the inelastic length or the thermal diffusion length. The dephasing length l_ϕ is thus the distance for which the electron transport has quantum characteristics. Systems in which electrons maintain coherence and remain in phase during transport are called mesoscopic systems, which have properties strongly dependent on the geometry of the sample, contacts, and quantum structures.

The theoretical analysis of various quantum-transport regimes is very complex and is outside the scope of this book. One of the simplest examples of quantum transport is electron transport in the absence of any scattering. A system with no scattering is a perfect crystalline solid where the equation of motion of the electron is

$$\frac{dp}{dt} = \hbar\frac{d\mathbf{k}}{dt} = e\mathcal{E} \tag{7.182}$$

The electron starts at the bottom of the energy band and moves along the E versus **k** curve until it reaches the Brillouin zone edge. Since we have a perfect crystal, the energy bands are periodic in **k**-space. Thus, when the electron reaches the zone edge, it is reflected and starts to lose its energy, and then continues the cycle under the influence of the electric field. The momentum of the electron changes direction as the electron passes through the zone edges, leading to oscillations in **k**-space (and consequently in the real space). These oscillations are called Zener-Bolch oscillations, and their frequency is given by

$$\omega = \frac{ae\mathcal{E}}{\hbar} \tag{7.183}$$

where a is the lattice constant. For an electric field on the order of 10^7 V/m, the frequency of the Zener-Bloch oscillations is $\sim 1.21 \times 10^{12}$ Hz. This frequency range is very important for high-speed devices. Experimentally, it is difficult to observe these oscillations due to various scattering mechanisms that prevent the coherent transport of electrons.

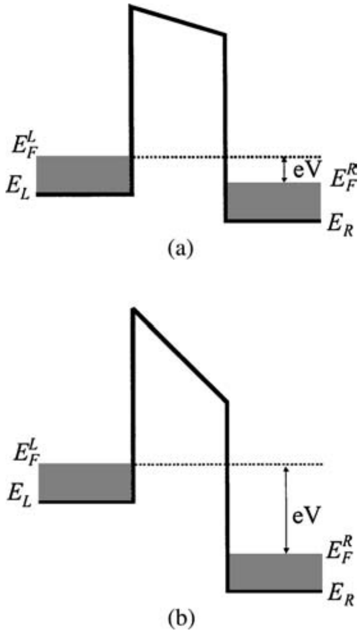


Figure 7.26 A single barrier with Fermi electron seas on both sides is shown for (*a*) small bias and (*b*) large bias.

An expression for the conductance of a system where the phase coherence is maintained can be derived for different contacts and sample geometries. Consider a one-dimensional simple barrier under bias voltage, as shown in Fig. 7.26. Under a small bias voltage, as shown in Fig. 7.26*a*, the electrons tunnel from both left to right and right to left. Under a bias voltage, each side of the barrier has its own Fermi energy level with a difference of $E_F^L - E_F^R = eV$. As the bias voltage is increased, as shown in Fig. 7.26*b*, the electron tunneling from right to left becomes negligible.

The electric current depends on the tunneling transmission coefficient and is given by

$$I_L = \frac{2e}{2\pi} \int_{E_L}^{\infty} f\left(E, E_F^L\right) \upsilon(\mathbf{k}) T(\mathbf{k})\, d\mathbf{k} = \frac{2e}{h} \int_{E_L}^{\infty} f\left(E, E_F^L\right) T(E)\, dE \tag{7.184}$$

where the factor 2 is for spin degeneracy, $\upsilon(\mathbf{k})$ is the electron group velocity, $f(E, E_F^L)$ is the Fermi-Dirac distribution function, and $d\mathbf{k}/2\pi$ is introduced to account for the $\mathbf{k}$ states. In this equation, we used $dE = (\hbar^2\mathbf{k}/m)d\mathbf{k} = \hbar\upsilon d\mathbf{k}$ so that the velocity is cancelled in the current expression. Similarly, the current from right to left under a small bias voltage can be approximated as

$$I_R = -\frac{2e}{2\pi}\int_{E_R}^{\infty} f\left(E, E_F^R\right)\upsilon(\mathbf{k})T(\mathbf{k})\,d\mathbf{k} = -\frac{2e}{h}\int_{E_R}^{\infty} f\left(E, E_F^R\right)T(E)\,dE \tag{7.185}$$

assuming that the transmission coefficient is the same from both sides of the barrier. The total current is the sum of I_L and I_R as follows:

$$I = I_L + L_R = \frac{2e}{h}\int_{E_L}^{\infty}\left[f\left(E, E_F^L\right) - f\left(E, E_F^R\right)\right]T(E)\,dE \tag{7.186}$$

where the lower limit of Eq. (7.185) is changed to E_L since the electrons in the range E_R to E_L do not contribute to the current. For a small bias voltage, the Fermi-Dirac distribution functions can be expanded to a first-order Taylor expansion to give

$$f\left(E, E_F^L\right) - f\left(E, E_F^R\right) = -eV\,\frac{\partial f(E, E_F)}{\partial E} \tag{7.187}$$

where E_F is the distribution function at equilibrium. In this equation, the Fermi energy levels on both sides of the barriers were approximated as $E_F^L = E_F + \frac{1}{2}eV$ and $E_R^L = E_F - \frac{1}{2}eV$. Substitute Eq. (7.187) into (7.186) to obtain

$$I = \frac{2e^2V}{h}\int_{E_L}^{\infty}\left[-\frac{\partial f(E, E_F)}{\partial E}\right]T(E)\,dE \tag{7.188}$$

From Eq. (7.188), one can obtain the conductance G as

$$G = \frac{I}{V} = \frac{2e^2}{h}\int_{E_L}^{\infty}\left[-\frac{\partial f(E, E_F)}{\partial E}\right]T(E)\,dE \tag{7.189}$$

At low temperatures, the Fermi-Dirac distribution function can be approximated as a step function and its derivative is simply a δ-function. Thus, $-\partial f(E, E_F)/\partial E \approx \delta(E - E_F)$ and Eq. (7.189) becomes

$$G = \frac{2e^2}{h}\int_{E_L}^{\infty}\delta(E - E_F)T(E)\,dE = \frac{2e^2}{h}T(E_F) \tag{7.190}$$

The factor e^2/h is known as the quantum unit of conductance, and the corresponding resistance is $R = h/e^2 \approx 25.829\,\mathrm{k\Omega}$. Equation (7.190) shows that the conductance is independent of the length of the sample and depends solely on the transmission coefficient. For $T(E_F) = 1$, the conductance is $2e^2/h$, which is independent of the sample geometry. For higher temperatures, the δ-function approximation is no longer valid and the integration of Eq. (7.189) should be performed.

The conductance result expressed in Eq. (7.190) is very simplistic since it is derived for only one mode or one path that the electron will take when traveling from one contact to another through the sample. In reality, one has to sum the electron contributions from all different paths the electron can take as it moves from one contact to another. For many different paths or propagating states, Eq. (7.190) can be written as

$$G = \frac{2e^2}{h} \sum_{n,m} T(E_F, m, n) = 2G_o \sum_{m,n} T(E_F, m, n) \tag{7.191}$$

where $G_o = e^2/h$ and the sum is over all electron states, m and n, with energy $E < E_F$. Equation (7.191) is called the *Landauer formula*. Each channel or mode has two quantum numbers, m and n, where, for example, m represents the mode or state of the electron when leaving the left contact and n represents the mode or state of the electron when arriving at the right contact, as illustrated in Fig. 7.27.

Landauer formalism provides a means to understand the transport in terms of the scattering process, as illustrated in Fig. 7.27, where, for mesoscopic structures, the electron waves can flow from one contact to maintain phase coherence. The phase coherence is maintained at low temperatures where scattering processes due to phonons are suppressed. Thus, Landauer formalism is valid only at low temperatures and small bias voltages. An important property of phase coherence transport is the fluctuation observed in the conductivity (resistivity) as a function of the magnetic field. In Eq. (7.191), the sum is over the electron contribution from all modes (channels) that the electron takes when traveling from the left contact to the right contact in Fig. 7.27. A fluctuation in the conductance is observed in a large number of experiments and systems, and it is found to be independent of the sample size. An example of the conductance fluctuation, or more precisely quantization, is shown in Fig. 7.28, obtained by van Wees et al. (1988) where

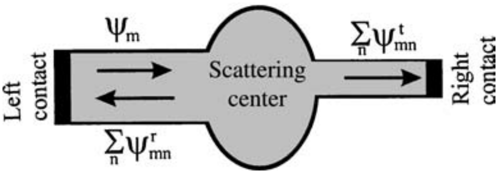


Figure 7.27 Illustration of coherent transport through a device with two leads. Each contact has many propagating states.

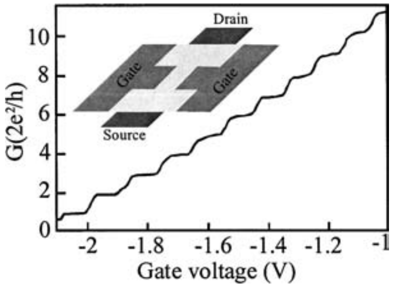


Figure 7.28 Conductance as a function of the gate voltage is plotted for a GaAs/AlGaAs high electron mobility transistor. The inset is a sketch of the MODFET showing the slit gate, drain, and source. (*After van Wees et al. 1988.*)

the conductance was measured as a function of the gate voltage for a GaAs/AlGaAs high electron mobility transistor. In the van Wees et al. experiment, the HEMT or MODFET structure was fabricated with a pair of contacts to produce a short channel of a one-dimensional electron gas with high mobility, as shown in the inset of Fig. 7.28. The Fermi level and the electron wave functions are altered by controlling the gate voltage.

The conductance is discussed briefly for a mesoscopic system of two leads and one electron path and for a system with two contacts and many electron paths. For a mesoscopic system with four contacts or probes, as shown in Fig. 7.29, the conductance can be derived as (see, for example, Singh 2003 and Davies 1998)

$$G_{4-\text{probe}} = \frac{2e^2}{h}\frac{T}{R} = \frac{2e^2}{h}\frac{T}{1-T} \tag{7.192}$$

where T is the transmission coefficient and R is the reflection coefficient. Recall that $T + R = 1$. It appears that there is a difference in the conductance obtained from two probes and four probes. For a weakly transmitting barrier, there is a small difference between the conductance obtained from two probes and that obtained from four probes. But when the barrier is transparent or the transmission coefficient is approaching unity, then the conductance expressed in Eq. (7.192) approaches infinity, while the conductance obtained for two probes and

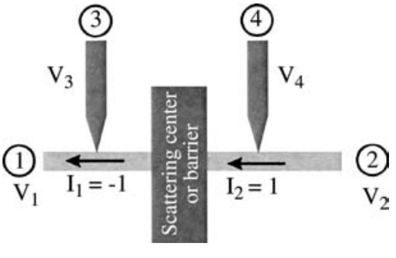


Figure 7.29 Four-probe measurements of the conductance of a scattering center (tunneling barrier) showing the four terminals at which the current and voltages can be measured.

expressed in Eq. (7.190) takes the value $2e^2/h$. This behavior may be explained as follows: For a system where the scattering center or the barrier is absent, the distribution of the electrons should be the same everywhere within the channel such that the voltage probe, in the case of a four-probe experiment, should read the same value at any point. Thus, the voltage difference between any two points is zero giving rise to an infinite value for the conductance. When a bias voltage is applied to the two-probe configuration, an extra voltage appears due to an extra contact resistance of $h/(2e^2)$ in series with the sample. The extra resistance exists even though the electrons are transmitted without any scattering.

Summary

The general formalisms of the charge carrier transport properties in semiconductors were presented with an emphasis on the basic concepts that allow one to investigate nanoscale materials and devices. The aim of presenting the formalisms here was to show the reader in a broader sense how the bulk materials are treated and to show the limitations of a classical treatment of the subject. Various characteristic lengths, including the de Broglie wavelength, and time scales of several physical processes were presented. These characteristic lengths show the transport regimes in bulk, mesoscopic, and low-dimensional semiconductor systems.

Hall-effect, quantum Hall-effect, and Shubnikov–de Haas measurements are widely used to investigate the transport properties in bulk and heterojunction semiconductors. These experimental techniques were discussed, and the basic formalisms to interpret the data were presented.

The discussion in this chapter started with the charge carrier transport in bulk materials, where the drift and diffusion current densities are dominant. The phenomenon of hot electron was briefly discussed, and the Gunn diode was presented as an example of the hot electron transport in bulk semiconductors. The Gunn diode operates in the gigahertz region where a negative differential resistance is observed at high electric fields in compound semiconductors, such as GaAs and InP. The negative differential resistance is the primary cause of the instability in the device, which leads Gunn oscillations with frequencies in the microwave region. In addition to drift and diffusion current densities, the generation and recombination processes in semiconductor materials were discussed. The continuity equation, which combines the drift, diffusion, recombination, and generation was derived for both n-type and p-type semiconductors. From the continuity equation, it was realized that the transport properties of devices that are based on carrier injections, such as bipolar transistors and pn-junction diodes, are governed by the minority carrier transports.

When a bias voltage is applied to semiconductor materials and devices, the carriers are no longer at equilibrium. To derive the transport parameters for the nonequilibrium case, the Boltzmann transport equation was introduced. This equation is very complicated to solve without extensive approximations. One of these approximations is called the relaxation time approximation, where the integral part of the Boltzmann transport equation is replaced by the difference between the nonequilibrium and equilibrium distribution functions divided by the relaxation time required for the system to relax from the nonequilibrium case to the equilibrium case. Examples of how to derive the transport coefficients using the Boltzmann transport equation were presented.

Charge carriers suffer many scatterings during transport in semiconductors. Expressions for the electron mobility and relaxation time were presented for various scattering mechanisms ranging from defect scattering to lattice scattering. Scattering mechanisms in two-dimensional systems were discussed. In the absence of scattering, the electron transport is called ballistic transport, and the electrons maintain their phase during transport. Coherent transport in mesoscopic systems was briefly discussed, and the conductance in two-and four-terminal mesoscopic structures with many channels was derived using Landauer formalism.

Problems

7.1 Calculate the coherence length l_ϕ for an electron traveling with a velocity of 3×10^4 m/s in a bulk GaAs semiconductor material. Assume that the lifetime (time between two successive inelastic collisions) is 1.0 ps. What would be the coherence length in GaAs quantum wells and quantum wires?

7.2 A Hall-effect device is fabricated from GaAs bulk material with $d = 0.5$ mm, $W = 3$ mm, and $L = 10$ mm. The electric current is $I_x = 5$ mA, the bias voltage is $V_x = 2.5$ V, and the magnetic field is $B_z = 0.1$ T. The Hall voltage was measured to be $V_H = -3.0$ mV. Calculate the majority carriers, the mobility and the resistivity. What is the conductivity type?

7.3 The filling factor ν is defined as the number of Landau levels lying below the Fermi energy level. Show that this factor is given by $\nu = n_s h/(eB)$. Assume that Landau levels are spin and valley degenerate.

7.4 Derive Eq. (7.34) and show that for the steady-state case, the magnetoconductivity tensor is give by Eq. (7.35).

7.5 Show that the magnetoresistivity tensor is given by Eq. (7.38).

7.6 A Shubnikov–de Haas experiment was performed on a GaAs/AlGaAs single quantum well. Two consecutive minima in ρ_{xx} were observed at $B_z = 1.56$ and 2.34 T. Calculate the density of the 2DEG in the GaAs well in the unit of cm^{-2}.

7.7 An n-type GaAs sample was subject to light illumination with α photon energy larger than the bandgap energy. Assume that the light illumination was turned off at time t_o and the minority carriers and the excess minority carriers are much smaller than the majority carriers. Derive an expression for the minority carrier concentration as a function of time.

7.8 Derive an expression for the indirect net recombination rate for an indirect bandgap semiconductor. Use Fig. P7.8 where the Fermi energy E_F is pinned at the recombination center; $\mathcal{E}_e$ and $\mathcal{E}_h$ are the emission rates of electrons and holes, respectively; and C_e and C_h are the capture rates of electrons and holes, respectively. The recombination center has electron and hole cross sections of σ_n and σ_p, respectively.

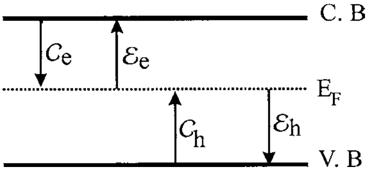


Figure P7.8

7.9 Derive Eq. (7.95) and then show that $D = \frac{D_p D_n (n+p)}{D_n n + D_p p}$.

7.10 Calculate the minority excess carrier concentration as a function of time t for an n-type semiconductor sample that is subject to a generation rate for $t \geq 0$. Assume that the sample is in equilibrium with a zero applied electric field at $t < 0$.

7.11 A p-type semiconductor sample is subject to a process where the excess minority carriers are generated only at $x = 0$ and then begin to diffuse in both the $-x$ and $+x$ directions. Derive an expression for the steady-state excess minority carrier concentration as a function of x. Plot the concentration as a function of x.

7.12 The maximum amplitude of the minority carrier pulse measured at $t_1 = 50\ \mu\text{s}$ in the Haynes-Shockley experiment for an n-type semiconductor sample is six times larger than the amplitude of the pulse measured at $t_2 = 150\ \mu\text{s}$. Calculate the minority carrier lifetime.

7.13 Derive an expression for the current in an n-type GaAs sample of length 1 mm and a cross section of $10^{-6}\ \text{mm}^2$. Assume that the electron concentration is $5 \times 10^{16}\ \text{cm}^{-3}$, the hole concentration is negligible, a bias voltage of 5 V is applied across the sample, the sample was illuminated uniformly with light at $t < 0$ where the generation rate $\mathcal{G}$ is 5×10^{21} cm^{-3}·s^{-1}, and the minority lifetime is 0.3 μs. The light was turned off at $t = 0$. The electron and hole mobilities are given as 1350 and 480 cm^2/(V · s), respectively.

7.14 Excess carriers are generated at $x = 0$ in an n-type GaAs sample, which was subject to a constant electric field $\mathcal{E}$. Derive an expression for the carrier concentration in the steady-state case.

7.15 Show that the average kinetic energy of electrons in a bulk semiconductor sample is given by $\langle E \rangle = \frac{3}{2} k_B T$. Assume that the electrons obey the Maxwell-Boltzmann distribution function.

7.16 Derive the electron Hall coefficient expression shown in Eq. (7.145). Notice that this expression is derived for a small magnetic field.

7.17 Assume that $\tau(E) = \tau_o E^{\beta}$, where β is a positive real number. Show that the electron Hall coefficient is given by$R_H = -\frac{1}{ne}\Gamma(\frac{5}{2} + 2\beta)\Gamma(\frac{5}{2})/[\Gamma(\frac{5}{2} + \beta)]^2$, where $\Gamma(n) = \int_0^{\infty} x^{n-1} e^{-x}\, dx$ and the distribution function is assumed to follow the Maxwell-Boltzmann statistics.

7.18 Use the relaxation time expression shown in Eq. (7.153), which is derived for the scattering of electrons from ionized impurities. Derive the mobility as shown in Eq. (7.154). Calculate the electron mobility in GaAs due to the ionized impurity scattering at 300 and 77 K. Assume that the electron concentration is 10^{16} cm^{-3}, the impurity concentration is 10^{17} cm^{-3}, the dielectric constant is 12.91, and the effective mass is $0.067 m_o$.

7.19 Consider the modulation-doped GaAs/AlGaAs structure shown in Fig. 7.22. The spacer is chosen as 5 nm, and the doped AlGaAs layer is replaced by a δ-doped layer located at the far edge of the spacer; i.e., the δ-doped layer is located at 5 nm from the interface. Assume that the two-dimensional dopant density is 5×10^{12} cm^{-2}. Calculate the momentum relaxation time due to the electron scattering from a remote ionized impurity. Obtain the mobility and compare your results to the data reported in Fig. 7.23. What would be the relaxation time and mobility for a spacer thickness of 20 nm?

7.20 A GaAs sample has an electron mobility of 10^5 cm^2(V · s) and an effective mass of $0.067 m_o$. Calculate the thermal diffusion length, elastic scattering length, inelastic scattering length, dephasing length, and de Broglie wavelength. Assume that the electron group velocity is 5×10^7 cm · s^{-1}.

7.21 A GaN thin film grown on sapphire was found to contain a considerable amount of dislocations, where the inelastic scattering dominates the electron transport. The room temperature electron mobility of this thin film was measured by the Hall effect and found to be 35 cm^2 · V^{-1} · s^{-1}. Calculate the phase coherence length. Compare your result to the interatomic distance.

7.22 Show that the nonequilibrium distribution, when an electric field is applied along the x direction, can be written as follows: $f_k(k_x, k_y, k_z) = f_k^o[(k_x - e\mathcal{E}_x\tau/\hbar), k_y, k_z]$, where f_o is the equilibrium distribution function.

7.23 Assume that only a temperature gradient is applied along the x direction in a semiconductor sample. Show that the steady-state nonequilibrium function can be written as $f_k(k_x, k_y, k_z) = f_k^o[(k_x + \Delta k_x), k_y, k_z]$ where $\Delta k_x = \frac{\tau\hbar\mathbf{k}_F}{m^*T}(\mathbf{k} - \mathbf{k}_F)\frac{dT}{dx}$ and $\mathbf{k}_F$ is the Fermi wavevector.

Chapter

8

Semiconductor Growth Technologies: Bulk, Thin Films, and Nanostructures

8.1 Introduction

The development of growth technologies to produce high-quality and high-purity crystals during the last century enabled the fabrication of electronic and optoelectronic devices down to the nanoscale region. High-performance devices can only be fabricated from high-quality materials that are grown under well-controlled conditions. This chapter is directed toward the discussion of semiconductor crystal growth using various techniques ranging from bulk crystal growth to the epitaxial growth of quantum dots. Bulk crystal growth techniques include liquid-encapsulated Czochralski (LEC), horizontal Bridgman (HB), liquid-encapsulated Kyropoulos (LEK), and vertical gradient freezing (VGF) methods. There are also many improved methods available for the growth of bulk semiconductor crystals. For example, magnetic LEC, direct synthesis LEC, pressure-controlled LEC, and thermal baffle LEC methods are all variations of the original LEC technique, but with improved growth conditions. Other bulk growth techniques include dynamic gradient freezing, horizontal gradient freezing, magnetic LEK, and vertical Bridgman methods. The widely used epitaxial growth techniques are molecular beam epitaxy (MBE), metal-organic chemical vapor deposition (MOCVD), and liquid-phase epitaxy (LPE). The term *epitaxy* is of Greek origin and is composed of two words, *epi* (placed or resting on), and *taxis* (arrangement). Thus, *epitaxy* refers to the formation of single-crystal films on top of a substrate.

Copyright © 2005 by The McGraw-Hill Companies, Inc. Click here for terms of use.

The growth techniques of bulk semiconductor crystals are designed to produce large-volume crystals under equilibrium conditions with almost no flexibility in the production of alloy composition. These techniques lack the ability to produce heterojunctions needed for advanced semiconductor devices. The ability of bulk semiconductor growth techniques to produce large single crystals that can be cut into submillimeter-thick wafers, subsequentially to be used for epitaxial growth, is invaluable. Silicon single-crystal boules as large as 12 inches ($\sim$300 mm) in diameter and over 3 ft (1 m) in length are currently produced by the LEC technique. Diameters of GaAs single-crystal boule are usually smaller than that of Si boules.

Preparing wafers from the boules is called the *wafering process* and includes slicing, lapping, polishing, and cleaning. Since most wafers are used as substrates for epitaxial growth, the wafering process technology and the bulk crystal growth are very important for successful epitaxial growth. For example, the surface orientation accuracy, which is determined during the slicing process, affects the morphology of the epitaxial layer surface. Wafer flatness is another important parameter for high-quality epitaxial growth. Single- or double-sided polished wafer flatness is defined by specific parameters, such as total thickness variation, total indicator reading, or focal plane deviation. These parameters are needed for precise photolithography. The surface roughness is another important aspect of the wafering process, since surface roughness on a subnanometer scale is required for many epitaxially grown quantum structures.

For many epitaxial growths, wafers that can be used without any treatment prior to the growth are required. There are many technologies available to prepare ready-to-use wafers. For example, thermal oxidation and/or ultraviolet/ozone oxidation processes have been effective in producing thin oxide layers, which protect the wafer surface. These oxide layers can be removed by heating prior to epitaxial growth. Packaging the wafers in nitrogen gas is an effective method used to reduce residual oxidation of polished surfaces during storage.

Semiconductor alloys (see Fig. 8.1), heterojunctions, and other quantum structures such as superlattices and quantum dots are currently grown by two main epitaxial growth techniques, namely, MBE and MOCVD. These growth techniques enable the synthesis of high-quality single-crystal thin films deposited layer by layer on suitable substrates.

Growth of bulk materials using any of the growth techniques mentioned earlier represent growth at equilibrium, while the epitaxial growth techniques are nonequilibrium growth methods. In both equilibrium and nonequilibrium growth, the production of semiconductors is based on chemical reactions. Thermodynamic analysis provides information about the feasibility of chemical reactions that can lead to the

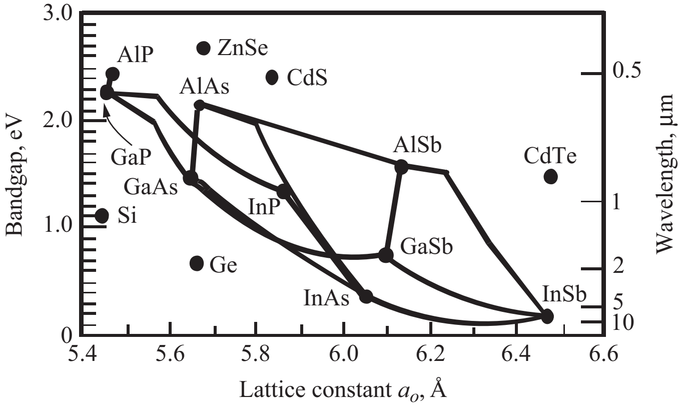


Figure 8.1 The bandgap as a function of the lattice constant plotted for several binary semiconductors. Silicon and germanium are also shown. The solid lines represent the ternary compounds.

production of compound semiconductors. For example, the free energy function G can be written as

$$G = H - TS \tag{8.1}$$

where H = enthalpy
S = entropy
T = temperature

For a system that is undergoing a chemical reaction, the change in the free energy can be written as

$$\Delta G = G_f - G_i = \Delta H - T\,\Delta S \tag{8.2}$$

where G_i and G_f are the free energy of the initial and final states of the reaction, respectively. According to the second law of thermodynamics, *In all energy exchanges, if no energy enters or leaves the system, the potential energy of the final state will always be less than that of the initial state*, or $G_f < G_i$. Hence, the system tends to minimize its free energy to a lower value than the initial state. For a forbidden process, we have $\Delta G > 0$, and for a system at equilibrium, the change in the free energy is zero ($\Delta G = 0$).

Let us consider the following chemical reaction between materials X and Y, which yields material Z,

$$x\mathrm{X} + y\mathrm{Y} \rightarrow z\mathrm{Z} \tag{8.3}$$

where x, y, and z are called the stoichiometric coefficients. If one assumes that the materials X, Y, and Z are at equilibrium, then the change in the free energy is given by

$$\Delta G = zG_z - xG_x - yG_y \tag{8.4}$$

The free energy of individual reactants is usually given as

$$G_j = G_j^o + RT \ln a_j \tag{8.5}$$

where j = reactant (X, Y, or Z)
a_j = activity, which reflects change in free energy when material is not in its standard state
G_j^o = free energy of jth reactant in its standard state
R = gas constant, 8.3143 J/(K·mol), 1.9872 cal/(mol·K), or 62.363 l/(mol·K).

The standard state is defined as 1 atm for a gas at 25°C, and a pure liquid or pure solid is the standard state of the relevant substance. Substituting Eq. (8.5) into (8.4) and solving for $\Delta G = 0$, we obtain

$$-\Delta G^o = RT \ln k \qquad \text{where} \quad k = \frac{(a_Z)^z}{(a_X)^x (a_Y)^y} \tag{8.6}$$

The values of $a_j (j = \mathrm{X, Y, Z})$ are usually taken while the system is at equilibrium.

The inspection of changes of composition of a material from one phase to another is usually accomplished by visualizing the phase diagram. The phase diagram helps us to understand the chemical and physical properties of the material and how a material forms microstructures. For example, when a material fails to perform, one can refer to the phase diagram and deduce what might have happened to cause the failure.

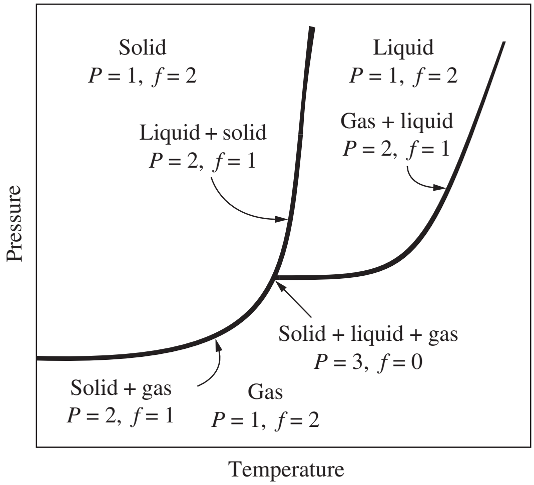


Figure 8.2 The pressure-temperature phase diagram of a single-component system illustrates the degrees of freedom for different numbers of phases.

Then one can revisit the thermodynamic laws that govern the phase diagram and extrapolate information.

Several rules associated with phased diagrams must be observed. The Gibbs phase rule is one of the most important rules. It describes the possible number of degrees of freedom in a closed system at equilibrium in terms of the number of separate phases and the number of chemical constituents in the system. This rule can be expressed as

$$f = C - P + 2 \tag{8.7}$$

where f = number of degrees of freedom
C = number of components
P = number of phases

The physical meaning of the number of degrees of freedom refers to the number of independent variables (temperature, pressure, concentrations, etc.), independent of the quantity of material, that need specified values to fully determine the state of the system. Figure 8.2 illustrates the pressure-temperature phase diagram of a single substance ($C = 1$). The Gibbs phase rule is also illustrated in this figure by the number of phases and the number of degrees of freedom shown. For example, a solid substance (one phase) with a fixed pressure and temperature, results in two degrees of freedom from Eq. (8.7).

Another example of a phase diagram is shown in Fig. 8.3, which is for GaN at a pressure of 1 atm of nitrogen where the temperature is plotted

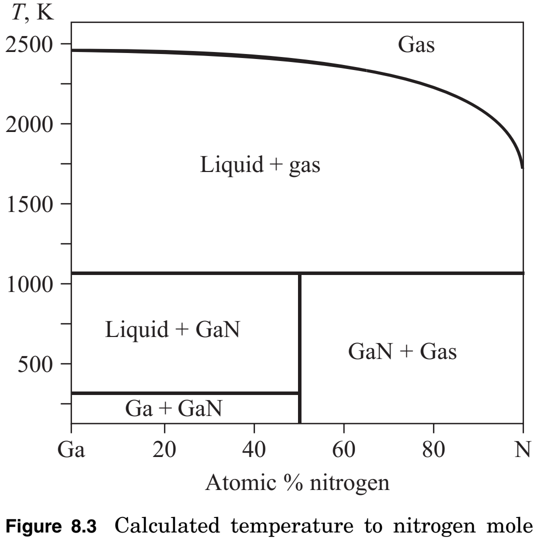


Figure 8.3 Calculated temperature to nitrogen mole fraction (composition) phase diagram for GaN obtained at atmospheric pressure. (*After Davydov et al. 2001.*)

as a function of the nitrogen atomic percentage. This figure indicates that solid GaN can exist at temperatures in the vicinity of 1000 K. The growth temperature for GaN is usually at ~1050°C in the case of the MOCVD growth technique. On the other hand, the MBE technique currently being used to grow GaN and related compounds uses a growth temperature on the order of 850°C.

8.2 Growth of Bulk Semiconductors

Single-crystal growth of various bulk semiconductor materials has been achieved by several methods, such as the LEC, modified LEC, VGF, and HB methods. In this section the most commonly used methods are discussed.

8.2.1 Liquid-encapsulated Czochralski (LEC) method

The LEC method of bulk growth was first developed by Czochralski in 1916. This method uses what is called a *crystal puller* as described in Fig. 8.4. The crystal puller consists of a high-purity quartz crucible filled with polycrystalline materials, which are heated above their

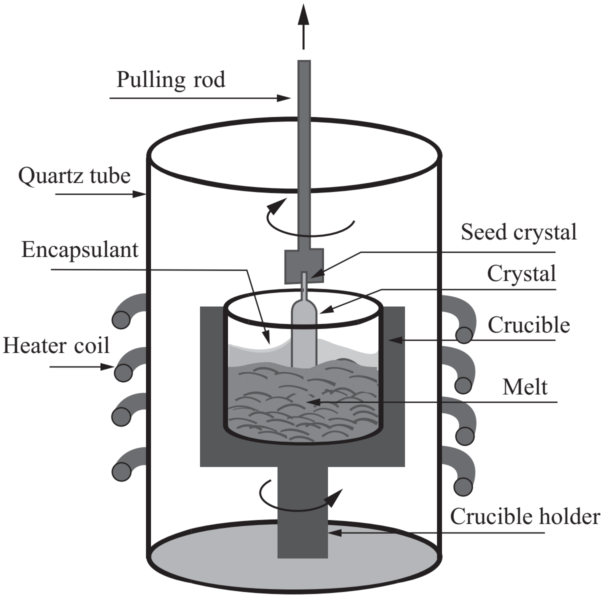


Figure 8.4 A cross section of a furnace (crystal puller) typically used in LEC single-crystal growth.

melting point by induction using radio-frequency (RF) energy. The crucible holder is usually made of graphite. A *seed* single crystal with a specific orientation is lowered into the molten material and then drawn upward using a pulling-rotation mechanism. The material in the melt makes a transition into a solid-phase crystal at the solid-liquid interface. The new solid-phase crystallographic structure is a replica of that of the seed crystal. During the growth process, the crucible rotates in one direction (12 to 14 rotations per minute), while the seed holder rotates in the opposite direction (6 to 8 rotations per minute). At the same time the boule is slowly pulled upward. The crystal diameter is usually monitored by an optical pyrometer, which is focused at the interface between the crystal (boule) and the melt. An automatic diameter-control system maintains the desired crystal diameter through a feedback loop control. An inert gas such as argon is usually used as the ambient gas during the crystal pulling process.

In the LEC crystal growth, boric oxide (B_2O_3) is used as an encapsulant to prevent the decomposition of the melt. Boric oxide is extracted as an $Na_2B_4O_7$ solution from minerals including boron and then precipitated as boric acid (H_3BO_3). Boric acid is refined by repeated recrystallization and dehydrated by heating. The purity of boric oxide is very important since impurities will contaminate the melt.

The molten semiconductor and solid are usually kept at the same pressure and have approximately the same composition. The crystallization results from a reduction in temperature; as the melt is pulled up, it loses heat by radiation and convection to the inert gas (for example, argon). The heat lost to the inert gas causes a substantial thermal gradient across the liquid-solid interface. Additional energy loss is due to solidification (latent heat of fusion). For a fixed volume, a one-dimensional energy balance for the interface can be expressed as

$$\left(-k_l A \left.\frac{dT}{dx}\right|_l\right) - \left(-k_s A \left.\frac{dT}{dx}\right|_s\right) = L\,\frac{dm}{dt} \tag{8.8}$$

where k_l, k_s = thermal conductivity, at the melting point, of the liquid and solid, respectively
A = boule cross-sectional area
T = temperature
L = latent heat of fusion
m = mass of growing solid
t = time

Generally speaking, the heat diffusion from the liquid is small as compared to that from the solid. Thus, Eq. (8.8) can be approximated by neglecting the first term on the left-hand side. With this approximation,

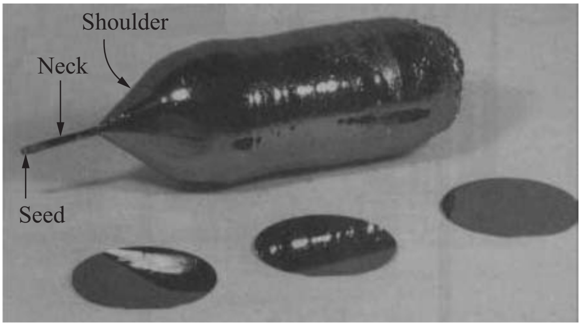


Figure 8.5 A 75-mm-diameter, 240-mm-long InP single crystal grown by the pressure-controlled LEC method. (*After Oda et al.* 2000.)

one can express the maximum velocity, $\upsilon_{\max}$ at which the solid can be pulled as

$$\upsilon_{\max} \approx \frac{k_s A}{L} \frac{dT}{dm} = \frac{k_s}{M_v L} \left. \frac{dT}{dx} \right|_s \tag{8.9}$$

where M_v is the solid density of the crystal being grown. If the crystal is pulled with a velocity larger than this maximum velocity, it will not conduct heat fast enough and the formation of a single crystal becomes difficult to achieve. In general, the pull rate of the seed crystal varies during the growth cycle. It is faster when growing the narrow neck so that the generation of dislocations is minimized and slower during the growth of the boule. Figure 8.5 shows a picture of a 75-mm-diameter, 240-mm-long InP single crystal grown by the pressure-controlled LEC method (Oda et al. 2000). The seed, neck, and shoulder of the crystal are indicated.

Crystals grown by the LEC technique are susceptible to the incorporation of unwanted impurities. For example, quartz is used as a crucible when growing silicon crystals, and the growth temperature is on the order of 1500°C. Thus, a small amount of oxygen will be incorporated into the boule. For extremely low concentrations of oxygen impurities in silicon, the boule can be grown under the influence of a magnetic field. The magnetic field is usually directed perpendicular to the pull direction, where the Lorentz force will change the motion of the ionized impurities in the melt in such a way as to keep them away from the liquid-solid interface. This results in a substantial decrease in incorporation of impurities in the crystal.

The LEC growth of compound semiconductors such as GaAs and InP is more difficult when compared to that of silicon crystals. For example, pyrolytic boron nitride crucibles are used for compound semiconductors instead of quartz crucibles, and B_2O_3 is used as encapsulant. The thermal conductivity of GaAs [$\sim$0.5 W/(cm·K)] is about one-third that of silicon [$\sim$1.4 W/(cm·K)]. Thus, GaAs cannot dissipate the latent heat of fusion as fast as silicon. Additionally, the sheer stress required to generate a dislocation in GaAs at the melting point is about one-fourth that in silicon. These thermal and mechanical properties only permit the growth of 125-mm-diameter GaAs boules as compared to 300-mm-diameter boules for silicon. Furthermore, the GaAs boules contain defect densities a few orders of magnitude larger than those of silicon boules.

In addition to the stated difficulties of LEC semiconductor growth, there are other issues worth mentioning. One of them is the *stacking fault energy*. Just a small amount of stacking fault energy will promote the formation of twins. This is the main reason why few compound crystals are as difficult to grow as single crystals. An example is InP where twinning, which is more common in InP as compared to GaAs, is the main growth obstacle. The stacking fault energy in InP is smaller than for GaAs or GaP. Another issue is the *resolved sheer stress*, which is a major factor in predicting the generation of dislocations. Most compound semiconductors have a lower resolved shear stress compared to that of silicon. Thus, the reduction of dislocation densities becomes a key issue for realizing high-quality materials.

Doping the single crystal during LEC growth is very important since it is desired to produce n-type, p-type, or semi-insulating substrates. For example, introducing boron or phosphorus into a silicon melt produces p-type or n-type silicon, respectively. In the case of GaAs, the undoped or Cr-doped material is usually semi-insulating. Adding silicon as a dopant to GaAs produces n-type materials, while the addition of carbon or beryllium produces p-type GaAs. During LEC growth, the dopant concentration in the boule is usually different from the dopant concentration in the melt. The ratio between the two concentrations is known as the *equilibrium segregation coefficient*, k_o, which is expressed as

$$k_o \equiv \frac{C_s}{C_l} \tag{8.10}$$

where C_s and C_l are the equilibrium dopant concentrations in the solid and liquid, respectively, in the vicinity of the interface. Usually k_o is less than unity. For example k_o values for B and P dopants in silicon are 0.80 and 0.35, respectively, while the values of k_o for Si and C dopants in GaAs are 0.185 and 0.8, respectively.

It is desired to obtain an expression for the dopant concentration in the solid crystal as it is pulled out of the melt during the LEC growth. Assume that the initial crystal weight and dopant concentration in the crystal are m_o and C_s, respectively. Also, assume that the amount of dopants by weight remaining in the melt is σ when the crystal weight increases to m during growth. When the crystal weight increases by the amount dm, the corresponding reduction of the dopant weight from the melt is

$$-d\sigma = C_s\, dm \tag{8.11}$$

On the other hand, the remaining weight of the melt is $(m_o - m)$. Hence the doping concentration in the liquid, C_l, is given by

$$C_l = \frac{\sigma}{m_o - m} \tag{8.12}$$

By combining Eqs. (8.10) to (8.12), the reduction of the dopant (by weight) in the melt can be written as

$$\frac{d\sigma}{\sigma} = -k_o \frac{dm}{m_o - m} \tag{8.13}$$

Integrating Eq. (8.13) using $C_o m_o$ and σ for the initial and final weights of the dopant in the melt, respectively, the initial and final weights of the crystal are 0 and m, respectively. The final result is given as

$$C_s = k_o C_o \left(1 - \frac{m}{m_o}\right)^{k_o - 1} \tag{8.14}$$

A plot of C_s versus m for different values of k_o is left as an exercise. Notice that C_o is the initial dopant concentration in the melt. For $k_o = 1$ we have a constant concentration profile. On the other hand, C_s increases as m is increased for $k_o < 1$, while C_s decreases as a function of m for $k_o > 1$. Equation (8.14) tells us that there is a concentration gradient along the length of the crystal. In other words, the dopant concentration near the seed is different than that near the tail of the crystal. There is also a radial gradient dopant concentration. In other words, the dopant concentration near the center of the boule is different than that near the rim. Usually, the mapping of dopant concentration across the wafer using techniques such as photoluminescence or absorption at a fixed wavelength is very helpful in determining the carrier concentration in wafers.

The segregation coefficient may not be constant for dopants in LEC-grown semiconductor materials. The segregation coefficient discussed earlier is derived for the system near the liquid-solid interface. Away from this interface the segregation coefficient can be different. To derive an expression for the effective segregation coefficient, let us assume that the dopant distributions in the solid and liquid phases are given by the profiles shown in Fig. 8.6 (see, for example, Ohring 1992). As mentioned earlier, the segregation coefficient at equilibrium is defined near

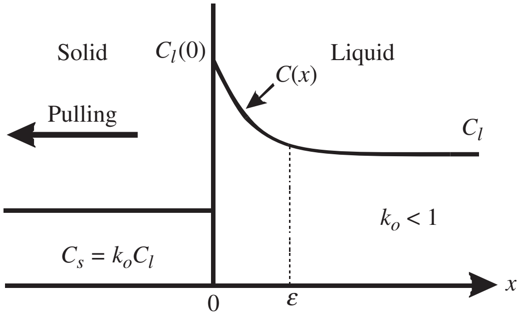


Figure 8.6 The distribution of a dopant near the solid-liquid interface in an LEC crystal puller.

$x = 0$ as $k_o = C_s/C_l(0)$. However, one can define the effective segregation coefficient k_e as the ratio between C_s and C_l, where C_l is the dopant concentration away from the solid-liquid interface. Now, let us consider a small layer of the melt with width ε, in which the only flow is that required to replace the crystal being withdrawn from the melt, as shown in Fig. 8.6. Outside this layer, the dopant concentration is almost constant with a value of C_l, while inside the layer, the dopant concentration $C(x)$ can be described by the steady-state continuity equation as

$$\upsilon \frac{dC(x)}{dx} + D_d \frac{d^2C(x)}{dx^2} = 0 \tag{8.15}$$

where υ is the velocity at which the crystal is being pulled out of the melt [see Eq. (8.9)] and D_d is the dopant diffusion coefficient. A possible solution of this equation is

$$C(x) = Ae^{-\upsilon x/D_d} + B \tag{8.16}$$

where A and B are constants that need to be determined from the boundary conditions. The first boundary condition is that $C(x) = C_l(0)$ at $x = 0$, which gives

$$C_l(0) = A + B \tag{8.17}$$

The second boundary condition is that the sum of the dopant fluxes at the interfaces must be zero, which yields

$$D_d \left. \frac{dC(x)}{dx} \right|_{x=0} + [C_l(0) - C_s]\upsilon = 0 \tag{8.18}$$

Substituting Eq. (8.16) into (8.18) gives

$$A = C_l(0) - C_s \tag{8.19}$$

Combining Eqs. (8.16), (8.17), and (8.19) yields

$$C(x) = [C_l(0) - C_s](e^{-\upsilon x/D_d} - 1) + C_l(0) \tag{8.20}$$

From Fig. 8.6, one can see that $C(x) \approx C_l$ at $x = \varepsilon$. Thus, Eq. (8.20) becomes

$$e^{-\upsilon\varepsilon/D_d} = \frac{C_l - C_s}{C_l(0) - C_s} \tag{8.21}$$

which yields

$$k_e \equiv \frac{C_s}{C_l} = \frac{k_o}{k_o + (1 - k_o)e^{-\upsilon\varepsilon/D_d}} \tag{8.22}$$

A plot of Eq. (8.22) is shown in Fig. 8.7 for different values of the growth parameters ($\upsilon\varepsilon/D_d$) ranging from 0 to 10. For low values of $\upsilon\varepsilon/D_d$ and $k_o < 1$, the effective segregation coefficient is approximately the same as k_o, but it is always larger than k_o and approaches unity for large values of $\upsilon\varepsilon/D_d$. Thus, a uniform doping distribution, where k_e approaches unity, in the crystalline solid can be achieved by increasing the pull maximum velocity and a low rotation speed. Because of the centripetal force, the rotation speed is inversely proportional to ε.

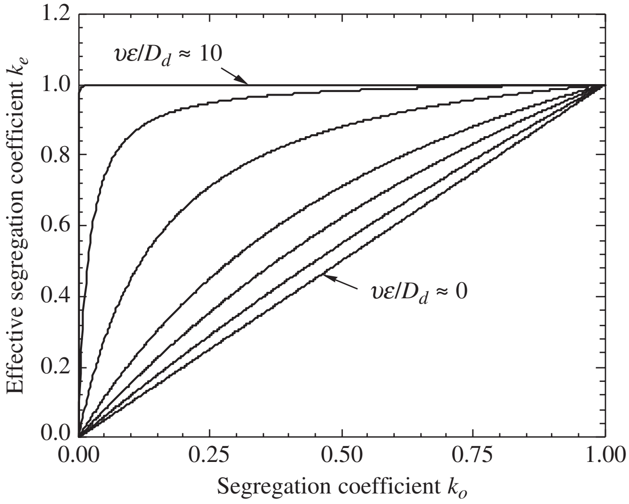


Figure 8.7 The effective segregation coefficient k_e is plotted as a function of k_o for different values of $\upsilon\varepsilon/D_d$.

8.2.2 Horizontal Bridgman method

The horizontal Bridgman and LEC growth methods are similar in that both depend on the presence of the crystal seed. Basically, both growth techniques consist of melt and solid phases. In the horizontal Bridgman method, however, the semiconductor material, including melt, crystal, and seed, is kept inside the crucible during the entire heating and cooling processes. This technique is illustrated in Fig. 8.8, where two variations of the same method are presented. In the case of silicon growth using the Bridgman method, a quartz crucible filled with polycrystalline material is placed inside a furnace tube, and the heater is pulled. As the heater is drawn slowly away from the seed (see Fig. 8.8*a*), the polycrystalline material is melted near the seed. The heater continues to move away from the seed's region, and the molten material solidifies into a single crystal with a crystallographic structure similar to that of the seed. The shape of the resulting crystal is determined by the shape of the crucible. Another variation of this growth technique is shown in Fig. 8.8*b*, where the crucible is pulled slowly from the heater region into a colder region. The seed crystal induces single-crystal growth.

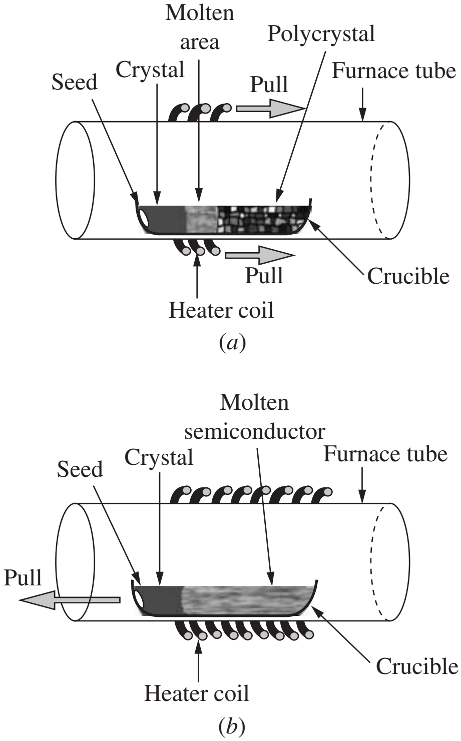


Figure 8.8 Schematic diagrams of the horizontal Bridgman growth method. (*a*) The heater is pulled over the polycrystalline material, causing it to melt. As the heater moves away from the seed's region, the molten material solidifies into a single crystal. (*b*) A variation of the Bridgman method, where the crucible is pulled away from the heater region.

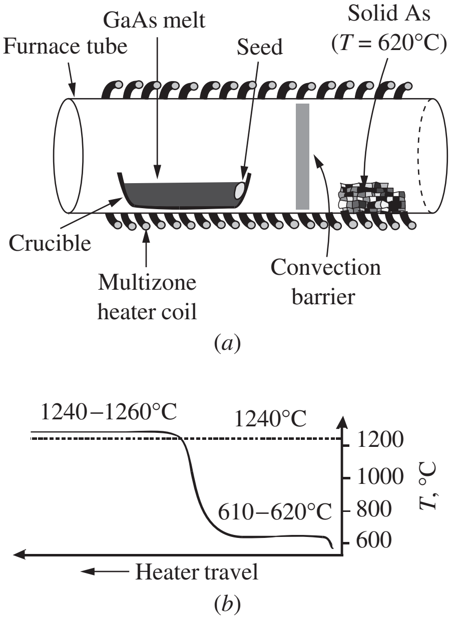


Figure 8.9 (*a*) A schematic diagram of the horizontal Bridgman growth method for GaAs and other compound semiconductors. (*b*) A sketch of the temperature for a two-stage furnace.

The drawback of the Bridgman growth method is that the material is constantly in contact with the crucible, which produces two effects. First, the silicon crystals tend to adhere to the crucible, and second, the crucible wall introduces stress in the solidifying crystal. The presence of stress causes deviations from the ideal crystal structure.

The growth of compound semiconductors using the Bridgman method is somewhat different than the growth of silicon crystals. For example, the growth of GaAs crystals is illustrated in Fig. 8.9*a*. In this growth method, the solid gallium and arsenic are loaded onto a fused silica ampoule, which is then sealed. The addition of the solid arsenic in the chamber provides the overpressure necessary to maintain stoichiometry. The furnace tube is slowly pulled past the charge (*charge* is a term used to describe the semiconductor components that are placed in the crucible). The temperature of the furnace is set to melt the charge when it is completely inside the furnace. As the furnace is pulled past the ampoule, the molten GaAs charge recrystallizes with a structure similar to that of the seed crystal.

The heater coil in the Bridgman growth method is actually a multizone furnace. Figure 8.9*b* shows the temperature as a function of the direction of the heater travel for a two-stage heater. The first stage (on the right) is kept at about 610 to 620°C to maintain the required overpressure of arsenic. The second stage (on the left) is held at about 1240 to 1260°C, which is just above the melting point of GaAs (~1240°C). It is possible to grow GaAs with this method using GaAs polycrystalline as the charge (the starting material) instead of solid gallium and arsenic components.

Compound semiconductor boules grown by the horizontal Bridgman method are usually ~50 mm in diameter, which is small compared to those boules grown by the LEC method. However, with precise control of the stoichiometry and the radial and axial temperature gradients, a large boule can be grown. The advantage of the horizontal Bridgman method is that the dislocation densities in materials, such as GaAs, are on the order of 10^3 cm^{-2}, which is about an order of magnitude smaller than the dislocation densities found in material grown by the LEC method.

8.2.3 Float-zone growth method

The float-zone growth method is mostly used to grow high-purity silicon boules directly from a high-purity rod of polycrystalline material obtained from other methods such as purification processes. With this growth method, a background carrier concentration lower than 10^{11} cm^{-3} can be easily achieved. Compound semiconductor materials are not generally grown by this technique. A schematic of the float-zone growth apparatus is shown in Fig. 8.10. A seed crystal is attached at the bottom of the polycrystalline rod in a vertical position. The rotating polycrystalline rod is enclosed in a quartz tube. An inert gas

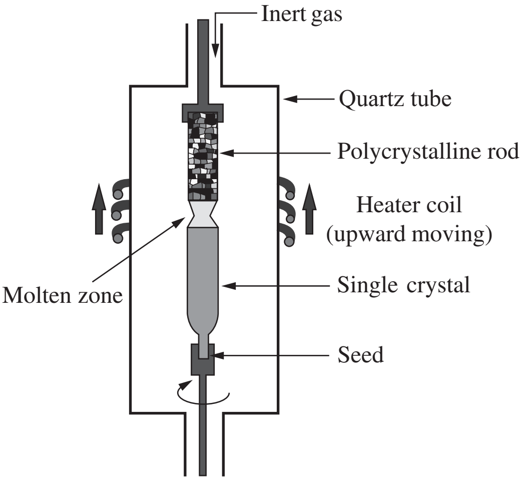


Figure 8.10 A cross section of the float-zone apparatus used to grow silicon single crystals.

(argon) flows in the tube such that 1-atm pressure is maintained during growth. A small region of the polycrystalline rod is melted by passing an RF heater, which is moved upward from the seed. A float-zone (a few centimeters in length) of melt is formed between the seed crystal and the polysilicon rod and retained by the surface tension between the melting and the growing solid phases. The molten zone that solidifies first remains in contact with the seed crystal retaining the same crystallographic structure of the seed. As the molten region is moved along the length of the polycrystalline rod, the rod melts and then solidifies throughout its entire length becoming a single crystal. The motion of the heater controls the diameter of the crystal.

Difficulties in preventing the molten zone from collapsing have limited the float-zone method to growing small-diameter crystals. The maximum crystal diameter is on the order of 70 mm. However, one of the advantages of this technique is that there is no crucible involved, so oxygen contamination is eliminated. Another advantage of this growth method is that the background impurities can be substantially reduced by passing the heater coil over the crystal several times. The background impurities can be reduced by a few orders of magnitude when the heater coil is passed over the crystal seven or eight times (see Pfann 1966).

The introduction of doping in the growth method is more difficult when compared to the LEC and Bridgman growth methods. There are four techniques used to introduce dopants in the float-zone growth method. First, core doping is based on the introduction of doped polysilicon as the starting material. Second, gas doping is based on the injection of gases, such as $AsCl_3$, BCl_3, or PH_3, into the polysilicon rod as it is being deposited or into the molten zone region during refining. Third, pill doping is based on the insertion of a small pill of dopant, such as gallium or indium, into a hole at the top of the polysilicon rod. Dopants with small segregation coefficients will diffuse into the rod as the melt passes over the polysilicon rod. Fourth, neutron transmutation doping is based on irradiating the silicon single crystal by thermal neutrons. This process produces a fractional transmutation of silicon into phosphorus, which leads to *n*-type silicon. The neutron transmutation equation is given by

$$\mathrm{Si}_{14}^{30} + \text{thermal neutron} \rightarrow \mathrm{Si}_{14}^{31} + \gamma\,\text{ray} \rightarrow \mathrm{P}_{15}^{31} + \beta\text{ray} \tag{8.23}$$

The lifetime of the intermediate Si_{14}^{31} is 2.62 hours. Neutron transmutation doping is very uniform since the penetration length of thermal neutrons is $\sim$100 cm in silicon.

The doping distribution in the float-zone process can be understood by considering the model illustrated in Fig. 8.11. Assume that the initial

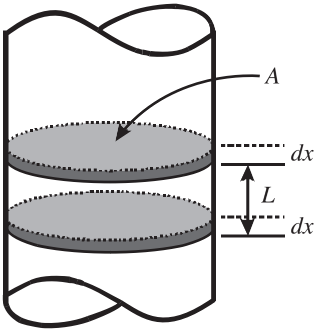


Figure 8.11 A sketch of a portion of a semiconductor boule used to illustrate the calculation of the doping profile.

uniform doping concentration in the rod is C_o, L is the length of the molten zone at a distance x along the rod, A is the cross-sectional area of the rod, ρ is the specific density of silicon, and σ is the dopant concentration in the molten zone. When the molten zone is moved a distance dx, the amount of dopant added to it at its advancing end can be expressed as $C_o \rho A\ dx$, while the amount of dopant removed at the retreating end is $k_e \sigma dx/L$. The differential amount of dopant, $d\sigma$, remaining in the molten zone as it moves a distance dx can be expressed as

$$d\sigma = C_o \rho A dx - \frac{k_e \sigma}{L} dx \tag{8.24}$$

which can be rewritten as

$$\int_0^x dx = \int_{\sigma_o}^{\sigma} \frac{d\sigma}{C_o \rho A - k_e \sigma / L} \tag{8.25}$$

where σ_o is the amount of dopant in the molten zone when it was first formed at the front end of the rod and is given by $C_o \rho A L$. An expression for the dopant concentration in the crystal at the retreating end is given by

$$C_s = \frac{k_e \sigma}{A \rho L} \tag{8.26}$$

An expression can be obtained for σ by integrating Eq. (8.25), which can be substituted into Eq. (8.26) to yield the following relation:

$$C_s = C_o[1 - (1 - k_e)e^{-k_e x/L}] \tag{8.27}$$

A plot of this equation is left as an exercise. For small values of $k_e x/L$, C_s is nearly constant.

8.2.4 Lely growth method

The growth techniques discussed earlier could not be employed to grow wide-bandgap materials such as GaN and SiC. This is due to the fact that these materials do not have a liquid phase under reasonable thermodynamic conditions. The growth conditions of the materials require a high-temperature, high-pressure environment. For example, SiC melt exists only at pressures in excess of 10^5 atm and at temperatures higher than 3200°C. Under these extreme conditions, the stoichiometry and stability of the melt are difficult to maintain. Silicon carbide material is grown by the Lely method, which is schematically shown in Fig. 8.12*a*. The growth process is driven by a temperature gradient, which is maintained between the outer and inner areas of the crucible. The system is kept near equilibrium with lower partial pressures of the SiC precursor in the inner and colder zone. The two areas are separated by a porous graphite, which provides nucleation centers. Since the inner region is colder, the chemical gradient causes a mass transport from the outer region to the inner region. Single crystals of SiC start to nucleate on the inner side of the porous graphite. As illustrated in Fig. 8.12*a*, the

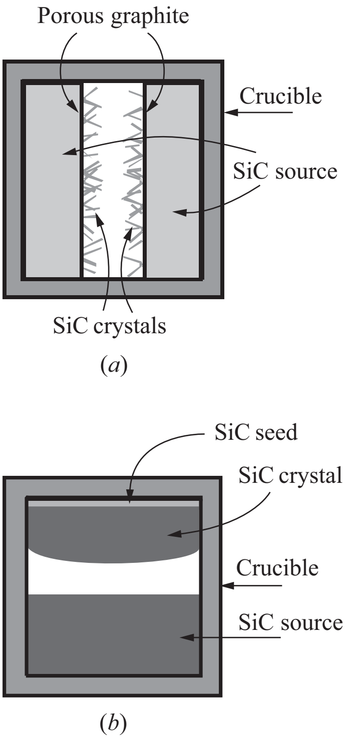


Figure 8.12 (*a*) A cross-sectional diagram of a cylindrical crucible for the Lely growth of SiC. (*b*) A cross-sectional diagram of the modified Lely growth technique.

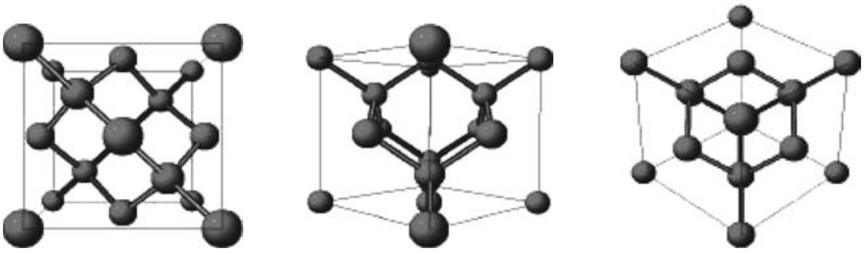
crystals are limited in size with random dimensions, but nonetheless they are of a high quality in terms of low-defect densities. A typical size of these crystals is ~1 cm, but they are used as seed crystals for other bulk SiC growth techniques including the modified Lely method described next.

The modified Lely method is based on a seeded sublimation growth or physical vapor transport technique. It is basically similar to the Lely method except that a SiC seed crystal is used as shown in Fig. 8.12*b* to achieve a controlled nucleation. According to Fig. 8.12, the cooler seed is placed at the top to minimize falling contaminations. The polycrystalline SiC source is heated to ~2600°C at the bottom of the crucible and sublimes at low pressure. Mass transport of SiC occurs and recrystallizes through supersaturation at the seed. The disadvantages of this technique include the poor control of the polytype and shape of the crystals, nonuniform doping, and high density of defects. Furthermore, the screw dislocations have been one of the long-standing problems of commercial bulk SiC wafers. This class of dislocations, which are also known as nano- or micropipes, can be closed or hollow. These micropipes have detrimental effects on SiC-based devices.

8.3 Growth of Semiconductor Thin Films

The growth of thin films requires finely polished substrates (wafers) cut from single-crystal boules grown by the bulk crystal growth methods described in Sec. 8.2. The growth of thin films, quantum wells, superlattices, quantum wires, and quantum dots requires a precise knowledge of the crystallographic structure of the substrates. Silicon substrates are used in the vast majority of silicon-based devices and technology. Silicon also has been used as a substrate for GaN-based devices due to its favorable physical properties, high quality, large sizes, and above all low cost. It has a diamond structure, which can be seen as two interpenetrating face-centered cubic sublattices with one sublattice displaced from the other by one quarter of the distance along the diagonal of the cube. Each atom in the lattice is surrounded by four equivalenly nearest neighbors that lie at the corners of a tetrahedron. The three commonly used structural orientations of Si are shown in Fig. 8.13.

Gallium arsenide–based technology, which includes both electronic and optoelectronic devices, has been advancing rapidly because the epitaxial growth is mature for many GaAs quantum structures. The reason for this advancement is the availability of GaAs substrates with many structural orientations. Gallium arsenide single crystals have a zinc-blende structure. A view of the main three structural orientations is shown in Fig. 8.14. Recently, GaAs substrates have been used for the growth of III-nitride materials. However, because of the relatively



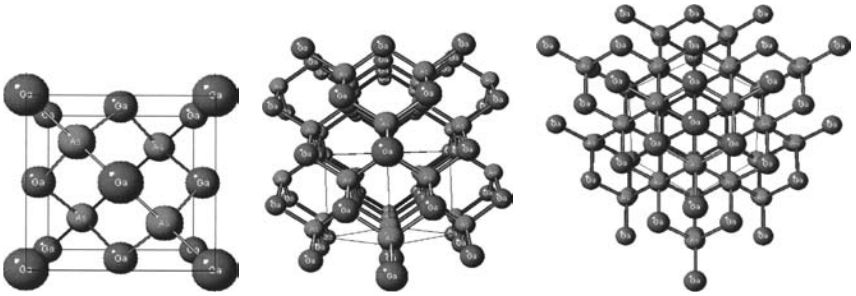
[100] [110] [111]

Figure 8.13 Views of the three commonly used crystallographic directions of Si wafers.

low melting point of GaAs, it is less stable when compared to SiC and sapphire substrates.

8.3.1 Liquid-phase epitaxy method

The liquid-phase epitaxy (LPE) growth method is basically a precipitation of materials from supercooled solution onto a substrate. The LPE reactor is shown in Fig. 8.15*a*, which consists of a horizontal furnace system and a sliding graphite boat. An enlarged illustration of the sliding graphite boat is shown in Fig. 8.15*b*, where two different melts A and B are used as an example of how a heterojunction can be made. If melt A is GaAs, melt B is AlGaAs, and the substrate is semi-insulating GaAs, then a GaAs/AlGaAs heterojunction can be grown with this technique. This simple reactor usually produces high-purity thin films. The epitaxial growth processes of this technique are usually maintained at thermodynamic equilibrium. The composition of the thin films depends mainly on the equilibrium phase diagram of the material and to a lesser extent on the orientation of the substrate. The molten material



[100] [110] [111]

Figure 8.14 A view of three different structural directions of GaAs, which has a zincblende crystal structure.

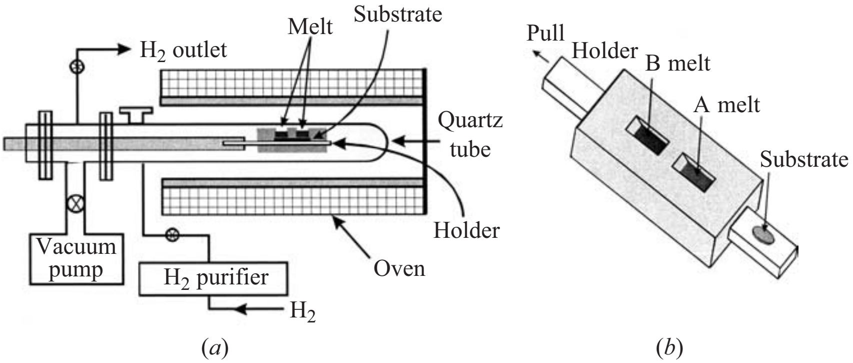


Figure 8.15 (*a*) A cross section of the liquid-phase epitaxy reactor. (*b*) An enlarged crucible showing the melts and the substrate on the holder, which slides under the melt.

is placed in a graphite boat and is slid inside the heated furnace of a suitable atmosphere. A subsequent cooling causes the solute to come out and deposit on the underlying substrate forming an epitaxially grown layer. Growth using the LPE method is affected by the melt composition, growth temperature, and growth duration.

The advantages of the LPE method are the simplicity of the equipment used, higher deposition rates, low defect concentrations, excellent control of stoichiometry, and high-purity materials. Background impurities are minimized by using high-purity melt materials and by the inherent purification process that occurs during the liquid-to-solid phase transition. Disadvantages, on the other hand, include poor thickness uniformity, high surface roughness, melt-back effect, and high growth rates, which prevent the growth of multilayer structures, such as multiple quantum wells and superlattices, with abrupt interfaces. Additionally, only small-size wafers can be used with the LPE method, which makes it a small-scale process. Contrary to materials grown from the melt, LPE-grown materials are temperature-independent and thermal gradients are usually neglected.

8.3.2 Vapor-phase epitaxy method

The term vapor-phase epitaxy (VPE) implies that the growth of thin films is based on reactive compounds in their gaseous form. The recent development of III-nitride materials renewed interest in the VPE growth method. This growth technique is performed at thermodynamic equilibrium. A generic sketch of the VPE reactor is shown in Fig. 8.16 (see, for example, Razeghi 1989). As shown in the figure, the reactor consists of a quartz tube (chamber), gas inlets, exhaust, and a furnace

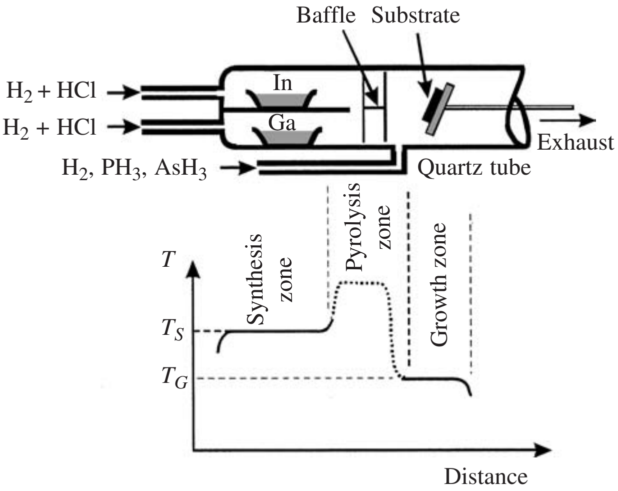


Figure 8.16 A cross section of a vapor-phase epitaxy reactor with a heater with multiple temperature zones.

with different temperature zones. Three zones (synthesis, pyrolysis, and growth) are shown in the temperature trace. The growth of InP and GaAs samples is taken as an example. The gaseous species for the group III source materials are synthesized by reacting hydrogen chlorine gas with a melted pure metal placed in crucibles. This process occurs in the first zone, the synthesis zone, which is maintained at temperatures T_s of 750°C and 850°C for GaAs and InP, respectively. The reaction between the metals and hydrogen chlorine results in group III–chloride vapor compounds as follows:

$$\mathrm{In}_{\mathrm{liquid}} + \mathrm{HCl}_{\mathrm{gas}} \rightarrow \mathrm{InCl}_{\mathrm{gas}} + \frac{1}{2}\mathrm{H}_{2,\mathrm{gas}} \tag{8.28a}$$

$$\mathrm{Ga}_{\mathrm{liquid}} + \mathrm{HCl}_{\mathrm{gas}} \rightarrow \mathrm{GaCl}_{\mathrm{gas}} + \frac{1}{2}\mathrm{H}_{2,\mathrm{gas}} \tag{8.28b}$$

The group V source materials are provided in the form of hydride gases, such as arsine (AsH_3) and phosphine (PH_3). The hydride gas are pyrolized in the second zone, which is maintained at temperatures $T > T_s$. The decomposition of group V can be described as follows:

$$\mathrm{AsH}_3 \rightarrow \frac{u}{4}\mathrm{As}_4 + \frac{1-u}{2}\mathrm{As}_2 + \frac{3}{2}\mathrm{H}_2 \tag{8.29a}$$

$$\mathrm{PH}_3 \rightarrow \frac{v}{4}\mathrm{P}_4 + \frac{1-v}{2}\mathrm{P}_2 + \frac{3}{2}\mathrm{H}_2 \tag{8.29b}$$

where u and v are the mole fractions of AsH_3 and PH_3 that are decomposed into As_4 and P_4, respectively.

The gas flow is cooled by a temperature gradient between the second and third zones. The cooling of reactants results in the growth of semiconductor materials, such as GaAs and InP, on the substrate in the growth zone. The growth zone is maintained at temperatures T_G of 680°C and 750°C for GaAs and InP, respectively. It is clear from Eqs. (8.28) and (8.29) that there are several chemical reactions taking place in the VPE reactor. These reactions can be classified as heterogeneous reactions, which occur between solids and liquids, solids and gases, and liquids and gases, and homogeneous reactions that occur in the gas phase. During the steady-state film growth, the overall growth process is limited by the heterogeneous reactions, whereas the changes in the composition of the grown semiconductor in the process (for example, switching the growth from InP to InGaAs) is limited by the mass transport in the gas phase.

The advantages of the VPE method include high deposition rate, multiwafer growth, flexibility in introducing dopants into the materials, and good control of the composition gradients by accurate control of the gas flows. The disadvantages include difficulties in growing multilayer quantum structures, potential formation of hillocks and haze, and interfacial decomposition during the preheat stage.

The renewed interest in the VPE method stems from its ability to achieve high deposition rates under reasonable growth conditions. This advantage has been used for the growth of thick GaN films, on the order of 100 μm or thicker, where native bulk substrates are not available. The idea here is to replace the current substrates, which are mainly sapphire and SiC, by producing thick GaN films that can be used as compliant (quasi) substrates. The thick GaN films can be grown on other substrates, such as sapphire, and then lifted and used as substrates in other growth techniques, such as MBE and MOCVD. The VPE growth of GaN utilizes hydrogen chlorine gas that passes over a crucible containing metallic gallium at a temperature of ~850°C to form gaseous GaCl. Ammonia (NH_3) and HCl are then injected into the hydride pyrolysis zone using N_2 as a carrier gas. Gallium chloride is injected through a showerhead into the growth zone, which is kept at temperatures in the range of 950 to 1050°C. Gallium chloride then reacts with NH_3 on the substrate surface to produce GaN according to the following reaction:

$$\mathrm{GaCl} + \mathrm{NH_3} \rightarrow \mathrm{GaN} + \mathrm{HCl} + \mathrm{H_2} \tag{8.30}$$

The $NH_3 : HCl$ ratio is typically 30 : 1 with a growth rate of ~0.3 μm/min. There are, however, problems associated with the VPE growth of GaN.

First, it is quite possible for NH_3 to dissociate and react with HCl to produce NCl_3, which is highly explosive. Second, HCl can potentially cause leaks in the reactor. Third, undesired by-products such as NH_3Cl and $GaCl_3$ can clog the exhaust system unless heated to temperatures higher than 150°C. Fourth, due to the exchange reactions with the quartz chamber walls of the reactor, AlGaN growth and p-type doped GaN are difficult to realize.

8.3.3 Hydride vapor-phase epitaxial growth of thick GaN layers

The hydride vapor-phase epitaxy (HVPE) method is essentially a variation of the VPE method, but it has been developed to grow thick GaN films (for further details on the subject see Paskova and Monemar 2003). The increasing interest in III-nitride materials and devices has led to the long-standing demand for GaN substrates for homoepitaxy of GaN, which has yet to be satisfied. There are substantial difficulties in growing large-volume GaN single crystals at the high-equilibrium vapor pressure of N_2 and the high-growth temperature needed in bulk growth from Ga solution. Currently, there are three promising techniques that can be used to obtain GaN bulk crystals: high-pressure crystal growth from Ga solution, the sublimation technique, and HVPE growth.

Hydride vapor-phase epitaxy growth of thick GaN layers was developed by several research groups to provide quasi-bulk thick GaN layers. GaN layers with a thickness of several hundred micrometers have been achieved in HVPE atmospheric pressure reactors at temperatures of about 1050°C, and at a reasonable cost. The deposition process of GaN for substrate application requires a high growth rate and the ability to produce low-defect material, since the threading defects are likely to extend into the subsequently grown epilayers. The growth rates in HVPE have been reported to be as high as 100 μm/h with a crystalline quality comparable to the best quality reported for metal-organic chemical vapor deposition (MOCVD) grown GaN films. Several substrate pretreatments such as a GaCl sapphire pretreatment, a sapphire nitridation pretreatment, different buffer layers such as ZnO, reactive sputtered AlN, MOCVD-grown GaN, and epitaxial lateral overgrowth technique have greatly improved the quality of thick HVPE-grown GaN films.

Despite rapid progress in the HVPE technique, a number of basic issues remain to be solved. One of them is the presence of a high density of extended defects such as dislocations, domain boundaries, and cracks. Efforts to further develop the HVPE-grown GaN thick layers for substrate use are concentrated on two main issues. The first focuses on the reduction of defect density and the control of the initial stage of the growth, which is the source for most defects. Second, an optimal procedure for subsequent removal of the foreign substrate from the GaN layer is far from complete, although very intense investigations of chemical, reactive ion etching, laser-induced liftoff, and polishing separation have been reported in the literature.

The basic HVPE reactions that describe the GaN deposition process can be written as follows:

$$x\mathrm{Ga}(l) + \mathrm{HCl}(g) \rightarrow x\mathrm{GaCl}(g) + (1-x)\mathrm{HCl}(g) + \frac{x}{2}\mathrm{H}_2(g) \qquad (8.31a)$$

$$\mathrm{GaCl}(g) + \mathrm{NH}_3(g) \rightarrow \mathrm{GaN}(s) + \mathrm{HCl}(g) + \mathrm{H}_2(g) \qquad (8.31b)$$

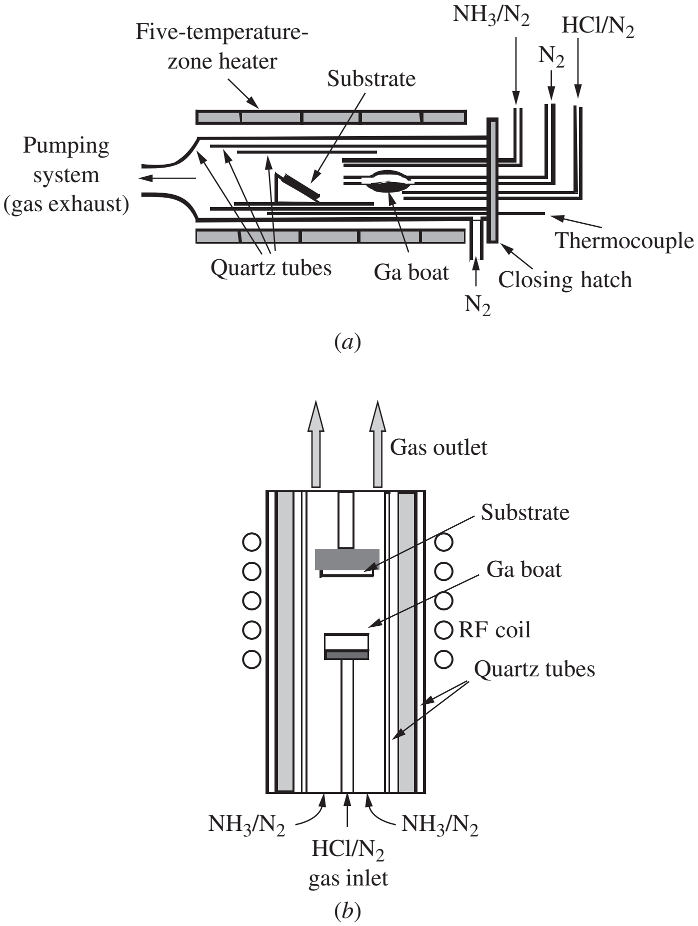
where l is liquid, g is gas, s is solid, and x is the mole fraction of HCl reacting in the process. Notice that Eq. (8.31*b*) is the same as Eq. (8.30). The GaN deposition is determined by the efficiency of both chemical reactions. Values of x in reaction (8.31*a*) were found to be in the range from 0.70 to 0.86 depending on the temperature, the position of the HCl inlet, the carrier gas ambient, and the liquid Ga surface exposed to the HCl gas. The chemical reaction (8.31*b*) depends on the fraction of ammonia not decomposed into nitrogen and hydrogen, since GaN cannot be formed by direct reaction between GaCl and N_2. It is known that ammonia is a thermodynamically unstable gas at the temperatures employed in the GaN growth. Fortunately, the thermal decomposition of NH_3 is a very slow reaction, and when no catalyst is present, no more than about 4 percent of the NH_3 is typically decomposed at temperatures higher than 950°C. Equation (8.31) is accompanied by GaN decomposition via the following two reactions:

$$\mathrm{GaN}(s) + \mathrm{HCl}(g) \rightarrow \mathrm{GaCl}(g) + \frac{1}{2}\mathrm{N}_2(g) + \frac{1}{2}\mathrm{H}_2(g) \qquad (8.32a)$$

$$\mathrm{GaN}(s) \rightarrow \mathrm{Ga}(l) + \frac{1}{2}\mathrm{N}_2(g) \qquad (8.32b)$$

These decomposition reactions are unlikely to occur in the growth temperature range of 950 to 1150°C.

The basic design of the HVPE reactor is similar to the VPE reactor with some modifications, as shown in Fig. 8.17. These modifications can be summarized into two groups: horizontal and vertical reactor design. The horizontal reactor shown in Fig. 8.17*a* typically has five main temperature zones. In the first upstream zone HCl reacts with metallic Ga forming GaCl and H_2. The area of the liquid Ga source is increased as much as allowed (typically 10 to 100 cm^2) to achieve a large reactive Ga surface area for efficient GaCl production. The optimum temperature in the first zone is about 850°C. The second zone may be used for other metallic sources, such as In or Al when needed, or for dopants. The temperature of the third zone is kept in the range of about 1000 to 1060°C where GaCl and NH_3 are introduced and mixed. The substrate



(*a*)

(*b*)

Figure 8.17 Schematic diagrams of (*a*) horizontal and (*b*) vertical HVPE reactors.

holder is placed in the fourth region of the reactor, where the temperature is kept at ~1080°C. The most common method of heating in this design is resistive heating. The horizontal reactors utilize a susceptor that is situated approximately parallel to the gas flow direction. Uniform growth can be improved by tilting the substrate holder to eliminate reactant depletion along the flow direction. Another approach used in some horizontal reactors is the rotation of the substrate holder.

In the vertical design, the reactants are typically introduced through the top. The substrate is held flat on a susceptor that is perpendicular to the gas flow direction. The vertical reactor design facilitates substrate rotation during the growth to improve film uniformity. Heating is accomplished by resistance or RF induction, and temperature monitoring is accomplished by an infrared pyrometer or a thermocouple. An alternative modification is an inverted vertical reactor as shown in Fig. 8.17*b*, where the process gases are supplied through the bottom inlet flange, while the top flange can be lifted for loading and unloading. The substrates are placed in the upper part where the gases are mixed. The inverted reactor keeps all advantages of the vertical design and also provides the possibility for raising the substrate holder. An additional advantage of the inverted vertical reactor is the significant reduction of solid particle contamination.

8.3.4 Pulsed-laser deposition technique

Pulsed-laser deposition (PLD) is a relatively new technique widely used for the growth of oxide thin films, such as ferroelectrics and superconductors (for a detailed discussion, see Huang and Harris 2003). There are, however, several advantages of PLD for depositing high-quality thin films that make it worthy of study as a method of growing III-nitride materials. One of these advantages is the simplicity of the technique. Pulsed-laser deposition is typically accomplished with a high-power pulsed-laser beam irradiating a bulk stoichiometric target. Through the interaction of the laser beam with the target, a forward-directed flux of material is ejected. A plasma is formed which is then transported toward a heated substrate placed directly in the line of the plume. This is illustrated in Fig. 8.18. The congruent ablation achieved with short laser pulses enables stoichiometric composition transfer between targets and films and allows deposition of multicomponent materials by employing a single target. This feature makes PLD the best initial investigation tool for complex materials because the stoichiometry control is vastly easier.

A useful feature of the PLD method is that multiple targets can be loaded inside the chamber on a rotating holder, which can then be used to sequentially expose different targets to the laser beam, thereby enabling in situ growth of heterostructures and superlattices with relatively clean interfaces. Virtually any material can be laser evaporated, leading to the possibility of multilayers of a variety of materials. Therefore, PLD is suitable for rapid exploration of new materials-integration strategies for developing heterostructures and performing basic studies at the laboratory scale. The growth rate achieved by PLD can be varied from subangstroms per second to a few microns per hour by adjusting the repetition rate and the laser fluence, which is useful for both atomic-level investigations and thick layer growth. Moreover, the strong nonequilibrium growth conditions of PLD may allow a much broader range of metastable materials to be grown, including the introduction

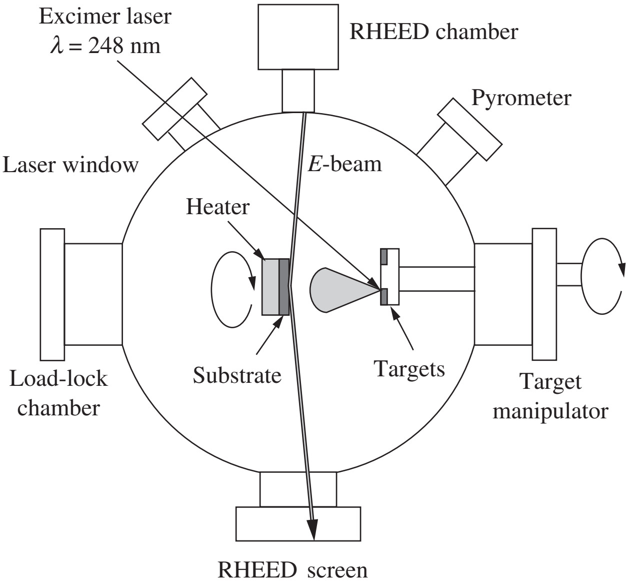


Figure 8.18 Schematic diagram of the pulsed-laser deposition chamber. (*After Huang and Harris 2003.*)

of higher dopant concentrations and alloy compositions that normally segregate in the equilibrium phase.

There are three main stages of the PLD process: laser-target interaction, laser-plume interaction, and subsequent deposition of the thin film. In the beginning of the laser pulse, the optical energy is largely absorbed by the surface of the target. Since the laser energy is supplied to a small volume of 10^{-13} m^3 in a short time (typically 30 ns), the local temperature of the target is frequently on the order of 10^4 K. Therefore, all the species of the target evaporate simultaneously, i.e., congruent evaporation. This condition ensures that the ejected materials have the same stoichiometry as the target which makes the PLD process particularly suitable for exploring binary, ternary, and more complicated systems without having to adjust fluxes from multiple sources. Continuous interaction of the laser beam with the plume results in the photo-dissociation and photo-ionization of the evaporated material. This interaction breaks molecular species and clusters and ionizes the evaporated material by a nonresonant multiphoton process, leading to the formation of expanding plasma above the target surface, and transport toward the substrate.

The evaporation of the materials from targets by laser irradiation depends on the laser parameters (such as laser fluence, pulse duration, and wavelength) and material properties (such as reflectivity, absorption coefficient, and thermal conductivity). According to Fig. 8.18, a KrF excimer laser operating at a wavelength of 248 nm and a pulse duration of 20 ns are used to ablate materials from the targets. The laser is incident at 45° from the target normal, and the substrate is centered along the target normal. The system is capable of holding six targets for multilayer growth. Each target is rotated about its axis to ensure uniform wear on the targets, and individual targets can be successively clocked into position for the ablation of multitargets. A load-lock chamber with a magnetically coupled transfer rod is equipped to facilitate the transfer of both targets and substrates without breaking the vacuum of the main chamber. The base pressure, on the order of 10^{-8} torr, is achieved by pumping the chamber with turbo and mechanical pumps. The target to substrate distance can be varied over 15 cm to operate in different pressure regions. The substrate is rotated to enhance the temperature and thickness uniformity during deposition. The substrate heater is capable of reaching 800°C in either an oxygen ambient for oxide growth or a nitrogen ambient for nitride growth. Another attractive feature of PLD apparatus is the capability of *in-situ* monitoring of the growth process, such as reflection high-energy electron diffraction (RHEED). The RHEED patterns offer abundant information on the crystal structure and the quality of the growing film. It also provides means to study surface structure and growth kinetics.

8.3.5 Molecular beam epitaxy growth technique

Despite the high price tag on the molecular beam epitaxy (MBE) reactor, it is one of the most versatile and widely used nonequilibrium growth techniques. While the MBE growth processes are under continuous development ranging from effusion cell shape to the addition of many *in-situ* diagnostic tools, they have been used to grow almost any kind of doped and undoped semiconductor materials ranging from thin films and quantum wells to quantum dots. MBE is capable of controlling the deposition of a submonolayer on substrates with various crystallographic structures. A schematic of the MBE growth chamber is depicted in Fig. 8.19, which shows the sources for the growth of GaAs and GaN with two different dopants (Si and Mg).

A thin-film deposition process is performed inside the MBE chamber in which thermal beams of atoms or molecules react on the clean surface of a single-crystalline substrate that is held at high temperatures under ultrahigh vacuum conditions ($\sim 10^{-10}$ to 10^{-11} torr) to form an epitaxial film. It turns out that this ultrahigh vacuum is a major advantage to the MBE growth. This is primarily due to this very low impurity environment and the fact that many *in-situ* tools can be added

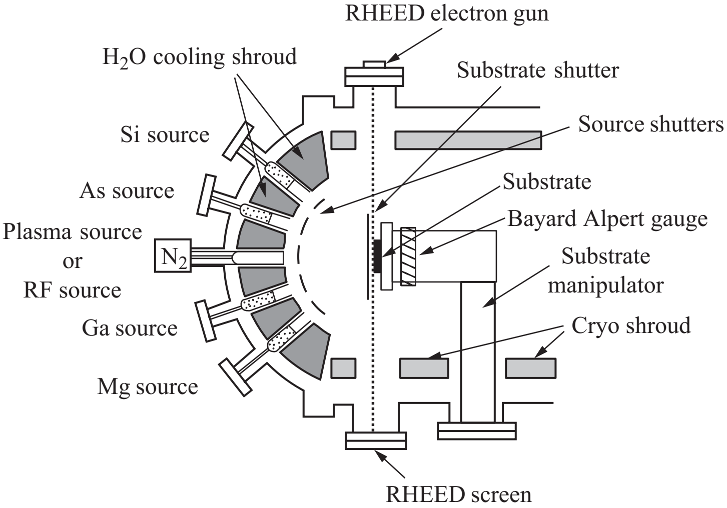


Figure 8.19 A sketch of a molecular beam epitaxy growth chamber showing the configuration of the sources and RHEED electron gun.

to the vacuum chamber. The most common way to create a molecular beam for MBE growth is through the use of Knudsen effusion cells. The crucibles employed in Knudsen cells are mostly made of pyrolytic boron nitride (PBN). The temperatures of different crucibles are usually independently controlled to within $\pm 1^\circ$C.

The material sources could be solid, gas, or metal-organic materials. Solid precursor sources are generally solids that are heated above their melting point in effusion cells known as Knudsen cells. The atoms of source material escape the cell into the vacuum chamber by thermionic emission. The beam flux is a function of its vapor pressure, which can be controlled by the source temperature. In the gas source MBE, method, group V of III-V semiconductors are connected through an injector or cracker. The gas source beam flux is controlled by using a mass flow controller. The metal-organic sources are either liquids or powders. An inert carrier gas is usually used to control the beam flux. The thickness and compositions of the epitaxial layers are controlled by the interruption of the unwanted atomic beam using shutters, which are usually remotely controlled by a computer. The beam of atoms and molecules will attach to the substrate forming the epitaxial layers. The growth rate is generally about a monolayer per second. The layers crystallize through the reaction between the atomic or molecular beams of the source materials and the substrate surface that is maintained at a certain temperature.

Another major difference between the MBE growth method and other growth techniques is that it is far from thermodynamic equilibrium

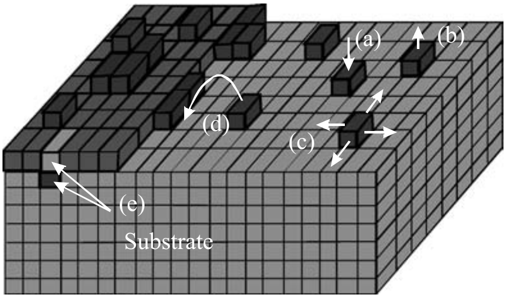


Figure 8.20 A schematic illustration of the kinetic processes that occur at the surface of the substrate during MBE growth.

conditions. It is mainly governed by the kinetics of the surface processes. The five major kinetic processes are illustrated in Fig. 8.20 where the blocks represent the atoms and molecules that reach the surface of the substrate. Process (*a*) is the adsorption of the atoms or molecules impinging on the substrate surface; process (*b*) is the thermal desorption of the atoms or molecules that are not incorporated in the epitaxial layer; process (*c*) is the surface migration and dissociation of the absorbed atoms and molecules; process (*d*) is the incorporation of atoms and molecules into the epitaxial layer or the surface of the substrate; and process (*e*) is the interdiffusion between the substrate and epitaxial layer.

In order to grow smooth surfaces, the atoms impinging on the substrate surface should be given enough time to reach their proper position at the edge before the formation of the next entire new layer. This also reduces the formation of defects and dislocations. The atoms in the MBE growth chamber have a long mean free path where collisions and scattering with other atoms are infrequent before reaching the surface of the substrate. This mean free path $\mathcal{L}$ can be written in terms of the atoms or molecules concentration $\mathcal{N}$ according to the following relation:

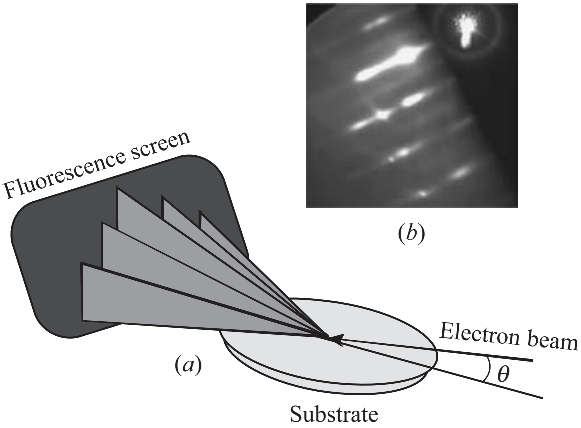
$$\mathcal{L} = \frac{1}{\sqrt{2}\pi \mathcal{N} d^2} \tag{8.33}$$

where d is the diameter of the species. The concentration $\mathcal{N}$ is determined by the vapor pressure P and temperature T inside the MBE chamber according to the following relation:

$$\mathcal{N} = \frac{P}{k_B T} \tag{8.34}$$

where k_B is Boltzmann's constant.

As mentioned earlier, one of the advantages of having an ultrahigh vacuum in the MBE growth chamber is that *in-situ* tools can be added to monitor the epitaxial layer during growth. One of these tools is



(*b*)

(*a*)

Figure 8.21 (*a*) The configuration of the RHEED inside the MBE chamber. (*b*) RHEED patterns formed a fluorescence screen. (*The RHEED pattern was obtained from G. J. Salamo.*)

reflection high-energy electron diffraction. The electron energy in the RHEED gun is typically 5 to 50 keV. The electrons are directed toward the substrate at a grazing angle $\theta \leq 1^\circ$. A schematic showing the RHEED configuration is shown in Fig. 8.21*a*. The electrons are then diffracted by the epitaxial layer formed at the substrate surface. This leads to the appearance of intensity-modulated streaks on a fluorescence screen. The results obtained using the RHEED gun are generally characterized as being in the static or dynamic mode. In the static mode, the atomic construction of the surface can be determined from RHEED diffraction patterns. These patterns (see Fig. 8.21*b*) usually provide information on the atomic surface construction, which is a function of the incoming electron beam flux, the substrate temperature, and the strain of the epitaxial layer. On the other hand, the dynamic mode is based on the change in intensity of the central diffraction streak as the wafer roughness changes over time. This process is illustrated in Fig. 8.22, where the formation of a single complete monolayer is shown. The fractional layer coverage is represented by the factor S. During the epitaxial growth process, the roughness of the epitaxial layer increases as a new atomic layer forms. When the surface coverage reaches 50 percent, or $S = 0.5$, the roughness is at a maximum and begins to decrease as the growing layer is complete, which corresponds to a minimum roughness ($S = 1.0$). The intensity of the main RHEED streak follows the period oscillation of the layer's roughness with the maximum intensity corresponding to the minimum roughness. This is illustrated in the RHEED oscillation signal depicted in the right-hand side panel. The time separation between two adjacent peaks in the RHEED

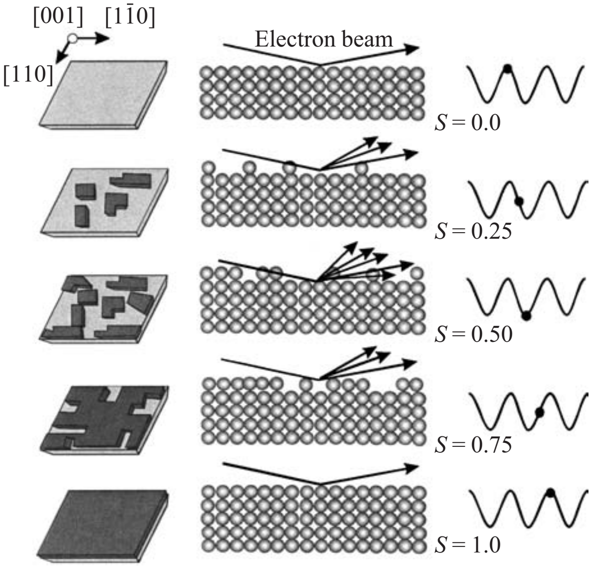


Figure 8.22 An illustration of the formation of a single monolayer as seen by a RHEED *in-situ* instrument. The corresponding RHEED oscillation signal is shown.

oscillations provides the time needed for the growth of a single layer of a crystal.

Another *in-situ* tool that is often used in the MBE growth chamber is Auger electron spectroscopy. This technique is based on the Auger effect of measuring the elemental composition surface. This technique uses an electron beam with energy ranging between 3 and 25 keV. The beam excites the atoms at the surface of the substrate by knocking a core level electron to a higher energy level. When the excited electrons relax, the atoms release the extra energy by emitting Auger electrons with characteristic energies. These energies are measured, and the quantity of Auger electrons is proportional to the concentration of atoms on the substrate surface. Thus, the Auger electron spectroscopy technique measures the planar distribution of elements on a surface. This technique can also be used to measure the depth profile when used in conjunction with an ion sputtering method.

8.3.5.1 Molecular beam epitaxy growth of III-nitrides. The growth of many semiconductor material thin films and quantum structures by the MBE technique, in particular III-V semiconductors, is approaching maturity in almost every aspect. Recently, the technique has been employed in growing GaN and related compounds. The MBE sketch shown in Fig. 8.18 is in fact configured to grow GaN structures. One of the major issues in the deposition of GaN and III-nitrides by MBE is the incorporation of an appropriate nitrogen source, since molecular nitrogen (N_2) does not chemisorb on Ga due to its large binding energy of 9.5 eV. To solve this problem, different approaches are currently being reported for the growth of cubic and hexagonal III-nitrides. The first approach is the use of gaseous sources like ammonia (NH_3) or dimethylhydrazine (DMH); this kind of MBE is also called chemical beam epitaxy (CBE) or reactive ion molecular beam epitaxy (RMBE). This compound is quite thermally stable and as a result limits the growth temperature significantly. Therefore, lower growth temperatures, such as those needed for low-temperature nucleation buffers or for layers containing indium, cannot be grown as easily with NH_3. DMH has a higher reactivity than NH_3 and is expected to produce better-quality crystals.

The second approach utilizes plasma-activated molecular nitrogen supplied via dc plasma sources, microwave plasma-assisted electron cyclotron resonance (ECR) plasma sources, or radio frequency (RF) plasma sources. However, due to the low growth rate of 10 to 30 nm/h imposed by the limited nitrogen flux of the dc source, the synthesis of a 1-μm-thick film would require approximately 50 hours, making it almost impossible to achieve stable growth conditions throughout such a run. Electron cyclotron resonance sources rely upon coupling microwave energy at 2.45 GHz with the resonance frequency of electrons in a static magnetic field. Such coupling allows for ignition of the plasma at low pressures and powers and produces a high concentration of radicals. In an ECR source, approximately 10 percent of the molecular nitrogen is converted into atomic nitrogen. Because these sources operate very efficiently at fairly low powers, they are usually cooled by air. A typical growth rate of an ECR source is about 200 nm/h. A detailed description of the design and principle of operation of microwave plasma-assisted ECR sources is given by Moustakas (1999). The physical properties of binary nitrides, namely GaN, AlN, and InN, are tabulated in Table 8.1.

Nitrogen plasmas are generated by inductively coupling RF energy at a frequency of 13.56 MHz into a discharge chamber filled with nitrogen at pressures of $>10^{-6}$ mbar. The discharge tube and the beam exit plate can be fabricated from pyrolytic boron nitride (BN) avoiding quartz, which may be a source of residual Si or O doping of GaN. The plasma sheath effect confines ions and electrons within the plasma discharge regions allowing only low-energy (<10 eV) neutral species to escape. Therefore these sources are believed to produce significant concentrations of atomic nitrogen. Because of the very high powers used in these sources, up to 600 W, the plasma chambers are usually water cooled. RF sources permit growth rates of up to about 1 μm/h.

All epitaxial growth requires bulk substrates. Bulk GaN is difficult to grow since high-pressure, high-temperature growth conditions are

TABLE 8.1 Physical Parameters of GaN, AlN, and InN

Parameter	Notation	Unit	GaN	AlN	InN
Lattice constant	a	Å	3.189	3.112	3.548
Lattice constant	c	Å	5.185	4.982	5.760
Thermal coefficient	$\Delta a/a$	$10^{-6}\,K^{-1}$	5.59	4.2	
Thermal coefficient	$\Delta c/c$	$10^{-6}\,K^{-1}$	3.17	5.3	
Bandgap, 300 K	E_g	eV	3.42	6.2	1.89?
Bandgap, 4 K	E_g	eV	3.505	6.28	
Electron effective mass	m_e	m_o	0.22		
Hole effective mass	m_h	m_o	>0.8		
Elastic constant	C_{13}	GPa	94	127	100
Elastic constant	C_{33}	GPa	390	382	392
Static dielectric constant	ε_r	ε_o	10.4	8.5	15.3
Spontaneous polarization	P_{spon}	C/m^2	−0.029	−0.081	−0.032
Piezoelectric coefficient	e_{31}	C/m^2	−0.49	−0.60	−0.57
Piezoelectric coefficient	e_{33}	C/m^2	0.73	1.46	0.97
Binding energy, exciton A	E_{xb}	meV	21		
Thermal conductivity	κ	W/(cm·K)	2.1	2.85	
Refractive index	n_r	—	2.2	2.15	
Melting point	T_m	K	>2573	>3000	

required (see Manasreh and Ferguson 2003). Thus, substrates other than GaN are currently used for the growth of epitaxial GaN thin films, heterojunctions, and quantum wells. The most commonly used substrates are sapphire (Al_2O_3) and SiC. The properties of these substrates are listed in Table 8.2. Although sapphire has a rhombohedral structure, a hexagonal cell can describe it as shown in Fig. 8.23. This is the same orientation as that of the grown GaN layer of wurtzite symmetry. The growth of GaN on sapphire suffers from the lattice mismatch of interatomic separation in the (0001) interface and from the mismatch of thermal expansion coefficients. The large lattice constant mismatch between GaN and sapphire causes the film to be completely relaxed (not strained). This large lattice constant mismatch must be improved by introducing various processing schemes, such as surface preparation, substrate nitridation, and the growth of buffer layers.

Surface preparation. Wet and *in-situ* methods of etching sapphire include phosphoric acid (H_3PO_4), sulfuric-phosphoric acid mixtures, H_2SO_4-H_3PO_4 fluorinated and chlorofluorinated hydrocarbons, tetrafluoro sulfur (SF_4), and sulfur hexafluoride (SF_6). However, the most common substrate preparation procedure prior to growth of GaN is to simply heat the sapphire under flowing hydrogen at high temperatures.

Sapphire nitridation. Sapphire is nitridated by exposure to nitrogen plasmas or thermally cracked ammonia in MBE systems. Under the conditions of temperature used for MBE growth, AlO_xN_{1-x} should be

TABLE 8.2 Physical Parameters of Sapphire and SiC

Parameter	Sapphire (Al_2O_3)	SiC	
Lattice constant, nm	$a = 0.4765$ @ 20°C	3C	$a = 0.43596$
	$c = 1.2982$ @ 20°C	2H	$a = 0.30753, c = 0.50480$
		4H	$a = 0.30730, c = 1.0053$
		6H	$a = 0.30806, c = 1.51173$
Melting point,* °C	2030	3C	2793
Density, g/cm^3	3.98 20°C	3C	3.166
		2H	3.214
		6H	3.211
Thermal expansion coefficients,† 10^{-6} K^{-1}	6.66 $\parallel$ to c axis @ 20–50 °C	3C	3.9
	9.03 $\parallel$ to c axis @ 20–10^3 °C	4H	4.46 a axis 4.16 c axis
	5.0 $\perp$ to c axis @ 20–10^3 °C		
Percent change in lattice constants between 293–1300 K for Al_2O_3 and 300–1400 K for SiC‡	$a/a_o = 0.83$	6H	$\Delta a/a_o = 0.4781$ $\Delta c/c_o = 0.4976$
	$c/c_o = 0.892$	3C	$\Delta a/a_o = 0.5140$
Thermal conductivity, W/(cm · K)	0.23 $\parallel$ to c axis @ 296 K	3C	3.2
	0.25 $\parallel$ to a axis @ 299 K	4H	3.7
		6H	3.8
Heat capacity, J/(K · mol)	77.9 @ 298 K	6H	0.71
Dielectric constant	8.6 $\parallel$ to c axis @ 10^2–10^8 Hz	3C	$\epsilon(0) = 9.75, \epsilon(\infty)$: 6.52
	10.55 $\parallel$ to a axis @ 10^2–10^8 Hz	6H	$\epsilon(0) = 9.66, \epsilon(\infty)$: 6.52 $\perp$ c axis $\epsilon(0) = 10.3, \epsilon(\infty)$: 6.70$\parallel$ c axis
Refractive index	1.77 @ $\lambda = 577$ nm	3C	2.6916 @ $\lambda = 498$ nm
	1.73 @ $\lambda = 2.33\,\mu$m	2H	2.6686 @ $\lambda = 500$ nm
		4H	2.6980 @ $\lambda = 498$ nm
		6H	2.6894 @ $\lambda = 498$ nm
Resistivity,§ Ω · cm	>10^{11}@ 300 K	4H	10^2–10^3
Young's modulus, GPa	452–460 in [0001] direction	3C	~440
	352–484 in [11$\bar{2}$0] direction		

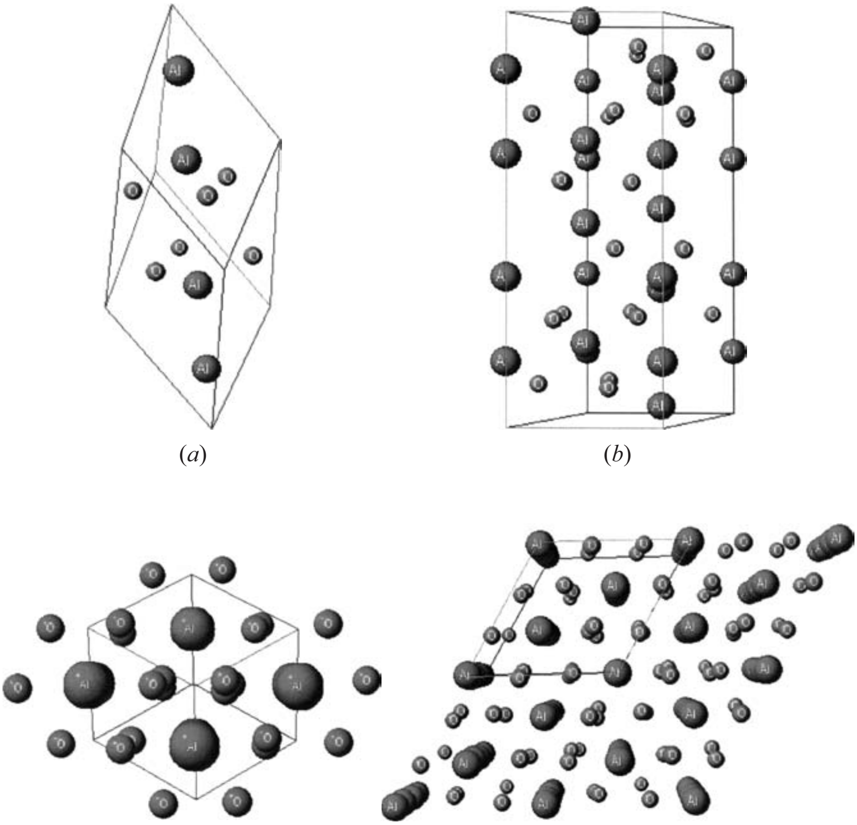
NOTE: The parameters of sapphire were obtained from Belyaev et al. (1980). Parameters for SiC were obtained from Harris (1995) unless otherwise specified.

* The melting point value for SiC obtained from Weast and Astle (1992).

† Thermal expansion coefficients for SiC obtained from Ambacher (1998).

‡ Percent change values for SiC obtained from Reeber and Wang (2000).

§ Resistivity value for SiC obtained from Siergiej et al. (1999).



(*a*) (*b*)

(*c*) (*d*)

Figure 8.23 The unit cell of sapphire. (*a*) Rhombohedral unit cell, (*b*) hexagonal unit cell, (*c*) a view in (2 × 2 × 1) unit cells along the longest diagonal in rhombohedral unit cells, and (*d*) a view in (2 × 2 × 1) unit cells along the (0001) direction in hexagonal unit cells.

unstable, and nitridation of sapphire results in the formation of AlN. The AlN layer promotes GaN nucleation and increases the wetting of the GaN overlayer from 550 to 820°C in MBE growth. The benefits of the nitridation layer are due to a change in the surface energy in the low-temperature GaN or AlN buffer layer. The nitridation of sapphire before the growth of a low-temperature buffer AlN or GaN is an important step for reducing the defect density, enhancing the electron mobility, and reducing the yellow luminescence in subsequently deposited films. The chemical alternation of surfaces of sapphire substrates using particle beams can be used as an alternative process to nitridation. The advantage of this method over nitridation is its simplicity and room temperature operation. The reactive ion (N_2^+) beam has also been used for pretreatment of sapphire substrates.

Buffer Layer. A low temperature GaN or AlN buffer (the growth temperature is usually about 400°C for MBE) is an important technique for III-nitride growth, since it can dramatically improve the surface morphology and crystalline quality of GaN sequentially deposited at high temperatures (700 to 850°C for MBE growth).

Polarity. Control of polarity of GaN film is critical in epitaxial growth. This is because it will change the surface morphology and doping character, and most importantly, it will determine the direction of the piezoelectric field which is crucial to the device performance.

Other substrates. In addition to sapphire substrate, III-nitride materials have been grown on other substrates such as SiC, GaAs, and Si (for more details, see Liu and Edgar 2002). The SiC substrates are second to sapphire for epitaxial growth of GaN thin films and quantum structures. The most common polytype SiC structures are illustrated in Fig. 8.24. These have several advantages over sapphire including a smaller lattice constant mismatch (3.1 percent) for [0001] oriented films, a much higher thermal conductivity [3.8 W/(cm·K)], and low resistivity, so electrical contacts to the back side of the substrate are possible.

The lattice constant mismatch for SiC is smaller than that for sapphire, but it is still sufficiently large to cause a large density of defects to form in the GaN layers. The crystal planes in epitaxial GaN parallel those of the SiC substrate, making facets for lasers easier to form by cleaving. It is available with both carbon and silicon polarities, potentially making control of the GaN film polarity easier. High-gain heterojunction bipolar transistors taking advantage of the discontinuity created at the GaN-SiC interface are possible. However, SiC does have its disadvantages. Gallium nitride epitaxy directly on SiC is problematic, due to poor wetting between these materials. This can be remedied by using an AlN or $Al_xGa_{1-x}N$ buffer layer, but at the cost of increasing the device resistance. This roughness and also remnant subsurface polishing damage are sources of defects in the GaN epitaxial layer. The screw dislocation density in SiC is 10^3 to 10^4 cm^{-2}, and these defects may also propagate into the GaN epitaxial layer and/or degrade device performance. The thermal expansion coefficient of SiC is less than that of AlN or GaN, and thus the films are typically under biaxial tension at room temperature. Finally, the cost of silicon carbide substrates is high.

Silicon substrate. Silicon substrates possess physical properties, including high quality and low cost, that are very attractive to GaN-based

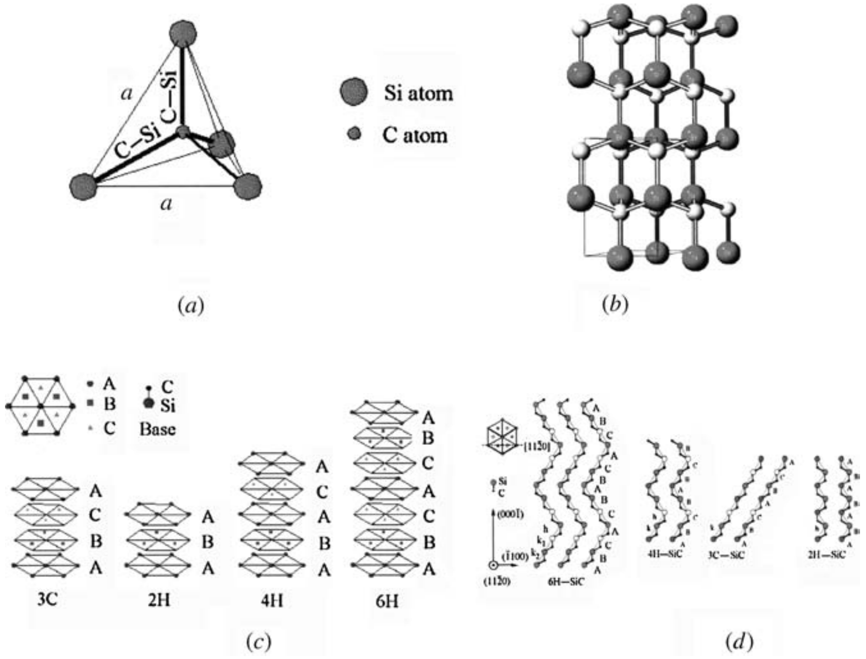


Figure 8.24 (*a*) The tetragonal bonding of a carbon atom with the four nearest silicon neighbors. The distance *a* and C–Si bond are approximately 3.08 and 1.89 Å, respectively. (*b*) The three-dimensional structure of 2H–SiC. (*c*) The stacking sequence of double layers of the four most common SiC polytypes. (*d*) The [11$\bar{2}$0] plane of the 6H–, 4H–, 3C–, 2H–SiC polytypes. (*After Gith and Petusky 1987.*)

devices. The physical parameters of bulk silicon are listed with those of GaAs in Table. 8.3. The crystallographic structure of silicon is shown in Fig. 8.13. Silicon wafers are very low priced and are available in very large sizes due to their mature development and large-scale production. Silicon has good thermal stability under GaN epitaxial growth conditions. The crystal perfection of silicon is better than any other substrate material used for GaN epitaxy, and its surfaces can be prepared with extremely smooth finishes. The possibility of integrating optoelectronic GaN devices with Si electronic devices is another advantage. To date, the quality of GaN epitaxial layers on silicon has been much poorer than that on sapphire or silicon carbide, due to a large lattice constant and thermal expansion coefficient mismatch, and the tendency of silicon to form an amorphous silicon nitride layer when exposed to reactive nitrogen sources. Gallium nitride and AlN grown on Si(111) are highly defective, and nonradiative carrier recombination channels severely limit the luminescence efficiency for device application.

TABLE 8.3 Physical Parameters of Silicon and GaAs.

Properties	Silicon	GaAs
Lattice constant, nm	0.543102	0.56536
Density, g/cm^3	2.3290	5.32
Melting point,°C	1410	1240
Heat capacity, J/(g·K)	0.70	0.327
Thermal conductivity, W/(cm·K)	1.56	0.45
Thermal diffusivity, cm^2/s	0.86	0.26
Thermal expansion (linear), $\times 10^{-6}\,\mathrm{K}^{-1}$	2.616	6.03
Percent change in lattice, 298 K to ~1311 K	$\Delta a/a_0 = 0.3995$	$\Delta a/a_0 = 0.5876$
Bulk modulus, GPa	97.74	75.0
Young's modulus, GPa	165.6	85.5
Refractive index	3.42	3.66
Dielectric constant	11.8	13.1

GaAs substrate. The crystallographic structure of GaAs is zinc-blende, which is shown in Fig. 8.14. Gallium nitride materials have been grown on zinc-blende GaAs (for more details on the subject, see As 2003), which is the most widely used III-V compound semiconductor as a substrate for zinc-blende GaN epitaxy since it is well developed and large-area substrates are commercially available. The properties of GaAs are listed in Table 8.3. In principle, zinc-blende structures of GaN possess superior electronic properties for device applications, such as a higher mobility, isotropic properties due to the cubic symmetry, and high optical gain. These advantages may not have been realized due to the difficulty in producing low-defect-content material. The growth of zinc-blende GaN requires (001)-oriented substrates having a four-fold symmetry. Several substrates, such as GaAs, Si, 3C-SiC, GaP, and MgO, can be used to grow zinc-blende GaN. The isoelectronic structure (i.e., both GaAs and GaN are III-V compounds), shared element (Ga), potential to convert the surface of GaAs to GaN, and cleavage planes parallel to the epitaxial GaN cleavage planes are the major material advantages of GaAs as a substrate for GaN epitaxy. Technological advantages include a well-established process technology, several readily available substrate orientations of both polar and nonpolar varieties, and low-resistance ohmic contacts. There are several disadvantages to GaAs substrates, including a large lattice constant and thermal expansion coefficient mismatch, a poor thermal conductivity, and perhaps most problematic, low thermal stability.

GaAs is much more readily wet etched than sapphire, making GaN films easier to separate from GaAs than sapphire. Thus, GaAs(111) substrates are considered a better template for creating freestanding thick GaN films for subsequent epitaxy and device fabrication, with the ultimate goal of eliminating the problems associated with heteroepitaxy. Since the decomposition rate of GaAs in NH_3 or an ultrahigh vacuum rapidly increases at temperatures above 700°C, this imposes limits on the epitaxial growth temperature of GaN and hence its maximum growth rate. Even a small amount of GaAs decomposition is detrimental to zinc-blende GaN epitaxy, as surface roughening or faceting enhances the onset of wurtzite growth. Since MBE is capable of depositing epitaxial GaN films at a lower temperature, it has been more commonly employed than MOCVD or HVPE when GaAs is the substrate.

8.3.5.2 Growth rate. The gas impingement flux Φ on the surface of a substrate is a measure of the frequency with which atoms and molecules impinge on, or collide with, the surface. This flux can be defined in one dimension as the number of molecules or atoms striking a surface per unit area and unit time, assuming that the surface is perpendicular to the direction of motion of the atoms or molecules, and can be expressed as (see Ohring 1992)

$$\Phi = \int_0^{\infty} \upsilon_x \, d\mathcal{N}_x \tag{8.35}$$

where

$$d\mathcal{N}_x = \mathcal{N} f(\upsilon_x)\, d\upsilon_x \tag{8.36}$$

The velocity distribution function $f(\upsilon_x)$ is given by the Maxwell-Boltzmann formula as

$$f(\upsilon_x) = \sqrt{\frac{M}{2\pi RT}} e^{-M\upsilon_x^2/(2RT)} \tag{8.37}$$

where M = atomic or molecular weight
R = gas constant
T = temperature
υ_x = velocity of the atoms or molecules

By combining Eqs. (8.35) to (8.37), the flux is obtained as

$$\Phi = \mathcal{N}\sqrt{\frac{M}{2\pi RT}} \int_0^{\infty} \upsilon_x e^{-M\upsilon_x^2/(2RT)} \, d\upsilon_x = \mathcal{N}\sqrt{\frac{M}{2\pi RT}} \frac{RT}{M} = \mathcal{N}\sqrt{\frac{RT}{2\pi M}} \tag{8.38}$$

Substituting the ideal gas equation, $P = \mathcal{N}RT/N_A$, into Eq. (8.38) yields

$$\Phi = \frac{PN_A}{\sqrt{2\pi MRT}} \approx 3.513 \times 10^{22} \frac{P}{\sqrt{MT}} \quad \text{molecules per } (\text{cm}^2 \cdot \text{s}) \tag{8.39}$$

where P is the gas vapor pressure in torr. Consider a gas escaping a container through an opening of area $\mathcal{A}$ into a region where the gas concentration is zero. Thus, the rate at which the molecules leave the container is $\Phi\mathcal{A}$ and the corresponding volume flow per second is $\dot{V} = \Phi\mathcal{A}/\mathcal{N}(\mathrm{cm^3/s})$, which can be rewritten as $\dot{V} = 3.64 \times 10^3 \sqrt{T/M}\mathcal{A}\,(\mathrm{cm^3/s})$.

Another aspect of the gas impingement flux is to calculate the time required for a surface to be coated with one monolayer of gas molecules. The characteristic deposition time $\mathcal{T}$ can be considered as the inverse of the impingement flux. The surface density of most semiconductor crystals is on the order of $\sim 7 \times 10^{14}$ atoms per cm^2. Hence, $\mathcal{T}$ can be obtained as

$$\mathcal{T} = \frac{7 \times 10^{14} \text{ atoms per cm}^2}{\Phi} = \frac{7 \times 10^{14}\sqrt{MT}}{3.513 \times 10^{22} P} \approx 2.0 \times 10^{-8} \frac{\sqrt{MT}}{P} \text{ s} \tag{8.40}$$

The pressure is measured in torr. For a gas with an atomic weight of 30 g/mol, the deposition time at $T = 300$ K in 1 torr of pressure is $\sim 1.9 \times 10^{-6}$ s. On the other hand, if the pressure is 10^{-10} torr, the deposition time is $\sim$5.3 hours.

Let us consider a substrate positioned at a distance l from the aperture of area $\mathcal{A}$ of a source in an MBE growth chamber. The number of molecules $\mathcal{G}$ striking the substrate per unit area per second can be expressed as (see Cho 1983)

$$\mathcal{G} = 3.513 \times 10^{22} \frac{P\mathcal{A}}{\pi l^2 \sqrt{MT}} \quad \text{molecules per } (\mathrm{cm^2 \cdot s}) \tag{8.41}$$

For a Ga source in a BN crucible with an opening of $\mathcal{A} = 10$ cm^2 and $l = 20$ cm, the deposition rate can be calculated as follows: Assume the source temperature is 970°C or 1243 K and the vapor pressure is 1×10^{-4} torr. With an atomic weight of 70 g/mol, the arrival rate of Ga atoms at the substrate can be calculated from Eq. (8.41) as $\sim 9.5 \times 10^{13}$ atoms per (cm^2·s). The average GaAs monolayer thickness is 2.83 Å and contains $\sim 6.3 \times 10^{14}$ Ga atoms per cm^2. Hence, the growth rate is $[(9.5 \times 10^{13})/(6.3 \times 10^{14})] \times 2.83 \times 60 \approx 25.5$ Å/min.

The layer thickness can also be measured using the optical interference method. Consider Fig. 8.25 where the incident light reaches the thin film at an angle. Part of the light will be transmitted through the thin film and the substrate, but a portion of the light will be reflected back and forth between the two interfaces of the thin film. As the photons are bounced between the interfaces, the intensity of the light decreases and an interference pattern is formed due to the difference in the phase of the electromagnetic wave. If the mth-order maximum occurs at wavelength λ_1 and the $(m+1)$th-order maximum occurs at λ_2,

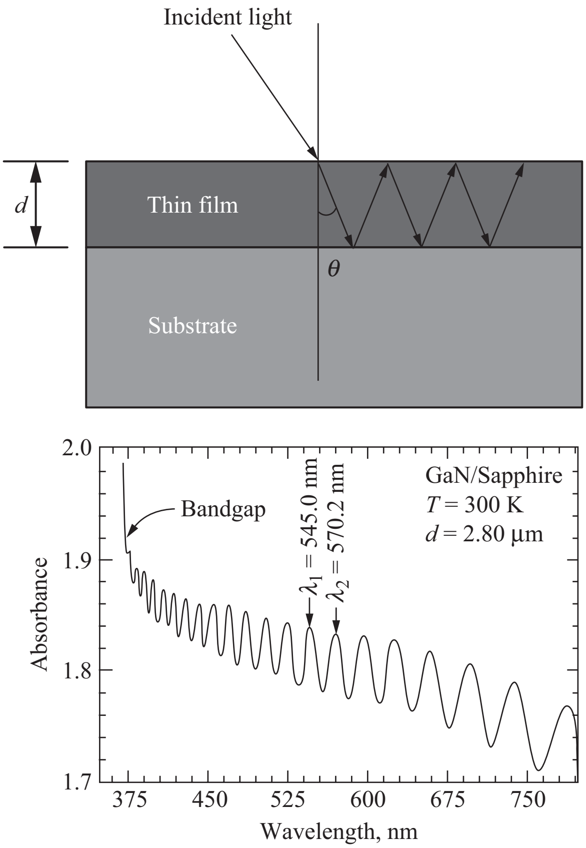


Figure 8.25 An optical interference pattern in a thin layer is illustrated as the electromagnetic wave bounces back and forth between the two layer interfaces.

we have

$$2n_r d \cos\theta = m\lambda_1 = (m+1)\lambda_2 \tag{8.42}$$

where d = thickness of layer
n_r = refractive index of layer material
θ = diffraction angle inside layer

Equation (8.42) can be rewritten in a more general form as

$$d = \frac{N_f}{2n_r\left(1/\lambda_2 - 1/\lambda_1\right)\cos\theta} \tag{8.43}$$

where N_f is the number of fringes between λ_1 and λ_2. The separation between the fringes decreases as the layer thickness is increased. The lower panel in Fig. 8.25 is an actual interference pattern observed in a GaN thin film grown by MOCVD on sapphire. By choosing any two adjacent peaks, the thickness can be calculated using Eq. (8.43). For normal incident light, θ is zero and $\cos\theta$ is 1.

Experimentally, the actual growth rate of thin films is determined by the measured layer thickness divided by the growth time. Finally, the mole fraction x of a ternary compound $\mathrm{A}_x\mathrm{B}_{1-x}\mathrm{C}$ can be determined from the growth rates as

$$x = \frac{\mathcal{G}(\mathrm{A}_x\mathrm{B}_{1-x}C) - \mathcal{G}(\mathrm{BC})}{\mathcal{G}(\mathrm{A}_x\mathrm{B}_{1-x}\mathrm{C})} \tag{8.44}$$

For example, if the growth rates of $\mathrm{Al}_x\mathrm{G}_{1-x}\mathrm{As}$ and GaAs are 36 Å and 25 Å, respectively, then $x = (36 - 25)/36 \approx 0.30$.

8.3.6 Metal-organic chemical vapor deposition growth technique

Metal-organic chemical vapor deposition (MOCVD), also known as metal-organic vapor-phase epitaxy (MOVPE) is becoming one of the most widely used techniques for the growth of various semiconductor films and structures. It is capable of mass production, where several wafers can be used at the same time for a single run. Thus, most industrial applications rely on this technique for mass production. For the growth of III-V semiconductor compounds, this technique relies on the pyrolysis of metal-organic compounds containing group III elements in an atmosphere of hydrides containing group V elements. Both the metal-organic compounds and the hydride gases are introduced in the reactor chamber in which a bulk semiconductor substrate is placed on a heated susceptor. The substrate has a catalyst effect on the decomposition of the gaseous products. The substrate temperature is usually higher than the temperature of the precursor sources. A sketch of the MOCVD reaction is shown in Fig. 8.26, and a picture of a modern MOCVD reactor is shown in Fig. 8.27. The gas handling system includes the metal-organic sources, hydride sources, valves, pumps, and any other instruments needed to control the gas flows. The most common carrier gases in MOCVD reactors are hydrogen, nitrogen, argon, and helium.

The metal-organic compounds are either liquids or powders contained in stainless-steel cylinders known as bubblers. The partial pressure of the source is regulated by controlling the temperature and total pressure inside the bubbler. Mass flow controllers are used to control the mass flow rate of hydride and carrier gases. By sending a controlled flow of carrier gas through the bubbler, the mass flow in a form of dilute vapors of the metal-organic compounds is obtained. The purity of the sources is of paramount importance in the growth of layered structures

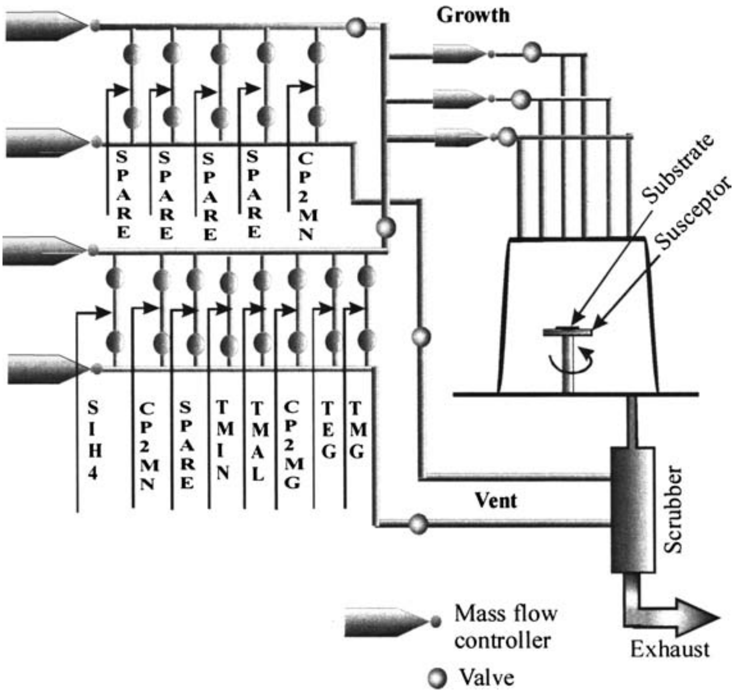


Figure 8.26 A sketch of an MOCVD reactor showing gas inputs with valves and mass flow controllers.

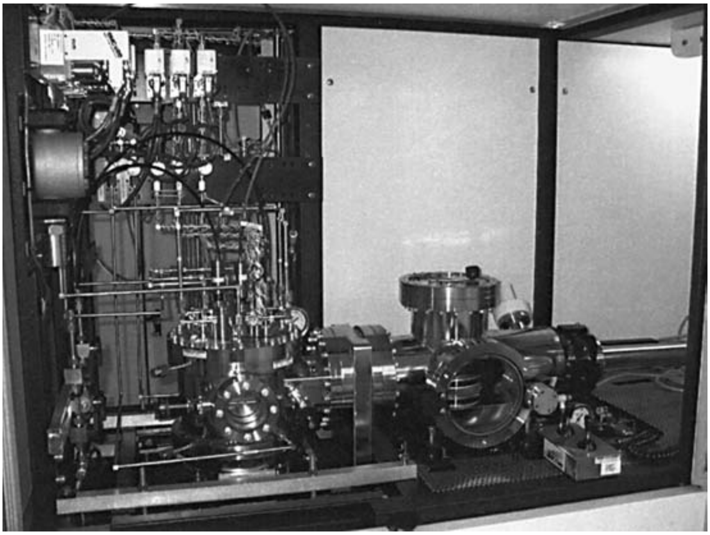
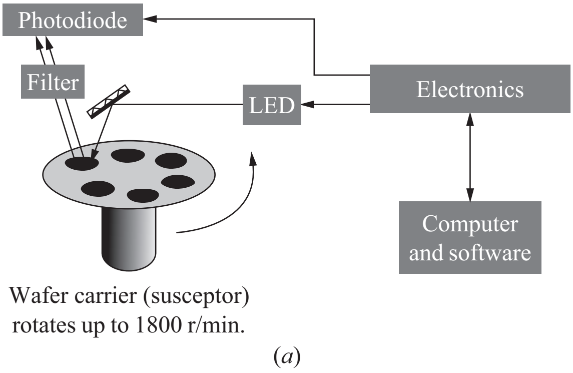


Figure 8.27 A picture of an MOCVD reactor showing the growth and add-on chambers. Analytical and structural characterization tools can be added to this reactor. (*Courtesy of I. T. Ferguson.*)

such as quantum wells and quantum dots. Thus, efforts are devoted to constantly purify source materials.

As seen from Fig. 8.27, the chamber is made of stainless steel containing the susceptor, which can hold one or several substrates. A commercial MOCVD reactor can hold up to sixteen 2-in wafers. The susceptor can be heated by different methods including RF induction heating, radiative heating (lamp), and resistance heating. Knowing and controlling the temperature of the substrate is extremely important for the growth of thin films and quantum structures. One of the recent schemes used to heat the substrate and control its temperature in MOCVD reactors is shown in Fig. 8.28.

In the MBE world, the substrate temperature is controlled by monitoring its bandgap absorption edge as a function of temperature. In the



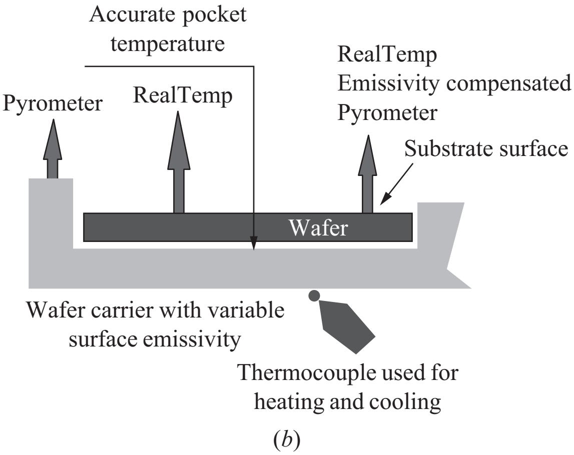

Accurate pocket temperature
Pyrometer
RealTemp
RealTemp Emissivity compensated Pyrometer
Substrate surface
Wafer
Wafer carrier with variable surface emissivity
Thermocouple used for heating and cooling
(b)

Figure 8.28 (*a*) A schematic of an *in-situ* substrate temperature measurement. (*b*) The substrate in the pocket is enlarged for clarity.

case of MOCVD, the temperature is basically controlled by measuring the wafer holder (pocket) temperature as shown in Fig. 8.28. The schematic in Fig. 8.28*a* consists of a light-emitting diode (LED), filter, photodiode, electronics, and software. An enlarged portion of the substrate is shown in Fig. 8.28*b* where the substrate temperature is actually the pocket temperature. Pyrometers are used to measure the pocket temperature. This technique is actually the only method that one can use to measure the substrate temperature very accurately in the case of transparent substrates such as sapphire, which is commonly used for the III-nitride heterostructures and nanostructures.

The bandgap of sapphire is over 7.3 eV ($\sim$170 nm), and the bandgap measurement as a function of temperature is difficult to implement, since light sources in this spectral region are difficult to find and thermal radiance from the wafer surface is below the detection level. As the nitride materials are grown on the substrate, the emissivity from the deposited material can be measured. The basic idea is that the temperature of the substrate can be determined accurately and repeatedly by accurate wafer carrier pocket temperature measurements. The thermocouple in Fig. 8.28*b* is used to give feedback on cooling and heating only, while the substrate temperature is measured by pyrometers as shown in the figure.

Two fundamental processes occur during epitaxial growth. The first, is the thermodynamic process which determines the overall epitaxial growth. The second is the kinetic process which defines the growth rates. The thermodynamic calculations provide information about the solid composition of multicomponent materials when vapor-phase compositions are known. The MOCVD growth is a nonequilibrium process and cannot provide any information about the time required to reach equilibrium. It also cannot provide information about the transition from the initial input gases to the final semiconductor solid.

While the MOCVD growth technique has been used extensively in the epitaxial growth of materials, such as III-nitride compounds, which require growth temperatures, over 1000°C it has its own limitations. In particular, many *in-situ* characterization tools, such as RHEED, scanning tunneling microscopy, and Auger electron microscopy, cannot be used in the MOCVD chamber due to the fact that the MOCVD growth occurs at atmospheric pressure. However, other techniques, such as photoreflectance, ellipsometry, and optical transmissions, have been used recently during the MOCVD growth to monitor growth rates and thin-film uniformity.

The growth kinetics of MOCVD layers depend on a few factors associated with the heterogeneous reaction (gas-substrate interface), such as the transport of reactants through the boundary layer to the substrate, adsorption of reactant at the substrate, atomic and molecular surface

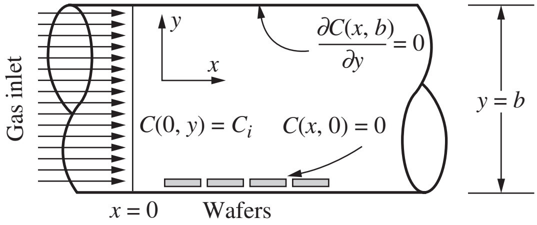


Figure 8.29 Horizontal reactor geometry used to obtain the growth rate.

diffusion, and transport of by-products away from the substrate through the boundary layer. While the microscopic details of these factors are difficult to model, the growth kinetics are often modeled in macroscopic terms and are capable of predicting the growth rate and uniformity of the grown layers. Following the discussion by Ohring (1992), the reactor configuration is shown in Fig. 8.29. By assuming that the gas has a constant velocity component along the axis of the furnace tube and a constant temperature and that the reactor extends a large distance in the z direction, the mass flux $\mathcal{J}$ can be written as

$$\mathcal{J} = C(x, y)\bar{\upsilon} - D\nabla C(x, y) \tag{8.45}$$

where the first term represents a bulk viscous flow where the source of concentration $C(x, y)$ is moving as a whole at a drift velocity $\bar{\upsilon}$. The second term represents the diffusion of individual gas molecules, with a diffusion coefficient D, along the concentration gradients. The flux source at the substrate surface is given by

$$\mathcal{J}(x) = D\left.\frac{\partial C(x, y)}{\partial y}\right|_{y=0} \quad \text{g/(cm}^2\cdot\text{s)} \tag{8.46}$$

where $C(x, y)$ is the solution of the steady-state continuity equation,

$$D\left[\frac{\partial^2 C(x, y)}{\partial x^2} + \frac{\partial^2 C(x, y)}{\partial y^2}\right] - \bar{\upsilon}\frac{\partial C(x, y)}{\partial x} = 0 \tag{8.47}$$

and is given by

$$C(x, y) = \frac{4C_i}{\pi}\sin\left(\frac{\pi y}{2b}\right)e^{-\pi^2 Dx/(4\bar{\upsilon}b^2)} \tag{8.48}$$

where C_i and b are defined in Fig. 8.29. Equation (8.47) is subject to three conditions that are shown in Fig. 8.29: (1) $C = 0$ for $y = 0$ and $x > 0$; (2) $C = C_i$ for $x = 0$ and $b \geq y \geq 0$; and (3) $\partial C/\partial y = 0$ for $y = b$ and $x \geq 0$. For an elemental semiconductor system, such as silicon, the resultant deposition growth rate $\mathcal{G}$ is related to $\mathcal{J}(x)$ according to the following relation:

$$\mathcal{G} = \frac{m_{\mathrm{Si}}}{\rho m_s}\mathcal{J}(x) \quad \mathrm{cm/s} \tag{8.49}$$

where m_{Si} and m_s are the molecular weight of the silicon and the source gas, respectively. Combining Eqs. (8.46), (8.48), and (8.49) yields

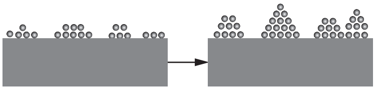
$$\mathcal{G} = \frac{2C_i m_{\mathrm{Si}}}{b\rho m_s} D e^{-\pi^2 Dx/(4\bar{\upsilon} b^2)} \ \mathrm{cm/s} \tag{8.50}$$

This equation predicts an exponential decay of the growth rate as a function of the distance along the reactor length, which is quite reasonable, since the input gases are progressively depleted of reactants. The expression of the growth rate provides design guidelines, although these guidelines are not always simple to implement.

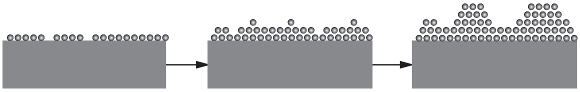
8.4 Fabrication and Growth of Quantum Dots

Semiconductor quantum dots have received significant attention in recent years. While the early techniques of producing quantum dots relied on lithography, such as optical lithography, x-ray lithography, and electron beam lithography, the preferred method today is epitaxial growth. The MBE and MOCVD growth of quantum dots is basically self-assembled growth. It is also possible to epitaxially grow quantum dots on prepatterned substrates. The starting material for the production of quantum dots by lithography techniques is multiple quantum wells. This production technique usually yields regular and uniform arrays of quantum dots where the charge carriers are confined in the three dimensions inside the dots.

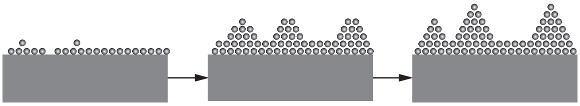
During the early stages of epitaxial thin-film formation, a small number of vapor atoms or molecules condense on the surface of the substrate. This stage is called nucleation. Modern *in-situ* techniques such as scanning tunneling microscopy and RHEED imaging provide useful information between the end of nucleation and the onset of nucleus growth. When the substrate is exposed to the incident vapor (atomic or molecular beams), a uniform distribution of small and highly mobile clusters or islands (3D structures) is observed. In this early growth stage, the prior nuclei incorporate impinging atoms to grow in size. As the growth continues, the islands merge together to form liquid-like materials especially at high substrate temperatures. Coalescence decreases the island density. Further deposition and coalescence causes the islands to connect, forming unfilled channels. Additional deposition fills the channel, and finally thin films are formed (2D structures). The idea of the



(*a*) Island (Volmer–Weber)



(*b*) Layer (Frank–van der Merwe)



(*c*) Stranski–Krastanov

Figure 8.30 The three common growth modes of heteroepitaxy. (*a*) Island or Volmer-Weber mode. (*b*) layer or Frank–van de Merwe mode, and (*c*) layer-island or Stranski-Krastanov mode.

quantum dot growth is to form islands (3D structures) and discontinue the vapor depositions before a thin film (2D structure) is formed.

There are three well-known modes of heteroepitaxial growth, which are illustrated in Fig. 8.30. The first is the island, or Volmer-Weber, mode (Fig. 8.30*a*), which is characterized by the island growth when the smallest stable clusters nucleate on the substrate and grow in three dimensions to form islands (quantum dots). This mode occurs when the atoms or molecules in the deposit are more strongly bound to each other than to the substrate. An example of this growth mode is the deposition of metals on insulators. The second is the layer, or Frank–van de Merwe, mode as demonstrated in Fig. 8.30*b*, which is opposite to the island mode during layer growth. The extension of the smallest stable clusters occurs primarily in two dimensions, resulting in the formation of planar film. The atoms in this mode are more strongly bound to the substrate than to each other. The first complete layer is then covered with a less tightly bound second layer. An example of this mode is the growth of single-crystal semiconductor thin films. The third is the Stranski-Krastanov (SK) mode shown in Fig. 8.30*c* (also known as the layer plus island mode) is an intermediate combination of both the island and layer modes. After forming one or two monolayers, subsequent layer growth becomes unfavorable and islands form. The layer composed of the first two or three monolayers formed at the surface of the substrate, or even the buffer layer, is often called the *wetting layer*.

The transition from 2D to 3D growth is still not well understood. However, any effect that disturbs the monotonic decrease in the binding

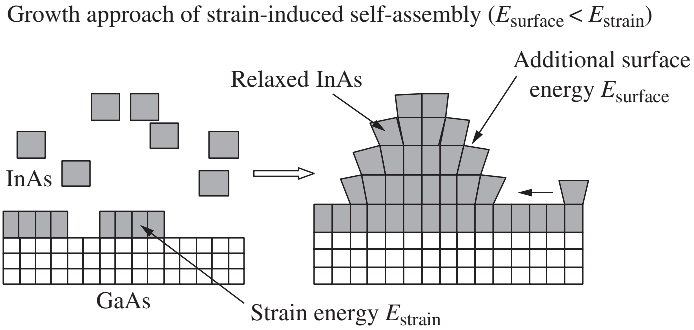

Growth approach of strain-induced self-assembly ($E_{\text{surface}} < E_{\text{strain}}$)
Relaxed InAs
Additional surface energy E_{surface}
InAs
GaAs
Strain energy E_{strain}

Figure 8.31 The formation of an InAs quantum dot (island) on GaAs substrate. (*Courtesy of Greg J. Salamo.*)

energy characteristic of the layer growth mode may cause the 2D-to-3D transformation. As an example, the film-substrate lattice mismatch causes strain energy to accumulate in the growing film. The release of this energy from the deposit-intermediate layer interface may trigger the formation of the islands. This process is illustrated in Fig. 8.31, where the formation of an InAs (lattice constant = 6.0564 Å) quantum dot on GaAs (lattice constant = 5.65321 Å) is illustrated.

8.4.1 Nucleation

There are a few theories dealing with nucleation. One is the capillarity theory, which is a simple qualitative model that describes the film nucleation. It does not provide quantitative information since it lacks the detailed atomistic assumption. However, it provides attractive broad generality, where useful connections between variables, such as substrate temperature, deposition rate, and critical film nucleus size can be deduced. Atomic nucleation processes theory, introduced by Walton et al. (1963), is based on the atomistic approach to nucleation. It treats clusters as macromolecules and applies concepts of statistical mechanics in describing them. Another useful model is based on cluster coalescence and depletion. Brief descriptions of these three models are presented in this section.

8.4.1.1 Capillarity theory. Island formation is assumed when atoms and molecules are impinging on the substrate. The change of the free energy accompanying the formation of islands of mean dimension r can be written as

$$\Delta G = \alpha_3 r^3 \Delta G_V + \alpha_1 r^2 \gamma_{vf} + \alpha_2 r^2 \gamma_{fs} - \alpha_2 r^2 \gamma_{sv} \tag{8.51}$$

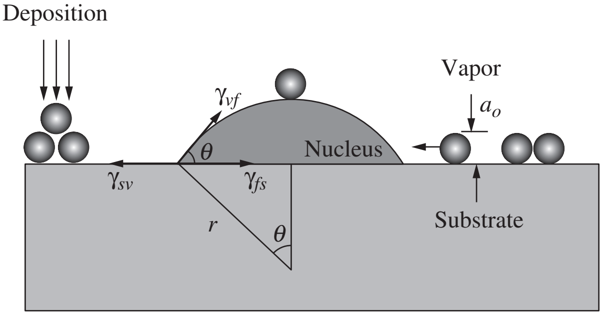


Figure 8.32 An illustration of the basic processes of vapor deposition on the surface of a substrate.

where ΔG_V = chemical free-energy change per unit volume which drives condensation reaction
γ_{vf} = interfacial tension between vapor and film
γ_{fs} = interfacial tension between film and substrate
γ_{sv} = interfacial tension between substrate and vapor

The parameters α_1, α_2, and α_3 are geometric constants given by $2\pi(1 - \cos\theta)$, $\pi \sin^2\theta$, and $\pi(2 - 3\cos\theta + \cos^3\theta)/3$, respectively, for the nucleus shape shown in Fig. 8.32. The curved surface area is $\alpha_1 r^2$ and the volume is $\alpha_3 r^3$. The projected circular area on the substrate is $\alpha_2 r^2$. Young's equation between the interfacial tensions at equilibrium yields

$$\gamma_{sv} = \gamma_{fs} + \gamma_{vf}\cos\theta \tag{8.52}$$

where the angle θ depends on the surface properties of the involved materials. The three growth modes described in Fig. 8.30 can now be distinguished according to the following relationships between the interfacial tensions. For the island (Volmer-Weber) growth mode, $\theta > 0$, which yields

$$\gamma_{sv} < \gamma_{fs} + \gamma_{vf} \tag{8.53}$$

For layer growth (Frank–van der Merwe), the deposit wets the substrate and $\theta = 0$; hence,

$$\gamma_{sv} = \gamma_{fs} + \gamma_{vf} \tag{8.54}$$

An ideal homoepitaxy implies that $\gamma_{fs} = 0$, which yields $\gamma_{sv} = \gamma_{vf}$. The Stranski-Krastanov growth mode fulfills the inequality

$$\gamma_{sv} > \gamma_{fs} + \gamma_{vf} \tag{8.55}$$

where the strain energy per unit area of film growth is larger than the interfacial tension between the vapor and film. This condition triggers the formation of quantum dots on the top of the wetting layer.

Figure 8.31 indicates that when a new interface appears there is an increase in the surface free energy. This implies that the second and third terms in Eq. (8.51) are positive. The loss of the circular substrate-vapor interface under the film nucleus indicates, as shown in Fig. 8.32, a reduction of the system energy. Thus the fourth term in Eq. (8.51) is negative. The energy barrier to a nucleation process ΔG^* can be obtained by first finding the critical radius of the film nucleus. This critical radius r^* is obtained by evaluating $\partial \Delta G/\partial r = 0$. Then, second, substitute r^* back into Eq. (8.51) to obtain

$$\Delta G^* = \frac{4(\alpha_1\gamma_{vf} + \alpha_2\gamma_{fs} - \alpha_2\gamma_{sv})^3}{27\alpha_3^2\Delta G_V^2} \tag{8.56}$$

By substituting the geometrical factors α_1, α_2, and α_3 into Eq. (8.56), the energy barrier, ΔG^* takes the following form:

$$\Delta G^* = \left(\frac{16\pi\gamma_{vf}^3}{3\Delta G_V^2}\right)\left(\frac{2 - 3\cos\theta + \cos^3\theta}{4}\right) \tag{8.57}$$

An island or aggregate smaller in size than r^* disappears by shrinking, thus lowering ΔG in the process. Equation (8.57) indicates that the heterogeneous nucleation depends on the angle θ. The second term in this equation is called the wetting factor. For $\theta = 0$ the wetting factor is zero and for $\theta = \pi$ the wetting factor is unity. When the deposited film wets the substrate, ($\theta = 0$), ΔG^* is zero and there is no barrier to nucleation. On the other hand, when the wetting factor is unity ($\theta = \pi$), ΔG^* is maximum and the growth is identical to that for homogeneous growth.

In the case where the strain energy per unit volume ΔG_s is considered in the analysis, the denominator of Eq. (8.57) is modified to $3(\Delta G_V + \Delta G_s)^2$. The chemical free energy per unit volume ΔG_V is usually a negative quantity, while ΔG_s is a positive quantity. Thus the overall energy barrier to nucleation is increased. If the substrate is initially strained, then release of this stress during nucleation would be indicated by a reduction in ΔG^*.

The nucleation rate $\dot{N}$ is one of the parameters that has to be considered during quantum dot growth. According to the capillarity model, the nucleation rate can be written as

$$\dot{N} = N^*A^*\Phi \quad \text{nuclei per } (\text{cm}^2 \cdot \text{s}) \tag{8.58}$$

where N^* = equilibrium concentration per square centimeter of stable nuclei

A^* = nucleus critical area

Φ = overall impingement flux

The equilibrium number of nuclei of critical size per unit area on the substrate is given by

$$N^* = n_s e^{-\Delta G^*/(k_B T)} \tag{8.59}$$

where n_s = total nucleation site density
k_B = Boltzmann's constant
T = temperature

A certain number of these sites are occupied by monomers (adatoms) whose surface density n_a is the product of the vapor impingement flux and the adatom lifetime τ_s, which is given by

$$n_a = \frac{\tau_s N_A P}{\sqrt{2\pi MRT}} \tag{8.60}$$

and τ_s is given by

$$\tau_s = \frac{1}{\nu} e^{E_{\text{des}}/(k_B T)} \tag{8.61}$$

where E_{des} is the energy required to desorb the adatom back into vapor, and ν is the vibrational frequency of the atom ($\sim 10^{12}\ \text{s}^{-1}$). The area of the nucleus, depicted in Fig. 8.32 can be expressed as

$$A^* = 2\pi r^* a_o \sin\theta \tag{8.62}$$

where a_o and θ are defined in the figure. The overall impingement flux is the product of the jump frequency and n_a, where the jump frequency is defined as the adatom diffuse jumps on the substrate with a frequency given by $\nu \exp(-E_s/k_B T)$ and E_s is the activation energy of the surface diffusion. Thus, Φ can be expressed as

$$\Phi = \frac{\tau_s P N_A \nu e^{-E_s/(k_B T)}}{\sqrt{2\pi MRT}} \quad (\text{cm}^{-2} \cdot \text{s}^{-1}) \tag{8.63}$$

Combine Eqs. (8.58), (8.59), (8.62), and (8.63) to obtain the following expression for the nucleation rate:

$$\dot{N} = 2\pi r^* a_o \sin\theta \frac{P N_A}{\sqrt{2\pi MRT}} n_s e^{(E_{\text{des}} - E_s - \Delta G^*)/(k_B T)} \quad \text{nuclei per } (\text{cm}^2 \cdot \text{s}) \tag{8.64}$$

The nucleation rate is a strong function of the desorption energy, the surface diffusion energy, and the nucleation energy.

8.4.1.2 Atomistic nucleation processes. The Walton et al. (1963) model of nucleation describes the role of individual atoms and the association of small numbers of atoms during the earliest stage of film formation.

The model introduces the critical dissociation energy E_{i^*}, which is defined as the energy required to dissociate a critical cluster containing i atoms into i separate monomers. The critical concentration of clusters per unit area, N_{i^*}, is given by

$$\frac{N_{i^*}}{n_o} = \left|\frac{N_1}{n_o}\right|^{i^*} e^{E_{i^*}/(k_B T)} \tag{8.65}$$

where n_o is the total number of adsorption sites; N_1 is the monomer density, which can be written as

$$N_1 = \Phi\tau_s = \Phi\nu^{-1}e^{E_{\text{des}}/(k_B T)} \tag{8.66}$$

where Φ and τ_s are as defined earlier. The critical monomer supply rate is given by the impingement rate and the area over which the monomers are capable of diffusing before desorbing. This area $\mathcal{L}$ can be defined as

$$\mathcal{L}^2 = 2D_s\tau_s \tag{8.67}$$

where $\mathcal{L}$ is the diffusing mean distance and D_s is the surface diffusion coefficient given by

$$D_s = \frac{1}{2}a_o^2\nu e^{-E_s/(k_B T)} \tag{8.68}$$

Thus, the critical monomer supply rate, $\Phi\mathcal{L}^2$, is given by

$$\Phi\mathcal{L}^2 = \Phi a_o^2\, e^{(E_{\text{des}}-E_s)/(k_B T)} \tag{8.69}$$

The critical nucleation rate $\dot{N}_{i^*}$ can now be obtained by combining Eqs. (8.65) to (8.69):

$$\dot{N}_{i^*} = N_{i^*}\Phi\mathcal{L}^2 = \Phi a_o^2 n_o \left|\frac{\Phi}{\nu n_o}\right|^{i^*} \exp\frac{(i^*+1)E_{\text{des}} - E_s + E_{i^*}}{k_B T} \quad \text{cm}^{-2}\cdot\text{s}^{-1} \tag{8.70}$$

This expression has been used extensively in determining the nucleation rates in many materials including metals and semiconductors. One of the advantages of this model over the capillarity model is that the uncertainties are in i^* and E_{i^*}, while the uncertainties in the capillarity model are in more parameters (ΔG^*, γ, and θ).

Equation (8.70) can be used to predict the thermally activated nucleation rate whose energy depends on the size of the critical nucleus. This means that there are critical temperatures where the nucleus size and orientation undergo change. As an example, let us consider the temperature $T_{1\rightarrow 2}$ at which there is a transition from a one-atom to a two-atom nucleus. This temperature can be obtained by equating the rates $\dot{N}_{i^*=1} = \dot{N}_{i^*=2}$, which leads to

$$T_{1\to 2} = -\frac{E_{\text{des}} + E_{21}}{k_B \ln[\Phi/(\nu n_o)]} \qquad \text{or} \qquad \Phi = \nu n_o \exp -\frac{E_{\text{des}} + E_{21}}{k_B T_{1\to 2}} \tag{8.71}$$

where $E_{21} = E_{i^*=2} - E_{i^*=1}$. According to Walton et al. (1963) when the cluster is composed of one atom, then $E_{i^*=1} = 0$. Thus, E_{21} is simply E_2, E_{31} is E_3, and so on. As an example, let us assume that a deposition growth rate of a compound semiconductor on a substrate was obtained as a function of temperature. An Arrhenius plot of the data yields an activation energy of $(E_{\text{des}} + E_{21}) = 2.1$ eV. The deposition rate was estimated to be 1.0×10^{14} atoms per $(\text{cm}^{-2} \cdot \text{s})$ and $\nu n_o = 7 \times 10^{27}$ atoms per $(\text{cm}^{-2} \cdot \text{s})$. The critical transition temperature can now be calculated from Eq. (8.71) to be ~764 K. The derivations of expressions for transitions from $i^* = 1$ to $i^* = 3$, $i^* = 1$ to $i^* = 4$, and $i^* = 2$ to $i^* = 4$ were left as an exercise.

Kinetic modeling of nucleation have been the subject of many complex mathematical and physical theoretical models in recent years. Detailed discussions of these models are outside the scope of this textbook. However, the general form of the rate equation for clusters with size i is

$$\frac{dN_i}{dt} = K_{i-1}N_1N_{i-1} - K_iN_1N_i \tag{8.72}$$

where N_i are the cluster densities and K_i are the rate constants. The first term on the right-hand side represents the increase in the clusters' size by attaching monomers to smaller $(i-1)$ sized clusters. The second term expresses the decrease in the cluster density when the smaller clusters react with monomers to produce larger $(i+1)$ sized clusters. There are i coupled rate equations to work with, each one of which depends directly on the impingement from the vapor as well as desorption through Eq. (8.66). The addition of diffusion terms $(\partial^2 N_i/\partial x^2)$ to these coupled equations allows one to account for the change in the cluster shape. A more complete nucleation event can be obtained by including the cluster mobility and coalescence.

Robinson and Robins (1974) presented a model for the nucleation and growth kinetics for a one-atom critical nucleus $(i^* = 1)$. They considered two temperature regimes separated by a characteristic temperature T_D given by

$$T_D = \left| \frac{2E_s - 3E_{\text{des}}}{k_B \ln[(C\alpha^2/\beta)(\Phi/\nu n_o)]} \right| \tag{8.73}$$

where C is a number of pair formation sites ($C = 4$ for a square lattice), and α and β are dimensionless constants with typical values of 0.3 and 4, respectively. For temperatures higher than T_D, the reevaporation rate from the surface will control the adatom density and exceed the rate of diffusive capture into growing nuclei. In this regime, the adsorption-desorption equilibrium is rapidly established where $N_1 = \Phi\tau_s$ and incomplete condensation is said to occur. The second regime, where the temperatures are lower than T_D, is characterized by high desorption energy (E_{des}) and insignificant reevaporation. Thus, the condensation is complete, and the monomer capture rate by growing nuclei exceeds the rate at which they are lost due to desorption (evaporation).

The analytical expressions for the time-dependent transient density of stable nuclei, $N(t)$ and the saturation value of $N(t = \infty) = N_s$ are given as follows:

$$N(t) = \begin{cases} N_s \tanh\left(\dfrac{\dot{N}(0)t}{N_s}\right) & \text{for } T > T_D \\ N_s[1 - e^{-3\eta^2\dot{N}(0)t/(N_s^3)}] & \text{for } T > T_D \end{cases} \tag{8.74}$$

$$N_s = \begin{cases} \sqrt{\dfrac{Cn_o}{\beta\nu}}\sqrt{\Phi}\, e^{E_{\text{des}}/(2k_BT)} & \text{for } T > T_D \\ \left(\dfrac{Cn_o^2}{\alpha\beta\nu}\right)^{1/3} \Phi^{1/3} e^{E_s/(3k_BT)} & \text{for } T < T_D \end{cases} \tag{8.75}$$

where

$$\dot{N}(0) = \left.\frac{\partial N(t)}{\partial t}\right|_{t=0} = \frac{C\Phi}{\nu n_o} e^{(2E_{\text{des}} - E_s)/(k_BT)} \tag{8.76}$$

and

$$\eta = \frac{n_o}{\alpha} e^{(E_s - E_{\text{des}})/(k_BT)} \tag{8.77}$$

These equations indicate that $N(t)$ increases with time and reaches saturation at the value N_s. Let us consider the deposition of clusters of a material where $E_s = 0.29$ eV and $E_{\text{des}} = 0.7$ eV. The rest of the parameters are $n_o = 5 \times 10^{15}$ cm^{-2}, $\nu = 1.65 \times 10^{12}$ s^{-1}, $C = 4$, $\alpha = 0.3$, $\beta = 4$, $\Phi = 8.5 \times 10^{14}$ nuclei per (cm$^{-2} \cdot$ s). The characteristic temperature is calculated from Eq. (8.73) to be $T_D = 441.0$ K. The cluster (nucleus) density $N(t)$ is calculated from Eq. (8.74) for the following temperatures: $T = 700$ K (larger than T_D) and $T = 430$ K (smaller than T_D). The results are shown in Fig. 8.33. It is clear from this figure that the number of nuclei (clusters) is larger when the deposition temperature T is smaller than the characteristic temperature T_D. When $T > T_D$, the desorption process (reevaporation rate) dominates over the condensation process. This leads to a lower cluster density as compared to the case when $T < T_D$.

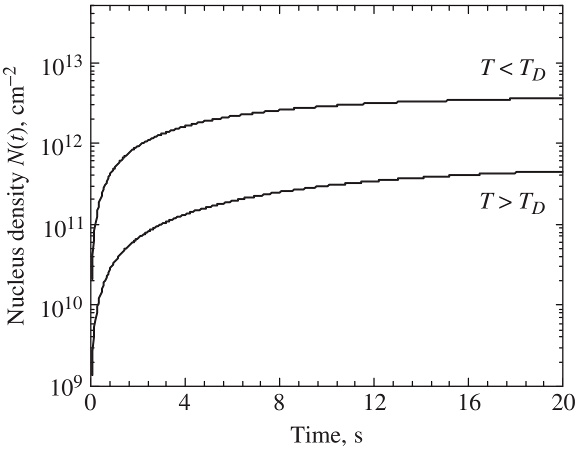


Figure 8.33 The nucleus density $N(t)$ is obtained as a function of deposition time for both $T > T_D$ and $T < T_D$ using Eq. (8.74).

For the case where i^* is any integer, the analysis becomes complicated. However, a review article on the subject was presented by Venables et al. (1984) in which the nucleation parameters for two and three dimensions were summarized. The stable cluster density is given by Venables et al. as

$$N_s = An_o \left| \frac{\Phi}{\nu n_o} \right|^p e^{E/(k_B T)} \tag{8.78}$$

where A is a dimensionless constant that depends on the substrate coverage. The parameters P and E depend on the condensation regimes which are summarized in Table 8.4. The complete and extreme incomplete regimes are similar to those discussed earlier. The extreme incomplete regime occurs when the reevaporation process (desorption) is dominant, and the complete regime occurs when the monomers capture rate exceeds the desorption rate. The initially incomplete regime

TABLE 8.4 Nucleation Parameters *P* and *E* in Eq. (8.78) for Different Regimes

Regime	3D	2D
Extreme incomplete	$P = 2i^*/3$ $E = \frac{2}{3}[E_{i^*} + (i^*+1)E_{\text{des}} - E_s]$	$P = i^*$ $E = E_{i^*} + (i^*+1)E_{\text{des}} - E_s$
Initially incomplete	$P = 2i^*/5$ $E = \frac{2}{5}(E_{i^*} + i^*E_{\text{des}})$	$P = i^*/2$ $E = \frac{1}{2}(E_{i^*} + i^*E_{\text{des}})$
Complete	$P = i^*/(i^* + \frac{5}{2})$ $E = (E_{i^*} + i^*E_s)/(i^* + \frac{5}{2})$	$P = i^*/(i^*+2)$ $E = (E_{i^*} + i^*E_s)/(i^*+2)$

* As reported by Venables et al. (1984).

can be thought of as an intermediate regime which is applicable to the Stranski-Krastanov growth mode.

8.4.1.3 Cluster coalescence, sintering, and migration models. According to the kinetic models, the initial stages of growth are characterized by an increase in the density of the stable clusters as a function of growth time until the density reaches a maximum level before starting to decrease (saturation effect). The process that describes the cluster behavior beyond saturation is called coalescence. This coalescence process is usually characterized by a decrease in the total number of clusters and an increase in the height of surviving clusters. Other features that describe the coalescence process include the following:

Clusters with well-defined crystallographic facets tend to become rounded.

Clusters take a crystallographic shape with time.

The process appears to be liquid-like in nature with clusters merging and changing shapes, where the crystallographic structure of the larger clusters dominate during the merger of smaller clusters.

Clusters are observed to migrate prior to their merger into one another.

There is a size variation when clusters are deposited on the surface of a substrate. The larger clusters tend to grow in size with time at the expense of the smaller ones. This is called the *ripening effect.* The time evolution of cluster distributions was investigated by Vook (1982) using both statistical models involving single-atom process and macroscopic surface diffusion-interface transfer models.

The coalescence process occurs by several methods. One method is called Ostwald ripening where the diffusion of adatoms proceeds from a smaller cluster to a larger cluster until the smaller cluster disappears completely. The diffusion of atoms occurs without the clusters being in direct contact. A second method is called sintering, where the clusters are in contact. A neck forms between the clusters and then thickens as the atoms are transported in the contact region. Cluster migration is another mechanism for coalescence where the clusters on the surface of the substrate actually migrate. Coalescence occurs in this mechanism as a result of collisions between separate clusters (droplets) as they randomly move around.

8.4.2 Fabrications of quantum dots

The production of low-dimensional semiconductor systems, where the charge carriers are confined in two directions (quantum wires) and/or three dimensions (quantum dots), is of interest to those who are involved in the basic understanding of the nature of these systems, as well as those who are interested in producing devices based on the novelty of these quantum structures. Since the early 1980s, many research groups throughout the world have been focused on the production of quantum wires and dots using lithography techniques. All the early efforts focused on processing quantum wells into quantum wires or quantum dots by patterning. While in many cases the patterned techniques have proven to be difficult or expensive to perform, they offer several advantages, such as good control on the lateral shape, size, and arrangement.

Optical lithography techniques using lasers and ultraviolet optics in conjunction with photoresists are used to produce quantum dots with a resolution as high as 100 nm. This technique may not be able to reach dimensions as low as 20 nm or so. However, x rays have the potential to mass-produce nanostructures, since they have a shorter wavelength.

Electron beam lithography is used to produce quantum dots and wires. The electron beam is usually emitted from a high-brightness cathode or a field emission gun. Since electrons are charged, it is very easy to focus them with a magnetic lens system. A resolution of 10 to 20 nm has been achieved by this technique. The electron beam is computer-controlled, and images can be defined on the substrate with the help of a deflection field system. The final resolution of the pattern is limited by the resists due to the finite length of the organic molecules and the grain size. Periodic nanostructures can be produced by using an electron beam interference technique.

In addition to electron beam lithography, focused ion beam lithography has been used in the production of patterned quantum dots. This technique has been used for maskless etching, maskless implantation of dopants, deposition of metallic structures, and patterning of resists. However, the resolution of the focused ion beam lithography is not as high as that of electron beam lithography. There are many other techniques used to pattern quantum dots or grow quantum dots on patterned substrate. For a detailed review of the subject, see Bimberg et al. (1999).

8.4.3 Epitaxial growth of self-assembly quantum dots

The growth of self-assembly quantum dots has been widely made by molecular beam epitaxy and metal-organic chemical vapor deposition techniques. Our knowledge of the quantum dot's structural characteristics has been obtained by tools, such as scanning tunneling microscopy (STM), atomic force microscopy (AFM), transmission electron microscopy (TEM), RHEED images, and x-ray diffraction. Atomic force microscopy has an atomic resolution, and its limiting factor is the size and shape of the microscope tip. Scanning electron microscopy, on the other hand, has the advantage of revealing the morphology of a surface on an atomic scale. It also allows the manipulation of the surface atoms by lining them up and forming shapes and figures.

One of the most important aspects of quantum dot growth is the starting surface of the substrate. The surface construction, surface strain, and crystallographic orientations of the substrate play a major role in the growth of self-assembly quantum dots. An example of the starting substrate surface is shown in Fig. 8.34, where STM images are displayed for GaAs (001) and InP (001) surfaces. As an example, let us consider the formation of InAs quantum dots on a GaAs (001) surface (Fig. 8.34*a*). In this case, the RHEED image exhibits a streaky

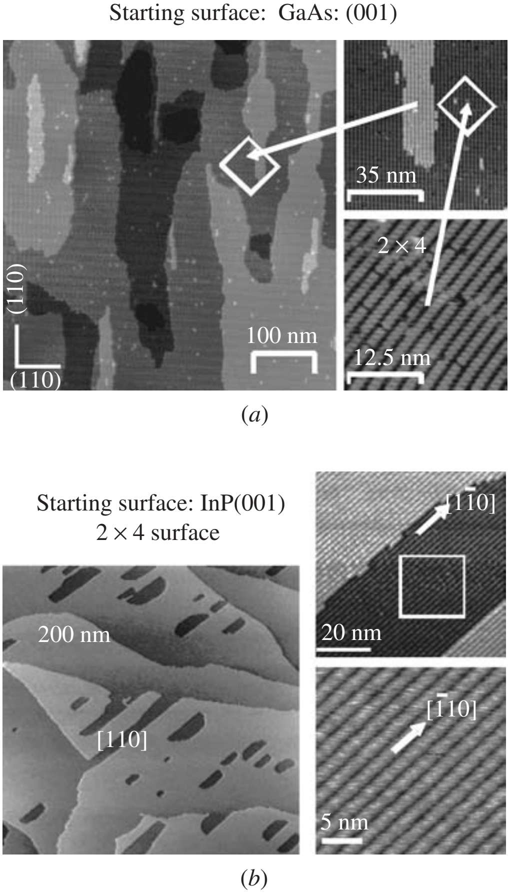


Figure 8.34 Scanning tunneling microscopy images of (*a*) GaAs (001) and (*b*) InP (001) substrate surfaces prior to the quantum dot growth process. (*Courtesy of Greg J. Salamo.*)

pattern before the deposition of InAs atoms indicating a flat surface. The RHEED pattern remains streaky before the 2D to 3D transition, which corresponds to a 1- to 1.5-monolayer deposition of InAs. When more InAs atoms are deposited, the RHEED pattern changes from streaky to spotty, which is an indication of 3D islands being formed. Generally speaking, a monolayer of InAs remains stable for a short growth interruption on the order of 10 s or so. As the growth interruption continues (more than 1000 s), the InAs transitions into 3D islands with anisotropic shapes and sizes.

A simple model was presented by Leonard et al. (1994), where the 3D islands start to develop on top of a 2D wetting layer above a critical thickness Θ_c. The planar density of the MBE-grown InAs 3D islands, ρ_{SAD}, was determined from AFM images versus the amount of InAs deposited, Θ. The relation between ρ_{SAD} and Θ was described as being similar to that of a first-order phase transition

$$\rho_{\mathrm{SAD}} = \rho_o(\Theta - \Theta_c)^{\alpha} \tag{8.79}$$

where α is the exponent and ρ_o is the normalization density. A fit of the results is shown in Fig. 8.35 using Eq. (8.79), which yields $\alpha = 1.76$ and $\rho_o = 2 \times 10^{11}\ \mathrm{cm}^{-2}$.

Since Leonard et al. (1994) reported their finding, several research groups have investigated the 2D-to-3D transition in many quantum dot systems, and their reports provide complex structures and behaviors

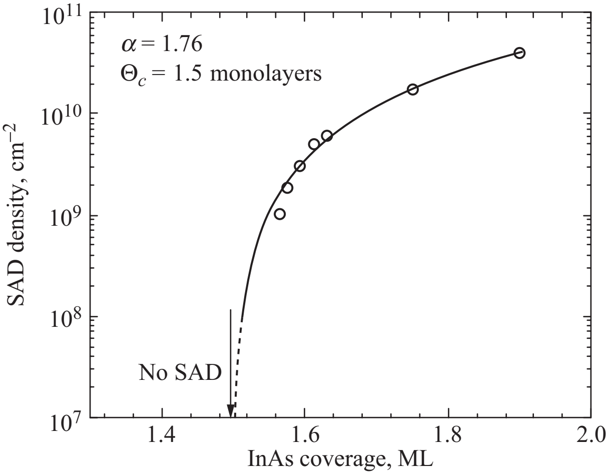
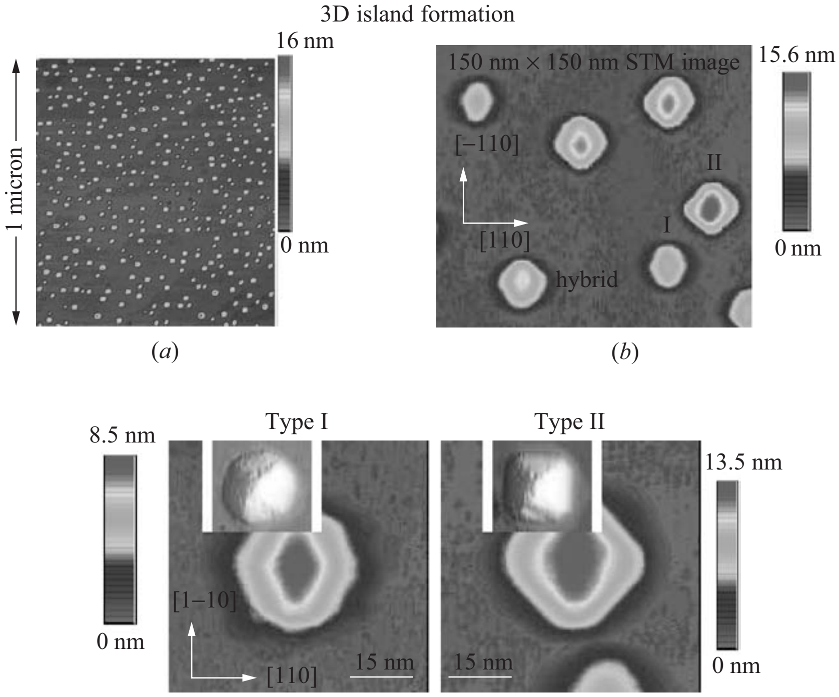


Figure 8.35 Density of self-assembled dots (SAD) versus InAs coverage. Treating these data as a first-order phase transition gives a critical thickness of 1.50 monolayers (ML). (*After Leonard et al. 1994.*)



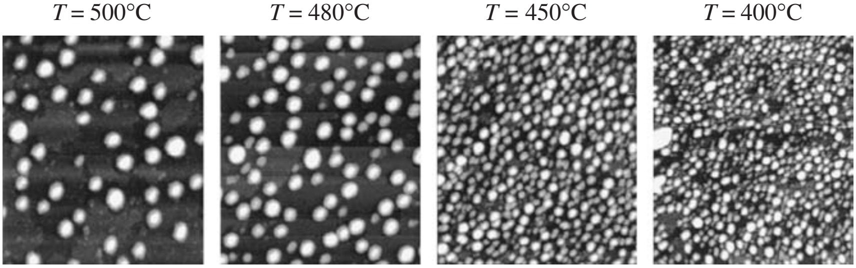
(*a*) (*b*)

(*c*) (*d*)

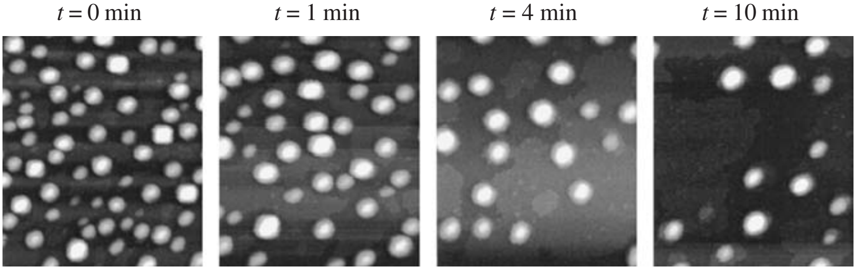
Figure 8.36 Scanning tunneling microscopy images of 2.1-monolayer MBE-grown InAs quantum dots on a GaAs (001) surface. The growth temperature was 495°C. (*Courtesy of Greg J. Salamo.*)

depending on the growth temperature, substrate starting surface, and crystallographic orientations. For example, STM images of 2.1-monolayer InAs self-assembled quantum dots grown on GaAs (001) substrate, with a growth temperature of 495°C, are shown in Fig. 8.36. In image (*a*) one can see many quantum dots that appear to be identical. Upon zooming in to an area of 150 nm^2 [image (*b*)], we note that the quantum dots actually have different shapes, heights, and volumes. It appears that there are two dominant shape types, labeled type I and type II, as shown in image (*b*). It is also noted that there is a third type, which is a hybrid between type I and type II. One can easily distinguish between the two quantum dot shapes by further zooming in on the dots with STM as shown in images (*c*) and (*d*).

The quantum dot size and density depend strongly on the growth temperature and postgrowth annealing. An illustration of these effects on InAs quantum dots is shown in Fig. 8.37, where the growth temperature (substrate temperature) varied from 400 to 500°C at a fixed deposition



(*a*) Effect of substrate temperature T ($t = 20$ s)



(*b*) Effect of annealing time t ($T = 500$°C)

Figure 8.37 Illustration of the growth temperature and postgrowth annealing effects on InAs quantum dots grown by MBE on an AlAs layer, which was grown on GaAs (001) substrate. (*a*) STM images obtained for different samples that were grown at different temperatures. (*b*) STM images that illustrate the effects of postgrowth annealing on the sizes and density of the quantum dots as the annealing time is increased. (*Courtesy of Greg J. Salamo*.)

time of 20 s. The quantum dots were MBE-grown on an ~0.20-μm-thick AlAs (lattice constant = 5.660 Å) layer, which was grown on a GaAs substrate. The InAs growth rate was ~0.30 Å/s, which yielded a total deposited material of ~2 monolayers. It can be seen that the quantum dot density decreases as the growth temperature increases. Accompanying the decrease in density is an increase in the quantum dot size, as shown in Fig. 8.37*a*. The postgrowth annealing has a similar effect. Figure. 8.37*b* shows STM images of the same sample that was grown at 500°C, but annealed as a function of time. In this case, the density of the quantum dots decreases and their size increases as the annealing time progresses from $t = 0$ to $t = 10$ min. The behavior of the quantum dots as a function of growth temperature and postgrowth annealing time can be understood in terms of the kinetics model discussed in Sec. 8.4.1.

The growth of quantum dots is highly influenced by the growth conditions. The growth temperature and postgrowth annealing just discussed are examples of how the density and size of the quantum dots are drastically changed. There are many other parameters that affect the structural and physical properties of quantum dots. Growth modes such as the simultaneous deposition and alternate deposition modes can yield different results. In the simultaneous deposition mode, the constituent atoms are deposited at the same time. This is accomplished by opening the source shutters at the same time. On the other hand, the alternative deposition mode relies on the alternative deposition of the constituents of the quantum dots. Quantum dots grown by the simultaneous deposition mode were found to be affected by the quantum dot arrangement on the surface of the substrate, while the alternative deposition mode produces higher densities of quantum dots grown on vicinal substrates.

Another example of how growth conditions affect quantum dots and their arrangement on the substrate surface is shown in Fig. 8.38. Image (*a*) shows an STM image of InGaAs quantum dots grown on a GaAs (001) surface. This quantum dot system is grown in an MBE chamber. The

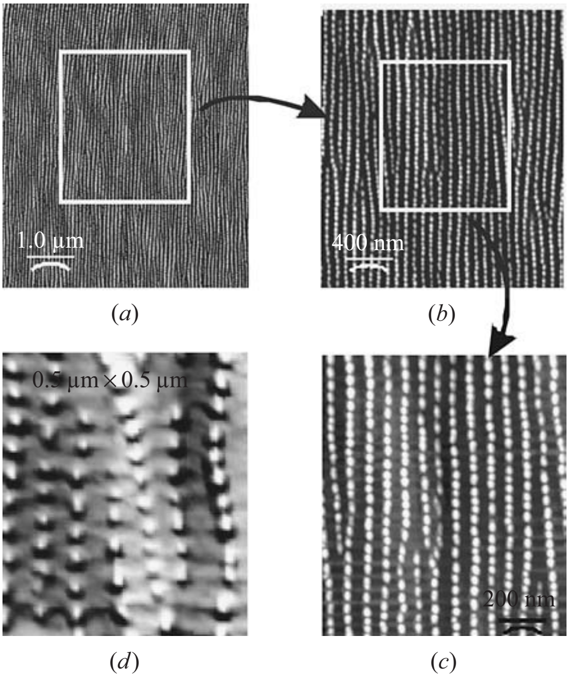


Figure 8.38 Scanning tunneling microscopy images of a InGaAs quantum dot system grown on a GaAs(001) surface. Image (*a*), (*b*), and (*c*) are different magnifications of the surface. The image shown in (*d*) is obtained for InAs quantum dots grown on the InGaAs template shown in image (*a*). (*Courtesy of Greg J. Salamo.*)

quantum dot growth started with a deposition of three monolayers of InGaAs followed by a growth interruption of 10 s. The cycle is repeated six times giving a total of 18 monolayers of deposited material. The quantum dot height is ∼6 monolayers. The remarkable thing about this growth is that the dots formed chains in one direction after the growth of 6 to 12 layers of InGaAs of quantum dots with a GaAs barrier. The length of some of these dot chains is as long as 6 μm. This STM image is further expanded in images (*b*) and (*c*). The surface shown in image (*c*) can used as a template to grow InAs dots. The STM image shown in Fig. 8.38*d* is obtained for InAs quantum dots grown on the above InGaAs template.

Arsenic pressure in the MBE growth of InAs affects the stability of the quantum dots. An increase of the arsenic pressure in the MBE chamber can dramatically affect the morphology of the quantum dots. The size of the quantum dots is reduced, and significant dislocation densities appear in larger InAs clusters. The arsenic pressure can also affect the transitions from 2D to 3D growth. The stoichiometry of (001) surfaces of III-V compound semiconductors that are in equilibrium with the gas is known to depend on the partial pressure of group V elements. Thus, a change in the arsenic pressure leads to a change in the surface reconstructions and to a change in the surface energy of the GaAs (001) surface. At a lower arsenic pressure, indium is known to segregate to the surface, which can be regarded as a quasi-liquid phase that shows no strain-induced renormalization of surface energy. Thus, an increase in the surface area due to the arsenic pressure reduction may lead to a large surface energy making the formation of 3D islands unfavorable.

Substrate orientation plays a vital role in the growth and formation of the quantum dot systems. It affects all parameters governing the quantum dot formation, such as strain energy and surface energy. The strain caused by the difference of the lattice constants of the quantum dots and the substrate is the primary reason for the formation of self-assembled quantum dots. Additionally, this strain difference affects the electrical and optical properties of the quantum dot systems. For homoepitaxial growth the lattice constants of the epitaxial film and the substrate are identical and strain does not exist in the grown film. For heteroepitaxial growth, the film being deposited on a substrate does not necessarily have a lattice constant similar to that of the substrate. Because of this lattice constant difference, one can envision three distinct epitaxial cases, as illustrated in Fig. 8.39. The first case (*a*) is when there is a lattice match or a very small lattice mismatch between the deposited film and the substrate. The strain in this case is almost zero, and the heterojunction growth is identical to the homoepitaxial growth. The second case (*b*) is when there is a lattice mismatch between the film and the substrate. This lattice mismatch causes the strain in the film as

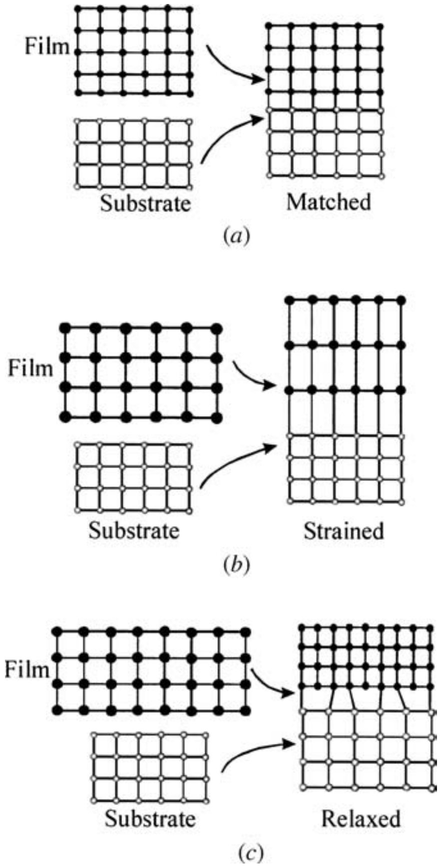


Figure 8.39 A sketch of heterojunction growth showing (*a*) lattice-matched, (*b*) strained, and (*c*) relaxed structures.

shown in Fig. 8.39*b*. A small lattice mismatch is actually beneficial for many heterojunction systems where the mechanical and optoelectronic properties are enhanced by the strain. Strain usually removes the degeneracy, which leads to an improvement in the electronic and optical properties of the film. In fact the presence of this strain in the deposited material is the driving mechanism for the formation of self-assembled quantum dots. The third case when the lattice mismatch is substantially large, the thin film is relaxed by the formation of the dislocations at the interfaces (*c*). The relaxed epitaxy is usually reached during a later film formation stage (thicker films) regardless of the crystal structures or lattice constant difference. For all quantum dot systems, the dots are strained.

The growth of quantum dots on high-index substrates produces quite different and intriguing systems. Several high-index surfaces were considered for the growth of InAs and InGaAs quantum dots on GaAs

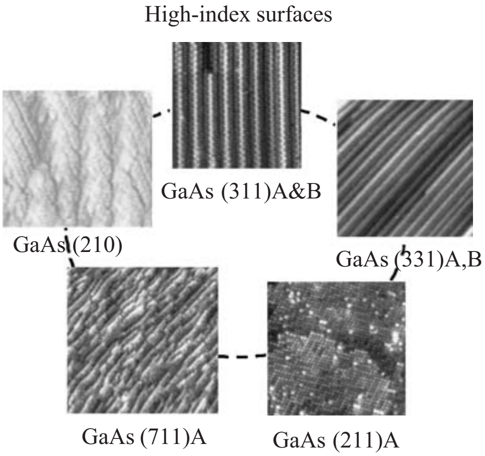
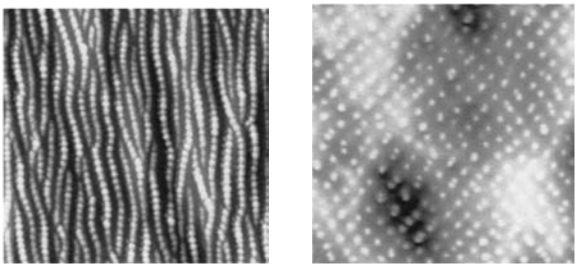


Figure 8.40 Scanning tunneling microscopy images of a few GaAs high-index surfaces. (*Courtesy of Greg J. Salamo.*)

substrates. Scanning tunneling microscopy images of a few of these surfaces are shown in Fig. 8.40 for GaAs substrates. Notice that when the substrate surface is arsenic-rich, the surface is called an A surface and when the surface is gallium-rich, it is called a B surface. Arrays or chains of quantum dots can be grown on high-index surfaces, such as GaAs (331) and (511), as shown in Fig. 8.41. Scanning tunneling microscopy images of InAs or InGaAs quantum dots grown on GaAs (001), GaAs (511)B, and GaAs (311)A surfaces are displayed in this figure. As a comparison, we have shown quantum dot chains that were formed on the (001) surface, where the diffusion of the dot constituents occurs along one direction during growth. For a (511)B surface, there are two directions for the quantum dots to take: (1) the diffusion direction and (2) the steps direction. These two directions are perpendicular to each other. This leads to self-organization in two dimensions as shown in Fig. 8.41*b*, where the quantum dots array is formed in checker-type geometries. Another remarkable property of the quantum dots grown on high index surfaces is that their shape is asymmetrical as shown in Fig. 8.41*c*.

High-index surfaces can be used to engineer quantum wires. An example is shown in Fig. 8.42*a*, where an STM image is obtained for a GaAs (331)B surface. This substrate is used as a template to grow an assembly of InGaAs quantum wires at 450°C, which are shown in the STM image (Fig. 8.42*b*). The formation of the quantum wires on corrugated surfaces, such as GaAs(331), can be understood by inspecting Fig. 8.42*c*. If one assumes that the deposited material does not wet the substrate, then an inhomogeneous growth occurs in which isolated clusters of the deposited material (darker panel) appear on the periodically corrugated substrate. Additional deposition of the material

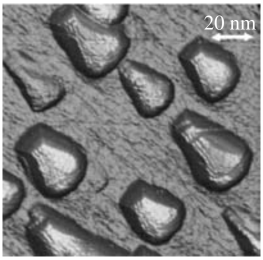


InAs quantum dots grown on a GaAs(001) surface

(*a*)

InGaAs quantum dots grown on a GaAs(511)B surface

(*b*)



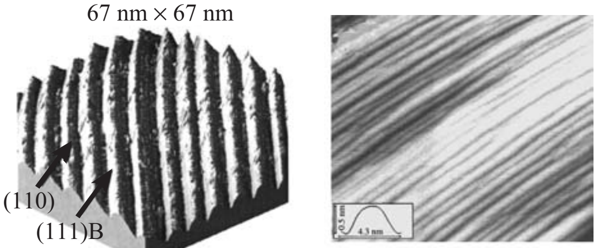
InAs quantum dots grown on a GaAs (311)A surface

(*c*)

Figure 8.41 Scanning tunneling microscopy images of three quantum dot systems grown on a (*a*) GaAs (001) surface, (*b*) GaAs (511)B surface, and (*c*) GaAs (311) surface. Self-organization of the quantum dots into intriguing geometries is apparent. (*Courtesy of Greg J. Salamo.*)

on the substrate produces a periodically modulated thickness as shown in Fig. 8.42*d*.

In addition to planar self-organization of quantum dots, it is possible to vertically stack layers of quantum dots. This vertical stacking is very important for devices, such as detectors and emitters, since vertical stacking increases the filling factor of quantum dots. Depending on the growth temperature and the spacer or barrier, the physical, structural, and morphological properties of the vertically stacked quantum dots can be different than those of the planar self-organized dot systems. An illustration of the vertical stacked quantum dots is shown in Fig. 8.43 for a InGaAs/GaAs system. The images were obtained using tunneling electron microscopy (TEM) for both the planar and cross-sectional (X-TEM) configurations. A remarkable property of the vertically stacked quantum dots is that the dots are self-aligned vertically as is clearly shown in the cross-sectional TEM images. The vertical correlation has



GaAs (331)B: template

(*a*)

$In_{0.2}Ga_{0.8}As$ quantum wires grown on a GaAs (331)B surface

(*b*)

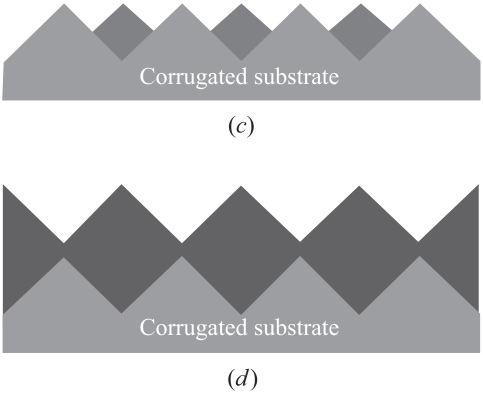


Figure 8.42 (*a*) An STM image of a corrugated GaAs (331)B template used as a substrate to grow self-organized quantum wires. (*b*) An STM image of the self-organized InGaAs quantum wires. (*c*) A presentation of clusters of the deposited material on corrugated substrate. (*d*) A periodically modulated thickness is formed by additional deposition of the material on the substrate. (*The STM images were obtained from Gregg J. Salamo.*)

been observed in many quantum dot systems, and it can be lost if the barrier thickness is too large. For vertical self-alignment to occur, it is very important that the quantum dot layer is successfully grown.

The planar TEM image shown in Fig. 8.43 clearly demonstrates the formation of the InGaAs quantum dot chains grown on the GaAs (001) surface. Inspection of the X-TEM image of the quantum dot stack shows that the shape of the quantum dots is not well defined in the first layer, but in subsequent layers the shape of the quantum dots are well defined and pyramidal in shape.

Additionally, the sizes of the quantum dots become uniform as the multiple layers of quantum dots are stacked. The production of quantum dots that are uniform in size is a very important and necessary aspect of nanotechnology, since the uniformity of the dots affects the device

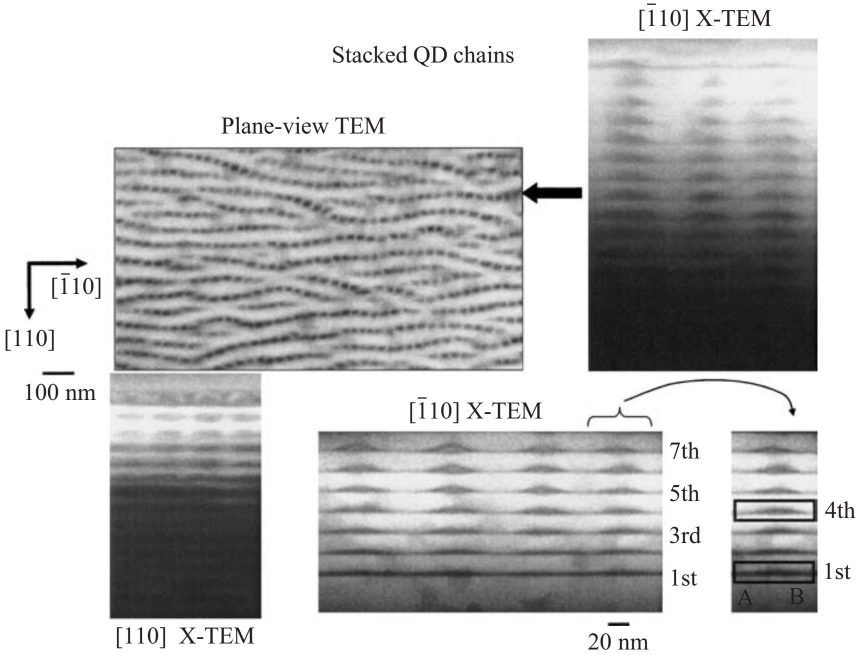


Figure 8.43 Tunneling electron microscopy images of InGaAs/GaAs multiple quantum dot layers. The planar view shows the formation of quantum dot chains, while the cross-sectional images show the vertical correlation of the quantum dots. (*Courtesy of Gregg J. Salamo.*)

performance. The size fluctuation impacts the quantized energy levels in the dots causing inhomogeneous energetic broadening. This broadening should be minimized by producing quantum dots with highly uniform size. For example, infrared detectors fabricated from quantum dots usually require structures that are composed of multiple layers of dots and barriers. This structure is required to obtain a minimum number of dots. The detection of light relies on the confined energy levels inside the quantum dots. The positions of these energy levels are very sensitive to the physical dimensions of the dots.

Summary

The basic principles of single-crystal growth ranging from bulk semiconductor materials to quantum dots was discussed. The introduction of this chapter focused on the importance of bulk materials and the wafering process. The growth of any materials, elemental or compound, depends on a set of thermodynamic conditions. The phase diagram is a critical aspect in understanding the thermodynamic conditions needed to grow single-crystal materials.

In addition to many bulk semiconductor applications, single-crystal wafers are vital to the growth of thin films and the epitaxial growth of all quantum structures including heterojunctions, quantum wells, quantum wires, and quantum dots. Thus, understanding the growth of bulk semiconductors is essential to this chapter. The most widely used bulk growth methods were discussed, including the liquid-encapsulated Czochralski, Bridgman, float-zone, and Lely methods. The segregation coefficient of dopants in bulk material was discussed. The elimination of background impurities or the introduction of a well-controlled dopant in the bulk materials is very important. The dopant distribution in bulk semiconductor crystals usually depends on the radius and the length of the boules.

There are several methods used to grow semiconductor thin films on substrates and wafers. The most common methods used in the growth of thin films are liquid-phase epitaxy, vapor-phase epitaxy, hydride vapor-phase epitaxy, pulsed-laser deposition technique, molecular beam epitaxy, and metal-organic chemical vapor deposition. Each of these techniques has advantages and disadvantages, which were briefly discussed in this chapter.

Epitaxial growth rate, nucleation, and growth kinetics of highly nonequilibrium growth were discussed. Several nucleation models were reported to explain the three-island nucleation. Finally, MBE-grown quantum dots and wires were discussed alongside examples and STM images to illustrate the self-organization along one and two directions. The self-organization in two directions can be achieved using high-index substrate surfaces. The vertical self-organization of quantum dots was also discussed.

Problems

8.1 Derive Eq. (8.6). Calculate ΔG^o for a chemical reaction at $T = 300$ K with $k = 2.3 \times 10^{-9}$.

8.2 Calculate the number of degrees of freedom for the various phases of a single substance as shown in Fig. 8.2. Repeat the calculations for Fig. 8.3.

8.3 Derive Eq. (8.9). Calculate the maximum velocity in inches per hour needed to grow silicon single crystals. Assume the following parameters: $L = 340$ cal/g, $M_v = 2.33$ g/cm^3, $dT/dx = 6$ K/cm, $k_s = 0.21$ W/(cm · K).

8.4 Derive Eq. (8.14) and plot C_s/C_o versus m/m_o for the following values of k_o: 0.01, 0.05, 0.3, 0.5, 0.9, 1, 2, and 3.

8.5 Derive Eq. 8.22.

8.6 Derive Eq. (8.27), and then plot C_S/C_o as a function of x/L for the following values of k_e: 0.01, 0.1, 0.5, 2, and 5.

8.7 Calculate the mean free path of an atom with a diameter of 2.5 Å in an MBE chamber with a vapor pressure of 5×10^{-2} torr. The substrate temperature is kept at 550°C. Compare your result to the typical source-substrate distance of 30 cm.

8.8 Show that the volume of flow of a gas escaping a container through an opening of an area $\mathcal{A}$ into a region where the gas concentration is zero is given by

$$\dot{V} = 3.64 \times 10^3 \sqrt{\frac{T}{M}}\,\mathcal{A} \quad \text{cm}^3/\text{s}$$

where T is the temperature and M is the atomic weight of the gas. Calculate the volume flow rate of a gas at 300 K and an atomic mass of 30 g/mol.

8.9 Show that the gas impingement flux Φ can be written as $\Phi = \frac{1}{4}\mathcal{N}\bar{\upsilon}$, where $\mathcal{N}$ is the concentration of molecules and $\bar{\upsilon}$ is the average velocity of the molecules. Consider the following form for the velocity distribution function: $f(\upsilon) = \frac{4}{\sqrt{\pi}}(\frac{M}{2RT})^{3/2}\upsilon^2 e^{-M\upsilon^2/(2RT)}$.

8.10 Derive Eq. (8.50). A plot of this equation as a function of distance x for an Si layer is shown in Fig. P8.10. Assume the following conditions: $\bar{\upsilon} = 10$ cm/s, $C_i = 3 \times 10^{-6}$ g/cm^3, $b = 1.5$ cm, $\rho = 2.3$ g/cm^3, $m_{\text{Si}}/\text{m}_s = 0.0205$. What would be the value of D needed to generate Fig. P8.10?

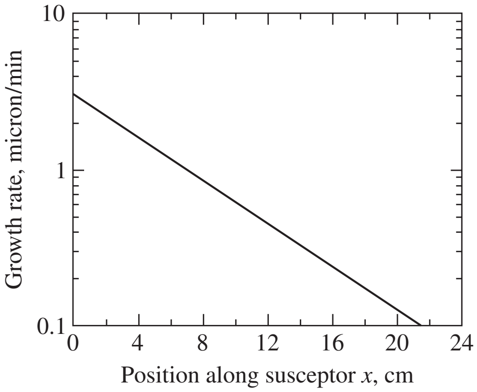


Figure P8.10

8.11 Derive the expression of the barrier energy ΔG^* given by Eq. (8.56). Show that this expression can be reduced to the form shown in Eq. (8.57). Plot ΔG^* in Eq. (8.57) as a function of the angle θ assuming that the first term in the right-hand side of the equation is unity.

8.12 The free-energy change per unit thickness for a cluster of radius r is given by $\Delta G = \pi r^2 \Delta G_V + 2\pi\gamma r + A - B \ln(r)$ where $A - B \ln(r)$ is the energy contributed from dislocations within the cluster. Determine the critical radius r^* and the nucleation barrier energy ΔG^*. Sketch ΔG versus r for the following parameters (assumed to unitless): $\pi \Delta G_V = -10, 2\pi\gamma = 50, A = 10,$ and $B = 0.1$. On your sketch show the locations of r^* and ΔG^*.

8.13 Derive Eq. (8.70). Then obtain expressions for the critical temperatures at which the following transitions occur: one- to three-atom nucleus, one- to four-atom nucleus, and two- to three-atom nucleus.

8.14 Consider the kinetics of a one-atom critical nucleus where $E_s = 0.22$ eV, $E_{\text{des}} = 0.48$ eV, $n_o = 5 \times 10^{15}\text{cm}^{-2}$, $\nu = 0.10 \times 10^{12}\,\text{s}^{-1}$, $C = 4, \alpha = 0.3, \beta = 4, \Phi = 1.0 \times 10^{14}$ nuclei/(cm$^2\cdot$ s). Calculate the characteristic temperature T_D using Eq. (8.73). Plot the cluster (nucleus) density $N(t)$ as a function of time for $T = 750$ K and $T = 530$ K.

8.15 Consider the following relations between the free energy $\Delta G(i^*)$ and the number of atoms, i^*, in 3D deposited clusters on the surface of a substrate. $\Delta G(i^*) = -i^* \Delta\mu + (i^*)^{2/3}\mathcal{X}$, where $\Delta\mu$ is the chemical potential energy and $\mathcal{X} = 3.9$ eV is the surface free energy. Derive expressions for r^* and $\Delta G^*(i^* = r^*)$ from $\Delta G(i^*)$, where r^* is the critical number of atoms in the 3D clusters. Plot $\Delta G(i^*)$ versus i^* for $\Delta\mu = -1$, 0,1, and 2 eV. From the graph, find the values of $\Delta G^*(i^* = r^*)$ and the critical number of atoms, r^*, in the clusters for each value of $\Delta\mu$. When you plot $\Delta G(i^*)$ versus i^*, use the limits of 0 to 100 atoms for i^*.

8.16 Use Fig. 8.36 to determine the density of the quantum dots in units of dots per cm^2. First, obtain the density from both images (a) and (b). Compare your results from both images.

8.17 Assume that the dimensions of the STM images of the InAs quantum dots in Fig. 8.37 are 0.5 μm × 0.5 μm. Estimate the density of the quantum dots, ρ, in cm^{-2} for all images. Plot ρ as a function of the growth temperature and as a function of the annealing time.

8.18 Estimate the density of InGaAs dots grown on a GaAs (001) surface as shown in Fig. 8.38. What is the best approach that you can take to obtain a reasonable estimate?

Chapter 9

Electronic Devices

9.1 Introduction

The electronic and transport properties of electronic devices were reviewed in Chap. 7. These devices are usually microelectronic devices based on homojunction and heterojunction structures. Electronic devices are divided into two classes depending on their operational mode. The first class is called potential-effect devices, where the transport properties are due to carrier injections. Bipolar transistors, which include heterojunction bipolar transistors, and hot electron transistors are examples of this class of devices. Hot electron transistors include both ballistic injection devices and real-space transfer devices. The second class is called field-effect or voltage-controlled devices. Metal-oxide-semiconductor field-effect transistors (MOSFETs), homogeneous field-effect transistors, and heterostructure field-effect devices all belong to the second class. There are several variations of MOSFETs, such as semiconductor on insulator, complementary MOSFETs, *n*-type MOSFETs, and *p*-type MOSFETs. Metal-semiconductor field-effect transistors (MESFETs) and junction field-effect transistors (JFETs) are examples of homogeneous field-effect devices. An example of heterojunction field-effect devices is modulation-doped field-effect transistors (MODFETs), which are also called high electron mobility transistors (HEMTs). This chapter focuses on heterojunction devices, and thus bipolar transistors and MOSFETs will not be discussed, since they are the subject of many textbooks.

The simplest electronic device is the ohmic contact that is based on metal-semiconductor interfaces. For any electronic device to be functional, ohmic contacts are required to allow the charge carrier to move with ease in and out of the device. In other words, the current-voltage (I-V) curve must be linear (nonrectifying) for both positive and negative

Copyright © 2005 by The McGraw-Hill Companies, Inc. Click here for terms of use.

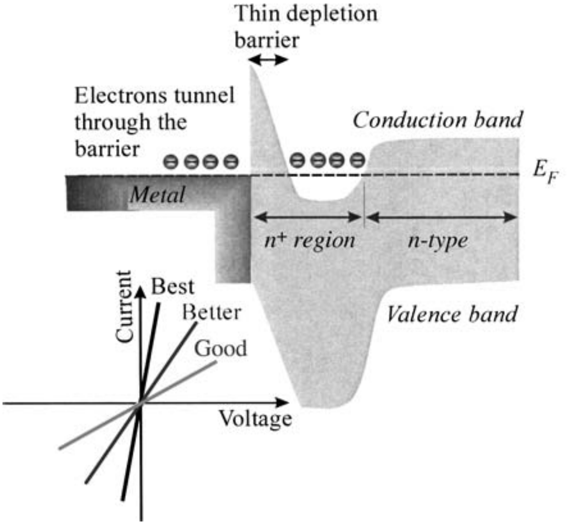


Figure 9.1 A band bending of metal-n^+-n-type semiconductor contact.

voltages with a very large slope. When a metal is brought in contact with a semiconductor, an energy barrier is formed. This barrier is referred to as the work function, which restricts the flow of charge carriers. One method to form an ohmic contact is to choose a metal that has a work function smaller than that of the semiconductor. But since the Fermi energy level lies within the bandgap of nondegenerate semiconductors, the formation of anohmic contact is difficult to obtain. This is due to the fact that the Fermi energy level is pinned at the metal semiconductor interface causing the formation of an energy barrier. Essentially, this is a Schottky contact. A practical solution to this problem is to heavily dope a small thickness of the semiconductor material near the surface before depositing the metal at the surface of the semiconductor. The metal-semiconductor interface is illustrated in Fig. 9.1 for an n-type semiconductor, where a small thickness near the surface of the semiconductor is heavily doped (n^+ region).

The heavily doped portion of the semiconductor reduces the depletion region such that the electron can easily tunnel through the barrier. The I-V curve for the ohmic contact is linear, as shown in the figure (the larger the slope, the better the ohmic contact). In the absence of an applied bias voltage the depletion width in Fig. 9.1 can be easily written as

$$W = \left(\frac{2\epsilon V_{\text{bi}}}{eN_d} \right)^{1/2} \tag{9.1}$$

where V_{bi} = built-in voltage
ϵ = permittivity of semiconductor material = product of dielectric constant ϵ_r and permittivity of space ϵ_o
N_d = donor concentration in semiconductor

The depletion width is inversely proportional to the square root of the dopant concentration. Assuming that each dopant atom contributes an electron to the conduction band, then the electron concentration is N_d. However, in many cases, the carrier concentration is smaller than N_d. For a built-in voltage of 0.3 V and for $N_d = 5 \times 10^{18}$ cm^{-3}, the depletion for GaAs with a refractive index of 3.5 is ~90 Å.

The specific contact resistance R_c, in units of $\Omega \cdot \mathrm{cm}^2$, of an ohmic contact is defined as the product of the contact resistance R and the area of the contact A or

$$R_c = RA = A\left(\frac{\partial I}{\partial V}\right)^{-1} = \left(\frac{\partial J}{\partial V}\right)^{-1} \tag{9.2}$$

where J is the tunneling current density. The tunneling current density is proportional to the tunneling probability of a triangular barrier given by Eq. (4.12). The built-in electric field $\mathcal{E}$ in Eq. (4.12) can be replaced by the built-in voltage divided by the depletion width. The tunneling current density becomes

$$J \propto \exp\left[-\frac{4}{3}\frac{W\sqrt{2m^*}}{\hbar e V_{\mathrm{bi}}}(\Delta E_c - E)^{3/2}\right] \tag{9.3}$$

Substituting Eq. (9.1) into (9.3) and assuming that $\Delta E_c \approx eV_{\mathrm{bi}}$, we have

$$J \propto \exp\left[-\frac{4}{3}\frac{\sqrt{2m^*}}{\hbar e V_{\mathrm{bi}}}\left(\frac{2\epsilon V_{\mathrm{bi}}}{eN_d}\right)^{1/2}(eV_{\mathrm{bi}} - eV)^{3/2}\right] \tag{9.4}$$

Thus,

$$R_c \propto \exp\left(\frac{V_{\mathrm{bi}}}{\sqrt{N_d}}\right) \qquad \text{or} \qquad \ln(R_c) \propto \frac{V_{\mathrm{bi}}}{\sqrt{N_d}} \tag{9.5}$$

where the electron energy E is taken as eV. This equation tells us that the specific contact resistance is minimized by using a metal with a small work function and by choosing N_d as high as possible. Table 9.1 lists the work function of a few metals.

A Schottky diode is formed between a metal and a semiconductor. Its *I*-*V* characteristic is similar to that of a homojunction *pn* diode except it has a much faster response. The forward current of the Schottky diode is due to the majority carrier injection from the semiconductor side. Both ohmic and Schottky contacts are the building blocks of

TABLE 9.1 A List of the Work Function ϕ_m for Several Metals That Are Common in the Metallization Used for Semiconductor Devices, and for the Electron Affinity χ for Four Semiconductors.

Metal	ϕ_m, V	Semiconductor	χ, V
Aluminum (Al)	4.28	AlAs	3.5
Chromium (Cr)	4.50	GaAs	4.07
Cobalt (Co)	5.00	Ge	4.13
Gold (Au)	5.10	Si	4.01
Molybdenum (Mo)	4.60		
Nickel (Ni)	5.15		
Osmium (Os)	5.93		
Palladium (Pd)	5.65		
Platinum (Pt)	5.65		
Rhenium (Re)	4.72		
Silver (Ag)	4.26		
Tantalum (TA)	4.25		
Titanium (Ti)	4.33		
Tungsten (W)	5.55		

metal-semiconductor field-effect devices. For example, the drain and source are ohmic contacts, while the gate is a Schottky contact (rectifying contact). Thus, it is very useful to discuss Schottky diodes in more detail.

9.2 Schottky Diode

The metal-semiconductor rectifying contact is called a Schottky barrier diode. The ideal energy band diagrams for both n-type and p-type Schottky barrier diodes are shown in Fig. 9.2. The electron affinity $e\chi$ is defined as the energy required to remove an electron from the surface of the semiconductor to the vacuum level. The electron affinity is approximately 4.01 and 4.07 eV for silicon and GaAs, respectively. At thermal equilibrium, the Fermi energy levels in both the metal and the semiconductor materials are equal. From this figure, one can find that the barrier height for the n-type semiconductor material, eV_{Bn}, can be written as

$$eV_{Bn} = e(\Phi_m - \chi) \tag{9.6}$$

The barrier height for the p-type semiconductor shown in Fig. 9.2d can be written as

$$eV_{Bp} = E_g - e(\Phi_m - \chi) \tag{9.7}$$

where E_g is the bandgap of the semiconductor. From Eqs. (9.6) and (9.7), one can realize that the sum of the n-type and p-type barrier heights of

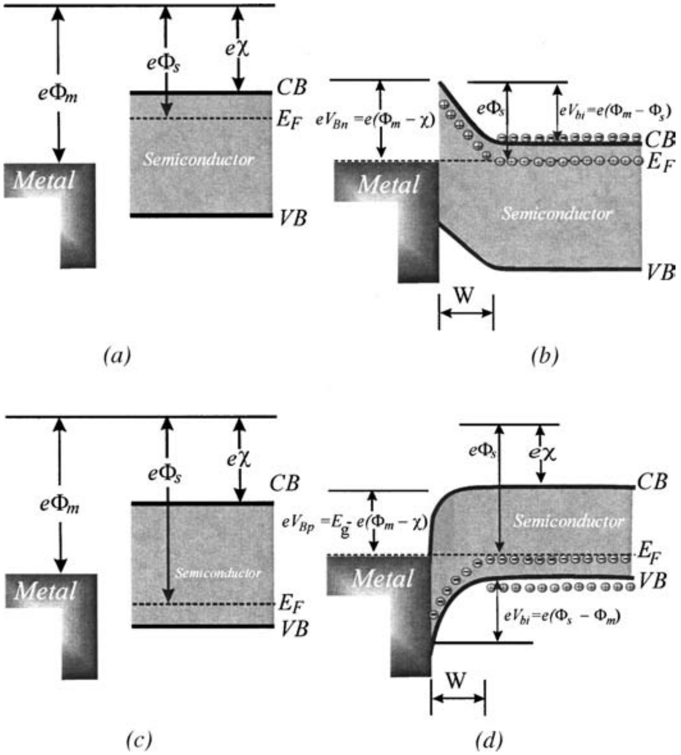


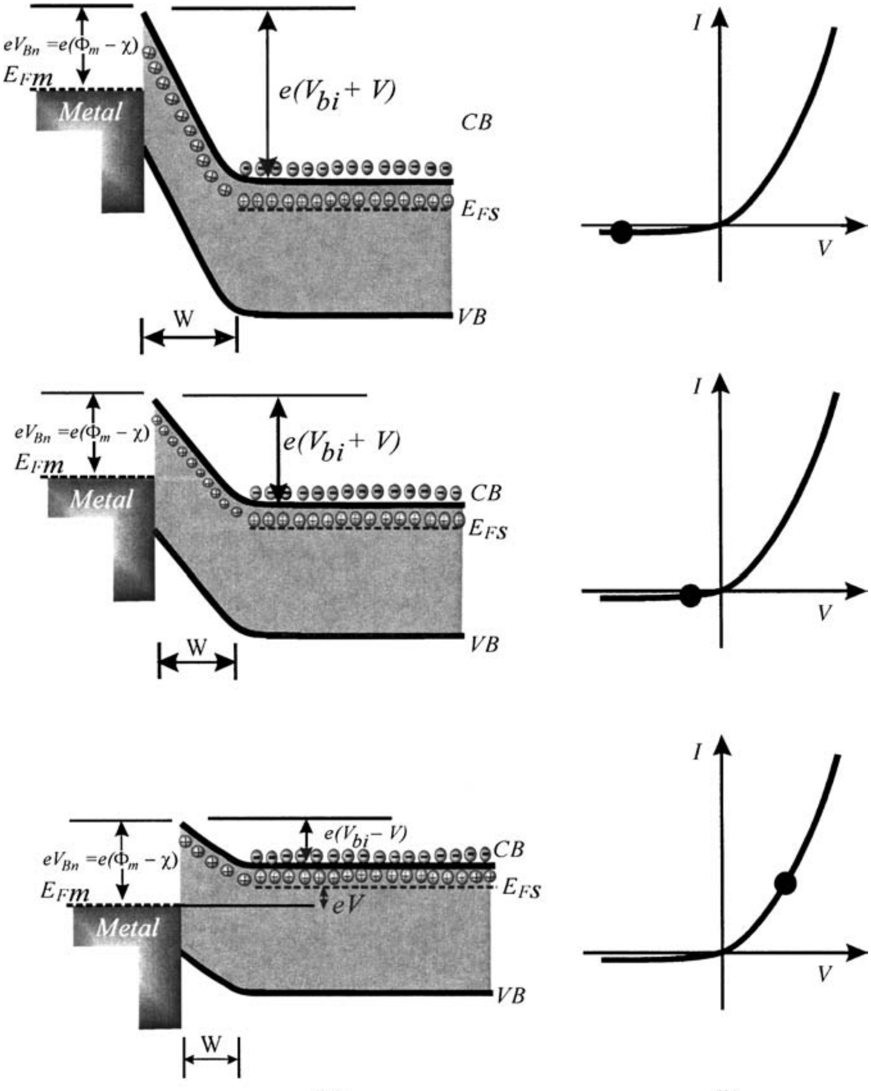
Figure 9.2 A band diagram of a Schottky barrier diode for an n-type semiconductor (a) before and (b) after contact. Similarly, the Schottky barrier diode for a p-type semiconductor is shown (c) before and (d) after contact.

any semiconductor is equal to the bandgap of the semiconductor regardless of the type of metal used for the Schottky barrier. Experimentally, however, the Schottky barrier height is usually smaller than the predicted values obtained from Eqs. (9.6) and (9.7). The discrepancy may be attributed to the presence of surface states.

The quantity V_{bi} shown in Fig. 9.2 is called the built-in potential, which according to Fig. 9.2b can be written for an n-type semiconductor as

$$eV_{\text{bi}} = e(\Phi_m - \Phi_s) = e(V_{Bn} - V_n) \tag{9.8}$$

where Φ_s is the work function of the semiconductor and V_n is the energy difference between the conduction band minimum and the Fermi energy level. Similarly, the built-in potential in p-type semiconductors is the same as for Eq. (9.8) except that V_n is replaced by V_p, where V_p is the



(*a*) (*b*)

Figure 9.3 (*a*) An energy band diagram of a Schottky barrier diode for three different bias voltages. (*b*) The *I*-*V* curves corresponding to the energy band diagrams.

energy difference between the Fermi energy level and the valence band maximum.

The energy band diagram of a Schottky barrier on an *n*-type semiconductor is shown in Fig. 9.3*a* for three different conditions. The corresponding *I*-*V* curve is shown in Fig. 9.3*b*. When a large reverse-biased voltage is applied (top panel in the figure), the barrier height on the semiconductor side increases, making it difficult for the electrons to flow from the semiconductor to the metal. Hence, the reverse current is very small. In the case of forward bias, the barrier height decreases as the forward bias increases, which permits electrons to flow as illustrated in the bottom panel of Fig. 9.3. The dot in the I-V curve represents the operational modes (reverse or forward) of the Schottky barrier diode. Notice that when the bias voltage is applied, the metal-semiconductor junction is no longer in equilibrium and the Fermi energy levels in the semiconductor (E_{FS}) and in the metal (E_{FM}) are necessarily the same.

For an ideal metal-semiconductor junction, the electric field $\mathcal{E}$ in the space charge region (depletion region) can be obtained from Poisson's equation, which relates the electric field to the space charge volume density $\rho(x)$ as follows

$$d\,\mathcal{E} = \frac{\rho(x)}{\epsilon_s}dx \tag{9.9}$$

where ϵ_s is the permittivity of the semiconductor material. The electric field is zero at $x = W$, which yields after integrating Eq. (9.9) the following relation:

$$\mathcal{E}(x) = -\frac{eN_d}{\epsilon_s}(W - x) \tag{9.10}$$

where W = depletion width
N_d = electron concentration in depletion region
$\rho(x) = eN_d$

For a uniformly doped semiconductor, $\mathcal{E}$ is linear as a function of distance. The depletion width can be obtained as

$$W = \left[\frac{2\epsilon_s(V_{\text{bi}} - V)}{eN_d}\right]^{1/2} \tag{9.11}$$

and the space charge density in the semiconductor region can be expressed as

$$\rho_{\text{SC}}(x) = eN_d W = [\epsilon_s eN_d(V_{\text{bi}} - V)]^{1/2} \quad \text{C/cm}^2 \tag{9.12}$$

Notice that the voltage V is positive for forward bias and negative for reverse bias. Differentiate Eq. (9.12) with respect to V to obtain the capacitance of the depletion region as

$$C = \left|\frac{\partial \rho_{\text{SC}}(x)}{\partial V}\right| = \left[\frac{\epsilon_s eN_d}{2(V_{\text{bi}} - V)}\right]^{1/2} = \frac{\epsilon_s}{W} \quad \text{F/cm}^2 \tag{9.13}$$

This equation is very useful in calculating the carrier concentration from the capacitance measurement as a function of bias voltage.

Recently, C-V profiling has been used to calculate the carrier concentration as a function of the depth in heterojunctions and multiple quantum wells.

The Schottky barrier height is usually much larger than the thermal energy k_BT. This barrier is usually obtained when the doping level in the semiconductor is smaller than the density of states in the conduction band or valence band. As mentioned earlier, the transport in a Schottky diode is due to the majority carriers, and the thermionic mechanism is dominant at room temperature. For a Schottky diode with an n-type semiconductor under a forward-biased voltage V_F, the electron concentration emitted by the thermionic mechanism can be expressed as

$$N_{\mathrm{th}} = N_c \exp\left[-\frac{e(\Phi_{Bn} - V_F)}{k_BT}\right] \tag{9.14}$$

Thus, the current density from the semiconductor to metal can be expressed as

$$J_{s\to m} = C_1N_c \exp\left(-\frac{e(\Phi_{Bn} - V_F)}{k_BT}\right) \tag{9.15}$$

Similarly, the current density from the metal to semiconductor is given as follows:

$$J_{m\to s} = C_1N_c \exp\left(-\frac{e\Phi_{Bn}}{k_BT}\right) \tag{9.16}$$

Notice that the forward-biased voltage is not included in Eq. (9.16), since the barrier height seen from the metal side is not affected by bias voltage. Under the influence of bias voltage, the current densities shown in Eqs. (9.15) and (9.16) are no longer the same. The net current density is the difference between the expressions in the two equations:

$$\begin{aligned} J &= J_{s\to m} - J_{m\to s} = C_1N_c \exp\left(-\frac{e(\Phi_{Bn} - V_F)}{k_BT}\right) \\ &\quad - C_1N_c \exp\left(-\frac{e\Phi_{Bn}}{k_BT}\right) \\ &= C_1N_c \exp\left(-\frac{e\Phi_{Bn}}{k_BT}\right)\left\{\exp\left(\frac{eV_F}{k_BT}\right) - 1\right\} \end{aligned} \tag{9.17}$$

where C_1 is a constant. The product C_1N_c is found to be equal to A^*T^2, where T is the temperature and A^* is called the *effective Richardson constant* taken in units of A/(K$^2 \cdot$ cm^2) and is expressed as

$$A^* = \frac{4\pi m^* k_B^2}{h^3} \tag{9.18}$$

where m^* is the electron or hole effective mass and h is Planck's constant. The values of A^* are 110 and 32 for n-type and p-type silicon, respectively, and 8 and 74 for n-type and p-type GaAs, respectively. A more familiar form of the current density is

$$J = J_s\left\{\exp\left(\frac{eV_F}{k_BT}\right) - 1\right\} \tag{9.19}$$

where

$$J_s = A^*T^2\exp\left(-\frac{e\Phi_{Bn}}{k_BT}\right) \tag{9.20}$$

The parameter J_s is called the saturation current density.

9.3 Metal-Semiconductor Field-Effect Transistors

Recent advances in materials growth using molecular beam epitaxy and metal-organic chemical vapor-phase deposition techniques have enabled the growth of high-purity thin films. With these growth techniques one can precisely control the doping level in semiconductor materials, which allows for the growth of semiconductor structures for device processing and fabrication. The metal-semiconductor field-effect transistor (MESFET) was first proposed by Mead (1966). These transistors are usually associated with GaAs. A cross section of an n-channel MESFET is shown in Fig. 9.4. The drain and source are ohmic contacts, while the gate metal is a Schottky contact. Ion plantation is usually used to produce the n^+ regions under the drain and source. The active region is an epitaxially grown n-type thin film, and the substrate

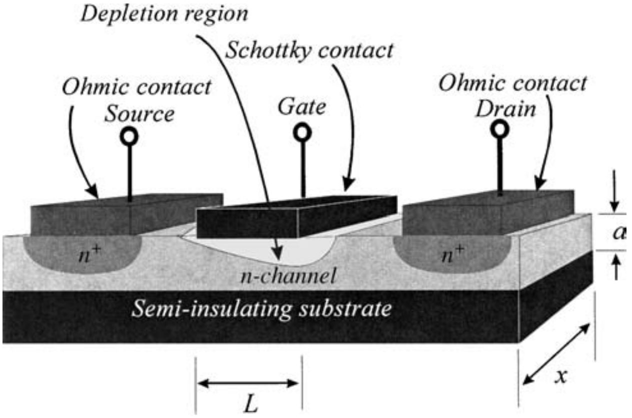


Figure 9.4 A sketch of a cross section of an n-channel MESFET.

is semi-insulating. The semi-insulating GaAs substrates are usually grown by LEC under high-pressure arsenic conditions or by doping the materials with chromium. The substrate resistivity is usually very high ($\rho > 10^7\ \Omega \cdot \text{cm}$), preventing the electric current from flowing through the substrate.

From Fig. 9.4 one can derive the resistance of the n-channel MESFET according to the relation

$$R = \rho \frac{L}{A} \tag{9.21}$$

where L is the length of the gate and A is the depletion region area given by $x(a - W)$, where x and a are shown in Fig. 9.4, and W is the width of the depletion region under the Schottky contact. The resistivity ρ is given by $(e\mu_n N_d)^{-1}$ where μ_n is the electron mobility in the n-channel MESFET and N_d is the electron concentration. Thus, the magnitude of the drain current I_D is given by V_D/R in the nonsaturation region.

A reverse-biased gate-source voltage induces a space charge region (depletion region) under the metal gate, which modulates the channel conductance. The operation of the MESFET is depicted in Fig. 9.5 in which only the gate section is sketched. The source is usually grounded, while the gate and drain voltages are measured with respect to the source. Under normal operation, the gate voltage V_G is either zero or reverse-biased ($V_G \leq 0$) and the drain voltage V_D is either zero or forward-biased ($V_D \geq 0$). For $V_G = 0$ and small V_D, a small drain current I_D flows in the channel, as shown in Fig. 9.5*a*. The dot on the curve represents the I_D-V_D coordinates associated with the sketch on the left-hand side. Notice that the channel is reverse-biased and the voltage is zero at the source and increases to V_D at the drain terminal. As the drain voltage increases, the width of the space charge region (depletion region) increases. Thus, the cross-sectional area for the current flow decreases causing a reduction of the current flow rate.

When $V_G = 0$ and the drain voltage is large enough such that the depletion region touches the substrate, as shown in Fig. 9.5*b*, the depletion region width is equal to the thickness of the n-type epitaxial layer at the drain side, i.e., $W = a$. The corresponding value of the drain voltage is called the *saturation voltage*, which can be obtained from Eq. (9.11) such as

$$V_D^{\text{sat}} = a^2 \frac{eN_d}{2\epsilon_s} - V_{\text{bi}} \tag{9.22}$$

where V_{bi} is the built-in voltage. Notice V_D^{sat} is larger than V_{bi}. The built-in voltage for an ideal MESFET is shown in Fig. 9.6 in which the energy band diagram is sketched. The condition at which the depletion region touches the substrate at the drain side is called *pinch-off*,

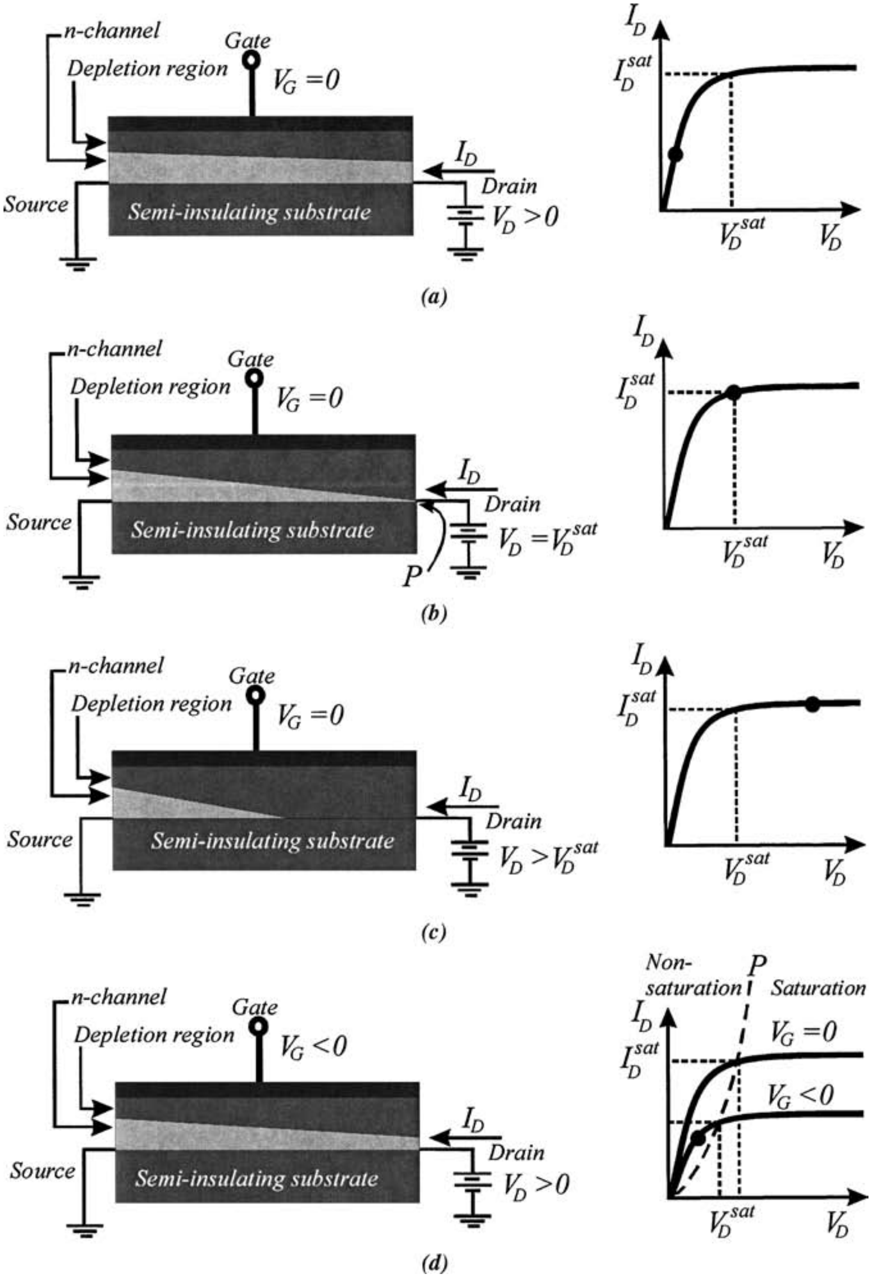


Figure 9.5 A schematic presentation of the gate cross-sectional area of a MESFET and the output I_D–V_D characteristic curves. (*a*) $V_G = 0, 0 < V < V_D^{\mathrm{sat}}$. (*b*) $V_G = 0, V = V_D^{\mathrm{sat}}$. (*c*) $V_G = 0, V > V_D^{\mathrm{sat}}$. (*d*) $V_G < 0, 0 < V < V_D^{\mathrm{sat}}$.

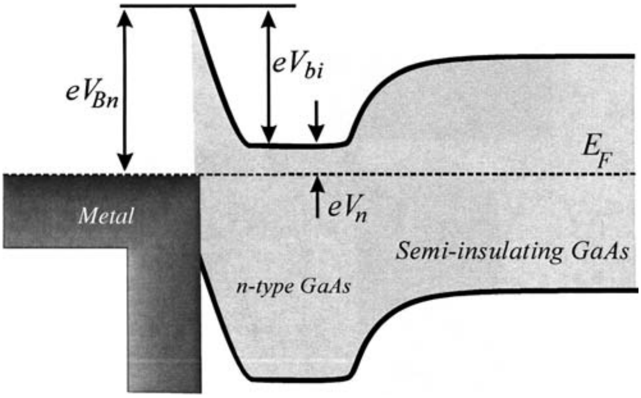


Figure 9.6 A sketch of a bandgap diagram of an ideal n-channel MESFET showing the barrier height, built-in voltage, and degenerate voltage V_n.

which is P in Fig. 9.5b. The current at point P is called the saturation current I_D^{sat}, and the electrons tunnel through the depletion region from the source to the drain. The voltage at point P remains the same, and the current flowing in the channel remains independent of V_D. This is because the potential drop from the source to point P does not change. Figure 9.5c illustrates this process where the current is constant for $V > V_D^{\text{sat}}$. Again, the processes described in Fig. 9.5 are for an ideal MESFET. In real devices, there are many parasitic effects that cause the drain current to be dependent on the drain voltage for $V_D > V_D^{\text{sat}}$.

Figure 9.5d illustrates the condition when the gate is reverse-biased, i.e., $V_G < 0$ and $0 < V < V_D^{\text{sat}}$. The depletion region is further increased as compared to the conditions shown in Fig. 9.1a. Thus, the drain current is further reduced as shown in the output characteristic curve. When the drain voltage is equal to V_D^{sat}, where the depletion region just touches the substrate, we have

$$V_D^{\text{sat}} = a^2 \frac{eN_d}{2\epsilon_s} - V_{\text{bi}} - |V_G| \tag{9.23}$$

where V_G values are taken as the absolute values. This equation tells us that V_D^{sat} is reduced by applying a reverse-biased voltage to the gate.

The current-voltage relation for a MESFET has been derived in many textbooks (see, for example, Sze 2002, Neamen 2003, and Mitin et al. 1999). The analyses by different authors are identical. The geometry of the depletion region under the gate contact is shown in Fig. 9.7. The voltage drop across an elemental section dy of the channel can be

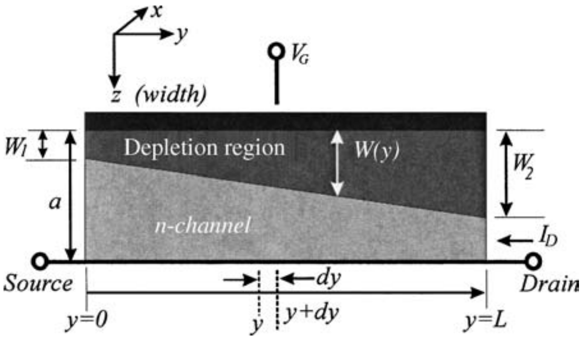


Figure 9.7 A sketch of the n-channel MESFET region under the gate contact.

written as

$$dV = I_D dR = I_D \rho \frac{dy}{A} = \frac{I_D dy}{e\mu_n N_d x[a - W(y)]} \tag{9.24}$$

where x is the length of the gate. By using Eq. (9.23), the depletion region width can be expressed as

$$W^2(y) = \frac{2\epsilon_s [V(y) + V_{\text{bi}} + |V_G|]}{eN_d} \tag{9.25}$$

Take the first derivative of the depletion region width with respect to the voltage and substitute the results into Eq. (9.24) to obtain the following expression for the drain current:

$$I_D dy = e\mu_n N_d x(a - W)\frac{eN_d}{\epsilon_s} W\, dW \tag{9.26}$$

Integrating this equation over the limits of $y = 0$ to $y = L$ and $W = W_1$ to $W = W_2$ yields

$$I_D = \frac{e^2 \mu_n N_d^2 x}{2\epsilon_s L}\left[a\left(W_2^2 - W_1^2\right) - \frac{2}{3}\left(W_2^3 - W_1^3\right)\right] \tag{9.27}$$

This equation can be rewritten as

$$I_D = I_p \left[\frac{V_D}{V_p} - \frac{2}{3}\left(\frac{V_D + |V_G| + V_{\text{bi}}}{V_p}\right)^{3/2} + \frac{2}{3}\left(\frac{|V_G| + V_{\text{bi}}}{V_p}\right)^{3/2}\right] \tag{9.28}$$

where

$$I_p = \frac{e^2 \mu_n N_d^2 x a^3}{2\epsilon_s L} \qquad \text{and} \qquad V_p = \frac{eN_d a^2}{2\epsilon_s} \tag{9.29}$$

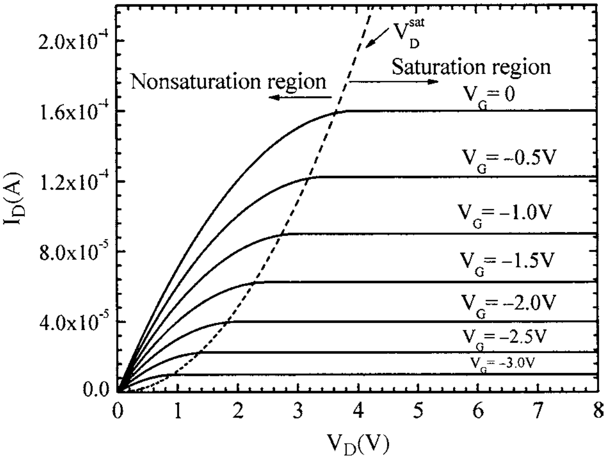


Figure 9.8 Drain current as a function of the drain voltage for an n-channel MESFET with different values of the gate voltage obtained using PSpice. The saturation and nonsaturation regions are indicated.

The quantities I_p and V_p are the pinch-off current and voltage, respectively. The saturation current can be obtained by setting $V_p = V_D + |V_G| + V_{\mathrm{bi}}$. Thus, the saturation drain voltage is given as

$$V_D^{\mathrm{sat}} = V_p - |V_G| - V_{\mathrm{bi}} \tag{9.30}$$

Substitute Eq. (9.30) into (9.28) to obtain the saturation current as

$$I_D^{\mathrm{sat}} = I_p \left[\frac{1}{3} - \left(\frac{|V_G| + V_{\mathrm{bi}}}{V_p} \right) + \frac{2}{3} \left(\frac{|V_G| + V_{\mathrm{bi}}}{V_p} \right)^{3/2} \right] \tag{9.31}$$

The I_D-V_D curves of a MESFET can be obtained by plotting Eq. (9.28) or by using PSpice. The results are shown in Fig. 9.8 for different gate voltages ranging between –3.0 and 0.0 V. The saturation voltage is shown as the dashed curve, which can be considered as the boundary between the saturation and nonsaturation regions of the transistor. For small V_D, the transistor is in the nonsaturation mode or linear mode. In this case, the channel behaves like an ohmic contact. When V_D is too large as compared to the saturation drain voltage, the current may substantially increase causing an avalanche breakdown.

The transconductance g_m of the transistor is a parameter that relates the output current to the input voltage and is a measure of the transistor gain. It is defined as

$$g_m = \left. \frac{\partial I_D}{\partial V_G} \right|_{V_D = \mathrm{constant}} \tag{9.32}$$

For small V_D, the first derivative of the drain current given by Eq. (9.28) with respect to V_G is

$$g_m = \frac{\partial I_D}{\partial V_G} \approx \frac{1}{2}\frac{I_p}{V_p}\frac{V_D}{\sqrt{V_p(|V_G| + V_{\text{bi}})}} \tag{9.33}$$

It is possible that the depletion region is large enough such that the device is normally off for $V_G = 0$. In this case, a forward-biased gate ($V_G > 0$) is needed to turn on the device. The voltage needed to turn on the MESFET is called the threshold voltage V_T, which is given by $V_T = V_{\text{bi}} - V_p$ or $V_{\text{bi}} = V_T + V_p$. Substituting this expression into Eq. (9.31) yields

$$I_D^{\text{sat}} = I_p\left[\frac{1}{3} - \left(1 - \frac{V_G - V_T}{V_p}\right) + \frac{2}{3}\left(1 - \frac{V_G - V_T}{V_p}\right)^{3/2}\right] \tag{9.34}$$

Notice that $|V_G|$ in Eq. (9.31) is replaced by $-V_G$. For $(V_G - V_T) \ll V_p$, the transconductance, when the device is in saturation mode, can be obtained as

$$g_m \approx \frac{x\mu_n\epsilon_s}{aL}(V_G - V_T) \tag{9.35}$$

The simplistic and ideal MESFET model presented describes the main characteristics of the transistor, such as the current-voltage relation and the transconductance. These FET devices are usually used as amplifiers and converters. They are also used in microwave and RF systems. For ac operation, the cutoff frequency f_T is a very important figure of merit. It is defined as the frequency at which the MESFET can no longer amplify the input signal. The generic ac small-signal equivalent circuit of an FET is shown in Fig. 9.9*a*. The ac equivalent circuit shows the parasitic resistances and capacitances related to the drain-gate and gate-source, such as the drain-gate capacitance, the gate resistance, and source resistance. The gate-source input resistance represents the leakage current through the input junction. For an ideal MESFET, the input small-signal current through the gate at a frequency $\omega = 2\pi f$ is mainly the displacement current i_{in} given by

$$i_{\text{in}} = 2\pi f C_G v_g \tag{9.36}$$

where C_G is the gate-to-channel capacitance, $C_G = C_{gs} + C_{gd}$, given by

$$C_G = \frac{xL\epsilon_s}{\overline{W}} \tag{9.37}$$

where $\overline{W}$ is the depletion region width at the gate Schottky contact. The small-signal output current is obtained from the definition of the

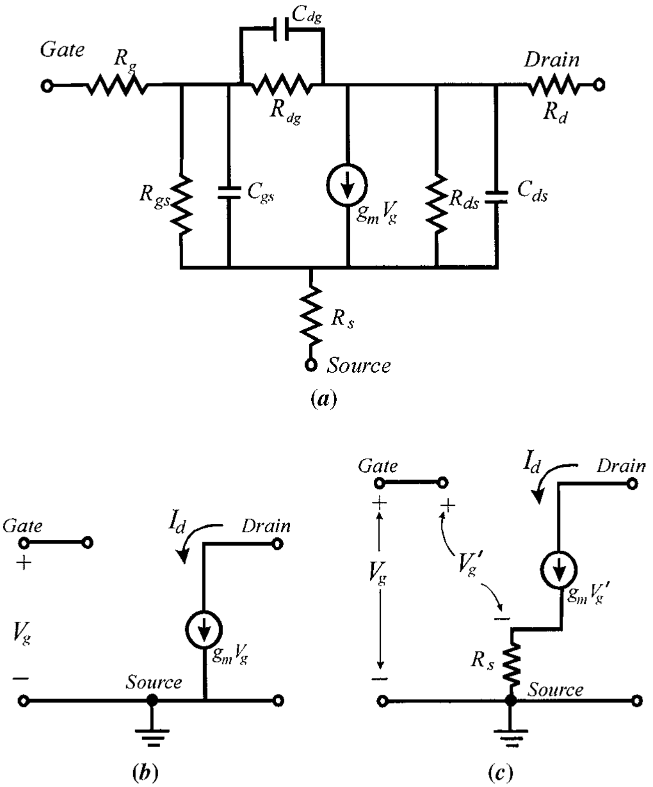


Figure 9.9 (*a*) An ac small-signal equivalent circuit model for an *n*-channel MESFET. (*b*) Ideal small-signal circuit. (*c*) Ideal small-signal circuit with R_s.

transconductance given by Eq. (9.32) such that

$$g_m = \frac{\partial I_D}{\partial V_G} = \frac{i_{\text{out}}}{v_g} \qquad \text{or} \qquad i_{\text{out}} = g_m v_g \tag{9.38}$$

From the definition of the cutoff frequency, i.e., the frequency at which the MESFET does not amplify, the output current must be equal to the input current. Thus, an expression for the cutoff frequency f_T can be obtained by equating Eqs. (9.36) and (9.38), which gives

$$2\pi f_T C_G v_g = g_m v_g \Rightarrow f_T = \frac{g_m}{2\pi C_G} \tag{9.39}$$

By using the maximum value of $g_m = I_p/V_p$ obtained from the drain current in the saturation mode of the normally on MESFET [see

Eq. (9.31)] and the definition of C_G given by Eq. (9.37), one can rewrite Eq. (9.39) as

$$\begin{aligned} f_T &= \frac{g_m}{2\pi C_G} = \frac{I_p}{V_p}\frac{1}{2\pi C_G} = \frac{e^2\mu_n N_d^2 x a^3}{2\epsilon_s L}\frac{2\epsilon_s}{eN_d a^2}\frac{\overline{W}}{2\pi x L\epsilon_s} \\ &= \frac{e\mu_n N_d a\overline{W}}{2\pi L^2\epsilon_s} = \frac{e\mu_n N_d a^2}{2\pi L^2\epsilon_s} \end{aligned} \tag{9.40}$$

Notice that in the saturation mode, the depletion region width is equal to a, where a is defined in Fig. 9.7 as the total thickness of the epitaxial layer. Equation (9.40) implies that to obtain a high cutoff frequency, the MESFET should be designed such that the electron mobility is as high as possible and the gate length is as short as possible.

The cutoff frequency can be written in terms of the transient time τ_{tr} required for the electron to travel through the channel as

$$f_T = \frac{1}{2\pi\tau_{\text{tr}}} = \frac{\upsilon_s}{2\pi L} \tag{9.41}$$

where υ_s is the carrier saturation velocity, which is essentially the drift velocity when the MESFET is operating in the saturation region. Again, the cutoff frequency can be increased by reducing the gate length and by choosing material with a high drift velocity. For example, if GaAs has a drift velocity on the order of 2.0×10^7 cm/s and the transistor gate length is 10 μm, the cutoff frequency is ~3.20 GHz. As a comparison, the electron drift velocity in silicon is about an order of magnitude smaller than that of GaAs, which leads to a cutoff frequency on the order of ~0.32 GHz for a transistor made of silicon.

The ideal ac small-signal circuit of the MESFET is shown in Fig. 9.9*b*, where all the diffusion parasitic resistances are infinite and the series resistances are zero. All the capacitances are open circuit for low frequency. In the small-signal analysis, the drain current becomes

$$I_d = g_m V_g \tag{9.42}$$

The effect of the source resistance on the drain current can be understood by examining Fig. 9.9*c*, where R_s is added to the ideal small-signal equivalent circuit. From this figure, the drain current is $I_d = g_m V_g'$ where V_g' and V_g are related through the following relation:

$$V_g = V_g' + g_m V_g' R_s = (1 + g_m R_s)V_g' \tag{9.43}$$

Thus, Eq. (9.42) can be rewritten as

$$I_d = g_m V_g' = \frac{g_m}{1 + g_m R_s} V_g = g_m' V_g \tag{9.44}$$

where g'_m is called the extrinsic transconductance and is smaller than g_m. The transconductance is decreased when the source resistance is added.

For high-power applications, it is desired to have V_D^{sat} as small as possible. This requirement can be achieved by choosing material with high electron mobility and very small source and drain resistances. It is also required that the breakdown voltage be as high as possible. The breakdown voltage V_B occurs at the drain end of the channel, where the reverse voltage is highest. Thus, $V_B = V_D + |V_G|$. The breakdown voltage is usually higher for larger bandgap materials.

9.4 Junction Field-Effect Transistor

The basic structure of the junction field-effect transistor (JFET) is shown in Fig. 9.10*a* for an *n*-channel device. The *n* region between the two p^+ regions is called the channel. The majority carriers in the channel flow from the source to the drain. The gate is the control terminal. A more realistic sketch of a cross section of a JFET is shown in Fig. 9.10*b*. A *p*-channel JFET can be fabricated in which the *p* and *n* regions are reversed. Usually the *p*-channel JFET is slower than the *n*-channel JFET due to the fact that the hole mobility is smaller than the electron mobility.

If the gate is kept at zero bias and a small voltage is applied to the drain terminal, a drain current is formed between the drain and source.

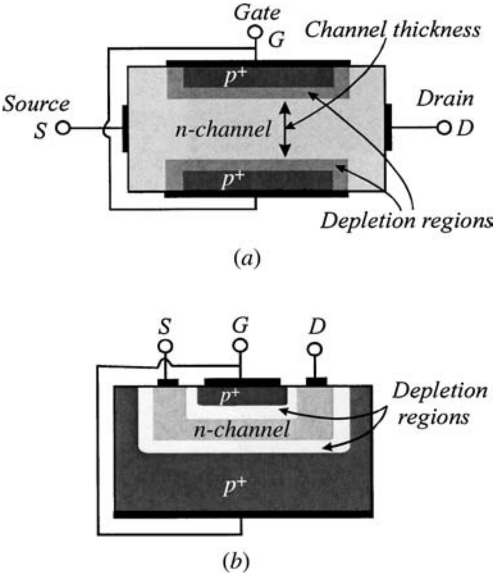


Figure 9.10 (*a*) The basic structure of the JFET. The drain and source metals are ohmic contacts, and the gate metal is a Schottky contact. (*b*) A sketch of a more realistic *n*-channel JFET. Notice that the gate metal is deposited on the p^+-type semiconductor and the drain and source contacts are made on the *n*-type semiconductor.

The source is usually grounded. The JFET under these conditions behaves as a resistance, and the *I-V* curve is linear so long as the drain voltage is small enough. If a reverse-biased voltage (negative voltage) is applied to the gate, the depletion region widens and the channel region becomes narrower, leading to an increase in the channel resistance. Thus, the slope of the *I-V* linear curve is reduced. For sufficiently large reverse bias applied to the gate, the two depletion regions touch each other causing the channel to be completely depleted. This condition is called the *pinch-off*.

If the drain voltage is increased, the JFET behaves differently. The operation of the JFET under increasingly forward-biased voltage applied to the drain and increasingly reverse-biased voltage applied to the gate is shown in Fig. 9.11. As seen from this figure, the *I-V* behavior of the JFET is similar to that of the MESFET. Starting from a zero-biased gate and $V_{D1} < V_D^{\text{sat}}$, the JFET acts like a resistor and the drain current is linear with the drain voltage. Again, the saturation voltage V_D^{sat} is the drain voltage required to produce a pinch-off condition. The dot on the curve represents the approximate coordinates of the drain current and voltage. The JFET remains in the linear region so long as $V_D < V_D^{\text{sat}}$. For example, when $V_G = 0$, and $V_{D2} < V_D^{\text{sat}}$ where $V_{D1} < V_{D2}$, the two depletion regions are still separated as shown in Fig. 9.11*b*. When $V_G = 0$, and $V_D = V_D^{\text{sat}}$, the two depletion regions meet at the drain terminal. This is the pinch-off condition, which is shown in Fig. 9.11*c*. For $V_G = 0$, and $V_D > V_D^{\text{sat}}$, the device reaches the saturation mode and the drain current becomes independent of the drain voltage. In the saturation mode, the electrons are injected into the depletion region and swept to the drain region under the influence of the electric field force. Ideally, the drain current remains constant in the saturation region until the avalanche breakdown voltage is reached, at which point the device acts as if it shorted.

When a reverse-biased voltage is applied to the gate, the depletion region will be increased even for a small drain forward-biased voltage. Thus, the drain current is reduced as shown in Fig. 9.11*e*. For a fixed gate voltage the drain current behavior as a function of the drain voltage is identical to the behavior observed in Fig. 9.11*a* to *d*, except that the current value is lower. Using PSpice, one can obtain the typical *I-V* characteristic curves of the JFET, which are similar to those obtained for the MESFET as shown in Fig. 9.8.

Figure 9.12 shows the drain current as a function of the gate voltage at a constant drain voltage (solid line) obtained using PSpice. The actual circuit used for PSpice simulation is shown in the inset of the figure. The I_D versus V_G curve is obtained for a pinch-off voltage of $V_p = -4.0$ V, and the drain voltage was fixed at $V_D = 15$ V. The source resistance was chosen as 50 Ω, which is not shown in the figure. The derivation

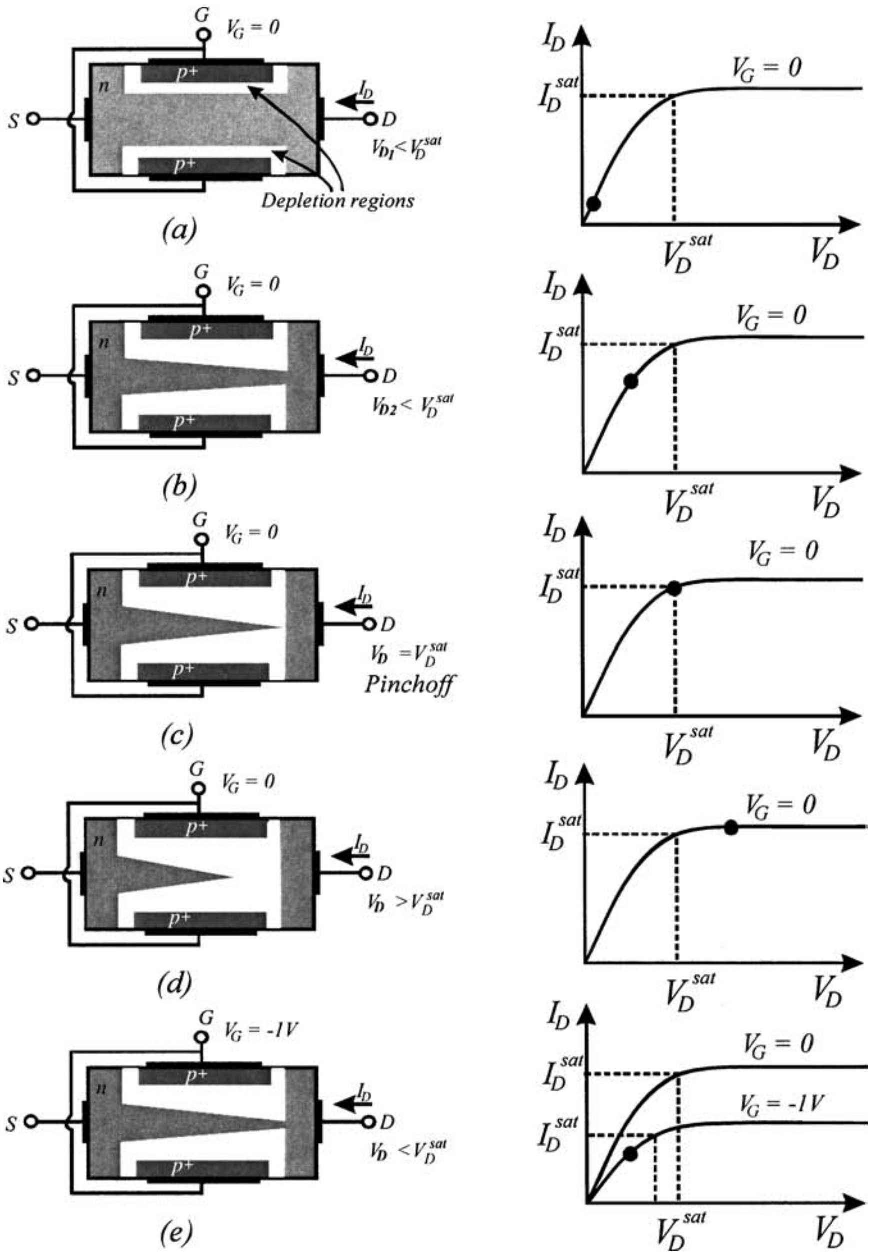


Figure 9.11 The basic operation of an n-channel JFET. The depletion regions and the I-V curves are shown for the following conditions: (a) $V_G = 0$, $V_{D1} < V_D^{\text{sat}}$, linear region; ($b$) $V_G = 0$, $V_{D2} < V_D^{\text{sat}}$ where $V_{D1} < V_{D2}$, linear region; (c) $V_G = 0$, $V_D = V_D^{\text{sat}}$, pinch-off condition; (d) $V_G = 0$, $V_D > V_D^{\text{sat}}$, saturation region; and (e) $V_G = -1.0$ V, $V_D < V_D^{\text{sat}}$, linear region.

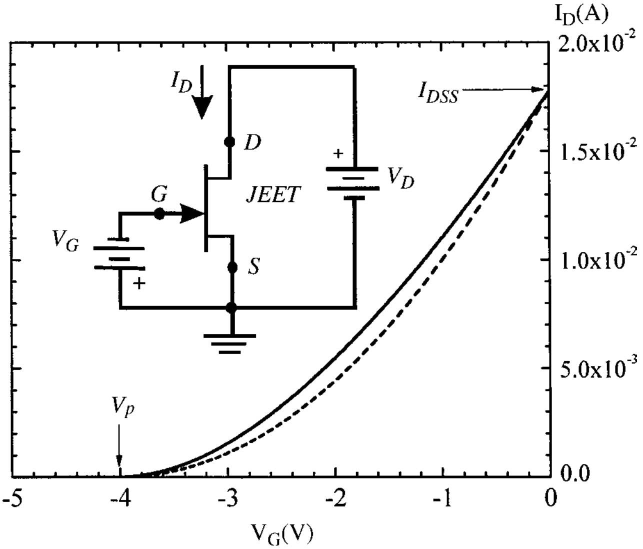


Figure 9.12 The drain current as a function of the gate voltage in the saturation region obtained by using PSpice (solid line). The pinch-off voltage was chosen as −4 V. The inset is the dc circuit of the JFET. The dashed line is obtained using Eq. (9.45).

of the drain current as a function of the gate voltage is quite similar to that of the MESFET. The drain current can be approximated as

$$I_D \approx I_{\mathrm{DSS}} \left(1 - \frac{V_G}{V_p}\right)^2 \tag{9.45}$$

where I_{DSS} is the drain-source current obtained when the drain voltage is in saturation. It is obvious from Eq. (9.45) that the drain current is a parabolic function of V_G. The current I_{DSS} is taken as the drain current at $V_G = 0$. Equation (9.45) is plotted in Fig. 9.12 as the dashed line, which is in good agreement with the result obtained from PSpice.

9.5 Heterojunction Field-Effect Transistors

The MESFETs and JFETs have their own limitations due to the fact that the high-level doping degrades the carrier mobilities and drift velocities. The next generation of FETs is based on epitaxially grown heterojunction structures such as GaAs/AlGaAs, and GaN/AlGaN. The epitaxial growth techniques, such as MBE and MOCVD, allow one to deposit stacks of different bandgap semiconductor layers. Heterojunction semiconductors are simple to grow with precise control of the

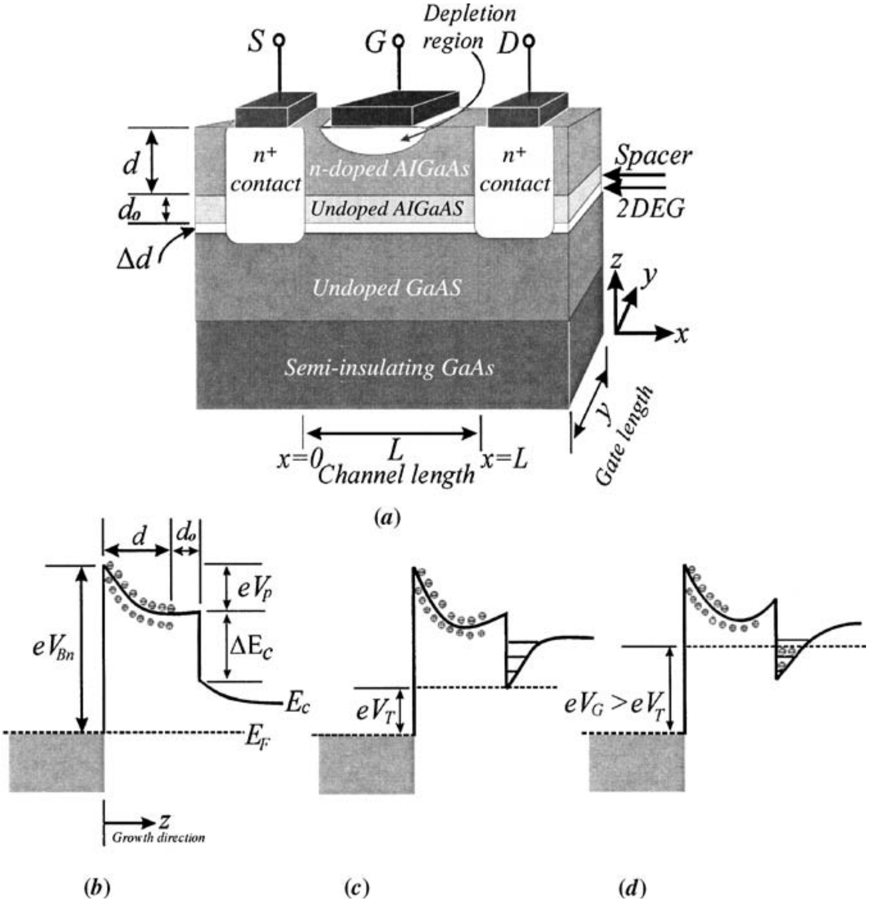


Figure 9.13 (*a*) A sketch of a MODFET based on a GaAs/AlGaAs heterojunction is shown. The conduction band diagram is shown for (*b*) $V_G = 0$, thermal equilibrium; (*c*) $V_G = V_T$, where V_T is the threshold voltage; and (*d*) $V_G > V_T$.

doping level, layer thickness, and alloy compositions. There are several types of heterojunction field-effect transistors (HFETs). The simplest structure is called the high electron mobility transistor (HEMT). This class of electronic devices is also called modulation-doped field-effect transistors (MODFETs). The conventional GaAs/AlGaAs MODFET structure is shown in Fig. 9.13*a*. The basic structure is composed of an undoped GaAs layer grown on semi-insulating GaAs substrate. Buffer layers are usually introduced between the substrate and the undoped GaAs layer to reduce the dislocation density as well as to prevent the propagation of the dislocation into the device structure. An AlGaAs layer (larger bandgap) is then grown on top of the undoped GaAs layer (smaller bandgap). The AlGaAs is usually n-type doped with a small portion of it left undoped at the GaAs/AlGaAs interface. The undoped AlGaAs portion is called the *spacer layer*. The term *modulation doping* comes from the fact that the AlGaAs barrier is doped instead of the GaAs layer. The n^+ layer under the drain and source is usually produced by ion implantation. Silicon is usually used as a donor dopant in both the AlGaAs barrier and in the n^+ layers.

The thickness of the spacer layer is usually between 20 and 200 Å, and it plays a major role in the determination of the carrier density and mobility. For example, the carrier density decreases while the mobility increases as the spacer thickness is increased. The effect of the spacer layer thickness can be understood in terms of electron-electron many-body interactions. When the barrier is doped, a triangular quantum well is formed at the GaAs/AlGaAs interfaces. This quantum well is populated by electrons forming what is called the two-dimensional electron gas (2DEG). The 2DEG is generated from the thermally ionized dopant atoms in the AlGaAs barrier. If the spacer thickness is larger, the probability of the electron being transferred to the quantum well is decreased. On the other hand, the electron-electron interactions (mainly exchange interactions) increase as the 2DEG density is increased. These interactions become significant when the carrier concentration exceeds 10^{18} cm^{-3}. The carrier mobility decreases as the magnitude of these interactions is increased, as discussed in Chap. 7.

The 2DEG in the triangular well forms the transistor channel. More complicated MODFET structures can be fabricated such that several channels can be formed by repeating the epitaxially grown structure in Fig. 9.13*a*; each period produces a channel. The conduction band diagram of the MODFET is shown in Fig. 9.13*b* to *d* under different gate voltages. In Fig. 9.13*b*, we sketched the conduction band diagram at thermal equilibrium. The thickness of the doped AlGaAs layer is d, and the spacer thickness is d_o. The conduction band offset is designated as ΔE_c, while the pinch-off voltage is shown as V_p. The gate metal is shown on the left-hand side of the diagram. The pinch-off voltage can be obtained by integrating the electric field over the total thickness of the doped AlGaAs barrier. Assuming that the dopant density N_d is uniform across the barrier, V_p can be written as

$$V_p = \frac{e}{\epsilon_s} \int_0^d N_d z \, dz = \frac{e N_d d^2}{2\epsilon_s} \tag{9.46}$$

where ϵ_s is the permittivity of the AlGaAs layer.

The threshold voltage V_T, shown in Fig. 9.13*c*, is defined as the voltage applied to the gate such that the Fermi energy level is touching the bottom of the GaAs conduction band. This corresponds to the beginning of the formation of the channel at the GaAs/AlGaAs interface. The threshold voltage can be written as

$$eV_T = eV_{\mathrm{Bn}} - \Delta E_c - eV_p \tag{9.47}$$

where eV_{Bn} is the Schottky barrier height. The threshold voltage can be adjusted for a MODFET by choosing a specific Schottky metal and by adjusting the pinch-off voltage. The latter can be adjusted by varying the dopant concentration in the barrier and the spacer thickness.

When V_T is positive, the MODFET is normally off. This is called the enhancement mode. On the other hand, the normally on MODFET corresponds to the depletion-mode or negative V_T. When the gate voltage is larger than V_T, as shown in Fig. 9.13*d*, a charge sheet or 2DEG density $\mathcal{N}_s$ is induced by the gate at the heterojunction interface, which can be given as

$$\mathcal{N}_s = \frac{C_i(V_G - V_T - V_x)}{e} \tag{9.48}$$

where the capacitance per unit area C_i is given as

$$C_i = \frac{\epsilon_s}{d + d_o + \Delta d} \tag{9.49}$$

where Δd is the thickness of the channel or the width of the triangular quantum well. The potential V_x is the channel potential with respect to the ground (source). Thus, V_x varies through the channel from zero at the source terminal to V_D at the drain terminal. According to Eq. (9.48), the 2DEG density increases as the positive gate voltage V_G (forward bias) increases. For negative V_G, $\mathcal{N}_s$ is reduced drastically, and the device is turned off.

The drain current in a MODFET can be obtained by assuming that the channel width varies along the x direction (see Fig. 9.13*a*) similar to the behavior of the channel in the MESFET and MOSFET. The only difference in the MODFET is that the channel is a two-dimensional system where the electrons are confined along the growth axis (z axis), but not confined in the xy plane. Thus, the drain current I_D at any point along the channel of the MODFET sketched in Fig. 9.13*a* can be written as

$$I_D = ye\mu_n\mathcal{N}_s\mathcal{E}_x \tag{9.50}$$

where $\mathcal{E}_x$ is the electric field at any point along the channel. Substitute Eqs. (9.48) and (9.49) into Eq. (9.50) to obtain

$$I_D = y\mu_n\epsilon_s\frac{V_G - V_T - V_x}{d + d_o + \Delta d}\frac{dV_x}{dx} \tag{9.51}$$

Integrating Eq. (9.51) from the source ($x = 0$ and $V_x = 0$) to the drain ($x = L$ and $V_x = V_D$) yields

$$I_D = \frac{y}{2L}\frac{\mu_n \epsilon_s}{d + d_o + \Delta d}\left[2(V_G - V_T)V_D - V_D^2\right] \tag{9.52}$$

For the linear region or nonsaturation region, $V_D \ll V_G - V_T$, which reduces Eq. (9.52) to the following:

$$I_D = \frac{y}{L}\frac{\mu_n \epsilon_s}{d + d_o + \Delta d}(V_G - V_T)V_D \tag{9.53}$$

For a large drain voltage, the depletion region reaches the channel and causes the pinch-off, where the density of the 2DEG is reduced to zero at $x = L$. Under this condition, V_D^{sat} can be obtained from Eq. (9.48) as

$$V_D^{\text{sat}} = V_G - V_T \tag{9.54}$$

The saturation current can now be obtained by replacing V_D in Eq. (9.52) with V_D^{sat} and by using Eq. (9.54) as follows:

$$I_D^{\text{sat}} = \frac{y}{2L}\frac{\mu_n \epsilon_s}{d + d_o + \Delta d}(V_G - V_T)^2 \tag{9.55}$$

The transconductance is obtained by using the definition given by Eq. (9.32), which yields

$$g_m = \frac{y}{L}\frac{\mu_n \epsilon_s}{d + d_o + \Delta d}(V_G - V_T) \tag{9.56}$$

Notice that this transconductance expression is similar to that of the MESFET given by Eq. (9.35).

The electric field along the channel reaches high values in high-speed devices, such as MODFETs, causing carrier velocity saturation. The drain current in the high-speed operation mode can be written as

$$I_D^{\text{hs}} = ye\mu_n \mathcal{N}_s \mathcal{E}_x = yev_s \mathcal{N}_s = yv_s C_i(V_G - V_T) \tag{9.57}$$

where the superscript hs is introduced to indicate "high speed" and v_s is the carrier saturation velocity. Using the definition of the cutoff frequency given by Eq. (9.39) and knowing that the capacitance C_i defined in Eq. (9.49) is a capacitance per unit area, one can express the cutoff frequency as

$$f_T = \frac{v_s}{2\pi[L + C_p/(yC_i)]} \tag{9.58}$$

where C_p is the total parasitic capacitance. To design a high-speed MODFET, one needs to choose a material with a high carrier saturation velocity, small channel length, and very small parasitic capacitances.

9.6 GaN/AlGaN Heterojunction Field-Effect Transistors

The unique material properties of GaN-based semiconductors have stimulated a great deal of research and development in materials growth and optoelectronic and electronic devices using this semiconductor system. Gate current collapse and device stability are among many other issues of interest in the research community. Recently, electron devices based on GaN/AlGaN have been the subject of various investigations by many groups throughout the world. The interest in GaN HFETs stems from high-power, high-temperature, and high-frequency applications. While the reported results are very encouraging thus far, there are several problems associated with GaN-related electronic devices. For example, the high dislocation densities can have a detrimental effect on the performance of the device. The gate leakage current, or the gate current collapse, is another problem facing the nitride HFETs.

The GaN/AlGaN heterostructures are normally grown by the MBE or MOCVD techniques. The MOCVD reactors require triethylgallium and ammonia as precursor gases to deposit a GaN layer in the conventional deposition regime where precursors enter the growth chamber simultaneously. $Al_xGa_{1-x}N$ layers are deposited in the atomic layer regime when precursors enter the chamber in a cyclic fashion using triethylgallium, triethylaluminum, and ammonia as precursors. The precursors are introduced into the chamber using hydrogen or nitrogen as a carrier gas. The common substrates used for HFETs are either sapphire or SiC substrates placed on a graphite susceptor, which is heated to the growth temperature by RF induction. Other substrates such as Si(111) have also been used for III-nitride HFETs.

Several device structures have been reported in the literature. However, the two most common structures are shown in Fig. 9.14. The 2DEG is formed at the GaN/AlGaN interface. The buffer layer varies. The most common buffer layer is the low-temperature-grown AlN layer. More elaborate buffer layers are composed of low-temperature AlN and

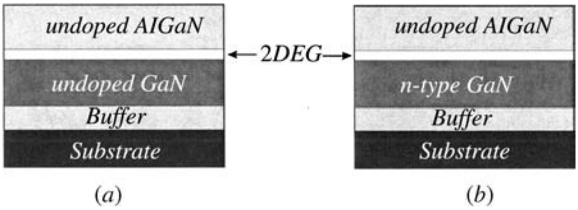


Figure 9.14 Two sketches of the most common GaN/AlGaN HFET structures.

GaN/AlGaN superlattices. One of the main functions of the buffer layer is to prevent the dislocations formed at the substrate surfaces from propagating in the HFET structure. It also acts as an insulator between the device and the substrate. The main difference between the two structures is that the GaN layer in the structure of Fig. 9.14*a* is undoped, while it is doped in the structure of Fig. 9.14*b*. Silicon is usually used as the *n*-type dopant in GaN. The formation of the 2DEG in the structure of Fig. 9.14*a* relies on the spontaneous polarization-induced charge sheet. This requires that the polarity of the GaN surface should be Ga-rich. The sheet carrier densities in nominally undoped GaN/AlGaN structures can, in fact, be comparable to those achievable in extrinsically doped structures, but without the degradation in mobility that can result from the presence of ionized impurities. A simple electrostatic analysis shows that the sheet carrier concentration $\mathcal{N}_s$ of the 2DEG at the GaN/$Al_xGa_{1-x}N$ heterojunction interface should be given approximately by (see Morkoç et al. 1999 and Yu 2003)

$$\mathcal{N}_s = \frac{\sigma_{\text{pol}}}{e} - \left(\frac{\epsilon_{\text{AlGaN}}}{de^2}\right)(e\phi_b + E_F - \Delta E_c) + \frac{1}{2}N_d d \tag{9.59}$$

where ϵ_{AlGaN} = dielectric constant of $Al_xGa_{1-x}N$

ϕ_b = Schottky metal work function

E_F = Fermi energy at heterojunction interface

The thickness of the intentionally doped layer is d, and N_d is the donor concentration, assumed to be uniformly distributed throughout the layer. For the structure in Fig. 9.14*a*, N_d is zero. The polarization-induced sheet charge density σ_{pol}, shown in Eq. (9.59), at the GaN/$Al_xGa_{1-x}N$ heterojunction interface is given, in the linear interpolation realm, approximately by (see Yu 2003)

$$\frac{\sigma_{\text{pol}}}{e} \approx -2\left[e_{31} - \left(\frac{c_{13}}{c_{33}}\right)e_{33}\right]\left(\frac{a_{\text{GaN}}}{a_{\text{AlN}}} - 1\right)x + P_{\text{sp},z}^{\text{GaN}} P_{\text{sp},z}^{\text{AlGaN}} \tag{9.60}$$

where $e_{31}, e_{33}, c_{13}, c_{33}$ = relevant piezoelectric and elastic constants for $Al_xGa_{1-x}N$

$a_{\text{GaN}}, a_{\text{AlN}}$ = lattice constants of GaN and AlN, respectively

$P_{\text{sp},z}^{\text{GaN}}, P_{\text{sp},z}^{\text{AlGaN}}$ = spontaneous polarizations of GaN and $Al_xGa_{1-x}N$, respectively

Using averages of the values (see Morkoç et al. 1999 and Simin et al. 2001 and 2002) of the spontaneous polarization fields, piezoelectric coefficients, and elastic constants, one can estimate that $\sigma_{\text{pol}} \approx (5 - 6.5) \times 10^{13}\ xe/\text{cm}^2$, where x is the Al mole fraction. Nonlinear models

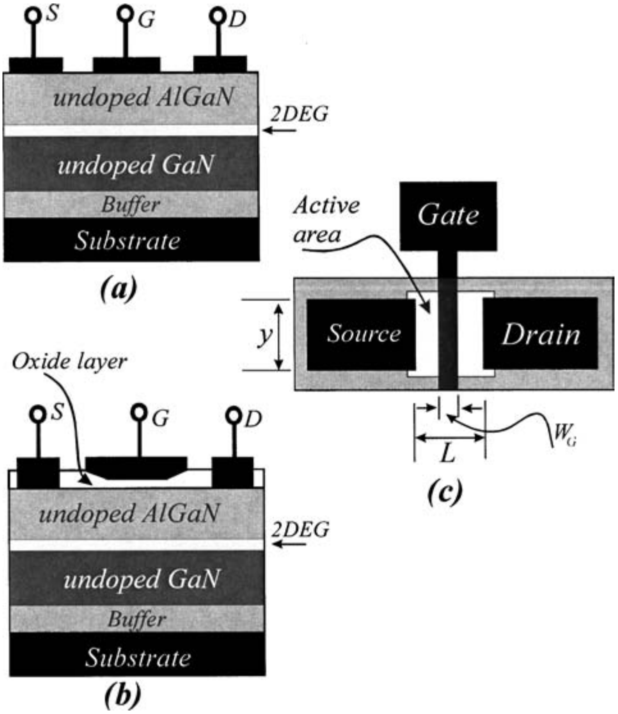


Figure 9.15 (*a*) A schematic structure of a GaN/AlGaN HFET showing the gate, drain, and source metals deposited directly on the surface of the AlGaN layer. (*b*) A schematic structure of a GaN/AlGaN HEFT with an additional oxide layer deposited underneath the gate metal. (*c*) Top view sketch of an HFET.

can be used for a more accurate estimation of the sheet charge in this manner.

There are two methods of fabricating GaN/AlGaN HFETs. The first method relies on depositing the ohmic contacts for the drain and the source and the Schottky contact for the gate directly on the AlGaN layer as shown in Fig. 9.15*a*. Ion implantation is not necessary to produce the n^+ region as it is in the case for GaAs/AlGaAs HFETs. This is because the metallic atoms from the ohmic contacts diffuse down to the channel. An example of this method is described by Gaska et al. (2003). The second method utilizes an oxide layer, such as SiO_2, Al_2O_3, and silicon oxynitrides, deposited underneath the gate metal, as shown in Fig. 9.15*b*. The acronym MOSHFET is given to this type of HFET, which indicates a metal-oxide semiconductor HFET. The advantage of the second method is that the gate current is reduced due to the presence of the oxide underneath the gate. An example of a top view of an HFET is shown in Fig. 9.15*c*.

There are several steps involved in the fabrication of the GaN/AlGaN HFETs. The main fabrication steps include:

Photolithography for ohmic contact openings

Ohmic contact metallization, which requires the deposition of a four-metal composition Ti/Al/Ti/Au

Rapid thermal annealing of ohmic contacts

Photolithography for device isolation level

Reactive ion etching or ion implantation for electric device isolation

Contact or electron-beam lithography for gate openings

Gate metallization, which requires the deposition of Ni/Au, Pd/Au or Pt/Au layers

Generally speaking, the source-to-drain spacing, or channel length, is $L \approx 2$ μm and the gate length is $y \approx 5$ μm. However, the total gate length could be as large as 50 to 200 μm. The width of the gate metal is $W_G \approx 0.2$ μm.

The drain current-voltage (I_D-V_D) characteristics of the GaN/AlGaN HFET are identical to those of the GaAs/AlGaAs HFET discussed in Sec. 9.5. A typical example is shown in Fig. 9.16 for $V_T = -5$ V and $0 \leq V_G \leq -4$ V. Notice that the gate voltage, channel current, and drain voltage were taken with respect to the source, which is usually grounded. In integrated circuits, the following notations are usually

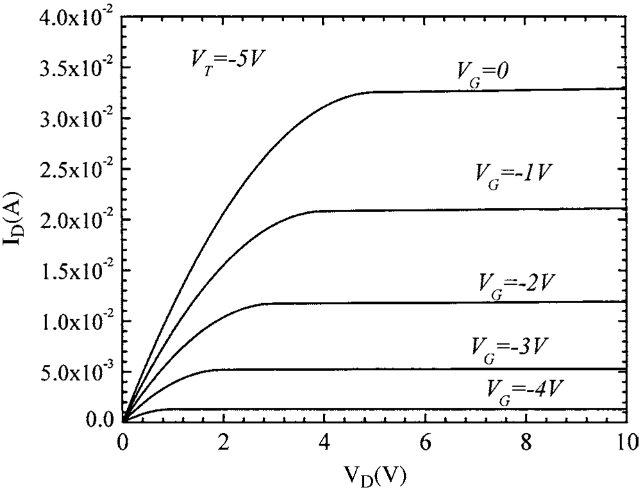


Figure 9.16 I_D-V_D characteristic curves of a GaN/AlGaN single-channel HFET with a threshold voltage of −5 V, channel length of 0.5 μm, and gate length of 20 μm.

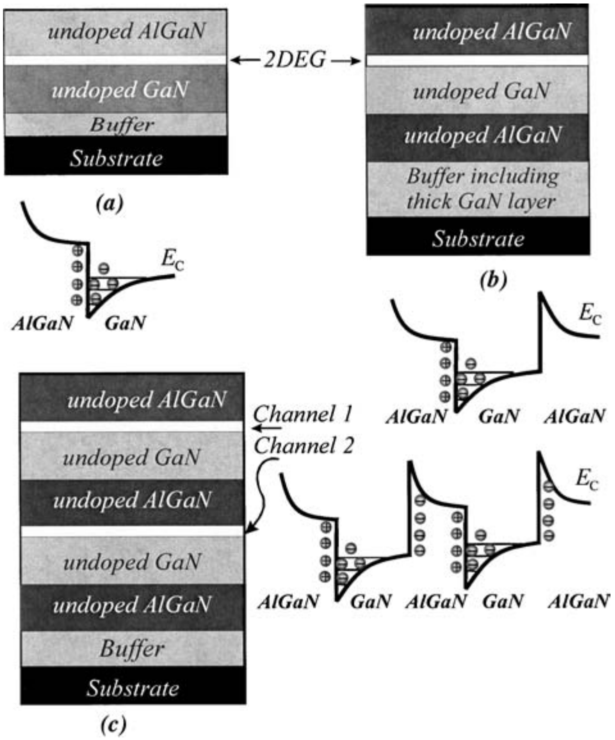


Figure 9.17 (*a*) A basic single heterojunction structure of a GaN/AlGaN HFET. (*b*) A double heterojunction structure. (*c*) Two-channel structures. The conduction energy band diagram is sketched for the three structures showing the spontaneous polarization-induced charge distribution.

used: V_{DD} is the drain bias voltage, V_{SS} is the source voltage, V_{DS} is the drain-source voltage, I_{DS} is the drain-source current, and V_{GS} is the gate-source voltage.

There are several structural variations to the GaN/AlGaN HFETs. For example, Fig. 9.17 shows three different variations. The first structure, Fig. 9.17*a*, was previously discussed, and it has only one channel formed at the GaN/AlGaN interface. It is characterized by a single heterojunction as shown in the conduction energy band diagram. The double-heterojunction HFET is shown in Fig. 9.17*b*. It is simply a single quantum well. The conduction energy band diagram shows the positive and negative charge carriers generated from the spontaneous polarization effect. A third type is shown in Fig. 9.17*c*, which is composed of two 2DEG channels. The 2DEG is almost doubled in this structure. Additional channels can be added to even further increase the 2DEG density.

The buffer layers in these structures vary from structure to structure. The most common buffer layer schemes are low-temperature-grown AlN, a thick undoped GaN layer, and GaN/AlGaN superlattices. More complicated buffers are composed of more than one scheme, such as GaN/AlGaN superlattices sandwiched between undoped GaN layers. The superlattice-GaN layer could be repeated several times to ensure the presence of high-quality surfaces on which the device structure can be deposited. The analysis for multichannel HFETs is usually more complicated. For example, each channel has its own resistance.

9.7 Heterojunction Bipolar Transistors

A vertical device that has been used in many applications is called the heterojunction bipolar transistor (HBT). The p-type base creates a barrier for electron diffusion from the n-type emitter to the n-type collector. One of the advantages of this electronic device is that the base thickness can be made too short and therefore the transient time for the electrons to tunnel through the base is very short. A typical example of an HBT based on GaAs/AlGaAs heterojunctions is shown in Fig. 9.18. A sketch of the energy band diagram is also shown. The device isolation is achieved by either deep ion implantation or by making mesa structures.

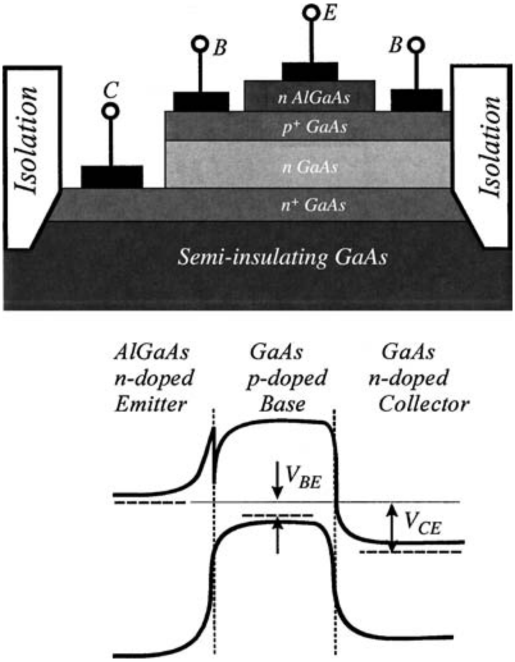


Figure 9.18 A schematic structure for a GaAs/AlGaAs HBT is shown with the emitter E, base B, and collector C. The energy band diagram is also sketched showing the base-emitter voltage V_{BE} and collector-emitter voltage V_{CE}.

Another advantage of HBTs over the regular bipolar transistors is that the junction between the emitter and base is abrupt on an atomic scale as compared to the macroscopic junction in the case of bipolar transistors. Since the base current is due to the carrier injection from the base into the emitter, the heterojunction barrier for holes reduces the base current and therefore enhances the current gain and allows a higher base doping. Higher doping in the base reduces the base resistance.

The formalism of the transport properties of HBTs is similar to that encountered in the bipolar transistor. The collector current density is given by

$$J_c = \frac{D_n e n_{\text{po}}}{W_B}\left(e^{eV_{\text{BE}}/k_B T} - 1\right) \tag{9.61}$$

where V_{BE} = base-emitter voltage
n_{po} = equilibrium electron density in base (minority carrier density)
D_n = electron diffusion coefficient
W_B = base width

For $eV_{\text{BE}} \gg k_B T$, the transconductance can be obtained as

$$g_m = \frac{\partial I_C}{\partial V_{\text{BE}}} = \frac{e}{k_B T} I_C \tag{9.62}$$

The exponential dependence of g_m on V_{BE} produces large transconductance values, which are advantageous for digital applications.

The capacitance due to the minority charge stored in the base is given by

$$C = \frac{\partial Q_B}{\partial V_{\text{BE}}} = \frac{e}{k_B T}\frac{e n_p W_B}{2} \tag{9.63}$$

where n_p is the minority carrier density at the base entrance. The time required for the electron to travel through the base, τ_{tr}, can be obtained as

$$\tau_{\text{tr}} = \frac{W_B}{\upsilon_d} = \frac{W_B}{2D_n/W_B} = \frac{W_B^2}{2D_n} \tag{9.64}$$

where the average diffusion velocity is taken as $\upsilon_d = 2D_n/W_B$. The time defined in Eq. (9.64) can be drastically reduced by minimizing the base width leading to a very high speed device.

In the GaAs/AlGaAs HBTs, the aluminum mole fraction in the emitter material is usually kept to less than 30 percent. Carbon is usually used as the dopant in the p-type GaAs base material. It has a low diffusion coefficient compared to other dopants, such as zinc and magnesium.

The dopant diffusion from the base to the emitter causes a significant problem in HBTs. Thus, there is a tradeoff between the doping level and the performance of the device. Heterojunction bipolar transistors have been fabricated from other semiconductor systems, such as Si/SiGe and InGaP/GaAs. The InGaP/GaAs system has a larger valence band offset as compared to the conduction band offset. This is an advantage since the holes see a larger barrier and hence the hole injection from the base to the emitter is reduced.

9.8 Tunneling Electron Transistors

As we have seen, the performance of the electronic devices is improved by the introduction of heterojunctions. The presence of potential barriers allows one to fabricate unipolar transistors where the structure is composed of only *n*-type layers. The simplest device of this class is called the hot electron transistor (HET). The tunneling hot electron transistor (THET) is another variation of the HET. The conduction energy band diagrams of both HET and THET are shown in Fig. 9.19. The HET structure utilizes a graded AlGaAs barrier between the emitter and the base as shown in Fig. 9.19*a*. When a high base-emitter voltage is applied, it causes the electrons in the emitter to gain enough kinetic energy and be swept across the base to the collector. These are hot electrons and can ballistically transverse the base region very rapidly, allowing high-frequency operation. By varying the base-collector voltage, it is possible to use the second AlGaAs barrier as an analyzer of the energy distribution of the electrons that arrive at the collector terminal.

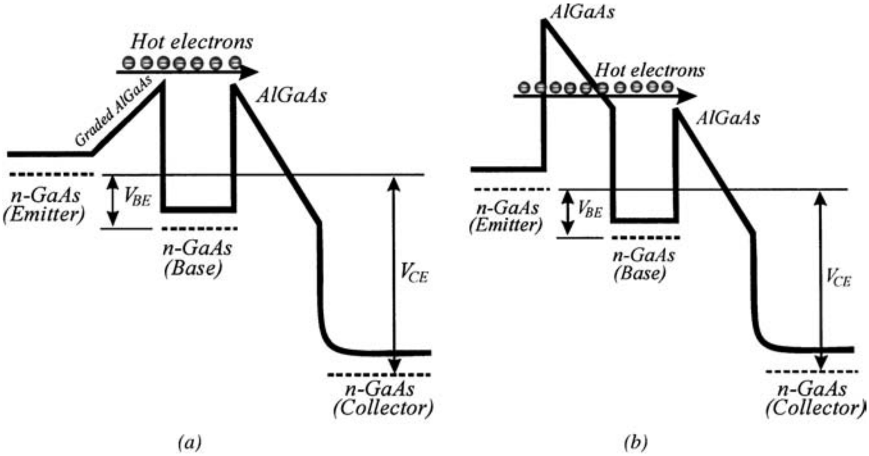


Figure 9.19 The conduction energy band diagram is shown for unipolar transistors. (*a*) Hot electron transistor with graded AlGaAs barrier and (*b*) a tunneling hot electron transistor.

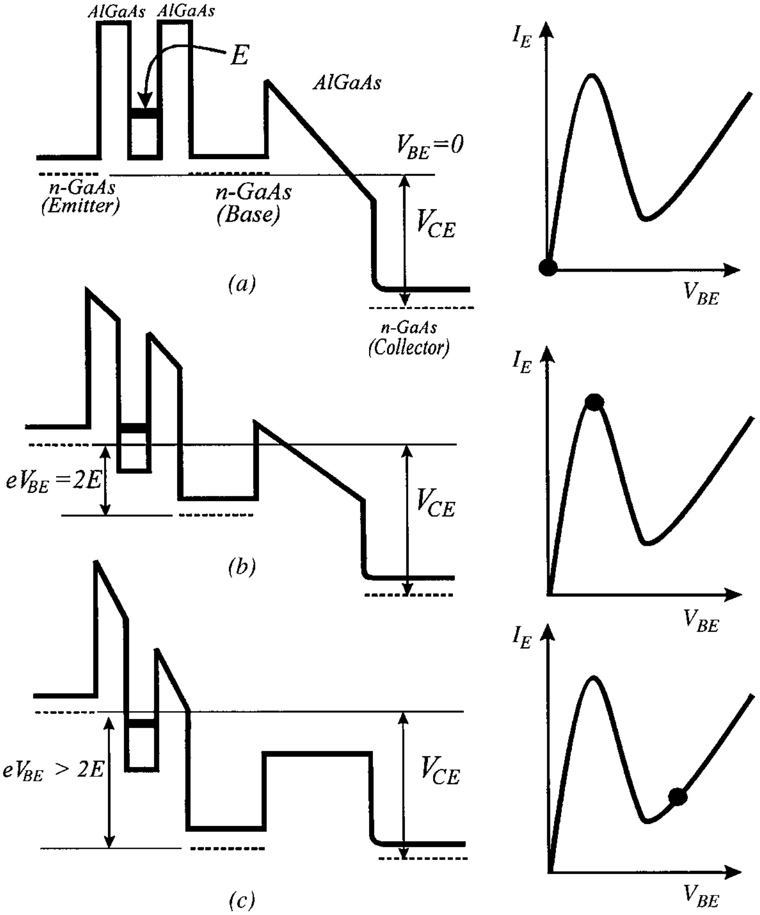


Figure 9.20 Conduction energy band diagram of a unipolar resonant-tunneling hot electron transistor (RHET) illustrated for a constant V_{CE}. The band diagram and the *I-V* curves were plotted for (*a*) $V_{\mathrm{BE}} = 0$, (*b*) $eV_{\mathrm{BE}} = 2E$, and (*c*) $eV_{\mathrm{BE}} > 2E$.

The tunneling hot electron transistor basically operates in a similar manner as the HET, except that the hot electrons tunnel through the AlGaAs barrier placed between the emitter and the base. This barrier serves to inject electrons from the emitter to the base. Most of these devices operate at temperatures lower than room temperature.

A resonant-tunneling hot electron transistor (RHET) was first proposed by Yokoyama et al. (1985). The principle of operation of the RHET is depicted in the conduction energy band diagram shown in Fig. 9.20. The structure is composed of a resonant-tunneling double-barrier structure inserted between the base and the emitter. The structural design is made such that there is a single bound state in the GaAs well. The base and collector are separated by a thick AlGaAs layer. In Fig. 9.20*a*, the collector-emitter bias voltage V_{CE} is kept constant while the base-emitter voltage V_{BE} is varied. This is a common emitter configuration. When V_{BE} is zero, there is no current flowing from the emitter to the base, and the emitter current I_E is zero as illustrated in the *I*-*V* curve on the right-hand side panel in the figure. The dot on each curve indicates the approximate current-voltage coordinates. A peak in the emitter current can be observed when eV_{BE} is equal to $2E$, where E is the bound energy level in the resonant tunneling structure as shown in Fig. 9.20*b*. For $eV_{\mathrm{BE}} > 2E$, the emitter current starts to decrease as V_{BE} increases, as shown in Fig. 9.20*c*. The *I*-*V* characteristic shown in this figure is similar to the negative differential resistance behavior encountered in the resonant tunneling diode discussed in Chap. 4.

While the first report on the RHET indicates that the device operates at cryogenic temperatures, it is possible to fabricate this class of device using different semiconductor systems, such as InAs/AlSbAs, that can operate at room temperature. The application of the RHETs can be realized in digital electronics and logic circuits. For example, if two inputs, A and B, are connected to the base, the output will be high (transistor is off) if both A and B are low or high. Otherwise, the output is low (transistor is on). This is an exclusive NOR logic circuit.

9.9 Coulomb Blockade and Single Electron Transistor

The investigation of low-dimensional systems has been a dominant force in recent years. The advances in semiconductor growth technology have enabled scientists and engineers to fabricate electronic and optoelectronic structures on a size scale where quantum effects, such as quantization, are the dominant factors in describing these structures' properties. Semiconductor quantum dots have dimensions much larger than the atomic scale, and a single dot may contain a few million atoms. Electrons in these dots are tightly bound and behave like waves with de Broglie wavelengths on the order of the dot size. The number of electrons occupying a quantum dot ranges between one and several hundred depending on the size of the dot. The energy required to add or remove an electron from the dot is called the *charging energy*. It is analogous to the ionization energy of an atom. While the atoms are investigated through their interaction with light, quantum dots are investigated using both light and by measuring their current-voltage characteristics.

The effect of a single electron was first observed by Millikan (1911) in his famous oil drop experiment, while electron tunneling was investigated by Gorter (1951), Giaever and Zeller (1968), and Lambe and Jaklevic (1969). Most of the theoretical aspects of electron charging effects and Coulomb oscillations were developed in the 1970s and 1980s. The invention of scanning tunneling microscopy (STM) had renewed the interest in Coulomb blockade since STM can both image the surface and measure the current-voltage characteristics of a single grain of size less than 10 nm. The lateral quantum dot defined by metallic surface gates has been widely investigated. The single-electron transistors operate at low temperatures ($T \ll 1$ K), but there are indications that these devices may operate at higher temperatures. According to Korotkov et al. (1995) the voltage gain is less than unity for temperatures as low as $0.015e^2/(k_BC) = (27.9 \text{ K})/(C \text{ aF})$. Thus, the unity gain condition places a stringent requirement on either the operating temperature or the size of the capacitance C. For $T = 300$ K, the capacitance is $0.015e^2/(300k_B) \approx 0.09$ aF. On the other hand, the lowest capacitance obtained experimentally is on the order of 0.1 fF.

The single-electron transistor describes a single electron transport through a quantum dot. There are many variations to the structure of the single-electron transistor. The main components of the single-electron structure are shown in Fig. 9.21. The island is the quantum dot which is connected to the drain and source terminals. Electron exchange occurs only with the drain and source terminals, which are connected to current and voltage meters. The gate terminal provides electrostatic or capacitive coupling. When there is no coupling to the source and drain, there is an integer number N of electrons in the quantum dot (island). The total charge on the island is quantized and equal to eN. If tunneling is allowed between the island, drain, and source terminals, then the number of electrons N adjusts itself until the energy of the total system is minimized. The tunneling junctions (barriers) are made thick enough so that the electrons exist in the island, source, or drain,

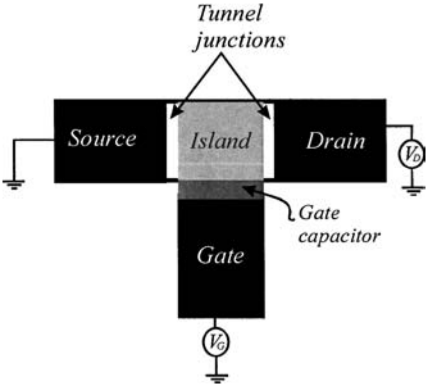


Figure 9.21 A sketch of a single-electron transistor showing the tunneling junctions, gate capacitor, island, and the three terminals (source, drain, and gate).

such that the quantum fluctuation in the number N due to tunneling through the barrier is much smaller than unity.

The electrostatically influenced electrons traveling between the source and the drain terminals need to tunnel through two junctions (barriers). The island is charged and discharged as the electrons cross it, and the relative energies of the island containing zero or one extra electron depends on the gate voltage. Thus, the charge of the island changes by a quantized amount e. The change in the Coulomb energy associated with adding or removing an electron from the island is usually expressed in terms of the island capacitance C. The charging energy E_c can be expressed as $E_c = e^2/C$. As mentioned in Chap. 4, this charging energy becomes important when it exceeds the thermal energy $k_B T$. The time Δt needed to charge or discharge the island can be expressed as $\Delta t = R_t C$, where R_t is the lower-bound tunnel resistance. From the Heisenberg uncertainty principle we have $\Delta E \Delta t = (e^2/C)(R_t C) > h$, or $R_t > h/e^2$. The quantity $h/e^2 = 25.813\ \text{k}\Omega$ is called the quantum resistance or quantum conductance ($G = 38.74\ \mu\text{S}$). Thus, two conditions must be met to observe the charge quantization:

$$R_t \gg \frac{h}{e^2} \qquad \text{and} \qquad \frac{e^2}{C} \gg k_B T \tag{9.65}$$

The capacitance can be made small by reducing the quantum dot size since $C = 4\pi\epsilon_s R$ for a sphere and $C = 8\pi\epsilon_s R$ for a flat disc, where R is the radius and ϵ_s is the permittivity of the material.

The gate voltage V_g is applied to change the island electrostatic energy in a continuous manner. The total gate voltage-induced charge on the island is expressed as $q = C_g V_g$. This charge is considered continuous. By sweeping the gate voltage, the buildup of induced charge will be compensated in a periodic interval due to the tunneling of discrete charges. The competition between the induced charges and the discrete compensation leads to so-called Coulomb oscillations. A typical example of these oscillations is shown in Fig. 9.22 where the conductance is measured as a function of the gate voltage for a fixed drain-source voltage. The sequence of the number of electrons residing in the dot (island) is $N \to N+1 \to N \to N+1 \to N \ldots$.

Consider that the gate voltage is fixed for the single-electron transistor while the drain-source voltage is varied. The current-voltage results exhibit a staircase-like behavior known as a Coulomb staircase. Figure 9.23 illustrates the Coulomb staircase behavior. A simple capacitor resistor circuit is shown in Fig. 9.23*a*, where the current as a function of applied voltage is a straight line and lacks the charge quantization. In Fig. 9.23*b*, we sketched a circuit that presents a single-electron transistor. The drain and source were presented as resistors and capacitors. In this circuit, the gate voltage is fixed while

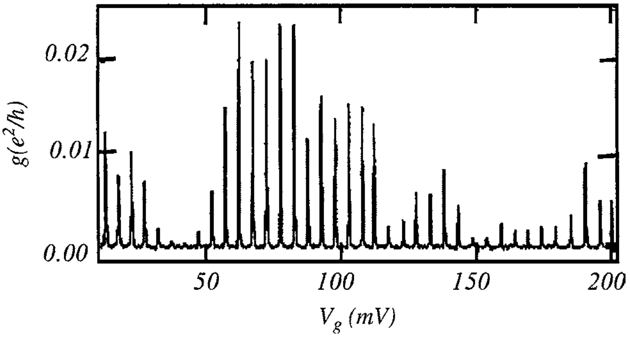


Figure 9.22 The conductance g as a function of the gate voltage V_g measured for a GaAs quantum dot showing Coulomb oscillations. (*After Folk et al. 1996*).

the drain-source voltage is varied. The drain current-source characteristic curve is shown as the steplike curve (gray line) in Fig. 9.23*c*. The black thin line in this figure is plotted for higher resistive tunnel junctions and lower temperatures. Notice that the drain voltage interval is e^2/C, where C is the sum of the gate, drain, and source capacitances. The steplike behavior in Fig. 9.23*c* is called the Coulomb staircase.

The basic operation of the single-electron transistor is shown in Fig. 9.24, where the tunneling junctions are presented as barriers. The island is represented as the well between the two barriers. The structure is symmetrical where the source and the drain can be exchanged without losing the transistor characteristic. A small drain-source voltage is applied, as shown in Fig. 9.24*a*. The solid lines in the island represent the occupied energy levels, while the dashed lines represent the empty energy levels. The energy required to move one electron from the full top energy level to the bottom empty level in the island can be derived as follows. The Fermi energy levels (chemical potentials) of the island (dot), the drain, and the source are shown in the figure as E_F^{dot}, E_F^D, and E_F^S, respectively. When a drain-source bias voltage is applied, these Fermi energy levels are no longer aligned, as shown in the figure. At zero temperature (zero thermal energy), the current is zero when the gap between the bottom empty state and the top full state is aligned with E_F^D and E_F^S. This is called Coulomb blockade. The minimum energy required to add an electron to the dot is $E_F^{\text{dot}} = E_N - E_{N-1}$ where N is the total number of electrons in the dot (island). For the linear response time, E_F^{dot} can be written as (Kouwenhoven et al. 1991).

$$E_F^{\text{dot}}(N) = E_N + \frac{e^2(N - N_o - 1/2)}{C} - e\frac{C_g}{C}V_G \tag{9.66}$$

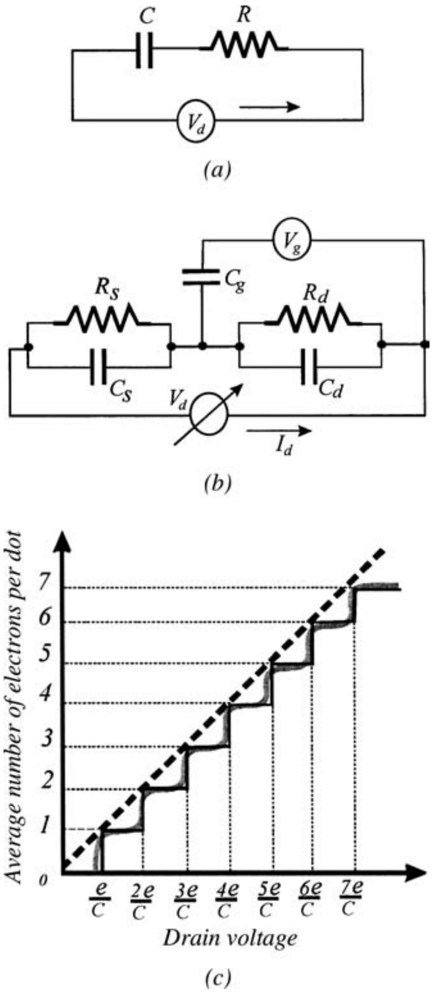


Figure 9.23 (*a*) A resistor-capacitor circuit with a bias voltage V_d. The dc current-voltage characteristic is shown as the dashed line in (*c*), which lacks charge quantization. (*b*) A representation of a single-electron transistor showing the resistances and capacitances of the drain and source. A fixed gate voltage is applied. (*c*) The drain current-voltage characteristic (gray line) of the circuit shown in (*b*). The steplike line is for more resistive tunnel junctions and at lower temperatures.

where N_o = number of electrons in island at $V_G = 0$
N = number of electrons in dot at gate voltage V_G
$C = C_g + C_s + C_d$

The energy E_N is a single particle energy for the Nth electron measured from the bottom of the conduction band. When an electron is added to the island at a constant gate voltage, the Fermi energy becomes

$$E_F^{\text{dot}}(N+1) = E_{N+1} + \frac{e^2\left(N - N_o + \frac{1}{2}\right)}{C} - e\frac{C_g}{C}V_G \qquad (9.67)$$

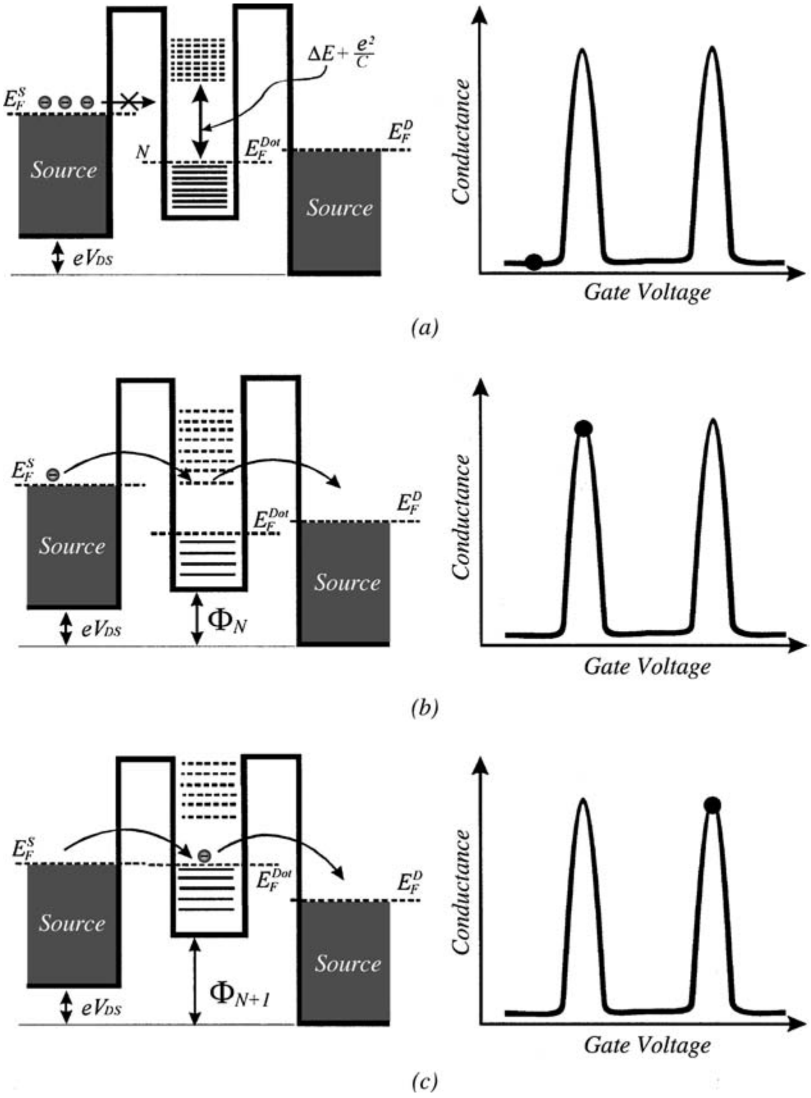


Figure 9.24 (*a*) A sketch of the single-electron transistor when Coulomb blockade exists. (*b*) An electron tunnel when the bottom energy level is aligned with E_F^S. (*c*) An electron tunnel when E_F^{dot} is aligned with E_F^S. Thus the sequence of electron tunneling is $N \to N+1 \to N \to N+1 \ldots$.

By taking the difference between Eqs. (9.66) and (9.67), one can find that

$$E_F^{\text{dot}}(N+1) - E_F^{\text{dot}}(N) = E_{N+1} - E_N + \frac{e^2}{C} = \Delta E + \frac{e^2}{C} \tag{9.68}$$

where $\Delta E = E_{N+1} - E_N$ is the energy separation between the energy levels in the quantum dot. Thus, the energy gap between the filled energy levels (lines in Fig. 9.24) and empty energy levels (dashed lines in Fig. 9.24) in the dot is composed of the energy separation between the energy levels that exist below $E_F^{\text{dot}}(N)$and the e^2/C. Notice that the e^2/C is a many-body contribution and it exists only at the Fermi energy level. Below the Fermi energy levels, the separation between the energy levels is only ΔE. If ΔE is too small, then the energy gap between the filled (solid lines) and empty (dashed lines) states is approximately e^2/C. The energy levels in the dot can be adjusted by applying a gate voltage as shown in Fig. 9.24*b* and *c* such that the electron can tunnel from the source to the drain. The conductance versus gate voltage is shown in the curve on the right in Fig. 9.24, where the dot on the curve represents the conductance and gate voltage coordinates. The minima of the conductance corresponds to the presence of the Coulomb blockade.

Another useful example is shown in Fig. 9.25 where a silicon-based single-electron transistor is sketched. The configuration of the single-electron transistor consists of a tellurium (Te) doped silicon material with SiO_2 between the aluminum and Si:Te. The aluminum layer forms the gate, source, and drain terminals. The gaps between the gate and drain and gate and source form the tunneling barriers. The island is formed under the gate metal. Contrary to the previous model where the energy levels in the island are assumed full and empty with an energy gap of $\Delta E + e^2/C$, the energy levels in the island in Fig. 9.25 are similar to the regular confined energy levels in a quantum well.

An electron tunnels from the source to the drain via the island when any of the energy levels in the island is aligned with the Fermi energy levels in the source. A small drain source voltage (V_{DS}) is applied such that the Fermi energy levels in both the drain and the source are almost the same. Thus, Coulomb blockade exists only when the Fermi energy level in the source is aligned between the energy levels in the island. The bottom of the potential in the island is adjusted by applying a gate voltage. When the gate voltage is increased, it forces one of the electrons on the Te atom to come close to the gate metal pushing the bottom of the island upward, as illustrated in Fig. 9.25. The conductance is increased when electrons tunnel from the source to the gate, as shown in the curves plotted on the right-hand side of the figure. To emphasize, the preceding description of single-electron tunneling is valid when the two conditions stated in Eq. (9.65) are satisfied (see Averin and Likharev 1991).

Inverter and complementary circuits based on single-electron transistors have been reported in the literature (see, for example, Korotkov et al. 1995). A typical example of an inverter is shown in Fig. 9.26. The V_{out} as a function of V_{in} is also shown in the figure with $C = C_g + C_s +$

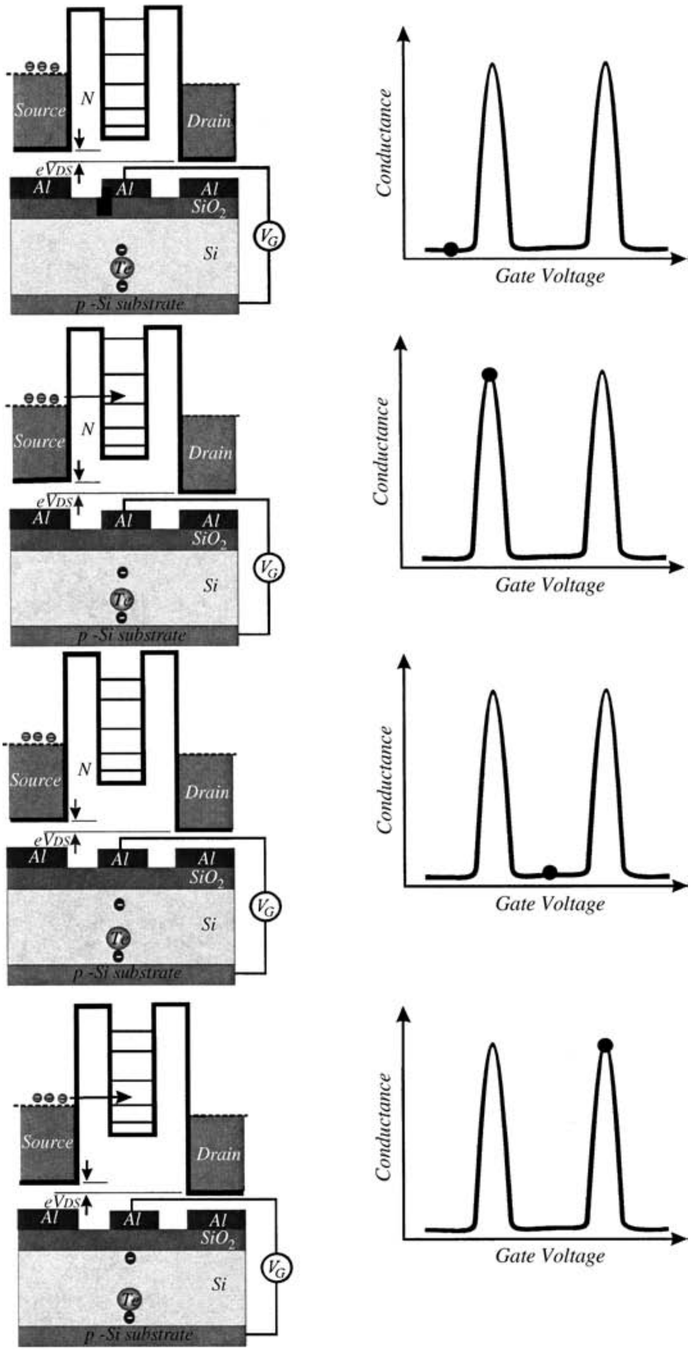


Figure 9.25 A schematic of a single-electron transistor based on Si:Te. The conductance as a function of the gate voltage at a fixed drain-source voltage is shown on the right-hand side where the dot represents the conductance-gate voltage coordinates.

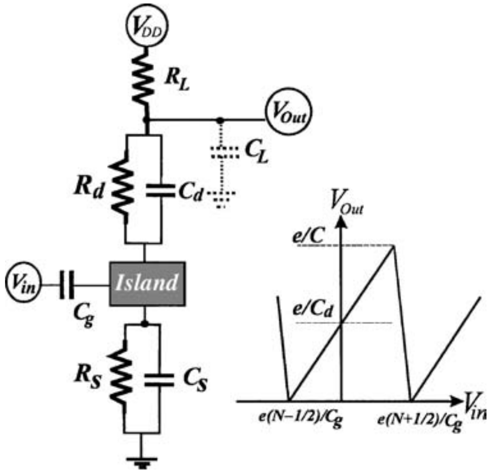


Figure 9.26 An illustration of a single-electron transistor as an inverter. The load is chosen as a resistance, but it can be an enhanced mode transistor. The V_{out} as a function of V_{in} is plotted as a sawtooth curve on the right-hand side.

C_d, $C_s = C_d$, $C_g = 8C_d$, $R_s = R_d = R_t$, and $V_{\text{DD}} \approx e/2C$. The maximum current gain (see Timp 1999) is approximately $1/(\omega C_G R_t)$, which could be very large depending on the frequency. The maximum theoretical voltage gain A_v is the slope of the falling portion of the sawtooth of the V_{out} curve as a function of V_{in}, which can be written as $A_v = -C_G/C_s = -8$. In practice this gain is much smaller.

The charge transport in a single-electron transistor is based on the quantum ballistic electron transport, which was briefly discussed in Chap. 7. This means that the single-electron transistor is a mesoscopic device where the coherence length is larger than the device dimension. The theory of ballistic transport is too complicated to be presented in this textbook. For a single-channel transport, the Landauer formula for the device conductance was derived in Chap. 7 and was given as

$$G = \frac{I}{V} = \frac{2e^2}{h} T(E_F) \tag{9.69}$$

where $T(E_F)$ is the electron transmission probability, E_F is the Fermi energy level, and the factor 2 is due to the spin degeneracy. This formula relates the quantum transmission probability directly to the conductance of the device. For a maximum conductance where $T(E_F) = 1$, we have $G = 77.44$ μS per spin.

The single-electron transistor can be considered as an electron waveguide which is schematically shown in Fig. 9.27. The gate voltage modulates the depletion region such that the penetration length L^* of the electron wave in the gate arm is changed. Thus, the electron transmission between the source and the drain is affected by changing the

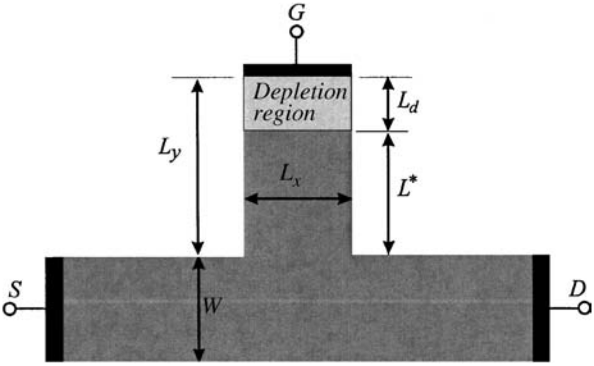


Figure 9.27 A planar view of a single-electron transistor as an electron waveguide.

length L^*. The drain current-voltage (see, for example, Sols et al. 1989) can be written as

$$I_D = \frac{2e^2}{h} T(E_F) V_D \tag{9.70}$$

The transconductance of the device is obtained by taking the first derivative of the drain current with respect to the gate voltage such that

$$g_m = \frac{\partial I_D}{\partial V_G} = \frac{2e^2}{h} V_D \frac{\partial T(E_F)}{\partial V_G} \tag{9.71}$$

The electron transmission probability is affected by the length L^*, which is modulated by the gate voltage. Thus, the transmission probability is also a function of L^* or $T(E_F, L^*)$, which means that Eq. (9.71) can be rewritten as

$$g_m = \frac{\partial I_D}{\partial V_G} = \frac{2e^2}{h} V_D \frac{\partial T(E_F, L^*)}{\partial L^*} \frac{\partial L^*}{\partial V_G} \tag{9.72}$$

The task now is to evaluate the terms $\partial T(E_F, L^*)/\partial L^*$ and $\partial L^*/\partial V_G$. The first term can be obtained by utilizing the tunneling probability reported by Glazman et al. (1988) for adiabatic constriction:

$$T(E_F, L^*) = \frac{1}{1 + \exp[-(k_F L^*/\pi - n)\pi^2 \sqrt{2R/L^*}]} \tag{9.73}$$

where k_F = Fermi wavevector
n = energy level quantum number
R = curvature of restriction

Using Eq. (9.73), Timp et al. (1999) obtained the following expression:

$$\frac{\partial T(E_F, L^*)}{\partial L^*} = \frac{0.35}{\sqrt{\mathcal{N}_s/(10^{12}\ \mathrm{cm}^{-2})}} \quad \mathrm{nm}^{-1} \tag{9.74}$$

where $\mathcal{N}_s$ is the density of the 2DEG in cm^{-2}. The second term, $\partial L^*/\partial V_G$, can be evaluated according to the following procedure (see Sols et al. 1989). The gate capacitance can be defined as

$$C_G = \left|\frac{\partial Q}{\partial V_G}\right| = e\mathcal{N}_s L_x \left|\frac{\partial L_d}{\partial V_G}\right| \tag{9.75}$$

where the lengths L_x and L_d are defined in Fig. 9.27 and $\mathcal{N}_s$ is the 2DEG density. From Fig. 9.27, one can write $L_y = L_d + L^* =$ constant, which yields

$$\frac{\partial L^*}{\partial V_G} = -\frac{\partial L_d}{\partial V_G} \qquad \text{or} \qquad \left|\frac{\partial L^*}{\partial V_G}\right| = \frac{C_G}{en_s L_x} \tag{9.76}$$

Substituting Eq. (9.76) into the transconductance Eq. (9.72) yields

$$g_m = \frac{2e^2}{h} V_D \frac{\partial T(E_F, L^*)}{\partial L^*} \frac{\partial L^*}{\partial V_G} = I_{Do} \frac{C_G}{e\mathcal{N}_s L_x} \frac{\partial T(E_F, L^*)}{\partial L^*} \tag{9.77}$$

where $I_{Do} = 2e^2 V_D/h$ is the drain current at $T(E_F, L^*) = 1$.

Using the definition of the cutoff frequency f_T given by Eq. (9.39), one can write the cutoff frequency of the single-electron transistor as

$$f_T = \frac{|g_m|}{2\pi C_G} = I_{Do} \frac{1}{2\pi e\mathcal{N}_s L_x} \left|\frac{\partial T(E_F, L^*)}{\partial L^*}\right| \tag{9.78}$$

The carrier density $\mathcal{N}_s$ can be written as

$$\mathcal{N}_s = \frac{\alpha}{W^2} \tag{9.79}$$

where W is given in Fig. 9.27 and α is a scaling constant that relates the Fermi energy levels to the zero-point energy E_o. According to Sols et al. (1989), $\alpha = E_F/E_o$ and $E_o = \hbar^2\pi^2/(2m^* W^2)$, which is related to higher subband energies E_n, such that $E_n = (n^2 - 1)E_o$. Substituting Eq. (9.79) into (9.78) yields

$$f_T = 7.693 \times 10^{13} \frac{V_D}{\alpha\beta} W \left|\frac{\partial T(E_F, L^*)}{\partial L^*}\right| \quad \mathrm{Hz} \tag{9.80}$$

where $\beta = L_x/W$. For $V_D = 30$ mV, $\alpha = 1.5$, $\beta = 2$, and $W|\partial T(E_F, L^*)/\partial L^*| = 1$, the cutoff frequency is 1.15 THz. This frequency is very high as compared with those of HFETs and HBTs.

The term $\partial L^*/\partial V_G = -\partial L_d/\partial V_G$ can be estimated by taking the first derivative of the depletion width L_d with respect to the gate voltage. The depletion length can be written for the single-electron transistor in a similar fashion as Eq. (9.25) by neglecting the drain voltage, i.e., $V(y) \approx 0$, or

$$L_d = \sqrt{\frac{2\epsilon_s(V_{\text{bi}} + V_G)}{eN_d}} \tag{9.81}$$

where the stub in Fig. 2.27 is terminated by a Schottky barrier and V_{bi} is the built-in voltage. Taking the derivative of Eq. (9.81) with respect to V_G yields

$$\left|\frac{\partial L^*}{\partial V_G}\right| = \left|\frac{\partial L_d}{\partial V_G}\right| = \frac{1}{2}\sqrt{\frac{2\epsilon_s}{eN_d}}\frac{1}{\sqrt{V_{\text{bi}} + V_G}} \tag{9.82}$$

If one assumes that $N_d = 10^{18}\ \text{cm}^{-3}$, this formula can be rewritten as

$$\left|\frac{\partial L^*}{\partial V_G}\right| = \frac{189.6}{\sqrt{V_{\text{bi}} + V_G}} \quad \text{Å}/V \tag{9.83}$$

where V_G and V_{bi} are given in volts. Thus, the stub length L^* can be modulated or controlled by the gate voltage. This implies that the basic operation of the single-electron transistor is determined by the control of the interference patterns of conducting electrons using an external gate voltage. The interference patterns are generated from traveling electrons in two or more channels.

To illustrate the formation of the interference patterns, consider the device shown in Fig. 9.28*a*. This device is simply an HFET with a channel short enough to permit ballistic transport without scattering; i.e., the coherence length is larger than the channel length. An embedded barrier is introduced parallel to the current, which splits the main channel into two channels. The symmetry of channels 1 and 2 is distorted when a gate voltage is applied. The widths of the channel and the split channels are small enough that the electron energy levels are quantized in the z direction. The first two subband energy levels are shown in Fig. 9.28*b*. Notice that the energy levels in the split channels are displaced from each other to indicate the broken symmetry between channels 1 and 2.

If an electron is injected from the source to the gate, its wave functions in the three regions (regions I, II, and III defined in Fig. 9.28*a*) of the

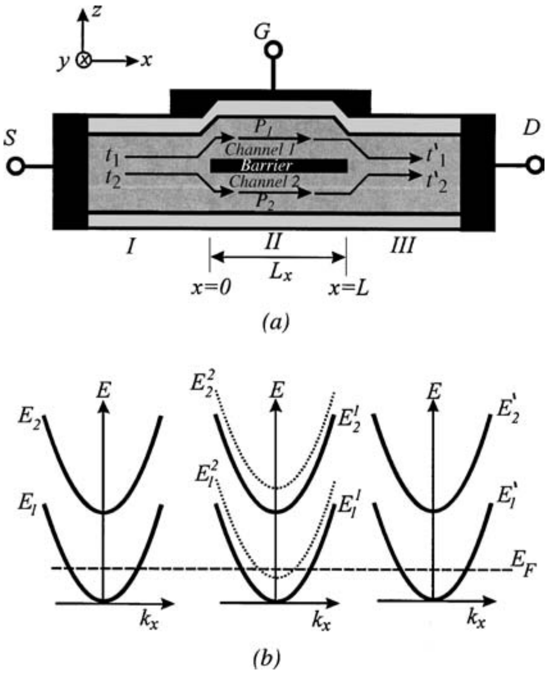


Figure 9.28 (*a*) A sketch of a single-electron transistor is shown with a barrier forming two electron channels. The parameters are defined in the text. (*b*) The electron subband energy levels are sketched for three different regions of the channel. The energy levels of the two channels in region II are displaced.

channel can be written according to Mitin et al. (1999) as

$$\psi = \begin{cases} u_{\mathrm{I}}(z)e^{i(k_x x+k_y y)} & \text{for region I} \\ t_1 u_{\mathrm{II}}^1(z)e^{i[k_x(x-L)+k_y y]} & \text{for region II (channel 1)} \\ t_2 u_{\mathrm{II}}^2(z)e^{i[k_x(x-L)+k_y y]} & \text{for region II (channel 2)} \\ (t_1 P_1 t_1' + t_2 P_2 t_2')u_{\mathrm{III}}(z)e^{i[k_x(L-x)+k_y y]} & \text{for region III} \end{cases} \tag{9.84}$$

where $u_{\mathrm{I}}(z)$, $u_{\mathrm{II}}^1(z)$, $u_{\mathrm{II}}^2(z)$, and $u_{\mathrm{III}}(z)$ are the wave functions of the subbands in their perspective regions. Quantities t_1 and t_2 are the amplitudes of the split waves in region II and t_1' and t_2' interfering waves in region III. Quantities P_1 and P_2 are the factors of the different phase shifts in the split channels (region II), and they can be taken as

$$P_1 = e^{ik_{x1}L} \qquad \text{and} \qquad P_2 = e^{ik_{x2}L} \tag{9.85}$$

where k_{x1} and k_{x2} are the wavevectors for channels 1 and 2, respectively. If multiple reflections between the source and drain are neglected and

$u_{\mathrm{I}}(z)$, $u_{\mathrm{II}}^1(z)$, $u_{\mathrm{II}}^2(z)$, and $u_{\mathrm{III}}(z)$ functions in Eq. (9.84) are normalized to unity, the transmission probability coefficient $T(E)$ can be written as

$$T(E) = 2(t_1 t_1')^2(1 + \cos\theta) \tag{9.86}$$

where θ is the relative phase shift in both channels and is given as

$$\theta = (k_{x1} - k_{x2})L \tag{9.87}$$

Notice that the channel is assumed to be symmetrical around $z = 0$ where the wave amplitudes satisfy the conditions $t_1 = t_2$ and $t_1' = t_2'$ for zero applied gate voltage.

If the average electron velocity υ_x in the split channels (region II in Fig. 9.28*a*) is expressed as

$$\upsilon_x = \frac{\hbar(k_{x1} + k_{x2})}{2m^*} \tag{9.88}$$

then the relative phase shift can be written as

$$\theta = \frac{L}{\upsilon_x}\frac{(E_1^1 - E_1^2)}{\hbar} \tag{9.89}$$

where L is the length of the channel as shown in Fig. 9.28*a*, and E_1^1 and E_1^2 are the ground energy levels for channels 1 and 2, respectively. Notice that the superscripts 1 and 2 are introduced to indicate channels 1 and 2. It is clear from Eq. (9.89) that if the ground states of the two channels are different, i.e., $E_1^1 \neq E_1^2$, then different phases of waves traveling from the source to the gate exist. The difference in the phases causes the quantum interference.

For a zero gate voltage the ground states of the two channels in region II of Fig. 9.28*a* are equal ($E_1^1 = E_1^2 = E_o$) which leads to a zero phase shift. When a nonzero gate voltage is applied, the potential energy $U(z)$ changes according to the following relation:

$$U(z) = U_o(z) - e\varphi(z) \tag{9.90}$$

where $U_o(z)$ is the potential energy at zero gate voltage and $\varphi(z)$ is the potential induced by the applied gate voltage. This induced potential causes a shift to the subband energy levels in the two channels according to the following relations:

$$\begin{aligned} E_1^1 &= E_o - e\langle u_{\mathrm{II}}^1|\varphi(z)|u_{\mathrm{II}}^1\rangle \\ E_1^2 &= E_o - e\langle u_{\mathrm{II}}^2|\varphi(z)|u_{\mathrm{II}}^2\rangle \end{aligned} \tag{9.91}$$

Substitute Eq. (9.91) into (9.89) to obtain

$$\theta = \frac{L}{\upsilon_x}\frac{e\varphi_{12}}{\hbar} \tag{9.92}$$

where φ_{12} is given by

$$\varphi_{12} = \langle u_{\text{II}}^2|\varphi(z)|u_{\text{II}}^2\rangle - \langle u_{\text{II}}^1|\varphi(z)|u_{\text{II}}^1\rangle \tag{9.93}$$

Thus, the electron transmission in the device is determined by the value of φ_{12}.

For the case of the device described in Fig. 9.28, the Landauer formula can be obtained by combining Eqs. (9.69) and (9.86) to obtain the following expression for the transconductance:

$$G = \frac{4e^2}{h}(t_1 t_1')^2(1 + \cos\theta) \tag{9.94}$$

The conductance is an oscillatory function of the relative phase shift. The transconductance g_m is maximum when $\theta = n\pi$, where $n = 0, 2, 4, \ldots$, and zero when $\theta = m\pi$, where $m = 1, 3, 5, \ldots$. When θ is zero, Eq. (9.92) implies that φ_{12} is also zero. But when $\theta = \pi$, then $e\varphi_{12} = \hbar\pi\upsilon_x/L$, which corresponds to destructive interference.

The transit time τ_{tr} required for the electron to cross the channel of length L can be defined as $\tau_{\text{tr}} = L/\upsilon_x$, which leads to a cutoff frequency similar to the frequency form given by Eq. (9.41) and is given by

$$f_T = \frac{1}{2\pi\tau_{\text{tr}}} = \frac{\upsilon_x}{2\pi L} \tag{9.95}$$

For a device with a gate length of $L = 100$ Å and an average velocity of $\upsilon_x = 10^7$ cm/s, the cutoff frequency is ~1.6 THz. This frequency is easily achieved in the single-electron transistor provided that the channel is undoped, the transport is ballistic, and the channel is made too short (shorter than the coherence length).

Summary

In this chapter, we have mainly discussed the electronic devices (transistors) based on heterojunctions, quantum wells, and quantum dots. From the discussion in this chapter, one can conclude that there are three generations of electronic devices. The first generation includes the metal-semiconductor field-effect transistors and junction field-effect transistors. These devices are based on doped semiconductors, and therefore they are the slowest of the three generations. The presence of dopants degrades the mobility and transport properties, and therefore the cutoff frequency is low. The second generation of field-effect transistors is based on quantum wells, where the channel of the device is the two-dimensional electron gas formed at the interfaces of the heterojunctions or in quantum wells. An example of this class of devices is the heterojunction field-effect devices, such as GaAs/AlGaAs and GaN/AlGaN HFETs. The frequency of these devices can reach the gigahertz range. This high cutoff frequency is due to the high mobility of the two-dimensional electron gas as compared to bulk mobility. Most of the HFETs contain a two-dimensional electron gas with a mobility larger than 10^5 $\text{cm}^2/(V \cdot \text{s})$. The GaN/AlGaN HFETs were discussed in more detail because of their potential use in high-frequency, high-temperature, and high-power applications.

The heterojunction bipolar transistors were also discussed since they can operate in the gigahertz frequency range. Another class of devices based on tunneling effects, which includes tunneling electron transistors, resonant tunneling transistors, and hot electron transistors, was discussed briefly. The latter class of devices is unipolar, which means they are based only on n-type materials. The disadvantage of this class of transistors is that most of them operate at a temperature lower than room temperature, but their high cutoff frequency and potential for digital applications make them attractive devices to investigate.

The third generation of the electronic devices discussed in this chapter is the single-electron transistor, which consists of an island or quantum dot, drain, source, gate, gate capacitance, and two barriers. The single-electron transistors operate in the terahertz frequency range, providing that the temperature is less than 1.0 K. The limiting factor of this operational temperature is the gate capacitance. Significant research is currently being conducted on single-electron transistors for their applications in quantum computing and molecular computing.

While the theory of single-electron transistors is very complicated, we presented simplistic ideas on how these devices operate. The operation of single-electron transistors is based on charge quantization. To observe this charge quantization, two important conditions must be met. First, the lower-bound tunneling resistance should be much larger than the quantum resistance (h/e^2) and second, the charging energy (e^2/C) must be much larger than the thermal energy $(k_B T)$.

Problems

9.1 Calculate the depletion width at the interface of a nickel film deposited at the top of an n-type doped GaAs film with a dopant concentration of 5×10^{18} cm^{-3}. The electron affinity in GaAs is $\chi = 4.07$ V. What would be the depletion width if silver is used instead of nickel? Does silver form a better ohmic contact than nickel?

9.2 Show that the width of the depletion region in the Schottky barrier diode is given by Eq. (9.11). A Schottky barrier diode is made of chromium metal deposited on a GaAs surface. Calculate the depletion width in GaAs at room temperature for the following bias voltage values: $V = -1.0, 0.0$, and $+0.5$ V.

Assume that the GaAs material is uniformly doped with $N_d = 10^{16}$ cm^{-3} and the dielectric constant of GaAs is 12.25.

9.3 Consider Fig. P9.3 where $1/C^2$ is plotted as a function of the bias voltage for a Schottky barrier diode made of tungsten and GaAs. Calculate the carrier concentration from the graph. Estimate the Schottky barrier height.

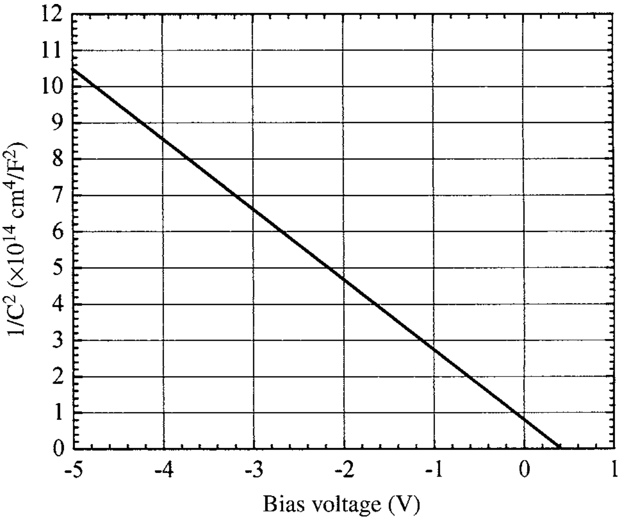


Figure P9.3

9.4 Show that the effective Richardson constant A^* is given by Eq. (9.18).

9.5 Show that the drain current of an n-channel MESFET is given by Eq. (9.28).

9.6 Show that the transconductance g_m for an n-channel MESFET is given by Eq. (9.33) for $V_D < V_D^{\text{sat}}$.

9.7 Derive the expression of the transconductance of an n-channel MESFET for the normally off condition, as shown in Eq. (9.35).

9.8 What is the cutoff frequency of an n-channel MESFET that is composed of a 1-μm-thick n-type epitaxial layer of GaAs grown on a semi-insulating GaAs substrate? The gate length is 20 μm, the dopant concentration is 7×10^{16} cm^{-3}, and the electron mobility is 2500 cm^2/(V · s).

9.9 High-power MESFETs usually operate under high drain-source voltages. Assume that the desired drain-source breakdown voltage is 75 V. Calculate the gate length and the corresponding cutoff frequency for the following materials: silicon (drift velocity is 5×10^6 cm/s and breakdown field is 5×10^5 V/cm), GaAs (drift velocity is 1×10^7 cm/s and breakdown field is

1×10^6 V/cm), GaN (drift velocity is 5×10^6 cm/s and breakdown field is 4×10^6 V/cm), and SiC (drift velocity is 2×10^7 cm/s and breakdown field is 5×10^6 V/cm).

9.10 Consider a GaAs/$Al_{0.3}Ga_{0.7}As$ MODFET, where the AlGaAs barrier is silicon-doped, producing a carrier concentration of $N_d = 1 \times 10^{18}$ cm^{-3}. Assume that the doped AlGaAs layer is 400 Å, the spacer is 100 Å, the Schottky barrier height is $V_{\text{Bn}} = 0.7$ V, the triangular quantum well width is 60 Å, the conduction band offset is 0.3 eV, and the AlGaAs refractive index is 3.5. Calculate the 2DEG density at zero gate voltage and zero channel potential.

9.11 Derive an expression for the transconductance g_m of a MODFET in the nonsaturation region.

9.12 Show that the cutoff frequency of a single-channel MODFET is given by Eq. (9.58). Consider a GaAs/AlGaAs MODFET where $d = 40$ nm, $d_o = 10$ nm, and $\Delta d = 6$ nm. Assume that the parasitic capacitance is small (1×10^{-15} F) and the carrier saturation velocity is 1×10^7 cm/s. Calculate the cutoff frequency for a channel of length 0.2 μm and of width 50 μm.

9.13 Calculate the density of the 2DEG in an undoped 100 Å GaN/$Al_{0.3}Ga_{0.7}N$ HFET. Assume that the band offset is 0.4 eV, the Fermi energy is 0.1 eV above the bottom of the conduction band, the Schottky barrier height is 0.75 eV, and the dielectric constant of $Al_{0.3}Ga_{0.7}N$ is 2.2.

9.14 Assume that the drift velocity in the base of an HBT is given by $v(x) = D_n/x$ for $0 < x < W$, where D_n is the electron diffusion coefficient and W is the width of the base. Show that the transient time required for an electron to cross the base is given by Eq. (9.64).

9.15 A single-electron transistor is fabricated from a GaAs/AlGaAs heterojunction. The island is assumed to be spherical with a radius of 0.1 μm. Calculate the charging energy assuming that the refractive index of GaAs is 3.5. What is the charging energy if the dot is a flat disc with the same radius? What is the radius of the spherical island needed for the transistor to operate at $T \leq 125$ K?

9.16 Consider a single-electron transistor with a gate capacitance that can be represented as two parallel plates with a material having a dielectric constant of 13 and an area of 100 nm^2. Calculate the separation between the plates that can yield a capacitance of 1.0 aF.

9.17 Consider a single-electron transistor where the condition to produce oscillations in the conductance as a function of the gate voltage V_G is given as $E_F^{\text{dot}}(N, V_G) = E_F^{\text{dot}}(N+1, V_G + \Delta V_G)$. (*a*) Derive an expression for the change in gate voltage between oscillations. (*b*) Use the initial conditions $E_F^{\text{dot}}(N) = -N_o e^2/C$ and $E_{\text{No}} = 0$ when $V_G = 0$ to derive an expression for the gate voltage for the Nth conductance peak.

9.18 Calculate the cutoff frequency of a single-electron transistor using the following parameters: $a = 1.5$, $b = 1$, $W|\partial T(E_F, L^*)/\partial L^*| = 4$, and $V_D = 45$ mV. Design a single-electron transistor that possesses a cutoff frequency of 2.0 THz for a drain voltage of 50 mV.

9.19 Show that the electron transmission coefficient for a single-electron transistor with a split channel, as shown in Fig. 9.28, is given by Eq. (9.86). Plot $T(E)$ as a function of the relative phase shift θ. Explain the results.

9.20 Show that when the ground energy levels of the split channels in region II of Fig. 9.28*a* are different, the relative phase shift θ is given by Eq. (9.89).

9.21 The difference between the average potential in the two channels (region II) in the single-electron transistor shown in Fig. 9.28*a* is given by $e\varphi_{12} = \hbar\pi\upsilon_x/L$ for a zero transconductance, where υ_x is the electron velocity and is given by $\upsilon_x = \hbar k_F/m^*$ where k_F is the Fermi wavevector. The voltage φ_{12} can be considered as the threshold voltage needed to destruct the interference pattern in the single-electron transistor. Show that $e\varphi_{12} = eV_T^{\text{FET}}\lambda_F/L$ where V_T^{FET} is the threshold voltage of a typical field-effect transistor and λ_F is the de Broglie wavelength associated with the Fermi energy. Consider that $eV_T^{\text{FET}} = E_F$ where E_F is the Fermi energy level.

This page intentionally left blank.

Chapter

10

Optoelectronic Devices

10.1 Introduction

Semiconductor heterojunctions and nanostructures have been investigated for their applications in electronic and optoelectronic devices. This chapter is directed toward the optoelectronic devices, such as detectors and emitters. There is a myriad of applications for the optoelectronic devices including 1.31- and 1.55-μm optical communications where the silica fibers exhibit the lowest losses, terahertz applications, infrared and long-wavelength infrared detectors, and multijunction solar cells. One of the mechanisms used to generate light from a semiconductor is the radiative recombination of electrons and holes across the fundamental bandgap, which gives rise to photon emission. Soon after the invention of the laser, *pn*-junction GaAs lasers were demonstrated with an emission in the 0.827- to 0.886-μm spectral range, which is basically limited by the bandgap of the GaAs material. This spectral range is outside the energy spectrum visible to the human eye. Substantial research has been performed since then toward the development and production of emitters in the visible, ultraviolet, and infrared spectral regions (see Fig. 10.1 for various spectral region limits). For example, recent research efforts in III-nitride semiconductor materials lead to the production of blue and green light-emitting diodes (LEDs) and diode lasers. Further research has pushed the performance of the III-nitride materials to the ultraviolet and far ultraviolet LEDs.

The development of infrared emitters goes back to the 1960s and continues to develop as new applications and needs emerge. In theory, the wavelength of emitters based on interband transitions in semiconductors can be constructed to cover long and very long wavelength spectral regions. This is due to the fact that one can grow ternary and quaternary alloys with a precise control on the composition using the latest

Copyright © 2005 by The McGraw-Hill Companies, Inc. Click here for terms of use.

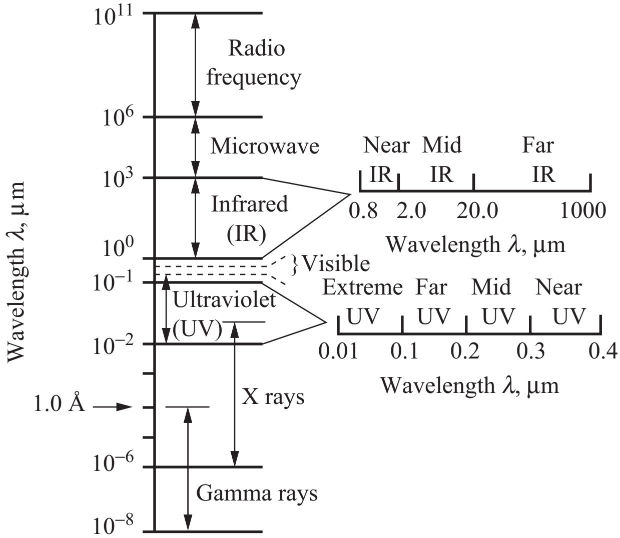


Figure 10.1 Wavelengths of various spectral regions. Notice that the spectral region visible to the human eye is very narrow.

growth technology, such as the MBE and MOCVD growth techniques. The limiting factors for III-V semiconductors, however, are the nonradiative recombination processes and the internal losses due to free carrier absorption. These factors become significant and detrimental to the device performance as the bandgap of the semiconductor decreases (or increases in wavelength). Advances in the long-wavelength infrared diode lasers continue in the spectral range of 3 to 30 μm using IV-VI compound semiconductors, such as PbTe, which have their nondegenerate direct bandgap located at the *L*-point of the first Brillouin zone. The nonradiative recombination processes at this point, such as Auger recombination processes, are less likely to happen.

There are many types of light sources that operate using different operational principles. The Gunn diode operates on the principle of fast charge oscillations, which act as a classical dipole source. This type of diode emits electromagnetic waves with wavelengths larger than 1000 μm (microwave source). Optically pumped gas lasers usually emit light in the 40- to 1000-μm spectral range. Generally speaking, the lasers can be categorized into four major classes: (1) gas lasers such as HeNe, CO_2, and argon-ion; (2) dye (liquid) lasers such as oxazine and polyphenyl; (3) solid-state lasers such as Ti:sapphire, Nd:YAG, and ruby; and (4) semiconductor lasers. This chapter, however, is focused on lasers based on interband and intersubband transitions in semiconductor quantum wells and quantum dots, such as edge-emitting lasers, vertical cavity surface emitting lasers, and quantum cascade lasers.

Another class of optoelectronic devices, which has a history going back to the 1800s, is the infrared detectors. The first infrared detector was the thermometer, discovered by Hershel in 1800. Nowadays, there are many types of infrared detectors ranging from a single element to large focal plane arrays and from thermal detectors to multiple-quantum-well infrared detectors. Again, the discussion here focuses on quantum detectors based on interband and intersubband transitions in semiconductor quantum wells and dots.

10.2 Infrared Quantum Detectors

The two major categories of infrared detectors are thermal and photonic (quantum) detectors. The principle of operation of the thermal detectors is based on the temperature change of the detector materials upon absorbing the photons. Accompanying this change in temperature is a change in at least one physical property of the material, which leads to generation of an electrical signal. In general, the electrical output signal of thermal detectors is independent of the incident photon wavelength, but of course depends on the radiant power. An example of thermal detectors is the bolometer, which could be made of metal, semiconductor, superconductor, or ferroelectric material. The physical parameter that changes when photons are absorbed is the electrical conductivity or electrical resistivity. Other examples of thermal detectors are as follows: (1) thermocouples, where voltage generation is caused by a change in the temperature of the junction of two dissimilar materials; (2) Golay cells, based on the change in the thermal expansion of the gas; (3) pyroelectric detectors, in which the spontaneous polarization is changed when absorbing photons; and (4) pyromagnetic detectors, where the name implies that the magnetic properties are changed.

Quantum detectors, on the other hand, operate on the principle of electron-photon interaction. Thus, these detectors are much faster than the thermal detectors. There are two basic processes involved in quantum detectors. First, conduction electrons or electrons bound to the lattice atoms or impurities absorb light and get excited to higher energy levels. Second, the excited electrons are swept by an applied bias voltage and collected as an electrical signal. Ideally all excited electrons can be collected, leading to 100 percent quantum efficiency. This means that each photon absorbed excites an electron and all the excited electrons are collected under bias voltage. In practice, 100 percent efficiency is very difficult to achieve due to the fact that many of the excited electrons recombine with the holes, are trapped by positively charged ions, or lose their energy as phonons. Thermal generation of charge carriers (dark current) can be significantly reduced by cryogenic cooling.

10.2.1 Figures of merit

There are too many variables, such as electrical, radiometric, and device design parameters, involved when measuring the photoresponse of a detector of regardless whether it is a thermal or photonic detector. Thus, it is difficult to measure the performance of the device. Several figures of merit have evolved over the years that are used to characterize and quantify infrared detectors. While some figures of merit that were developed may not be of any use to quantify many of the quantum detectors, the currently accepted figures of merit are briefly discussed in this section.

10.2.1.1 Responsivity. Responsivity of a detector is difficult to quantify especially when comparing various detectors. For example, the photoresponse of thermal detectors is independent of the photon wavelength, but the responsivity of a quantum detector is a linear function of the wavelength. For quantum detectors, the responsivity is defined as the ratio between the detector output electrical signal and the input radiant optical power. The detector output signal is either voltage or current. For the voltage output signal, the spectral responsivity is given as

$$\mathcal{R}_v(\lambda, f) = \frac{V_s}{P_{\text{in}}(\lambda)} \tag{10.1}$$

where $\mathcal{R}_v(\lambda, f)$ is the voltage spectral responsivity which is a function of the incident photon wavelength λ and the operating electrical chopping frequency f, V_s is the output signal voltage, and $P_{\text{in}}(\lambda)$ is the spectral radiant input power given by

$$P_{\text{in}}(\lambda) = \frac{A_d \Phi_{\text{ph}} hc}{\lambda} \tag{10.2}$$

where A_d = area of detector
Φ_{ph} = incident photon flux density, photons/($\text{m}^2 \cdot$ s)
λ = incident light wavelength.

The current responsivity $\mathcal{R}_i(\lambda, f)$ is similar to Eq. (10.1) and can be written as

$$\mathcal{R}_i(\lambda, f) = \frac{I_{\text{ph}}}{P_{\text{in}}(\lambda)} \tag{10.3}$$

where I_{ph} is the photocurrent. The spectral responsivity expressed in Eqs. (10.1) and (10.3) should be multiplied by the photoconductive gain, which is defined as the ratio between the recombination time and the transit time. For now, we assume the photoconductive gain is maximum (unity).

The blackbody responsivity is a very useful parameter and is defined as the output of a detector produced in response to a watt of input optical radiation from a blackbody at temperature T modulated at electrical frequency f. The blackbody source is usually calibrated and standardized at specific temperatures. For example, infrared blackbody sources are calibrated at 500 K, while the near infrared and visible sources are calibrated at 2856 K (see Dereniak and Boreman 1996). When measuring the blackbody responsivity, the radiant power on the detector contains all wavelengths of radiation regardless of the spectral response curve of the detector. Thus, the current responsivity shown in Eq. (10.3) can be modified as

$$\mathcal{R}_i(T, f) = \frac{I_{\text{ph}}}{\int_0^{\infty} P_{\text{in}}(\lambda)\, d\lambda} \tag{10.4}$$

Notice that the blackbody responsivity depends on temperature T and f. Notice that when Fourier-transform spectroscopy is used to measure the responsivity of the detector, the chopping frequency becomes irrelevant, since choppers are not used. Furthermore, the entire flux incident on the detector appears in the calculations of blackbody responsivity. The responsivity is usually a function of the bias voltage, f, and λ.

10.2.1.2 Noise equivalent power. The responsivity is a good figure of merit used to estimate an expected signal level for a given radiant power on the detector. However, it does not provide useful information regarding the sensitivity of the detector. In addition to the signal level, the noise level is important. The question that one may ask is what is the minimum radiant flux level a detector can measure? The detector output photocurrent due to an input power must be larger than the noise current level. The signal ($\mathcal{S}$) to noise (N) ratio can be defined as

$$\frac{\mathcal{S}}{\mathcal{N}} = \frac{\mathcal{R}_i \Phi_e}{i_n} \tag{10.5}$$

where Φ_e = radiant flux, W

i_n = noise current

$\mathcal{R}_i$ = current responsivity, A/W

The noise-equivalent power (NEP) is the radiant power incident on the detector that produces $\mathcal{S}/N = 1$. Setting Eq. (10.5) equal to unity yields the NEP:

$$\Phi_e = \text{NEP} = \frac{i_n}{\mathcal{R}_i} \tag{10.6}$$

Similarly, the NEP can be written in terms of voltage responsivity $\mathcal{R}_v$ as

$$\Phi_e = \text{NEP} = \frac{V_n}{\mathcal{R}_v} \tag{10.7}$$

where V_n is the noise voltage. When the responsivity is a spectral responsivity, the NEP is called spectral NEP. The term blackbody NEP is used when the blackbody responsivity is used. The unit of NEP is the watt, and a more sensitive detector has a lower NEP. In this sense, the NEP is a defect function rather than a figure of merit. In addition to the NEP, a detector performance is measured by other parameters, such as the optimum bias voltage, operating temperature, and noise-equivalent bandwidth.

10.2.1.3 Detectivity. The normalized spectral detectivity, known as D^*, is another important figure of merit, and is defined as

$$D^* = \frac{\sqrt{A_d\,\Delta f}}{\text{NEP}} \tag{10.8}$$

where Δf is the noise equivalent bandwidth and A_d is the area of the detector. The larger D^* is, the better the detector. Usually D^* is normalized to $A_d = 1\ \text{cm}^2$ and $\Delta f = 1$ Hz, and it can be interpreted as the signal-to-noise ratio (SNR) of the detector when 1 W of radiant power is incident on a 1-cm^2 active area of the detector given a noise-equivalent bandwidth of 1.0 Hz. The units of D^* is Jones = cm $\sqrt{\text{Hz}}$/W. The detectivity can be written in different forms such as

$$D^* = \frac{\sqrt{A_d\,\Delta f}}{\text{NEP}} = \frac{\sqrt{A_d\,\Delta f}}{V_n/\mathcal{R}_v} = \frac{\sqrt{A_d\,\Delta f}}{i_n/\mathcal{R}_i} = \frac{\sqrt{A_d\,\Delta f}}{\Phi_e}\frac{\mathcal{S}}{\mathcal{N}} \tag{10.9}$$

This equation implies the following definition of D^*: It is the root mean square (rms) signal-to-noise ratio of a detector of 1.0 cm^2 area in a 1-Hz bandwidth per unit rms incident radiant power.

The blackbody detectivity $D^*(T)$ can be obtained from the spectral detectivity according to the following relation:

$$D^*(T) = \frac{\int_0^\infty D^*\Phi_e(T,\lambda)\,d\lambda}{\int_0^\infty \Phi_e(T,\lambda)\,d\lambda} = \frac{\int_0^\infty D^* Q_B(T,\lambda)\,d\lambda}{\int_0^\infty Q_B(T,\lambda)\,d\lambda} \tag{10.10}$$

where $\Phi_e(T,\lambda) = Q_B(T,\lambda)A_d$ is the incident blackbody radiant flux in watts and $Q_B(T,\lambda)$ is the blackbody irradiance in W/cm^2.

When the detector noise is low compared to the photon noise, the detector is said to have reached its maximum performance. The photon noise arises from the detection process as a result of the discrete nature of the radiation field. For most infrared detectors, the practical operating limit is determined by the background fluctuation limits and not by the signal fluctuation limits. Thus, when the background photon flux is much larger than the signal flux, the photon flux is the dominant noise source. This condition is called background limited infrared performance (BLIP).

The spectral detectivity can be derived under the BLIP conditions by considering the photon irradiance from the signal source Q_s and the background Q_B. The rms noise current i_n can be established by assuming that the shot noise in the detector is due to dc-photogenerated current I_{ph} flowing across a potential barrier as

$$i_n = \sqrt{2e\bar{I}\,\Delta f} \tag{10.11}$$

It was first shown by Schottky in 1918 that the random arrival of electrons on the collecting electrode of a vacuum tube was responsible for what is called shot noise. The total photocurrent $\bar{I}$ caused by both background and signal photons is given as

$$\bar{I} = e\eta\,(Q_s + Q_B)\,A_d \tag{10.12}$$

where η is the quantum efficiency defined as the ratio between the electron-hole pairs generated per incident photon. Substituting Eq. (10.12) into (10.11) and assuming that $Q_s \ll Q_B$ yields

$$i_n = \sqrt{2e^2\eta Q_B A_d\,\Delta f} \tag{10.13}$$

The assumption that $Q_s \ll Q_B$ is valid since we are trying to determine the minimum detectable signal. The SNR can now be written as

$$\frac{\mathcal{S}}{\mathcal{N}} = \frac{I_{\mathrm{ph}}}{i_n} = \frac{e\eta Q_s A_d}{\sqrt{2e^2\eta Q_B A_d\,\Delta f}} \tag{10.14}$$

where I_{ph} is the photocurrent due to the signal alone. In order to detect the minimum detectable signal irradiance, the SNR in Eq. (10.14) is set to unity, which yields

$$Q_s = \sqrt{\frac{2Q_B\,\Delta f}{\eta A_d}} \tag{10.15}$$

This equation can be considered as the noise-equivalent irradiance on the surface of the detector. Using the definition of the power given by

Eq. (10.2), one can write the signal power as

$$\Phi_{e,\text{signal}} = Q_s \frac{hc}{\lambda} A_d \tag{10.16}$$

Since Q_s is determined by setting Eq. (10.14) to unity, it is defined as the noise equivalent irradiance. This implies that the power defined in Eq. (10.16) is the noise-equivalent power (NEP) or

$$\text{NEP}(\lambda) = \sqrt{\frac{2Q_B \Delta f}{\eta A_d}} \frac{hc}{\lambda} \tag{10.17}$$

Combining Eqs. (10.8) and (10.17), we obtain a definition of the detectivity at the BLIP conditions:

$$D^*_{\text{BLIP}} = \frac{\sqrt{A_d \Delta f}}{\text{NEP}(\lambda)} = \frac{\lambda}{hc} \sqrt{\frac{\eta}{2Q_B}} \tag{10.18}$$

The background photon irradiance Q_B (or the background photon flux density) reaching the detector can be written as (see, for example, Dereniak and Boreman 1996)

$$Q_B = \sin^2\left(\frac{\theta}{2}\right) \int_0^{\lambda_c} Q(\lambda, T_B)\, d\lambda \tag{10.19}$$

where $\sin^2(\theta/2)$ = numerical aperture
λ_c = cutoff frequency
T_B = blackbody temperature
$Q(\lambda, T_B)$ = Planck's photon emittance, photons ($\text{cm}^{-2} \cdot \text{s}^{-1} \cdot \mu\text{m}^{-1}$)

Planck's photon emittance is given by (see, for example, Hudson 1969)

$$Q(\lambda, T_B) = \frac{2\pi c}{\lambda^4 [e^{hc/(\lambda k_B T_B)} - 1]} = \frac{1.885 \times 10^{33}}{\lambda^4 [e^{14,388/(\lambda T_B)} - 1]} \tag{10.20}$$

where λ is given in microns. A plot of $Q(\lambda, T_B)$ is shown in Fig. 10.2 for different values of T_B. The integral in Eq. (10.19) can be obtained numerically. For additional analysis of the detectivity at the BLIP conditions see Rogalski (2000) and Dereniak and Boreman (1996).

10.2.1.4 Noise-equivalent temperature difference. Many of the long-wavelength infrared detectors are used as thermal imagers. One of the figures of merit used to describe the performance of the thermal imaging system is the noise-equivalent temperature difference (NETD).

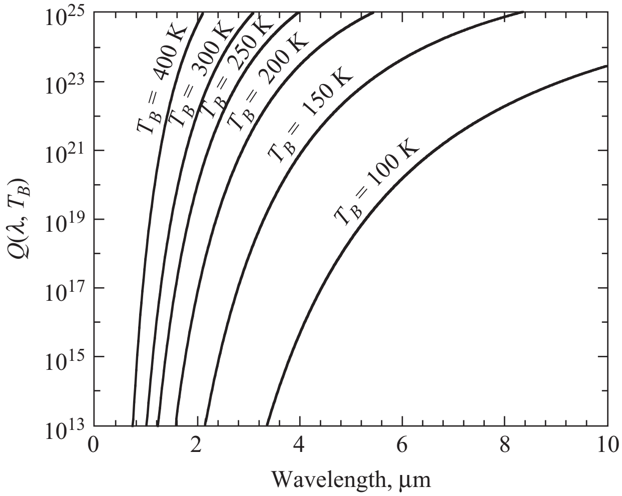


Figure 10.2 Planck's photon emittance is plotted as a function of the wavelength for different blackbody temperatures.

Thermal imager systems are used to map the temperature difference related to spatial flux and emissivity differences across an extended object. Thus, thermal sensitivity is concerned with the minimum temperature difference that can be distinguished above the noise level. The thermal imaging systems have been discussed and presented by Lloyd (1975). The NETD is derived by many authors (see, for example, Lloyd, 1995, Dereniak and Boreman 1996, and Rogalski 2000) and is given as

$$\text{NETD} = \frac{4}{\pi}\left[\frac{(F/\#)^2\sqrt{\Delta f}}{D*(\partial L/\partial T)\sqrt{A_d}}\right] \tag{10.21}$$

where $F/\#$ is known as the f-number, which is the distance between the pupil and the detector divided by the entrance-pupil area. The parameter L is the radiance at the location instantaneous field of view.

Another useful parameter is the minimum resolvable temperature difference (MRTD), which is useful as a summary measure of performance and design criterion. It combines both thermal sensitivity and spatial resolution. Smaller it is better when both NETD and MRTD are smaller.

10.2.2 Basic concepts of photoconductivity

The electrical conductivity due to the excess carrier generated by illuminating the sample with photons is called photoconductivity. The

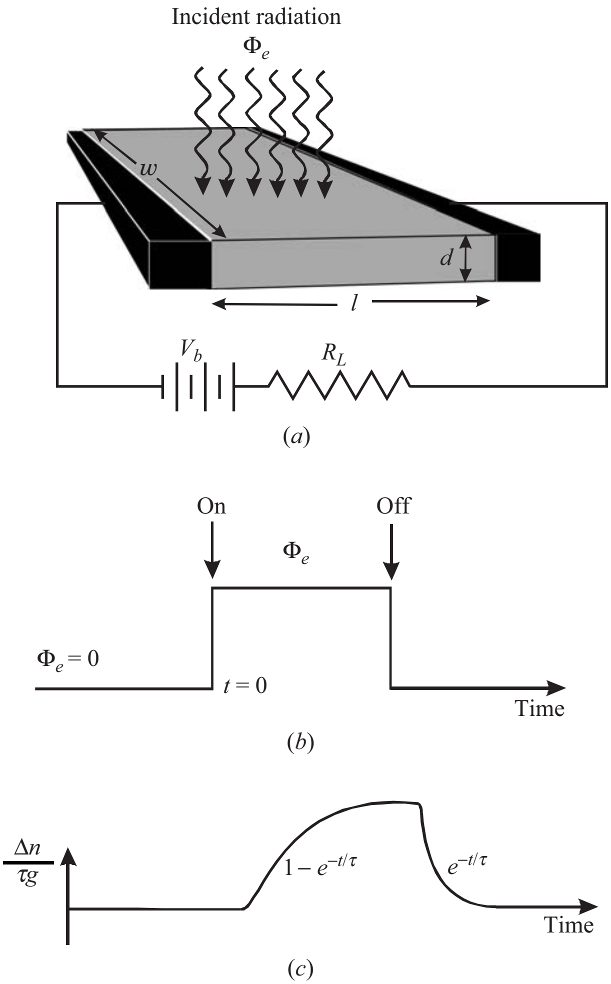


Figure 10.3 (a) Geometry of a biased photodetector showing the incident photons and the load resistance. (b) A radiation input pulse as a function of time. (c) The detector output signal as a function of time.

photoconductivity is measured by attaching electrodes to the sample as shown in Fig. 10.3a. For an undoped semiconductor photodetector, the dc short-circuit photocurrent is given as

$$I_{\mathrm{ph}} = e\eta wl\Phi_e G \tag{10.22}$$

where Φ_e = optical irradiant flux
G = photoconductive gain
wl = active detector area shown in Fig. 10.3a

Since the incident photons generate electron-hole pairs in semiconductor materials, the photoconductivity is a two-carrier process where the photoconductive current can be written as

$$I_{\mathrm{ph}} = ewd(\mu_n \Delta n + \mu_p \Delta p)\frac{V_b}{l} \tag{10.23}$$

where wd = detector cross section
V_b = bias voltage
l = distance between two electrodes
μ_n, μ_p = mobility of electrons and holes, respectively

The excess carrier concentration Δn and Δp are given by

$$n = n_o + \Delta n \qquad \text{and} \qquad p = p_o + \Delta p \tag{10.24}$$

where n_o and p_o are the thermal equilibrium carrier concentrations.

The transport properties of photodetectors are usually dominated by electrons. The current balance-equation can be constructed from the photogeneration, recombination, drift, and diffusion processes as follows:

$$\frac{\partial \Delta n}{\partial t} = \mathcal{G} - \frac{\Delta n}{\tau} - \frac{1}{e}\nabla \cdot \mathbf{J} \tag{10.25}$$

where $\mathcal{G}$ = generation rate, unit number/$(\mathrm{cm}^{-3} \cdot \mathrm{s})$
τ = recombination time
$\mathbf{J}$ = current density due to drift and diffusion

The diffusion and drift current can be neglected as compared to the generation recombination terms. Thus, Eq. (10.25) can be reduced to

$$\frac{\partial \Delta n}{\partial t} = \mathcal{G} - \frac{\Delta n}{\tau} \tag{10.26}$$

The generation rate can be written as

$$\mathcal{G} = \frac{\Phi_e \eta}{d} \tag{10.27}$$

For the steady-state case we have

$$\frac{\partial \Delta n}{\partial t} = 0 = \mathcal{G} - \frac{\Delta n}{\tau} \Rightarrow \tau = \frac{\Delta n}{\mathcal{G}} = \frac{\Delta n d}{\eta \Phi_e} \tag{10.28}$$

Additionally, the photoconductive gain can be obtained by equating Eqs. (10.22) and (10.23), such as

$$\mathcal{G} = \frac{d\mu_n \Delta n V_b}{\eta \Phi_e l^2} \tag{10.29}$$

Substituting Eq. (10.28) into (10.29) yields

$$\mathcal{G} = \frac{\tau \mu_n V_b}{l^2} = \frac{\tau}{l^2/(\mu_n V_b)} = \frac{\tau}{\tau_{\mathrm{tr}}} \tag{10.30}$$

where τ_{tr} is the transit time of electrons between the two electrodes, which is given as $\tau_{\mathrm{tr}} = l/\upsilon_d = l/(\mu_n \mathcal{E}) = l^2/\mu_n V_b$. Thus, the photoconductive gain is the ratio between the recombination time and the transit time.

When the excess carrier density is time-dependent, then the steady-state solution is not valid any more. The solution of Eq. (10.26) can be obtained by multiplying both sides by $e^{t/\tau}$ as follows:

$$e^{t/\tau}\left(\frac{\partial \Delta n}{\partial t} + \frac{\Delta n}{\tau}\right) = \mathcal{G}e^{t/\tau} \qquad \text{or} \qquad \frac{\partial}{\partial t}(\Delta n e^{t/\tau}) = \mathcal{G}e^{t/\tau} \tag{10.31}$$

Integrate Eq. (10.31) to obtain

$$\Delta n e^{t/\tau} = \int_0^t \mathcal{G}e^{t/\tau}\partial t = \tau\mathcal{G}(e^{t/\tau} - 1) \Rightarrow \Delta n = \tau\mathcal{G}(1 - e^{-t/\tau}) \tag{10.32}$$

A plot of the excess carrier as a function of time is sketched in Fig. 10.3*c*.

When the incident photon flux is turned off, Eq. (10.26) is reduced to the following

$$\frac{\partial \Delta n}{\partial t} = -\frac{\Delta n}{\tau} \tag{10.33}$$

The solution of this equation is

$$\Delta n = C_1 e^{-t/\tau} \tag{10.34}$$

where C_1 is a constant, which is determined from the initial conditions. If one assumes that the photon flux is turned off when Δn is maximum (i.e., $\Delta n = \tau\mathcal{G}$), as shown in Fig. 10.3*c*, then $C_1 = \tau\mathcal{G}$. By taking the Fourier transform of Eq. (10.34), one can express the spectral current responsivity in the modulation frequency domain as

$$\mathcal{R}_i(\omega) = \frac{\mathcal{R}_i(0)}{\sqrt{1 + \omega^2\tau^2}} \tag{10.35}$$

where $\mathcal{R}_i(0)$ is the responsivity at zero frequency given by

$$\mathcal{R}_i(0) = \frac{e\eta\lambda}{hc}\mathcal{G} \tag{10.36}$$

The expression for the photoconductive gain $\mathcal{G}$ is shown in Eq. (10.30), which depends on the bias voltage. Thus, the responsivity is a function of the wavelength, quantum efficiency, and the photoconductive gain. The simplistic responsivity model discussed excludes many effects, such as surface recombination and sweep-out effects. Furthermore, the responsivity depends on the product of the quantum efficiency and the photoconductive gain. It is difficult to measure them separately.

The internal quantum efficiency η_o is usually close to unity due to the fact that most of the photons are absorbed and contribute to the photoconductivity. On the other hand, the external quantum efficiency η depends on the reflection coefficients of both the top and bottom detector surfaces. It also depends on the optical absorption coefficient α. The derivation of the external quantum efficiency depends on the detector type and configuration. For example, the external quantum efficiency derived for infrared charge transfer devices is given by (see Nelson 1977)

$$\eta = \frac{\eta_o(1-r_1)(1+r_2e^{-2\alpha d})(1-e^{-\alpha d})}{1-r_1r_2e^{-2\alpha d}} \tag{10.37}$$

where d is the detector thickness (see Fig. 10.3*a*) and r_1 and r_2 are the reflection coefficients of the top and bottom surfaces. If $r_1 = r_2 = r$ and if the absorption coefficient is sufficiently large, Eq. (10.37) can be rewritten as

$$\eta \approx \eta_o(1-r) \tag{10.38}$$

If the photodetector is made such that the reflection coefficient from the top surface is zero and the reflection coefficient from the back side is unity, the quantum efficiency is reduced to

$$\eta \approx \eta_o(1+e^{-2\alpha d})(1-e^{-\alpha d}) \tag{10.39}$$

For highly absorptive detector material, the external quantum efficiency described by Eq. (10.39) is reduced to the internal quantum efficiency. This is an ideal design where the external quantum efficiency approaches unity. In general, the external quantum efficiency can take different forms depending on the type of the detectors and basic physical principles of operation.

10.2.3 Noise in photodetectors

There are different sources of noise in semiconductor photodetectors. The most important noise sources are Johnson noise, $1/f$ noise, generation-recombination noise, and preamplifier noise. Johnson noise is also called thermal noise and is associated with the finite resistance of the device. It is due to the random thermal motion of charge carriers in the detector material. This noise is present, in the absence of bias voltage, as a fluctuation in the current or voltage and is due to the random arrival of the charge carriers at the device electrodes. The Johnson noise is generated in both the detector and the load resistance. The root mean square of the Johnson current noise can be expressed as

$$i_J = \sqrt{\frac{4k_B T_d \,\Delta f}{R_d} + \frac{4k_B T_L \,\Delta f}{R_L}} \tag{10.40}$$

where T_d and T_L are the temperatures of the detector and load resistor, respectively, and R_d and R_L are the detector resistance and load resistance, respectively. If the temperatures of the detector and load resistor are the same ($T_d = T_L = T$), then the thermal fluctuation current can be written as

$$i_J = \sqrt{\frac{4k_B T \,\Delta f}{R_{\text{eq}}}} \qquad \text{or} \qquad V_J = R_{\text{eq}} i_J = \sqrt{4k_B T \,\Delta f \, R_{\text{eq}}} \tag{10.41}$$

where $R_{\text{eq}} = R_d R_L/(R_d + R_L)$ and V_J is the Johnson noise voltage.

The intrinsic noise mechanism of a photodiode is called shot noise, which is the noise in the current passing through a diode. The general form of the noise in the current, i_s, in an ideal diode is expressed as

$$i_s = \sqrt{[2e(I_D + 2I_s) + 4k_B T (G_j - G_o)]\Delta f} \tag{10.42}$$

where I_D = diode current = $I_s(e^{eV/k_B T} - 1)$
I_s = saturation reverse current
G_j = conductance of junction
G_o = value of G_j at low frequency

For zero bias voltage and low frequency, shot noise reduces to Johnson noise.

The origin of the $1/f$ noise is not well understood, but it appears to be associated with potential barriers. The general expression for the $1/f$ noise current is

$$i_{1/f} = \sqrt{\frac{A i_b^{\alpha} \,\Delta f}{f^{\beta}}} \tag{10.43}$$

where A = constant
$\alpha \approx 2$
$\beta \approx 1$
i_b = dc bias current

The $1/f$ noise is dominant at low frequency.

The current generation-recombination noise results from the random number of free carriers due to the background photons and thermal excitations in the detector. The general form of this noise is expressed as

$$i_{\mathrm{gr}} = 2eG\sqrt{\eta\Phi_b A_d\,\Delta f + \mathcal{G}_{\mathrm{th}}\Delta f} \tag{10.44}$$

where G = photoconductive gain
Φ_b = background flux density
$\mathcal{G}_{\mathrm{th}}$ = thermal generation rate of carriers

Many long-wavelength infrared detectors operate at temperatures lower than room temperatures where the thermal noise is negligible. In this case, the current recombination-generation noise can be expressed as the noise due to background photons, which is simply the first expression in Eq. (10.44):

$$i_{\mathrm{gr}} = 2eG\sqrt{\eta\Phi_b A_d\,\Delta f} \tag{10.45}$$

The expression given by Eq. (10.45) is further modified when $h\nu/k_B T < 1$ according to the following relation:

$$i_{\mathrm{gr}} = 2eG\sqrt{\eta\Phi_b A_d\,\Delta f(1 + b_{\mathrm{Bose}})} \tag{10.46}$$

where b_{Bose} is the Bose factor, which is not negligible at long wavelengths.

There are other sources of noise, such as the preamplifier noise, which will not be discussed here. Summing all the sources of noise, one can write the total noise as

$$i_{\mathrm{noise}} = \sqrt{i_J^2 + i_s^2 + i_{1/f}^2 + i_{\mathrm{gr}}^2} \tag{10.47}$$

A sketch of Johnson noise, $1/f$ noise, generation-recombination noise, and the total noise as a function of frequency is shown in Fig. 10.4. The dominant noise at low frequencies is $1/f$ noise, while the generation-recombination noise is dominant at midfrequencies. Note that the recombination-generation noise given by Eq. (10.46) is the low-frequency noise.

For higher frequencies, the generation-recombination noise can be obtained as

$$i_{\mathrm{gr}}(f) = \frac{i_{\mathrm{gr}}(0)}{\sqrt{1 + (\omega\tau)^2}} \tag{10.48}$$

where $\omega = 2\pi f$ and τ is the recombination lifetime. At high frequencies, the generation-recombination noise is rolled off and the dominant noise is Johnson noise.

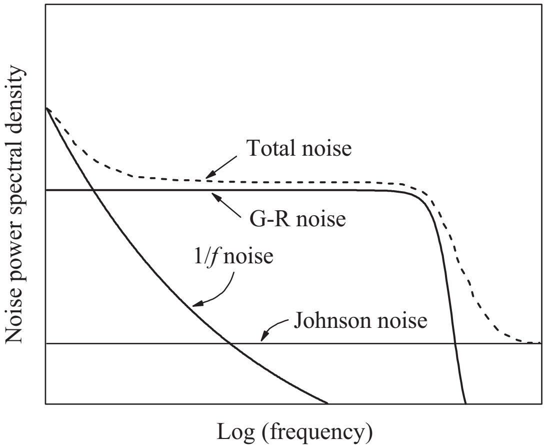


Figure 10.4 A sketch of $1/f$, Johnson, generation-recombination, and total noise plotted as a function of frequency.

10.2.4 Multiple-quantum-well infrared photodetectors

There are many types of quantum photodetectors, such as pn-junction photodiodes, Schottky barrier photodiodes, and metal-insulator-semiconductor photodiodes. This textbook focuses on heterojunctions and nanostructures, and therefore we will limit our discussion to detectors based on quantum wells, superlattices, and quantum dots. Different quantum well designs were investigated for long-wavelength infrared detectors. The basic principle of operation is based on photon absorption by electrons that exist in the quantum well ground state and are then excited to a higher energy level. These excited electrons are collected under a bias voltage to give a photoconductive signal in the infrared regions. The quantum well infrared photodetector (QWIP) is based on intersubband transitions, which are shown in Fig. 10.5 for different quantum well designs. The first design shown in Fig. 10.5a is a typical quantum well structure with two bound states. The applied bias voltage causes the conduction band to bend where the excited state is now located near the edge of the barrier conduction band. Electrons in the ground state can be excited by illuminating the sample with infrared light and then collected under the influence of the bias voltage. The detector output signal is usually called a photocurrent or photoresponsivity.

The second quantum well design is shown in Fig. 10.5b where the quantum well is sandwiched between two thin layers with a bandgap larger than the barrier materials. A typical example is a GaAs quantum well and $\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{As}$ barriers. The thin layer between the well and barrier is $\mathrm{Al}_y\mathrm{Ga}_{1-y}\mathrm{As}$ where $y > x$. The addition of the $\mathrm{Al}_y\mathrm{Ga}_{1-y}\mathrm{As}$ thin

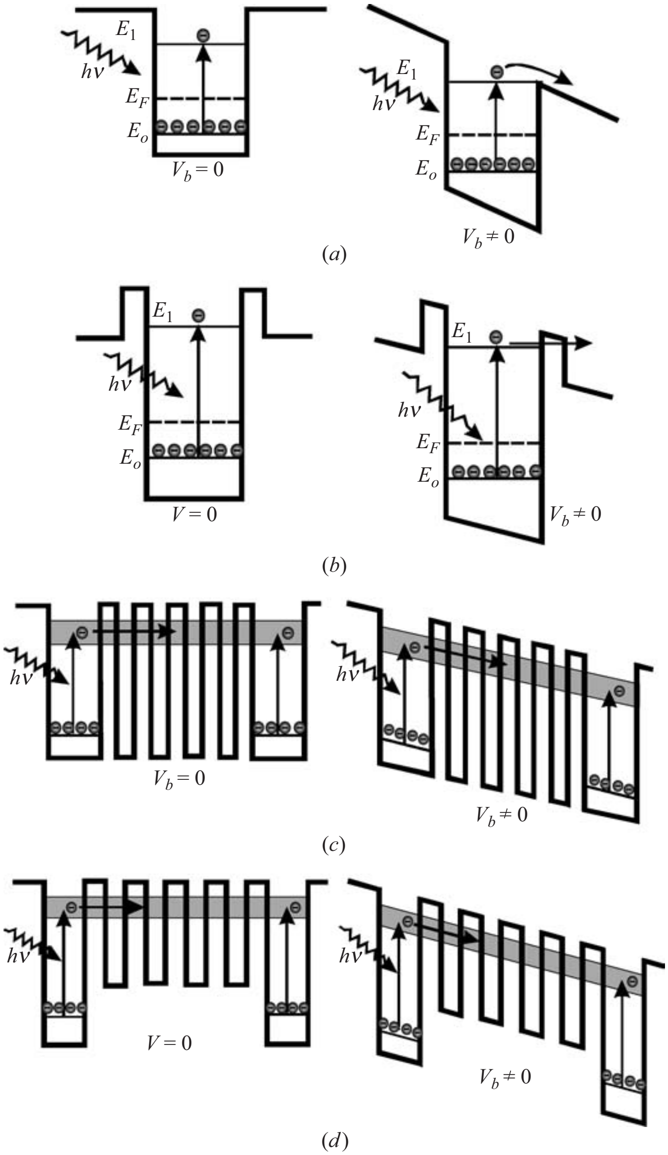
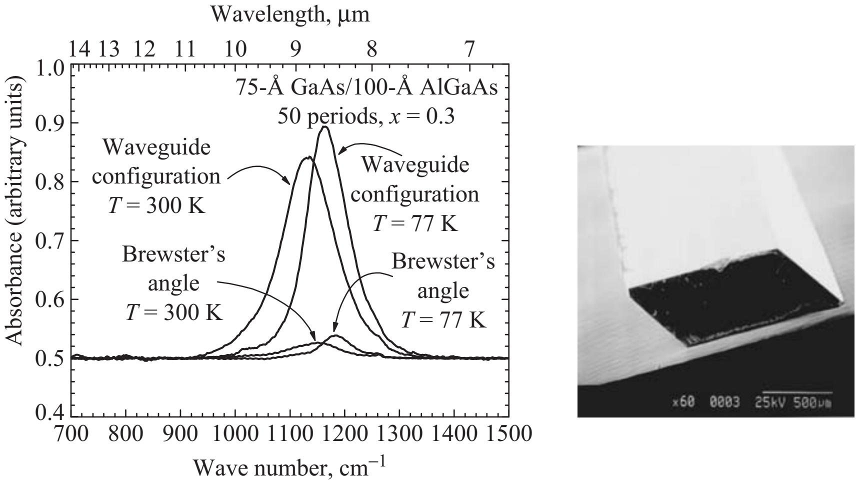


Figure 10.5 Intersubband transitions in multiple quantum wells with different designs. (*a*) Typical quantum well design structure with two bound states, for example, GaAs/AlGaAs. (*b*) Quantum well structure with additional thin barrier, for example, GaAs/$Al_yGa_{1-y}As$/$Al_xGa_{1-x}As$, where $y > x$. (*c*) Quantum well structure where the barrier is composed of superlattice, for example, GaAs (well)/(GaAs/AlGaAs) (superlattice barrier). (*d*) Embedded quantum well structure with superlattice barrier, for example, InGaAs (well)/(GaAs/AlGaAs) (superlattice barrier). The left panel is for zero bias voltage, and the right panel is sketched for applied bias voltage.

layers reduces the dark current and produces a narrower photoresponse as compared to the structure shown in Fig. 10.5*a*. The excited electrons in the latter structure tunnel through the thin barrier and then collect at one of the electrodes.

An alternative structure, commonly used with low dark current devices, is shown in Fig. 10.5*c*, where the barrier is composed of superlattices rather than a simple bulk barrier. A miniband is formed in the superlattice barrier, which is extended into the quantum well itself. When the sample is irradiated with photons, the electrons are excited from the ground state to the miniband and then transported via the miniband when the sample is biased. The dark current is further reduced by designing a quantum well, such as InGaAs, embedded between the GaAs/AlGaAs superlattices as shown in Fig. 10.5*d*.

Since the quantum wells in these structures are required to be populated with electrons, the wells are usually doped with silicon. Modulation doping in the barriers yields similar results. To increase the optical length, the quantum well/barrier repetition is usually chosen between 20 to 50 periods. A typical example of the optical absorption of an intersubband transition in *n*-type 75-Å GaAs/ 100-Å $Al_{0.3}Ga_{0.7}As$ multiple quantum wells grown on semi-insulating GaAs wafers is shown in Fig. 10.6*a*. In this figure, we plotted the optical absorption spectra obtained at room temperature and 77 K for both the Brewster's angle and



(*a*) (*b*)

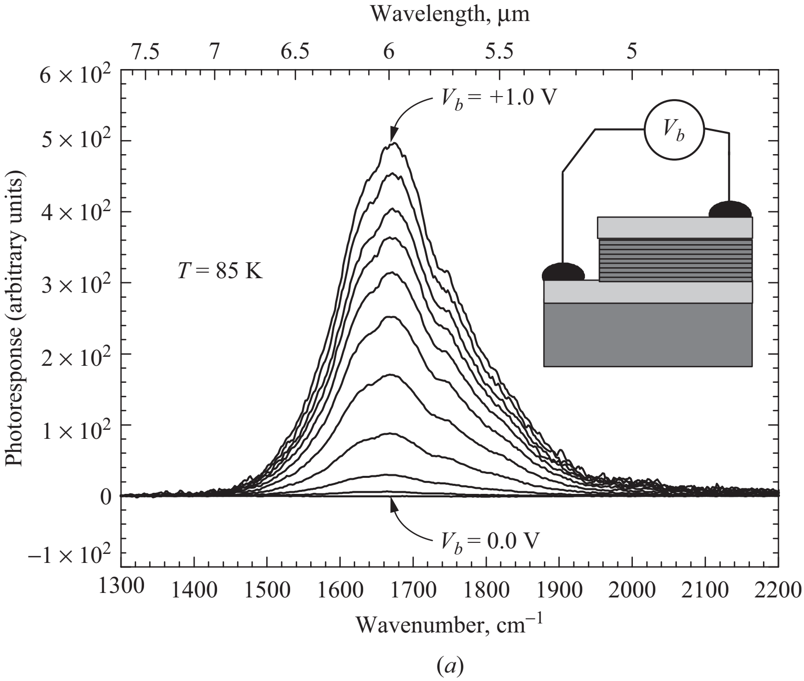
Figure 10.6 (*a*) Optical absorption spectra of the intersubband transition in GaAs/AlGaAs multiple quantum wells obtained at room temperature and 77 K using both Brewster's angle and waveguide configurations. (*b*) A scanning electron microscopy image of a waveguide made of GaAs/AlGaAs multiple quantum wells grown in a semi-insulating GaAs wafer with a thickness of ~0.450 mm.

waveguide configurations. The intensity of the spectra collected using the waveguide configuration is much larger than those obtained using the Brewster's angle configuration because the incident light makes multiple passes in the waveguide configuration. A typical waveguide is shown in Fig. 10.6*b*, where the width of the sample is ~2.0 mm and the length is ~5 mm. The waveguide is usually cut from the wafer and then the facets along the 5-mm edges are polished at an angle of 45°. The optical absorption coefficient and the selection rules were discussed in more detail in Chap. 6.

Photoconductivity measurements of a photodetector required to fabricate a mesa where electrodes are attached to the mesa and a bias voltage is applied. A typical mesa structure is shown in the inset of Fig. 10.7*a* where the quantum well structure is sandwiched between contact layers. The photoresponse spectra shown in Fig. 10.7*a* were collected under different bias voltages ranging between $0 \leq V_b \leq 1.0$ V with a step of 0.1 V. The structure of the multiple quantum wells is shown in Fig. 10.7*b*. It is composed of five periods of $In_{0.25}Ga_{0.75}As$ quantum wells and $GaAs/Al_{0.3}Ga_{0.7}A$ superlattice barriers. Thin layers of AlAs were also inserted as shown in the figure to reduce the dark current. The photoresponse spectra show the dominant transition between ground state E_1 and the miniband E_2 around 6.0 μm and the weaker transition between the ground state E_1 and the excited state E_3 at around 5.7 μm.

Multicolor detectors can be easily fabricated by growing different stacks of multiple quantum wells separated by contact layers. An example of a two-color detector is sketched in Fig. 10.8*a* where *n*-type doped GaAs layers were grown as contact layers. The top stack of the multiple-quantum-well structure is shown in Fig. 10.8*b*, which is essentially the same structure as shown in Fig. 10.7*b*. The structure of the bottom stack is shown in Fig. 10.8*c*, which is designed such that the photoresponse is at ~10.5 μm. From the design of the bottom stack, one can see that the dominant transition is between the ground state E_1 and the miniband E_2. The transition between the ground state and the continuum E_c is also allowed, but with a smaller oscillator strength.

The optical absorption of the intersubband transitions in the two-stack sample described in Fig. 10.8 was obtained at 77 K using a 45° polished facet waveguide configuration. The incident light makes two to three passes inside the waveguide before exiting the sample. The result is shown as the gray spectrum in Fig. 10.9. A bias voltage is not needed for the optical absorption measurement. The two dominant transitions in both stacks are associated with the transition from the ground state to the miniband in each stack. The miniband is formed from the ground states in the superlattice barrier, which is extended into the quantum wells. Each period in the superlattice barrier contributes one energy



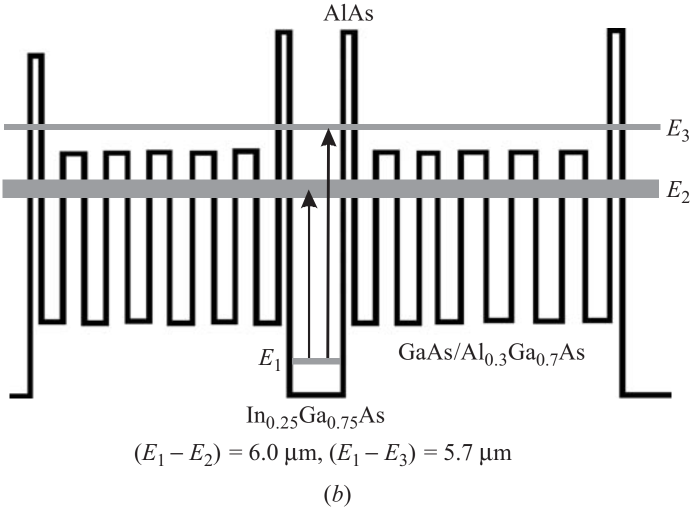

AlAs
E_3
E_2
E_1
GaAs/Al$_{0.3}$Ga$_{0.7}$As
In$_{0.25}$Ga$_{0.75}$As
$(E_1 - E_2) = 6.0$ μm, $(E_1 - E_3) = 5.7$ μm
(b)

Figure 10.7 (*a*) The photoresponse spectra of the multiple-quantum-well infrared photodetector obtained for different bias voltages applied to the mesa (shown in the inset). (*b*) The structure of the multiple quantum wells used to fabricate a photodetector device.

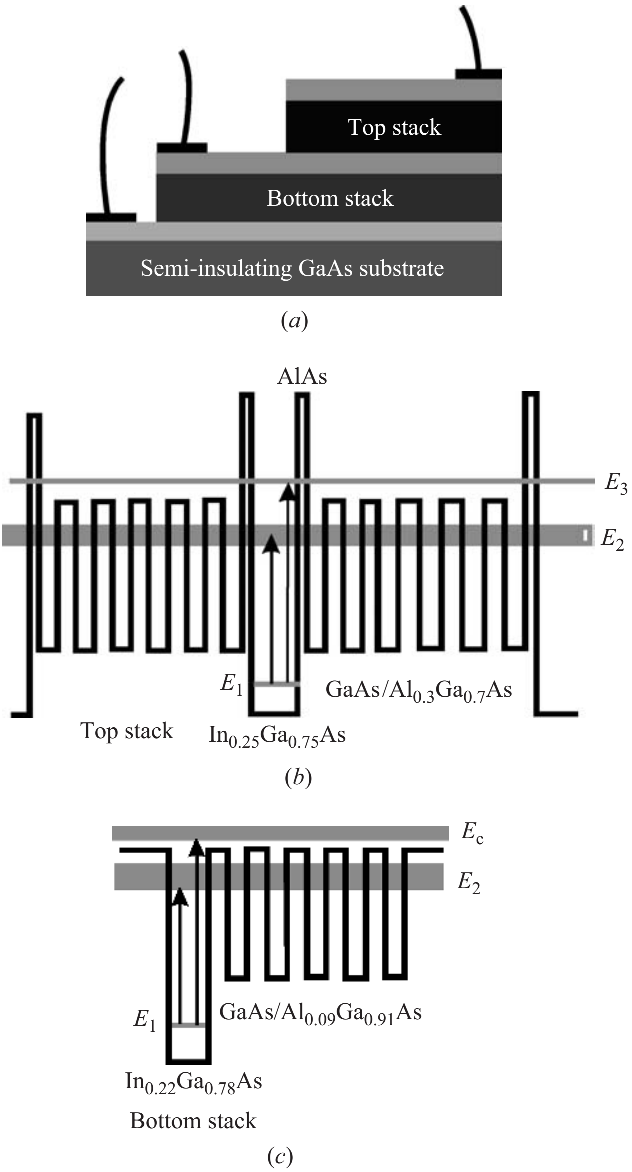


Figure 10.8 (*a*) A sketch of two stacks of multiple quantum wells separated by *n*-type doped GaAs layers, which act as contact layers. The contact pads are shown. (*b*) The multiple-quantum-well structure used for the top stack. (*c*) The multiple-quantum-well structure used for the bottom stack.

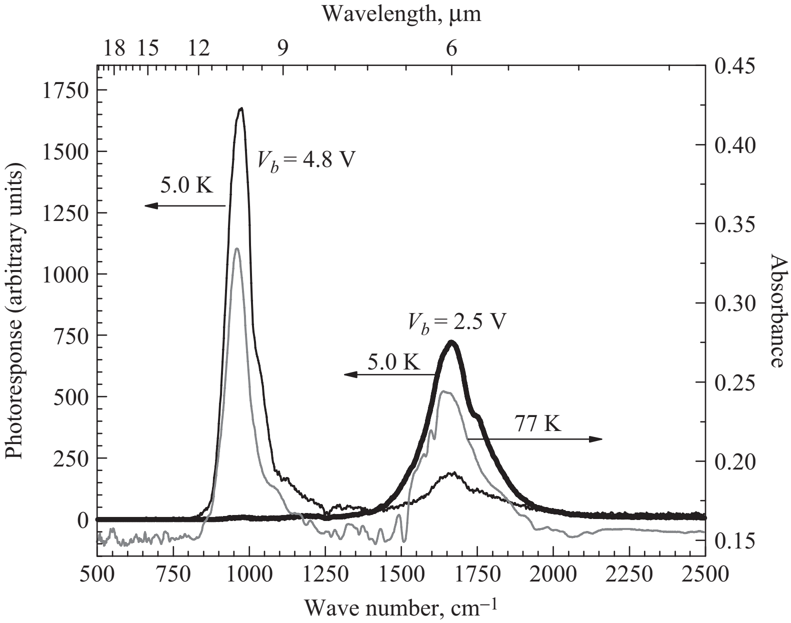


Figure 10.9 The optical absorption spectrum (gray line) obtained at 77 K for a two-color infrared photodetector. The photoresponse spectra obtained at 5.0 K for two bias voltages of (*a*) 2.5 V, thick black line, and (*b*) 4.8 V, thick black line.

level to the miniband. The optical absorption spectrum in Fig. 10.9 exhibits two peaks. The first peak is observed around 5.7 μm, which is originated from the top stack, and the second peak is observed around 10.5 μm, which is originated from the bottom stack.

The device photoresponse is measured at 5.0 K under different bias voltages. For this test, the bias voltage is applied between the contact layer deposited on top of the top stack and the contact layer deposited underneath the bottom stack (see Fig. 10.8*a*). When the voltage is low (V_b = 2.5 V), the peak from the top stack (~5.7 μm) is dominant while the peak from the bottom stack is below the detection limit as shown in the photoresponse spectrum (thick black line). However, when the voltage is increased to V_b = 4.8 V, the photoresponse peak (~10.5 μm) from the bottom stack is dominant (thin black line). Both peaks coincide with the peaks observed in the optical absorption measurements.

The multiple-quantum-well detectors described in Figs. 10.8 and 10.9 are called voltage tunable photodetectors. They have the ability to respond to different photon wavelengths under different bias voltages. Further illustration of how the photoresponse depends on the bias voltage is shown in Fig. 10.10. Several photoresponse spectra were recorded for different bias voltages applied across points *A* and *C* in the figure

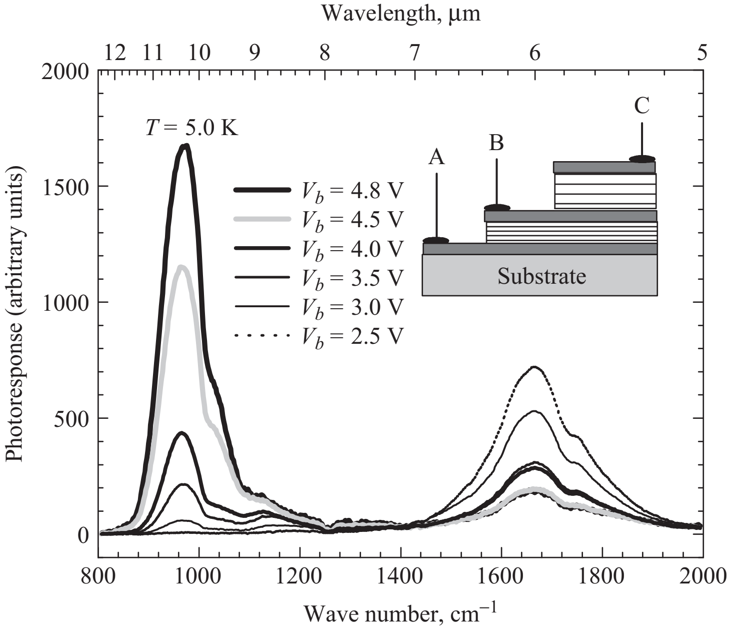


Figure 10.10 Several photoresponse spectra obtained as a function of the bias voltage for a two-stack photodetector. The inset is a sketch of the two stacks with A, B, and C electrodes.

inset. The measurements were obtained at 5.0 K. Similar results were obtained at 77 K, but with lower peak intensities. It is clear from this figure that different stacks can be turned on or off depending on the bias voltage. If a bias voltage is applied between points A and B, one can only observe the peak at ~10.5 μm. Similarly the peak around 5.7 μm is observed when a bias voltage is applied between points B and C.

The internal quantum efficiency of the QWIPs can be obtained from the optical absorption coefficient. If the light incident intensity is I_i, then the quantum efficiency is the fraction of the light incident intensity that is absorbed by the electrons that undergo the intersubband transition. The light intensity absorbed by the quantum wells can be defined as

$$\begin{aligned} I_1 &= I_i(1-\alpha_o) && \text{for a single quantum well} \\ I_2 &= I_i(1-\alpha_o)^2 && \text{for two quantum wells} \\ &\vdots \\ I_N &= I_i(1-\alpha_o)^N && \text{for } N \text{ quantum wells} \end{aligned} \tag{10.49}$$

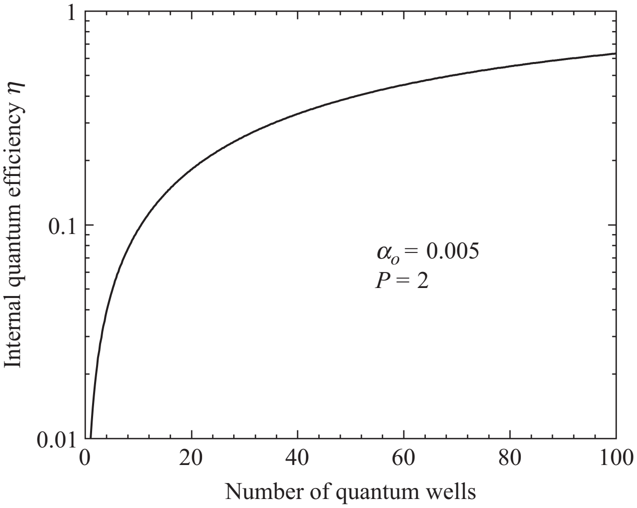


Figure 10.11 The internal quantum efficiency plotted as a function of the number of quantum wells assuming $\alpha_o = 0.005$ and $P = 2$.

where α_o is the fractional optical absorption coefficient due to a single quantum well. If a waveguide configuration is used instead of a single-pass (Brewster's angle) configuration, Eq. (10.49) can be rewritten as

$$I_N = I_i(1 - \alpha_o)^{NP} \tag{10.50}$$

where P is the number of passes that the light makes inside the waveguide. With the aid of Eq. (10.50), the internal quantum efficiency η can be defined as

$$\eta = \frac{I_i - I_N}{I_i} = \frac{I_i - I_i(1 - \alpha_o)^{NP}}{I_i} = 1 - (1 - \alpha_o)^{NP} \tag{10.51}$$

The internal quantum efficiency is plotted as a function of the number of quantum wells in Fig. 10.11. In this plot, the fractional optical absorption due to a single quantum well is assumed to be 0.005 and the total number of passes in the waveguide is $P = 2$.

Theoretically, α_o can be calculated using the formalism presented in Chap. 6. Experimentally, the fractional optical absorption due to a single quantum well can be obtained from the following expression:

$$A_{\max} = -\log_{10}(1 - \alpha_o)^{NP} \tag{10.52}$$

where $A_{\max}$ is the amplitude of the optical absorbance of the intersubband transition. For example, the two peaks of the optical absorbance shown in Fig. 10.9 yield $A_{\max} = 0.18$ and $A_{\max} = 0.1$ for the two peaks observed at 10.5 and 5.7 μm, respectively. The total number of quantum wells is $N = 10$ and the number of passes is $P \approx 2$. Substituting these values in Eq. (10.52), one can obtain $\alpha_o = 0.0205$ and 0.0114 for the two peaks. Substituting the values of α_o into Eq. (10.51) yields internal quantum efficiencies of ~34.0 and ~20.0 percent for the two peaks observed at 10.5 and 5.7 μm, respectively.

Another important aspect of the quantum well infrared photodetectors is the dark current, or the measured electric current when the device is kept in the dark. There are different mechanisms that contribute to the dark currents, such as thermionic emission and thermionic assisted tunneling. However, several methods can be used to reduce the dark current and enhance QWIP performance. For example, the barrier width can be grown thick enough to reduce the tunneling of charger carriers. Reduction of the operating temperature reduces the thermionic emission significantly.

Several models presented for the derivation of the dark current (see, for example, Levine 1993, Razeghi 1996, and Rogalski 2000) can be obtained for QWIPs. The general form of the dark current I_{dark} can be written as (Levine 1993)

$$I_{\text{dark}} = N^*(\mathcal{E}) e v_d(\mathcal{E}) A \tag{10.53}$$

where $N^*(\mathcal{E})$ is the effective number of electrons that are thermally excited out of the well into the continuum transport state (which is a function of the applied electric field $\mathcal{E}$), $v_d(\mathcal{E})$ is the average drift velocity (which is also a function of the applied electric field, $\mathcal{E}$), and A is the area of the detector. The effective number of the thermally excited electrons, $N^*(\mathcal{E})$ can be written as

$$N^*(\mathcal{E}) = \left(\frac{m^*}{\pi\hbar^2 L_p}\right) \int_{E_1}^{\infty} f(E) T_r(E, \mathcal{E})\, dE \tag{10.54}$$

where m^* is the electron effective mass, L_p is the total period length (the sum of the well and barrier thicknesses), $f(E)$ is the Fermi-Dirac distribution function given by

$$f(E) = (1 + e^{(E-E_1-E_F)/(k_B T)})^{-1} \tag{10.55}$$

where E = electron energy
E_1 = bound ground state in the well
E_F = Fermi energy level

All these parameters are illustrated in Fig. 10.12. The quantity $T_r(E, \mathcal{E})$ is the tunneling transmission coefficient for a single barrier, which can

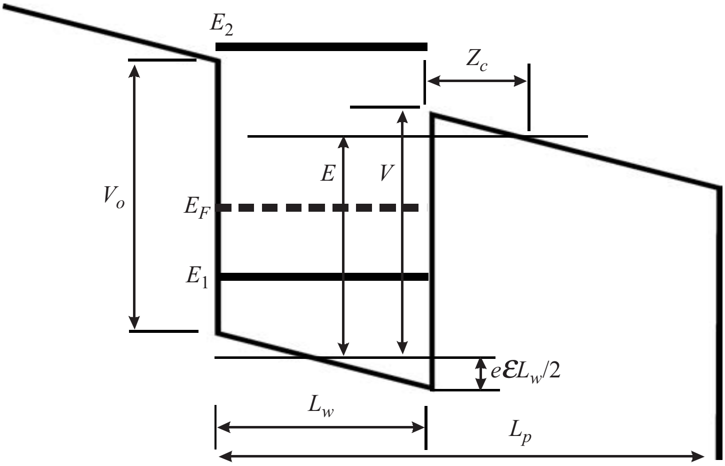


Figure 10.12 A sketch of a quantum well showing the energy levels, Fermi energy level, and other parameters used in the derivation of the dark current.

be written according to the WKB approximation as

$$T_r(E,\mathcal{E}) = \exp\left[-2\int_0^{Z_c} \sqrt{2m_b^*(V - E - e\mathcal{E}z)}\,/\hbar\,dz\right] \tag{10.56}$$

where m_b^* is the effective mass of the charge carrier in the barrier material, V is given by

$$V = V_o - \frac{e\mathcal{E}L_w}{2} \tag{10.57}$$

where V_o is the barrier height at zero applied electric field and L_w is the well thickness, and $Z_c = (V - E)/(e\mathcal{E})$ defines the semiclassical turning point, as illustrated in Fig. 10.12. The average drift velocity in Eq. (10.53) can be expressed as

$$\upsilon_d(\mathcal{E}) = \frac{\mu\mathcal{E}}{\sqrt{1 + (\mu\mathcal{E}/\upsilon_{\text{sat}})^2}} \tag{10.58}$$

where μ is the charge carrier mobility and υ_{sat} is the saturation drift velocity.

For low bias voltages, the tunneling transmission coefficient can be set as $T_r(E,\mathcal{E}) = 1$ for $E > V_o$ and $T_r(E,\mathcal{E}) = 0$ for $E < V_o$. In this case, Eq. (10.54) can be rewritten as

$$N^*(\mathcal{E}) \approx \left(\frac{m^* k_B T}{\pi \hbar^2 L_p}\right) e^{-(E_{\text{cut}}-E_F)/(k_B T)} \tag{10.59}$$

where $E_{\text{cut}} = V - E_1$ is the spectral cutoff energy and the Fermi energy can be obtained according the formalism in Chap. 5 as

$$N_d = \left(\frac{m^* k_B T}{\pi \hbar^2 L_w}\right) \ln(1 + e^{E_F/(k_B T)}) \tag{10.60}$$

where N_d is the doping density in the well. Thus, the dark current can be written in terms of the temperature as

$$I_{\text{dark}} \propto T e^{-(E_{\text{cut}}-E_F)/(k_B T)} \tag{10.61}$$

where E_{cut} is the cutoff energy taken as $E_{\text{cut}} = V_o - E_1$.

The detectivity of QWIPs is derived in a similar fashion as the detectivity described earlier in this chapter, which can be written as

$$D^* = \mathcal{R}_i \frac{\sqrt{A \Delta f}}{I_n} \tag{10.62}$$

where $\mathcal{R}_i$ is the spectral current responsivity, A is the area of the detector, and I_n is the shot noise introduced earlier in this chapter, which is given for a QWIP as (see Beck 1993)

$$I_n = \sqrt{4 e I_{\text{dark}} G \Delta f} \tag{10.63}$$

where G is the photoconductive gain and I_{dark} is the dark current derived previously. This expression is obtained assuming that the capture probability of the electron by the quantum well is much smaller than unity. Substituting the expression of the dark current given by Eq. (10.61) into (10.63) gives the following general expression for the detectivity:

$$D^* = D_o e^{[E_{\text{cut}}/(k_B T)]} \tag{10.64}$$

where D_o is a constant. Based on experimental measurements reported by Levine (1993) the best of the results yields

$$D^* = \begin{cases} 1.1 \times 10^6 e^{[E_{\text{cut}}/2k_B T]} & \text{for } n\text{-type QWIPs} \qquad (10.65a) \\ 2.0 \times 10^5 e^{[E_{\text{cut}}/2k_B T]} & \text{for } p\text{-type QWIPs} \qquad (10.65b) \end{cases}$$

The cutoff energy can be rewritten as $E_{\text{cut}}(eV) = hc/\lambda_{\text{cut}} = 1.24/(\lambda_{\text{cut}}\,\mu\text{m})$. A plot of Eq. (10.65*a*) is shown in Fig. 10.13 for different temperatures.

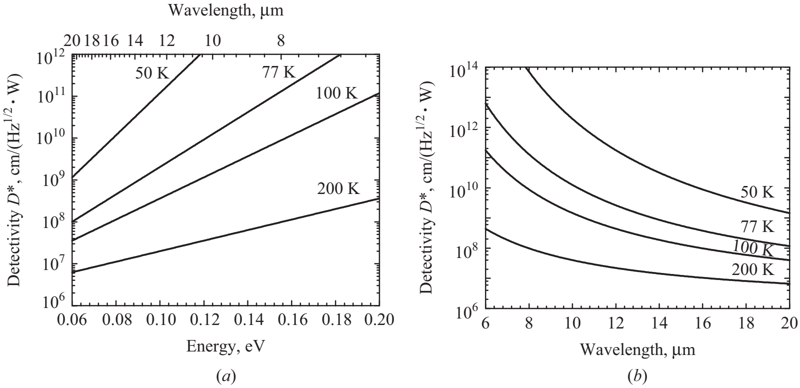


Figure 10.13 The detectivity plotted as a function of (*a*) energy and (*b*) wavelength for different temperatures.

The difference between the *n*-type and *p*-type QWIPs is that the quantum wells are doped with either donors or accepters. In case of *n*-type QWIPs, the wells are doped with silicon, and the quantum wells are populated with electrons. For the *p*-type QWIPs, the quantum wells are doped with either beryllium or carbon. Thus, the charge carriers are holes. A typical example of the photoresponse of a *p*-type QWIP is shown in Fig. 10.14, where the intersubband transition is between the ground state of the hole and the continuum in the valence band.

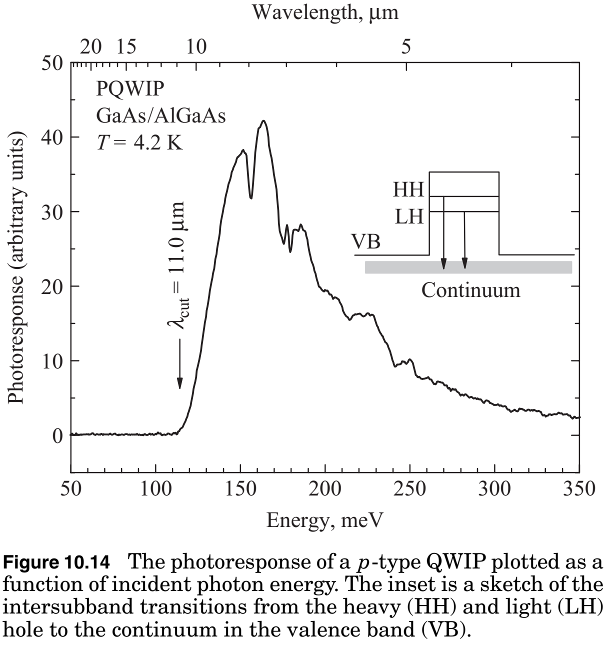


Figure 10.14 The photoresponse of a *p*-type QWIP plotted as a function of incident photon energy. The inset is a sketch of the intersubband transitions from the heavy (HH) and light (LH) hole to the continuum in the valence band (VB).

The advantage of the p-type QWIP is that normal incident photons can be absorbed by the charge carriers. The disadvantage is that the heavy hole effective mass is much larger than the electron effective mass, causing a reduction in the detectivity as shown in Eq. (10.64*b*). Further discussion about p-type quantum well infrared photodetectors is presented by Brown and Szmulowicz (1996).

10.2.5 Infrared photodetectors based on multiple quantum dots

In the case of n-type multiple quantum wells, the electron-photon coupling occurs when there is a polarization light component (TM) parallel to the growth axis. This requires that the photons illuminate the sample at an incident angle different than the normal. The maximum photon-electron coupling occurs at the Brewster's angle. Grating layers are usually fabricated to scatter the light at an angle. Because of the lack of energy dispersion in zero-dimensional electron systems such as quantum dots, the photon-electron selection rules are absent and the electron can absorb normal incident photons. This property makes quantum dots attractive as potential alternative systems to multiple quantum dots.

Now, one can design a quantum dot system by growing several layers of dots to increase the optical path length, as shown in Fig. 10.15. In this figure we sketched a structure depicting an infrared detector made of InAs quantum dots and a GaAs barrier. The wetting layer is also shown, which is the result of the layer-island growth mode known as

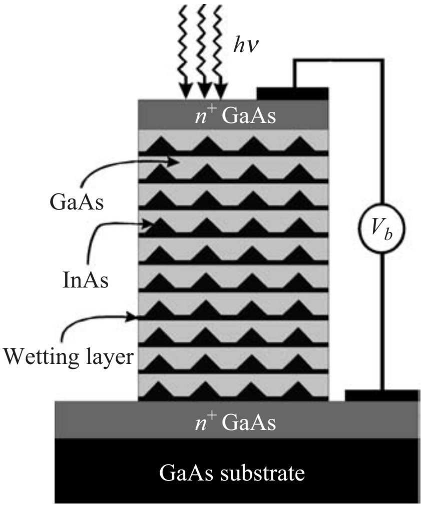


Figure 10.15 A simple sketch of an InAs/GaAs multiple-quantum-dot detector showing the wetting layers and pyramidal-shaped quantum dots.

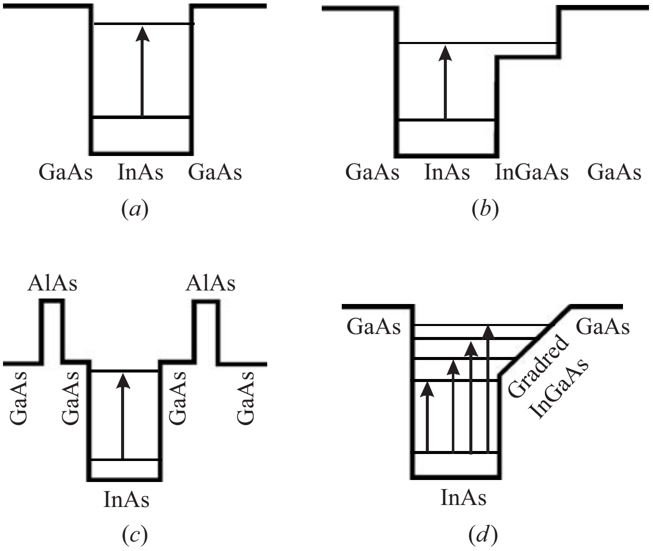


Figure 10.16 The conduction band structure of (*a*) an InAs/GaAs quantum dot, (*b*) an InAs/InGaAs/GaAs dot-well structure, (*c*) an InAs/GaAs quantum dot with an AlAs blocking layer, and (*d*) an InAs/graded InGaAs well/GaAs barrier structure.

the *Stranski-Krastanov* mode. Ohmic contacts are fabricated on top of the contact layers of n^+ GaAs material as shown in the figure to bias the device.

Several quantum dot systems have been enreported as infrared detectors. The most common system is the InAs/GaAs multiple quantum dots. Different structures have been designed and grown to mainly reduce the dark current. Unlike the doped multiple quantum wells, where the charge carriers are needed to populate the quantum wells, the multiple quantum dots in many cases do not even need to be doped, which leads to a significant reduction in the dark current. For undoped quantum dot detectors, the contact layers labeled n^+ GaAs provide the charge carriers under bias voltage. A few designs of the multiple-quantum-dot infrared detectors are shown in Fig. 10.16. The sketches in this figure do not show the wetting layer. There is a general consensus that the wetting layer is usually on the order of one monolayer of InAs intermixed with the GaAs barrier. A simple InAs/GaAs quantum dot band structure is shown in Fig. 10.16*a* where two bound states exist in the structure and the intersubband transition is indicated by the arrow.

A different structure composed of an InAs quantum dot embedded in an InGaAs quantum well is shown in Fig. 10.16*b*. This structure is usually referred to as a dot-well structure. Since the charge carriers often tunnel out of the dot especially when a bias voltage is applied, the dark current can be reduced by adding a thin layer of AlAs, as shown in Fig. 10.16*c*. The AlAs layer is usually referred to as the blocking layer. This layer can be chosen as AlGaAs with an Al mole fraction of ~30 percent. The presence of the blocking layer has proven to reduce the dark current and increase the responsivity of the quantum dot infrared photodetectors. Short-period superlattices, such as one monolayer of AlAs and two monolayers of GaAs, can also be used as blocking layers. The fourth structure shown in Fig. 10.16*d* is referred to as a dot-graded well structure. Several bound states are present in the dot-graded well structures.

The photoconductivity measurements of multiple-quantum-dot infrared detectors have been reported by many authors (see, for example, Madhukar et al. 2004). An example of the photoconductivity measurements obtained for multiple-quantum-dot infrared detectors is shown in Fig. 10.17. In this figure, the photoconductivity spectrum of an InAs/InGaAs/GaAs (dot-well) structure is shown as trace (*a*). It has a cutoff wavelength around 9.0 μm, and it peaks around 7.5 μm. As a comparison, the photoconductivity spectrum of a simple InAs/GaAs multiple-quantum-dot detector is shown as trace (*b*) in the figure. Both devices

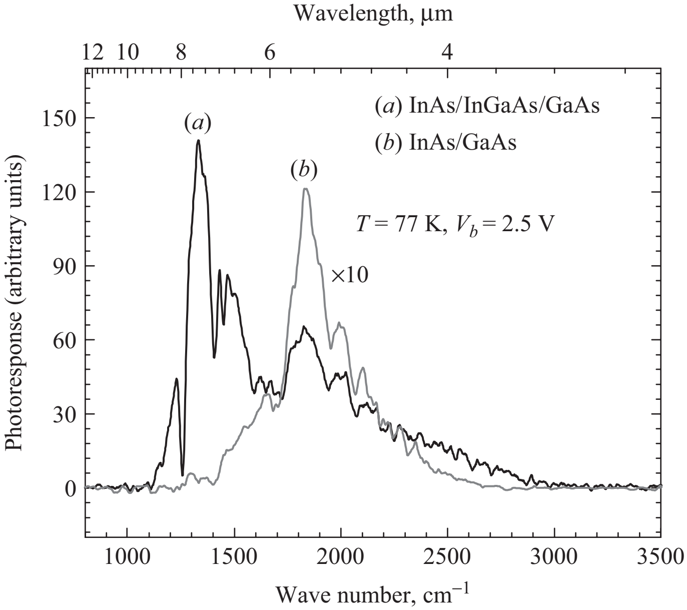


Figure 10.17 Photoconductivity measurements of (*a*) an InAs/InGaAs/GaAs dot-well detector and (*b*) an InAs/GaAs multiple-quantum-dot detector. The bias voltage was 2.5 V.

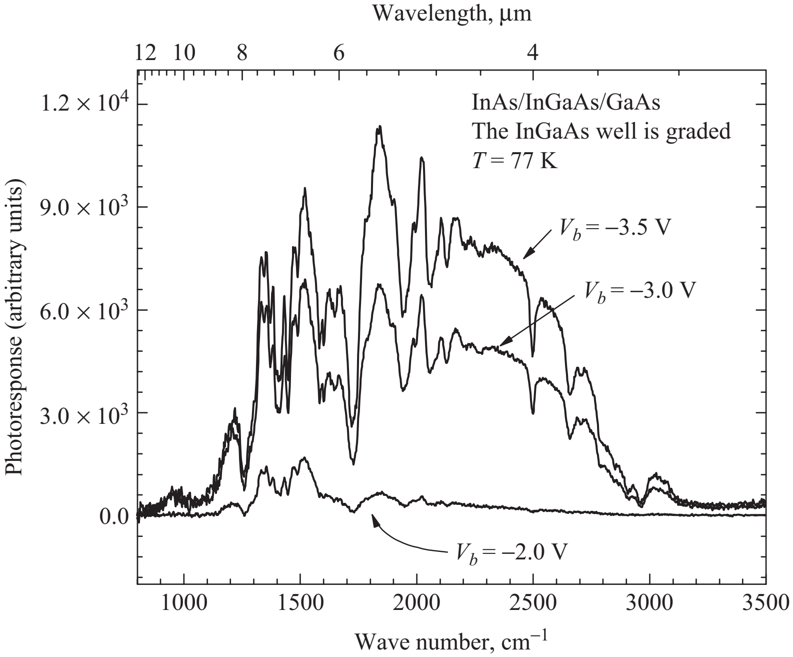


Figure 10.18 Photoresponse spectra measured at 77 K under three different reverse-biased voltages for InAs multiple quantum dots embedded in an InGaAs graded quantum well with a GaAs barrier.

were biased under the same condition (bias voltage $V_b = 2.5$ V). It peaks around 5.5 μm. The growth conditions are identical for both devices. In addition, the number of periods of the undoped multiple quantum dots is the same for both structures. The number of InAs monolayers deposited by the MBE growth system is also the same for both structures, but a significant redshift in the peak position is observed in the case of the InAs/InGaAs/GaAs dot-well device. This redshift is attributed to the lowering of the excited state, which exists in the quantum well. Thus, with the proper design, the intersubband transition in quantum dots can be tuned to cover a range of wavelengths. Finally, a quantum dot system, as schematically presented in Fig. 10.16*d*, exhibits a rich structure in the photoresponse as shown in Fig. 10.18. The three spectra were recorded at 77 K under three different reverse-biased voltages. The structure is composed of 10 periods of InAs quantum dots embedded in an InGaAs graded quantum well with a GaAs barrier. There are several excited states in the InGaAs graded quantum well such that a few intersubband transitions are present. The maximum photoresponse is relatively high compared to other quantum dot systems examined here, and the structure in the

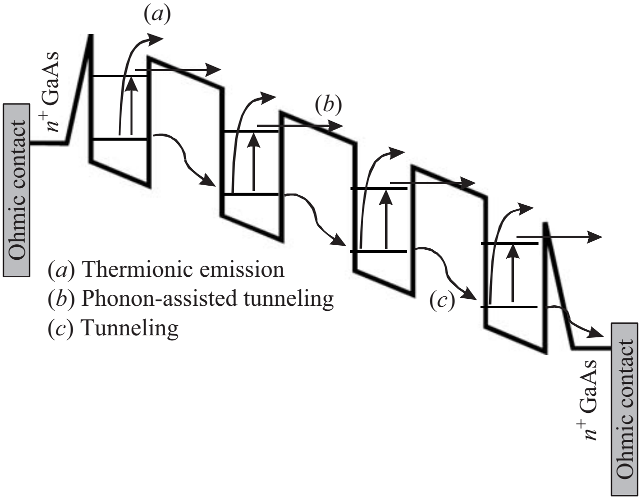


Figure 10.19 An illustration of the three major processes that contribute to the dark current in a multiple-quantum-dot infrared detector.

spectra is repeatable. The fine structure observed in the spectra could be a combination of intersubband transitions in different size quantum dots.

The dark current in multiple-quantum-dot infrared detectors is similar to that observed in multiple-quantum-well infrared detectors, which are dominated by the thermionic emission and tunneling processes. The three major processes that contribute to the dark current are labeled as (*a*), (*b*), and (*c*) in Fig. 10.19: (a) *thermionic emission*, where electrons are thermally excited out of the quantum dot to the continuum and are then swept by the bias voltage; (*b*) *phonon-assisted tunneling*, where electrons are thermally excited from the ground state to the excited state, followed by tunneling through a trapezoidal barrier to the continuum; (*c*) *tunneling process*, by which the electrons in the ground state tunnel when these energy levels are aligned. Similarly, the current noise in quantum dot detectors is governed by the same processes discussed previously in this chapter, which are mainly the $1/f$ noise, Johnson noise, and generation-recombination noise. The sketch shown in Fig. 10.18 is for undoped multiple quantum dots. The quantum dots, however are populated by the charge carriers that originated from the n^+ GaAs contact layer when the bias voltage was applied.

The external quantum efficiency η_{ex} of a multiple-quantum-dot infrared detector can be defined as the ratio between the number of collected output carriers to the number of incident photons according to the following expression:

$$\eta_{\text{ex}} = \frac{i_{\text{ph}}/e}{P_{\text{in}}/h\nu} = \mathcal{R}\frac{1.24}{\lambda} \tag{10.66}$$

where $\mathcal{R}$ = responsivity defined by Eq. (10.3)
i_{ph} = photocurrent defined by Eq. (10.22)
P_{in} = spectral radiant input power defined by Eq. (10.2)
λ = photon wavelength, μm

The external quantum efficiency is different than the internal quantum efficiency η_{in} defined by the optical absorption coefficient as shown in Eq. (10.51). The internal and external quantum efficiencies can be related according to the following approximate relation:

$$\eta_{\text{in}} = \frac{\eta_{\text{ex}}}{G(1-r)} \tag{10.67}$$

where G is the photoconductive gain given by Eq. (10.30) and r is the reflectivity of the semiconductor surface (in this case GaAs). If the photoconductive gain is unity, Eq. (10.67) is reduced to the relation given by Eq. (10.38). If the photoconductive gain is approximately equal to the noise gain, then one can write

$$G = \frac{i_n^2}{4eI_{\text{dark}}} \tag{10.68}$$

where i_n is the current noise and I_{dark} is the dark current. These current components can be measured experimentally from which the photoconductive gain can be estimated using Eq. (10.68).

The transport mechanisms in multiple-quantum-dot infrared detectors are similar to those in multiple-quantum-well detectors since the basic modes of operation are identical in both systems. Additional effects are present in the quantum dot systems. For example, planar transports between quantum dots through tunneling exist. Furthermore, the space between the quantum dots is comparable to the space occupied by the dots. Thus, the fill factor is less than unity and the photoconductive gain of a quantum dot detector can be written as

$$G = \frac{1-p}{pFN} \tag{10.69}$$

where F is the fill factor, N is the number of quantum dot periods, and p is the trapping or capture probability given by

$$p = \frac{\tau_{\text{esc}}}{\tau_{\text{life}} + \tau_{\text{esc}}} \tag{10.70}$$

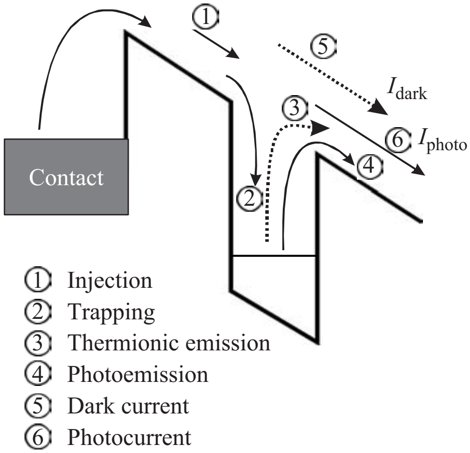


Figure 10.20 The major transport mechanisms in multiple-quantum-dot infrared detectors. The dashed arrows indicate the thermionic emission process that contributes to the dark current.

where τ_{esc} is the escape time for a photoexcited electron from the quantum dot and τ_{life} is the lifetime of the carrier.

The major transport mechanisms in a multiple-quantum-dot detector under a bias voltage are shown in Fig. 10.20. The photocurrent is usually a few orders of magnitude larger than the dark current. The dark current components are the thermionic emission and the injected carriers that reach the collector contact without being trapped. The total photocurrent can be derived in a fashion similar that of the multiple-quantum-wells photodetectors presented by Liu (1996) and is given by

$$I_{\text{photo}} = i_{\text{photo}} + \delta i \tag{10.71}$$

where i_{photo} is the total photocurrent due to only the photoexcited carriers. Taking into account the capture probability, i_{photo} can be written as

$$i_{\text{photo}} = i^1_{\text{photo}} \sum_{n=1}^{N} (1-p)^{n-1} = i^1_{\text{photo}} \frac{1-(1-p)^N}{p} \tag{10.72}$$

where $i^1_{\text{photo}}$is the photocurrent due to the photoexcited carriers from a single quantum dot and is given as

$$i^1_{\text{photo}} = e\Phi\eta \frac{1-p}{FN} \tag{10.73}$$

where Φ is the incident number of photons per second, and η is the absorption quantum efficiency.

The second quantity δi in Eq. (10.71) is the fraction of the extra injected current that reaches the collector contact and can be written as

$$\delta i = (1-p)^N \delta I_{\text{inject}} = i^1_{\text{photo}} \frac{(1-p)^N}{p} \tag{10.74}$$

The quantity δI_{inject} is the fraction of the injected current that is needed to balance the net loss of electrons in the quantum dot due to the photoemission and is given as

$$\delta I_{\text{inject}} = \frac{i^1_{\text{photo}}}{p} \tag{10.75}$$

Combining Eqs. (10.71), (10.72), and (10.74) yields

$$I_{\text{photo}} = i_{\text{photo}} + \delta i = \frac{i^1_{\text{photo}}}{p} = e\Phi\eta\frac{1-p}{pFN} \tag{10.76}$$

This expression verifies the photoconductive gain given by Eq. (10.69) or

$$G = \frac{I_{\text{photo}}}{e\Phi\eta} = \frac{1-p}{pFN} \tag{10.77}$$

The relationship tells us that if the capture probability is unity, the photoconductive gain is zero. As the capture probability is decreased, the photoconductive gain increases.

10.3 Light-Emitting Diodes

Light-emitting diodes (LEDs) were discovered early in the twentieth century. As the name implies, a semiconductor diode is subject to a small forward-biased voltage, where the electrons are injected in a normally empty conduction band. The injected electrons recombine in the holes in the valence band by emitting their energy as photons. This process is called electroluminescence or spontaneous emission. Since an optical cavity is not needed to provide photon feedback, the emitted photons have random phases, and therefore the LEDs are incoherent light sources. Furthermore, the emitted photon energy is close to the bandgap of the semiconductor material. A sketch of an LED pn junction is shown in Fig. 10.21 where the injected electrons in the conduction band recombine with the holes in the valence by emitting photons. A small fraction of the injected electrons recombine with holes in the valence band without emitting photons (nonradiative recombination process). Furthermore, the LEDs are usually made of direct bandgap semiconductor materials, since the conversion efficiency of the electrical signal to the photon signal in indirect semiconductor materials is very low. This is due to the involvement of the phonon in the interband transitions.

For pn-junction LEDs, the wavelength of the emitted light depends on the bandgap of the material. For narrow-bandgap materials, the wavelength is in the infrared region, and for wide-bandgap materials, the wavelength is in the visible and ultraviolet spectral regions. In addition

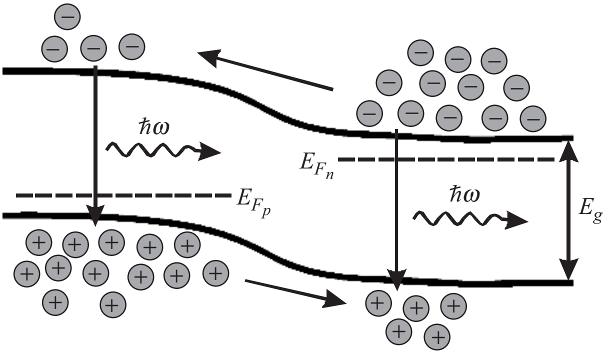


Figure 10.21 A sketch of a pn junction under the influence of a small forward-biased voltage. The spontaneous emission occurs when the electrons combine with the holes.

to the direct bandgap requirements, the semiconductor material should be easily doped with both n- and p-type dopants to form the junction. For example, GaAs is a direct semiconductor material and it can be easily doped with donors or acceptors. The wavelength of the LEDs made of GaAs is usually around 0.855 μm at room temperature. On the other hand, GaN is a direct semiconductor material with a bandgap of ~3.40 eV (0.365 μm). The wavelength of the LEDs can be easily tuned by choosing ternary materials such as AlGaAs, InGaN, and AlGaN.

In addition to the charge carrier injection and the radiative recombination processes, the generated photon should be able to exit the device and be used for the intended applications. The latter process is called the extraction processes. Each of these processes has its own efficiency. The overall device efficiency η_o (also known as the external efficiency) can be expressed as follows:

$$\eta_o = \eta_{\text{inj}} \eta_r \eta_e \tag{10.78}$$

where η_{inj} = injection efficiency
η_r = recombination efficiency
η_e = extraction efficiency

It is difficult to measure the individual efficiencies. But the external efficiency can be measured by taking the ratio between the output optical power and the input electric power. For pn-junction LEDs, the external efficiency is usually in the range of 1 to 10 percent.

For parabolic electron-hole bands, the LED spontaneous emission rate r_{sp} can be written as

$$r_{\text{sp}} = P_{\text{em}} N_j(E) f(E_e)[1 - f(E_h)] \tag{10.79}$$

where $f(E_e)$ and $f(E_h)$ are the Fermi-Dirac distribution functions of the electrons in the conduction band and the holes in the valence band, respectively; P_{em} is the emission probability, which is the inverse of the recombination lifetime τ_r; and $N_j(E)$ is the joint density of state given by

$$N_j(E) = \frac{(2m_r^*)^{3/2}}{2\pi^2\hbar^3}\sqrt{E - E_g} \tag{10.80}$$

where m_r^* is the reduced effective mass and E_g is the bandgap. The unit of the emission rate is $\mathrm{s}^{-1}\cdot\mathrm{eV}^{-1}\cdot\mathrm{cm}^{-3}$. The total spontaneous emission rate R_{sp} per unit volume can be defined as

$$R_{\mathrm{sp}} = \int r_{\mathrm{sp}}\,dE \tag{10.81}$$

For $E \ll k_BT$, the Fermi-Dirac distribution function can be approximated as the Boltzmann distribution function or $f(E_e)[1 - f(E_h)] \approx e^{-E/k_BT}$. With this approximation, the spontaneous emission rate can be

$$r_{\mathrm{sp}} = \frac{(2m_r^*)^{3/2}}{2\pi^2\hbar^3\tau_r}\sqrt{E - E_g}\,e^{-E_g/(k_BT)}e^{-(E-E_g)/(k_BT)} \tag{10.82}$$

For a weak forward-biased voltage (weak injection), the quasi-Fermi energy levels in the LED are still in the bandgap (nondegenerate case). Thus, Eq. (10.82) can be modified to include the quasi-Fermi energy levels:

$$r_{\mathrm{sp}} = \frac{(2m_r^*)^{3/2}}{2\pi^2\hbar^3\tau_r}\sqrt{E - E_g}\,\exp\left(\frac{E_{Fn} - E_{Fp} - E_g}{k_BT}\right)\exp\left(\frac{-E - E_g}{k_BT}\right) \tag{10.83}$$

The total photon flux Φ_o emitted from the LED is obtained by integrating over r_{sp}:

$$\begin{aligned}\Phi_o = V\int_0^\infty r_{\mathrm{sp}}\,dE &= \frac{(2m_r^*)^{3/2}V}{2\pi^2\hbar^3\tau_r}\exp\left(\frac{E_{Fn} - E_{Fp} - E_g}{k_BT}\right)\\ &\quad\times\int_0^\infty \sqrt{E - E_g}\,\exp\left(\frac{-E - E_g}{k_BT}\right)dE\\ &= \frac{V}{\sqrt{2}\hbar^3\tau_r}\left(\frac{m_r^*k_BT}{\pi}\right)^{3/2}\exp\left(\frac{E_{Fn} - E_{Fp} - E_g}{k_BT}\right)\end{aligned} \tag{10.84}$$

where V is the volume of the active region. In this expression, the recombination lifetime is assumed to be independent of the charge carrier energy, which may not be true. It is clear from Eq. (10.84) that the photon flux depends on temperature and on the positions of the quasi-Fermi energy levels. By increasing the injection level, the quasi-Fermi energy levels are moved closer to the conduction and valence bands for the electrons and holes, respectively, and the separation between them $(E_{Fn} - E_{Fp})$ increases. Thus, the photon flux increases as the charge carrier injection level is increased.

Another important parameter is the responsivity $\mathcal{R}_{\mathrm{LED}}$ of the LED, which is defined as the ratio between the emitted optical power to the injection current:

$$\mathcal{R}_{\mathrm{LED}} = \frac{\Phi_o \hbar\omega}{I_{\mathrm{inj}}} = \frac{hc}{e\lambda}\eta_o = \frac{1.24}{\lambda\ \mu\mathrm{m}}\eta_o \tag{10.85}$$

where the emitted optical power is defined as the product of the photon flux given by Eq. (10.84) and the photon energy $\hbar\omega$, and I_{inj} is the injection current. The units of responsivity are W/A. For an external efficiency of 5 percent of a GaAs LED at room temperature, the responsivity is $P_{\mathrm{LED}} = (1.24)(0.05)/(0.855) = 72.50$ mW/A=72.5 μW/mA. The optical power can be defined as the product of the emitted optical power and the injected current or

$$\mathcal{P}_{\mathrm{LED}} = \Phi_o \hbar\omega I_{\mathrm{inj}} = \eta_o \frac{hc}{e\lambda} = \frac{1.24 I_{\mathrm{inj}}\eta_o}{\lambda\ \mu\mathrm{m}} \tag{10.86}$$

For an injected current of 5 mA and an external efficiency of 5 percent, an LED emitting light at 0.532 μm has a power of $\mathcal{P}_{\mathrm{LED}} = (1.24)(5 \times 10^{-3})(0.05)/(0.532) = 0.58$ mW.

The surface nonradiative recombination and the carrier injections from the n region to the p region through the depletion region are two processes that reduce the external efficiency of the LED. These two disadvantages can be overcome by fabricating a heterojunction LED, where the active region can be made as thin as possible. Figure 10.22 shows a sketch of a GaAs/AlGaAs LED where the p^- GaAs is the active region. The n^+ AlGaAs provides an interface with a state density much lower than the free-surface states, which reduces the nonradiative recombination significantly. It should be pointed out that the bandgap of AlGaAs is larger than that of GaAs, and thus the top AlGaAs layer acts as a window through which the emitted photons exit with minimum reabsorption.

The LED frequency response can be derived from the continuity equation that takes into account the excess electron loss due to the spontaneous recombination and diffusion. If the drift current is neglected, the continuity equation can be written as

$$\frac{\partial n(x)}{\partial t} = -\frac{n(x)}{\tau_r} + D_e \frac{\partial^2 n(x)}{\partial x^2} \tag{10.87}$$

where τ_r = recombination lifetime throughout active region
D_e = diffusion coefficient
$n(x)$ = excess carrier concentration

A possible solution of this equation is

$$n(x) = n_0(x) + n_1(x)e^{i\omega t} \tag{10.88}$$

where the first term is due to the dc bias steady state and the second term is due to the time-dependent small-signal modulation of the diode. By separating the dc and the frequency-dependant parts, Eq. (10.86) can be rewritten as

$$D_e \frac{\partial^2 n_0(x)}{\partial x^2} - \frac{n_0(x)}{\tau_r} = 0 \quad \text{and} \quad D_e \frac{\partial^2 n_1(x)}{\partial x^2} - \frac{n_1(x)(1 + i\omega\tau_r)}{\tau_r} = 0 \tag{10.89}$$

The diffusion length can be defined as

$$L_e^2 = \begin{cases} D_e \tau_r & \text{for the dc part} \\ \dfrac{D_e \tau_r}{\sqrt{1 + \omega^2 \tau_r^2}} & \text{for the ac part} \end{cases} \tag{10.90}$$

Substituting Eq. (10.90) into (10.89) yields

$$\frac{\partial^2 n_0(x)}{\partial x^2} - \frac{n_0(x)}{L_e^2} = 0 \quad \text{and} \quad \frac{\partial^2 n_1(x)}{\partial x^2} - \frac{n_1(x)}{L_e^2(\omega)} = 0 \tag{10.91}$$

where $L_e(\omega)$ is taken as the second expression in Eq. (10.90).

If the width of the active region d of the LED is much larger than the diffusion length L_e, and the carrier concentration at the p-AlGaAs/p$^-$-GaAs interface (see Fig. 10.22) is zero, i.e., $n_1(x = d) = 0$, then the solution to the frequency-dependent part of Eq. (10.91) can be expressed as

$$n_1(x) = n_1^o e^{-x/L_e(\omega)} \tag{10.92}$$

where n_1^o is the initial electron concentration injected in the active region. The frequency response of the LED, $r(\omega)$, can be defined as

$$r(\omega) = \frac{\Phi_1(\omega)}{J_1(\omega)/e} \tag{10.93}$$

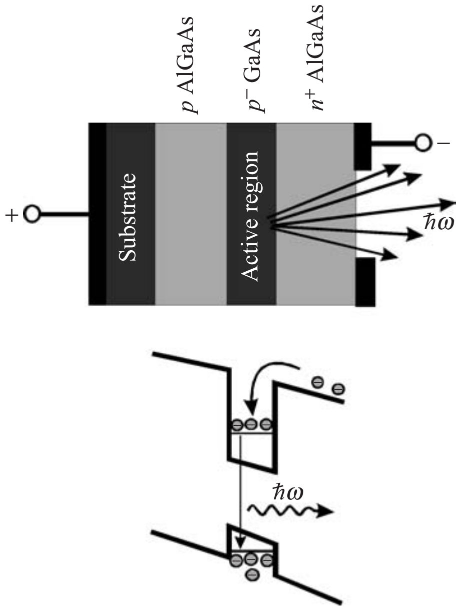


Figure 10.22 Sketch of a GaAs/AlGaAs heterojunction LED. The bandgap of the LED is also sketched under a forward-biased voltage.

where $\Phi_1(\omega)$ is the ac photon flux given by

$$\Phi_1(\omega) = \frac{1}{\tau_r}\int_0^d n_1(x)\,dx = \frac{n_1^o}{\tau_r}\int_0^d e^{-x/L_e(\omega)}\,dx$$

$$= \frac{n_1^o L_e(\omega)}{\tau_r} e^{-x/L_e(\omega)}\Big|_0^d = \frac{n_1^o L_e(\omega)}{\tau_r} \tag{10.94}$$

Notice that from the initial conditions we assumed that $n_1(x = d) = 0$. The ac current density, $J_1(\omega)$ shown in Eq. (10.93) can be obtained from the following equation:

$$J_1(\omega) = eD_e \frac{\partial n_1(x)}{\partial x} = -eD_e \frac{n_1^o}{L_e(\omega)} \tag{10.95}$$

Substituting Eqs. (10.94) and (10.95) into Eq. (10.93) gives

$$r(\omega) = \frac{L_e^2(\omega)}{\tau_r D_e} \tag{10.96}$$

Substitute the expression in Eq. (10.90) for the frequency-dependent diffusion length into Eq. (10.96) to obtain

$$r(\omega) = \frac{1}{\sqrt{1 + \omega^2\tau_r^2}} \tag{10.97}$$

This expression implies that for a high-frequency response, the recombination lifetime should be very small.

10.4 Semiconductor Lasers

Light amplification by stimulated emission of radiation (laser) is an optical waveguide terminated by mirrors or facets to form a resonant cavity. Einstein described the stimulated emission process in 1917, and Townes demonstrated this process at microwave frequencies. In 1990, Maiman demonstrated stimulated emission at optical frequencies in a ruby crystal. In 1962, the semiconductor laser was introduced. Propagation of electromagnetic waves in waveguides has been the subject of many textbooks, and it will not be discussed here.

10.4.1 Basic principles

There are different types of lasers based on the type of medium used. Semiconductor lasers depend on the design of the band structures and the confined energy levels. To illustrate the stimulated emission, consider a two-energy-levels system with energy levels E_1 and E_2 that are populated with electrons of densities N_1 and N_2, respectively, as shown in Fig. 10.23. At thermal equilibrium, the electron populations in the energy levels follow the Maxwell-Boltzmann distribution function, assuming that the two energy levels have the same degeneracy:

$$\frac{N_2}{N_1} = e^{-(E_2-E_1)/(k_B T)} = e^{-\hbar\omega_{12}/(k_B T)} \tag{10.98}$$

Three different processes can occur when a photon of energy $\hbar\omega_{12} = E_2 - E_1$ interacts with the system. The first process is called *absorption*, where the photon is absorbed by an electron causing the electron to jump from the ground state (E_1) to the excited state (E_2). This process is illustrated in Fig. 10.23*a*. The second process describes an electron relaxing from the excited state to the ground state releasing its energy as a photon. This process is called *spontaneous* emission, and is illustrated in Fig. 10.23*b*. The time that an electron spends in the excited state before relaxing to the ground state is called the spontaneous lifetime. While the energy of the photons resulting from the spontaneous emission is the same, they have random propagation directions and random phases. The third process illustrated in Fig. 10.23*c* can be thought of as the inverse of the absorption. When a photon of an energy $\hbar\omega_{12}$ interacts with the two-energy-level system, it passes the system without being absorbed and forces an electron to relax from the excited state to the ground state emitting another photon with the same energy ($\hbar\omega_{12}$). This process is called *stimulated* emission. This process generates a photon in a time τ_{st}, called the stimulated emission time. The propagation direction and the phase of the stimulated photon are identical to those of the passing (stimulating) photon.

In 1917 Einstein showed that the three processes of absorption, spontaneous emission, and stimulated emission are related. When the two

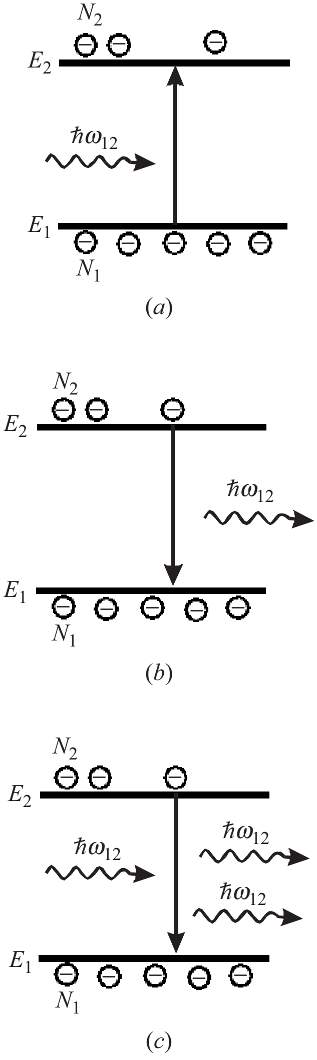


Figure 10.23 Interaction of a photon with an electron in a two-energy-level system showing the three different processes. (*a*) Absorption process, (*b*) spontaneous emission, and (*c*) stimulated emission.

energy levels in Fig. 10.23 have different degeneracy, Eq. (10.98) takes the following form:

$$\frac{N_2}{N_1} = \frac{g_2}{g_1} e^{-(E_2-E_1)/(k_B T)} = \frac{g_2}{g_1} e^{-\hbar\omega_{12}/(k_B T)} \tag{10.99}$$

where g_1 and g_2 are the degeneracies of the two energy levels. The upward transition rate R_{12} of an electron from E_1 to E_2 can be expressed as

$$R_{12} = N_1 \Phi(\omega_{12}) B_{12} \tag{10.100}$$

where B_{12} is the Einstein coefficient for absorption or stimulated upward transition and $\Phi(\omega_{12})$ is the radiation density of the energy density of the radiation field at frequency ω_{12}.

The spontaneous emission rate R_{21}^{sp} is defined as

$$R_{21}^{\mathrm{sp}} = N_2 \frac{1}{\tau_{21}} = N_2 A_{21} \tag{10.101}$$

where τ_{21} is the spontaneous emission lifetime, which is equal to the spontaneous recombination lifetime, and A_{21} is the Einstein coefficient for spontaneous emission ($A_{21} = 1/\tau_{21}$). Finally, the stimulated emission rate R_{21}^{st} of an electron from E_2 to E_1 is given by

$$R_{21}^{\mathrm{st}} = N_2 \Phi(\omega_{12}) B_{21} \tag{10.102}$$

where B_{21} is the Einstein coefficient for stimulated emission. Combining Eqs. (10.101) and (10.102) yields the total transition rate R_{21} from E_2 to E_1:

$$R_{21} = N_2 A_{21} + N_2 \Phi(\omega_{12}) B_{21} \tag{10.103}$$

At thermal equilibrium, the upward transition rate is equal to the downward transition rate, $R_{12} = R_{21}$, or

$$N_1 \Phi(\omega_{12}) B_{12} = N_2 A_{21} + N_2 \Phi(\omega_{12}) B_{21} \tag{10.104}$$

It follows from this equation that

$$\Phi(\omega_{12}) = \frac{A_{21}}{B_{21}} \frac{1}{N_1 B_{12}/(N_2 B_{21}) - 1} \tag{10.105}$$

By using Eq. (10.99), we can rewrite Eq. (10.105) as

$$\Phi(\omega_{12}) = \frac{A_{21}}{B_{21}} \frac{1}{[g_1 B_{12}/(g_2 B_{21})]\, e^{\hbar\omega_{12}/k_B T} - 1} \tag{10.106}$$

Without going through the detailed analysis, Einstein relations are given as

$$B_{21} = \frac{g_2}{g_1} B_{21} \tag{10.107a}$$

$$\frac{A_{21}}{B_{21}} = \frac{8\pi \nu^3 n_r^3}{c^3} \tag{10.107b}$$

where c = speed of light
n_r = refractive index of medium
$\nu = \omega/(2\pi)$

When the spontaneous emission lifetime τ_{21} is equal to the recombination lifetime τ_r, then $A_{21} = 1/\tau_r$ and $B_{21} = c^3/(8\pi \nu^3 n_r^3 \tau_r)$.

The ratio of the stimulated emission rate to the spontaneous emission rate is an important parameter for the laser action, which can be obtained by substituting Eq. (10.107*a*) into (10.106) such that

$$\frac{B_{21}\Phi(\omega_{12})}{A_{21}} = \frac{1}{e^{\hbar\omega_{12}/(k_B T)} - 1} \tag{10.108}$$

For stimulated emission to be dominant, this ratio must be made large. This could be accomplished by letting $\Phi(\omega_{12})$ be very large and having N_2 be larger than N_1. The latter condition is called the population inversion. To achieve the population inversion condition, a large amount of energy is required to excite the carrier from the ground state to the excited state. For a two-energy-level system, the population ratio is unity at best ($N_1 = N_2$) which does not produce significant optical gain. Many of the laser systems, in particular gas and solid-state systems, are based on three or four energy levels.

Lasing and optical gain in a semiconductor system are different than for two- or three-energy level systems. It is based on the creation of nonequilibrium conditions of the charge carriers in both the conduction and valence bands. Figure 10.24 illustrates the absorption and lasing action of a semiconductor material. The initial conditions at $T = 0$ K are that the valence band (VB) is completely full and the conduction band (CB) is completely empty. Incident photons with an energy $E = \hbar\omega_{12}$ generate electron-hole pairs such that the quasi-Fermi energies are

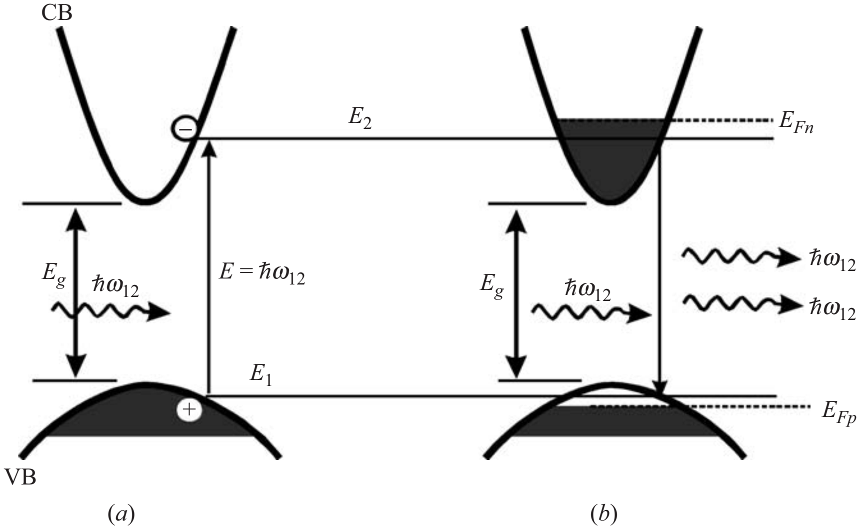


Figure 10.24 (*a*) Absorption and (*b*) stimulated emission in a direct semiconductor material at 0 K.

above E_2 and below E_1, as shown in Fig. 10.24*b*. Under this condition, photons with energies between E_g and $(E_{Fn} - E_{Fp})$ cannot be absorbed. On the other hand, photons with this energy, i.e., $E_g < \hbar\omega < (E_{Fn} - E_{Fp})$, can induce stimulated emission as shown in Fig. 10.24*b*. The absorption rate R_{abs} of photons with energy E lies between E_g and $(E_{Fn} - E_{Fp})$ can be expressed as

$$R_{\mathrm{abs}} = \mathcal{W}[1 - f_n(E_2)]f_p(E_1)N_p(E) \tag{10.109}$$

where $\mathcal{W}$ is the transition probability discussed in Chap. 6, $N_p(E)$ is the density of photons of energy E, and $f_n(E_2)$ and $f_p(E_1)$ are the Fermi-Dirac distribution functions of the electrons and holes, respectively. Similarly, the stimulated emission rate R_{st} can be written as

$$R_{\mathrm{st}} = \mathcal{W}[1 - f_p(E_1)]f_n(E_2)N_p(E) \tag{10.110}$$

To achieve optical gain and population inversion the following conditions must be satisfied:

$$R_{\mathrm{st}} > R_{\mathrm{abs}} \qquad \text{and} \qquad f_n(E_2) > f_p(E_1) \tag{10.111}$$

When the photon energy satisfies the condition $\hbar\omega = (E_{Fn} - E_{Fp})$, the photons are not absorbed by the charge carriers and they pass through without losing their energy. This is known as the transparency point. For a voltage-driven semiconductor laser, the applied bias voltage V_b must satisfy the following condition in order for the lasing to occur:

$$E_g < eV_b = E_{Fn} - E_{Fp} \tag{10.112}$$

From this formalism, the optical gain is positive when $\hbar\omega = E_g$ and continues to increase until it reaches a maximum. Then it starts to decrease and approaches zero as $\hbar\omega$ approaches $(E_{Fn} - E_{Fp})$. The condition (10.112) tells us that the lasing action occurs when the gain is equal to or larger than the loss. This can be expressed as

$$g(E) \geq -\alpha(E) \tag{10.113}$$

where $g(E)$ is the optical gain and $\alpha(E)$ is the absorption coefficient.

The recombination lifetime and the optical gain are two important parameters for lasing action. Another important parameter is the threshold conditions for lasing. If the semiconductor structure is made as a waveguide and if the waveguide is cleaved at both ends, the two parallel facets form a resonant cavity that acts as Fabry-Perot cavity, as shown in Fig. 10.25*a*. The length of the cavity L is given by

$$L = \frac{m\lambda}{2} = \frac{m\lambda_o}{2n_r} \tag{10.114}$$

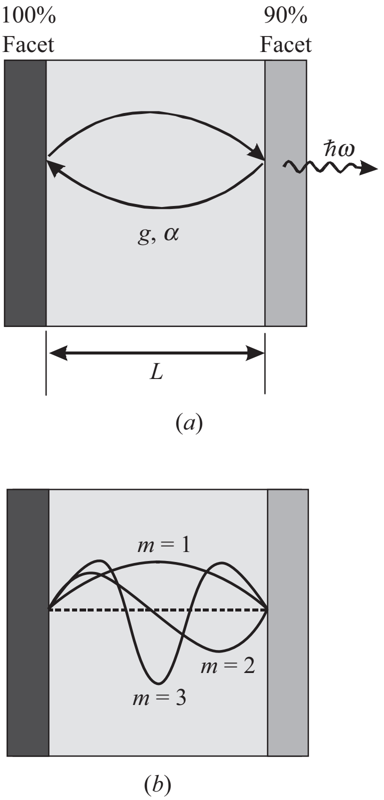


Figure 10.25 (a) A waveguide and two parallel facets form the laser cavity or the Fabry-Perot cavity. A round-trip of the light in the cavity is shown. (b) Three longitudinal normal modes of the laser are sketched for parallel facets.

where m = integer representing number of allowed modes in cavity
n_r = refractive index
λ = photon wavelength within cavity
λ_o = photon wavelength in vacuum = λn_r

The allowed wavelengths within the cavity are called longitudinal optical modes. Three longitudinal modes are shown in Fig. 10.25b. The transverse optical modes can be observed when the facets or the mirrors at the end of the cavity are not quite parallel. This leads to a special intensity distribution at the exit facet (or mirror). Notice that the reflectivity of one facet is made 100 percent, while the reflectivity of the other facet is made less than 100 percent so that the laser will exit the cavity.

The amplifying medium between the two mirrors or the two facets is characterized by a well-defined wavelength region in which the stimulated emission occurs. If the medium is characterized by an optical gain

g and absorption (loss) coefficient α, and if the reflectivities of two facets are represented by R_1 and R_2, the intensity of the light I that made a round-trip ($2L$) within the cavity (see Fig. 10.25*a*) can be expressed as

$$I = I_o R_1 R_2 e^{2(g-\alpha)L} \tag{10.115}$$

where I_o is the initial intensity. When $g > \alpha$, the light intensity increases cause amplification of light.

The threshold gain g_t can be obtained when the optical gain is equal to the optical loss. In this case $I = I_o$ and Eq. (10.115) becomes

$$1 = R_1 R_2 e^{2(g_t-\alpha)L} \tag{10.116}$$

or

$$g_t = \alpha + \frac{1}{2L} \ln\left(\frac{1}{R_1 R_2}\right) \tag{10.117}$$

The second expression on the right-hand side of Eq. (10.117) represents the useful laser output. In general, a few round-trips of light between the two mirrors are needed before the lasing reaches a steady state and the gain reached the threshold gain g_t.

Let us reexamine Eq. (10.114). The integer m represents the number of longitudinal modes. For a large value of m, one can write Eq. (10.114) as

$$\partial m = 2L\partial\left(\frac{n_r}{\lambda_o}\right) = \frac{2L}{\lambda_o}\partial n_r - \frac{2Ln_r}{\lambda_o^2}\partial\lambda_o \tag{10.118}$$

For discrete changes in m and λ_o, one can rewrite Eq. (10.118) as

$$-\partial\lambda_o = \left(\frac{2L}{\lambda_o}\frac{dn_r}{d\lambda_o} - \frac{2Ln_r}{\lambda_o^2}\right)^{-1}\delta m = \frac{\lambda_o^2}{2Ln_r}\left(\frac{\lambda_o}{n_r}\frac{dn_r}{d\lambda_o} - 1\right)^{-1}\delta m \tag{10.119}$$

Thus, the wavelength change between adjacent longitudinal modes is

$$\partial\lambda_o = \frac{\lambda_o^2}{2Ln_r}\left(1 - \frac{\lambda_o}{n_r}\frac{dn_r}{d\lambda_o}\right)^{-1} \tag{10.120}$$

The frequency separation between adjacent longitudinal modes ($\Delta m = \pm 1$) is given by

$$\delta\nu = \frac{c}{2Ln_r} \tag{10.121}$$

This separation is illustrated in Fig. 10.26*a*, where several longitudinal modes are sketched. The optical gain is sketched in Fig. 10.26*b*. A large number of laser longitudinal modes exist. However, the only modes that are amplified are those within the optical gain region, as

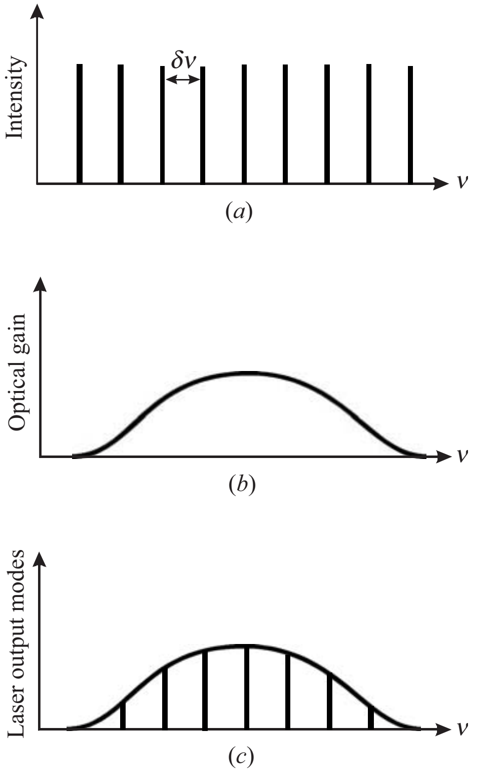


Figure 10.26 (*a*) Longitudinal modes of a Fabry-Perot cavity. (*b*) The output optical gain profile, which is essentially the spontaneous emission profile of the laser. (*c*) Only the longitudinal modes that are within the optical gain can be amplified.

shown in Fig. 10.26*c*. In general, the optical gain profile resembles the spontaneous emission region. A semiconductor laser can be designed to produce a few longitudinal modes. This can be accomplished by reducing the optical gain linewidth.

10.4.2 Semiconductor heterojunction lasers

The earliest semiconductor lasers were made of degenerate *pn* junctions. In these lasers, a large forward-biased voltage is applied to the junction, where a large density of electrons and holes are injected. The lasing action usually takes place in a narrow region in the *p*-type side, as shown in Fig. 10.27. Semiconductor lasers do not need external mirrors to form the cavity. The Fabry-Perot cavity is formed by cleaving the planes at the end of the waveguide. The small region in which the radiative electron-hole recombination occurs is called the active region. There are several disadvantages associated with the *pn* junction laser, such as the active region is not well defined, the cavity loss is large, and the threshold current density is large.

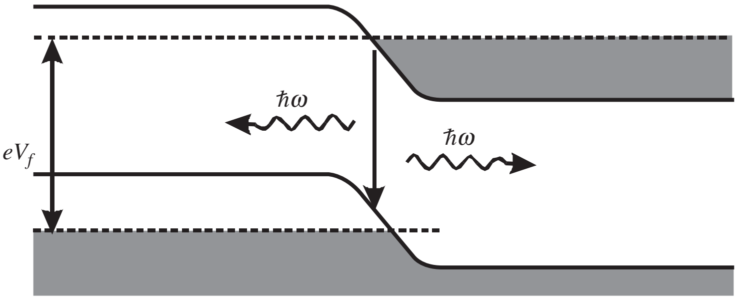


Figure 10.27 Degenerate pn-junction laser under forward-biased voltage V_f.

Semiconductor heterojunction lasers provide a better confinement for both the charge carriers and the longitudinal optical modes. The basic structure of a heterojunction laser is shown in Fig. 10.28*a* for a GaAs/AlGaAs heterojunction at equilibrium. It consists of n-type GaAs, p-type GaAs, and p-type AlGaAs. The band diagram for large forward bias is shown in Fig. 10.28*b* where the carriers are injected in both sides of the junction. Population inversion is achieved, and lasing action is

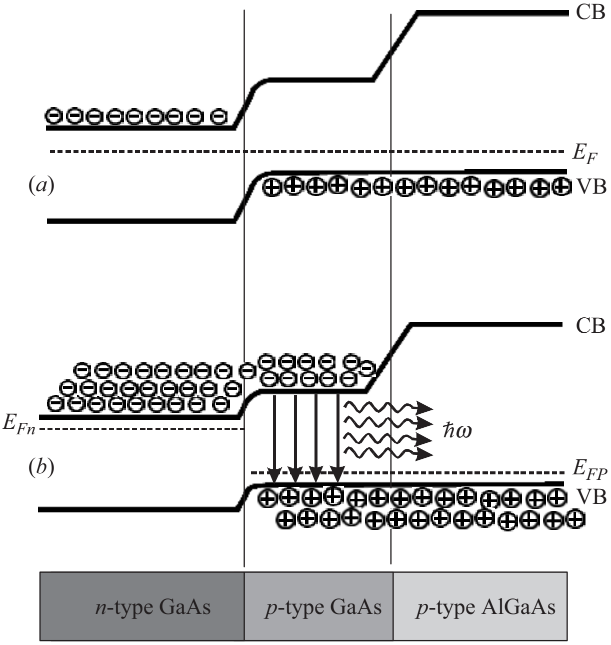


Figure 10.28 A GaAs/AlGaAs heterojunction used as a diode laser (*a*) at equilibrium and (*b*) under the influence of a large forward-biased voltage. Notice that the valence band offset between GaAs and AlGaAs is approximated to zero.

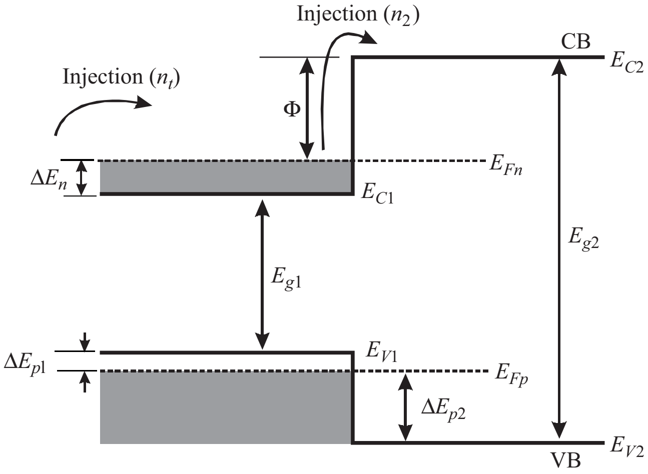


Figure 10.29 Sketch of the band diagram of a GaAs/AlGaAs heterojunction diode laser used to derive the threshold current density.

obtained. Notice that the valence band offset between p-type GaAs and p-type AlGaAs is approximated to zero. This approximation is realistic only when a small aluminum mole fraction is used in the AlGaAs layer. For a more realistic GaAs/AlGaAs heterojunction diode laser, the total band offset is usually divided into a 60 to 40 percent ratio between the conduction band and valence band offsets, respectively.

The current density needed to turn on the diode laser is called the threshold current density J_t, which is due to the injection and diffusion processes. The p-type GaAs active region thickness is usually smaller than the diffusion length of the carrier, but in the following analysis, the diffusion current density is not neglected. Consider the heterojunction, shown in Fig. 10.29, where the injection level is assumed to be high. The effective confinement barrier Φ shown in the figure can be written as

$$\Phi = \Delta E_g - \Delta E_n - \Delta E_{p1} - \Delta E_{p2} \tag{10.122}$$

where ΔE_g is the total band offset ($\Delta E_g = E_{g2} - E_{g1}$). For the steady-state case, the threshold current density can be written as

$$J_t = J_{\text{inj}} + J_{\text{diff}} = \frac{en_t\, d}{\tau} + \frac{eD_{e2}n_2}{L_{e2}} \tag{10.123}$$

where n_t = injected carrier density in active region (p-type GaAs)
d = thickness of active region
D_{e2} = electron diffusion coefficient in barrier region (p-type AlGaAs)
L_{e2} = electron diffusion length in barrier region
n_2 = injected minority carrier density in barrier

Using the formalisms of the density of states presented in Chap. 5 and Fermi-Dirac distribution functions, the carrier densities n_t and n_2 can be written as

$$n_t = \frac{4\pi}{h^3}(2m_{e1}^*)^{3/2} \int_{E_{c1}}^{\infty} \frac{\sqrt{E - E_{c1}}}{1 + e^{(E-E_{Fn})/(k_B T)}} dE \qquad (10.124a)$$

$$n_2 = \frac{4\pi}{h^3}(2m_{e2}^*)^{3/2} \int_{E_{c2}}^{\infty} \frac{\sqrt{E - E_{c2}}}{1 + e^{(E-E_{Fn})/(k_B T)}} dE \qquad (10.124b)$$

where E_{c1} and E_{c2} are the conduction band minima of GaAs and AlGaAs, respectively, and m_{e1}^* and m_{e2}^* are the electron effective masses in GaAs and AlGaAs, respectively. For $(E_{c2} - E_{Fn}) \gg k_B T$, the Fermi distribution functions can be approximated as the Boltzmann distribution function. Thus, Eq. (10.124b) can be rewritten as

$$\begin{aligned} n_2 &= \frac{4\pi}{h^3}(2m_{e2}^*)^{3/2} \int_{E_{c2}}^{\infty} \frac{\sqrt{E - E_{c2}}}{1 + e^{(E-E_{Fn})/(k_B T)}} dE \approx \frac{4\pi}{h^3}(2m_{e2}^*)^{3/2} \int_{E_{c2}}^{\infty} \frac{\sqrt{E - E_{c2}}}{e^{(E-E_{Fn})/(k_B T)}} dE \\ &= \frac{2}{h^3}(2\pi m_{e2}^* k_B T)^{3/2} e^{-(E_{c2}-E_{Fn})/(k_B T)} = N_{c2} e^{-(E_{c2}-E_{Fn})/(k_B T)} \qquad (10.125) \end{aligned}$$

where N_{c2} is the effective density of state given by $N_{c2} = (2/h^3)(2\pi m_{e2}^* k_B T)^{3/2}$. The diffusion current density, which is also known as the leakage current density, can be written as

$$J_{\text{diff}} = \frac{eD_{e2}n_2}{L_{e2}} = \frac{eD_{e2}}{L_{e2}} N_{c2}\, e^{-(E_{c2}-E_{Fn})/(k_B T)} = \frac{eD_{e2}}{L_{e2}} N_{c2}\, e^{-(\Phi)/(k_B T)} \qquad (10.126)$$

Notice that $\Phi = (E_{c2} - E_{Fn})$, as illustrated in Fig. 10.29. Substituting Eq. (10.122) into (10.126) yields

$$J_{\text{diff}} = \frac{eD_{e2}}{L_{e2}} N_{c2} \exp\left[\frac{-(\Delta E_g - \Delta E_n - \Delta E_{p1} - \Delta E_{p2})}{k_B T}\right] \qquad (10.127)$$

For a high injection rate, $\Delta E_{p1} \approx 0$ and the majority carriers in the barrier can be written as

$$p_2 = N_{v2}\, e^{-\Delta E_{p2}/(k_B T)} \qquad (10.128)$$

where N_{v2} is the effective density of states in the valence band of the barrier (p-AlGaAs). Combine Eqs. (10.127) and (10.128) to obtain

$$J_{\text{diff}} \approx \frac{eD_{e2}N_{c2}}{L_{e2}} \frac{N_{v2}}{p_2}\, e^{-(\Delta E_g - \Delta E_n)/(k_B T)} \qquad (10.129)$$

For high p-type doping, it can be assumed that $N_{v2} \approx p_2$, which yields

$$J_{\text{diff}} \approx \frac{eD_{e2}N_{c2}}{L_{e2}} e^{-(\Delta E_g - \Delta E_n)/(k_B T)} \tag{10.130}$$

For a heterojunction such as GaAs/AlGaAs, ΔE_g can be made high enough such that the leakage current density described in Eq. (10.130) is negligible. It is very important to keep the aluminum mole fraction in the barrier material less than 40 percent so that the bandgap of the barrier material remains direct. The effective density of states N_{c2} increases for indirect AlGaAs since the effective mass in the X and L minima are usually higher that that of the Γ minimum.

By neglecting the diffusion current density, the threshold current density given by Eq. (10.123) can now be approximated as

$$J_t \approx \frac{en_t d}{\tau} \tag{10.131}$$

The laser output power P_{out} is related to the photon density N_{ph} according to the following expression:

$$P_{\text{out}} = V\hbar\omega \upsilon_g \alpha_m N_{\text{ph}} \tag{10.132}$$

where V = volume of active region
υ_g = light group velocity
α_m = optical loss due to facets

The photon density N_{ph} is given by

$$N_{\text{ph}} = \eta_i \frac{J - J_t}{eD} \tau_{\text{ph}} \tag{10.133}$$

where η_i is the internal quantum efficiency, J is the current density ($J > J_t$), and τ_{ph} is the photon lifetime, which is given by

$$\tau_{\text{ph}} = \frac{1}{\upsilon_g(\alpha_m + \alpha_i)} \tag{10.134}$$

where α_i is the optical loss due to the active region. By combining Eqs. (10.132) to (10.134) and converting the current density to current ($J = I/A$, where A is the area of the active region), we obtain

$$P_{\text{out}} = \frac{\hbar\omega}{e} \frac{\alpha_m}{\alpha_m + \alpha_i} \eta_i (I - I_t) \tag{10.135}$$

The typical laser power output as a function of applied bias current is shown in Fig. 10.30, where P_{out} and the voltage are sketched as a function of the applied bias current. The threshold current I_t and the turn-on voltage V_t for the laser diode are indicated.

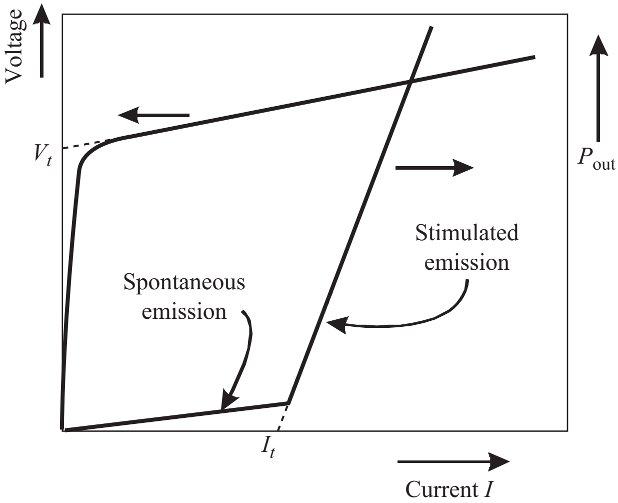


Figure 10.30 The characteristic behavior of the semiconductor laser power output as a function of the applied bias current. The spontaneous emission and stimulated emission regions are indicated.

The laser diode resistance is obtained from the slope of the linear portion of the I-V curve. The slope of the P_{out} curve in the stimulated emission region is usually referred to as the slope efficiency. The external differential quantum efficiency is defined as the increase in the light output due to the increase in the applied bias current and is given as

$$\eta_d = \frac{dP_{\text{out}}/\hbar\omega}{d(I - I_t)/e} = \eta_i \frac{\alpha_m}{\alpha_i + \alpha_m} \tag{10.136}$$

The facet loss α_m can be determined from the following relation:

$$1 = R_1 R_2 e^{-\alpha_m 2L} \qquad \text{or} \qquad \alpha_m = \frac{1}{2L} \ln\left(\frac{1}{R_1 R_2}\right) \tag{10.137}$$

For $R_1 = R_2 = R$, the external differential efficiency can be written as

$$\eta_d = \frac{dP_{\text{out}}/\hbar\omega}{d(I - I_t)/e} = \eta_i \frac{\ln(1/R)}{\alpha_i L + \ln(1/R)} \tag{10.138}$$

For most III-V semiconductor heterojunction laser diodes, the internal quantum efficiency is close to unity and the optical loss in the active region is on the order of 10 to 100 cm^{-1}. For a GaAs/AlGaAs heterojunction diode laser with a cavity length of 0.5 mm and a reflectivity on the order of $R \approx 0.3$, then $\eta_d \approx 0.7\eta_i$ for $\alpha_i = 10$ cm^{-1}.

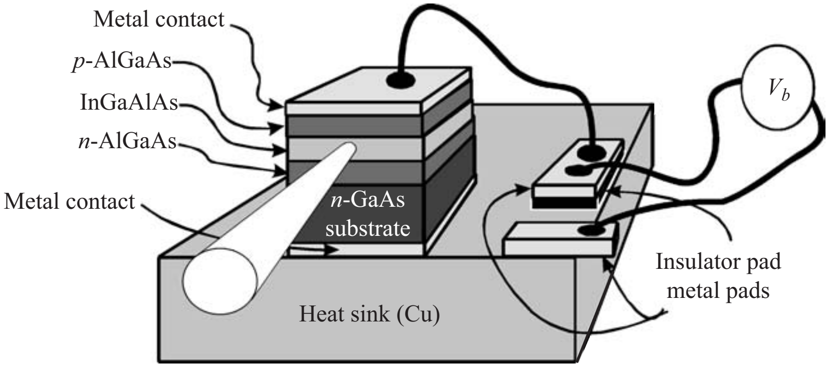


Figure 10.31 A sketch of an edge-emitting laser based on an InGaAlAs/AlGaAs single quantum well.

10.4.3 Quantum well edge-emitting lasers

A single quantum well edge-emitting laser diode is sketched in Fig. 10.31. The structure of the laser consists of an undoped InGaAlAs quantum well, which is the active region, sandwiched between the barrier layers of n-type AlGaAs and p-type AlGaAs. These layers also act as the waveguide for the photons. The sample is cleaved, and the facets act as the mirrors that form the laser cavity. The heat sink is a copper block. The laser output power is plotted as a function of the applied current as shown in Fig. 10.32. This figure shows the two characteristic regions: the spontaneous emission region below the threshold current and the stimulated emission region above the threshold current.

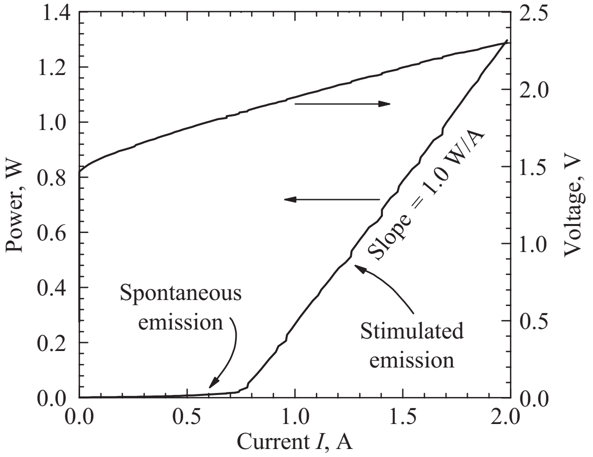
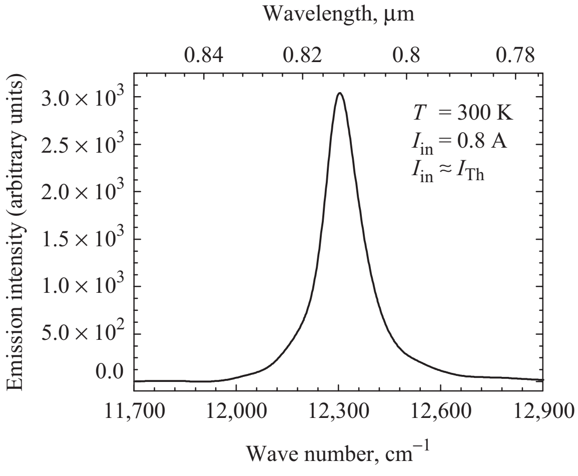
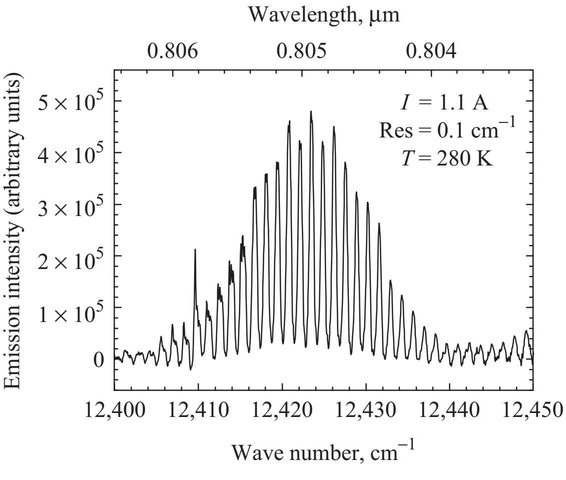


Figure 10.32 The output power of the edge-emitting laser, shown in Fig. 10.31, plotted as a function of the applied current. The I-V curve is also shown.



(*a*) Spontaneous emission



(*b*) Stimulated emission

Figure 10.33 (*a*) A Fourier-transform spontaneous emission spectrum obtained for the laser described in Fig. 10.31 when the applied current is approximately the same as the threshold current. (*b*) Stimulated emission spectrum obtained when the applied current is increased above the threshold current. The spectrum shows many longitudinal modes.

When the current is below the threshold current, the laser operates in the spontaneous emission mode, which is identical to the LED operation mode. The spontaneous emission spectrum shown in Fig. 10.33*a* is obtained using the laser as the source for a Fourier-transform spectrometer. The applied current is approximately equal to the threshold current.

Even when the applied current is just below the threshold current, the spectrum still resembles that of the LED spontaneous emission mode. The laser longitudinal modes are observed as shown in Fig. 10.33*b* when the applied current is increased above the threshold current. The longitudinal modes are observed at a spectral resolution as low as 0.1 cm^{-1}.

Ideally, the full width at half maximum of the longitudinal modes should be zero. The spectrum in Fig. 10.33*a* shows that these modes have a finite width, which is due to the photon attenuation and arises from several factors such as the finite facet transmission and the absorption in the cavity. Thus, instead of infinitely narrow longitudinal modes, we have modes with a finite full width at half maximum, $\Delta\omega$, given by

$$\Delta\omega = \frac{\pi c}{L\mathcal{F}} \tag{10.139}$$

where L is the length of the cavity, c is the speed of light, and $\mathcal{F}$ is a factor called the finesse or the contrast parameter given by

$$\mathcal{F} = \frac{\pi\sqrt{\gamma}}{1-\gamma} \tag{10.140}$$

where γ is the attenuation or loss factor in the cavity. For known γ and known full width at half maximum, one can calculate the cavity length using Eq. (10.139). The cavity length can also be calculated using the separation between the longitudinal modes given by Eq. (10.121). A more accurate method of obtaining the cavity length of the laser is by obtaining the interferogram from the Fourier-transform spectrometer as shown in Fig. 10.34. The interferogram contains several interference packets separated by the cavity length. In this case the packets are separated by 0.911 mm (911 μm). Each packet consists of an interference pattern as shown in the inset of the figure.

The edge-emitting laser structure described in this section is considered a high-power laser, since the output power is on the order of watts rather than milliwatts. The quantum well is compressively strained due to the lattice mismatch. The actual structure is comprised of a 70-Å $\mathrm{In_{0.06}Ga_{0.86}Al_{0.08}As}$ quantum well (active region) and $\mathrm{Al_{0.3}Ga_{0.7}As}$ barriers (waveguide). Because of the difference in the refractive indices of the quantum well (n_{rw}) and barriers (n_{rb}), the photons are confined in the well as shown in Fig. 10.35.

In the edge-emitting laser, the feedback needed for lasing action is achieved by the cavity facets formed by cleaving. For higher mode purity, the stimulated emission can be obtained by introducing a *distributed Bragg reflector* (DBR). This type of reflectors usually has a corrugated structure or grading introduced in the waveguide.

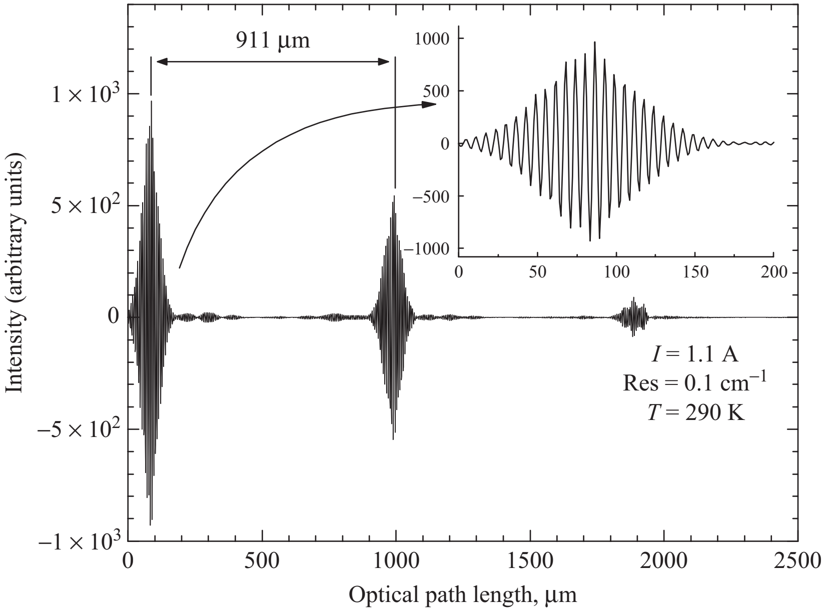


Figure 10.34 The interferogram of the edge-emitting laser described in Fig. 10.31 operating in the stimulated emission mode. The inset is the enlarged first packet.

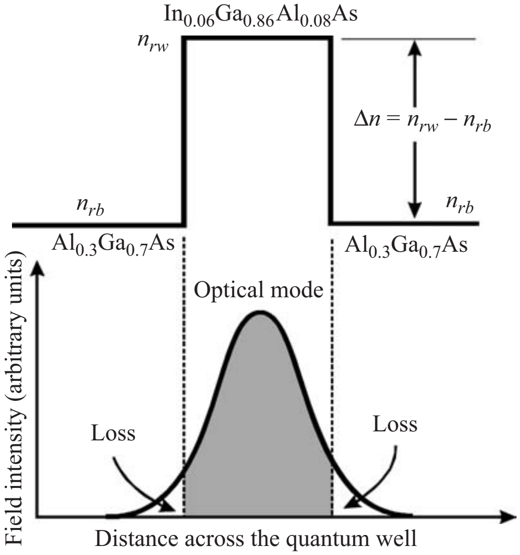

$In_{0.06}Ga_{0.86}Al_{0.08}As$
n_{rw}
$\Delta n = n_{rw} - n_{rb}$
n_{rb}
n_{rb}
$Al_{0.3}Ga_{0.7}As$
$Al_{0.3}Ga_{0.7}As$
Optical mode
Field intensity (arbitrary units)
Loss
Loss
Distance across the quantum well

Figure 10.35 Optical mode distribution in the single-quantum-well laser due to the step discontinuity in the refractive indices of the well (n_{rw}) and barriers (n_{rb}).

10.4.4 Vertical cavity surface-emitting lasers

Optical interconnects between chips, large arrays of sources, narrow beam widths, and generation of high power are a few applications of vertical cavity surface-emitting lasers (VCSELs). Optical interconnects based on VCSELs may play a major role at different levels of the interconnection hierarchy for data communication networks. Optical interconnect technology emerged in the late 1970s using AlGaAs/GaAs-based or InGaAsP/InP-based surface-emitting LEDs as the optical source.

One of the major research advances in VCSEL technology is the discovery of the selective wet oxidation of AlGaAs layers with high aluminum contents. This technique allows the fabrication of very small microcavity devices with low threshold current to be realized by eliminating the high carrier loss that occurs due to the defects created by ion implantation. Two-dimensional arrays of VCSELs are useful for implementing very dense optical interconnects spanning short distances that can potentially alleviate the peripheral pin-out limit of the very large scale integrated chips. These arrays can significantly reduce the power dissipation of parallel optical links. Parallel optical interconnects represent a compact and potentially low cost approach for transmitting a large volume of data across longer distances, thus replacing the large-profile shielded electrical cables that are needed to interconnect computer processors across local-area networks.

Another important factor in VCSEL technology is the development of the high-reflectivity semiconductor distributed Bragg reflectors, which are made possible with the advancement of the crystal growth using the MBE and MOCVD growth techniques. A typical example of a VCSEL structure is shown in Fig. 10.36. It consists of a GaAs/AlGaAs multiple-quantum-well active region. This active region is referred to as the multiple-quantum-well graded index separate confinement heterostructure (MQW-GRINCH), which is bounded by two distributed Bragg reflector (DBR) mirrors. The active region is composed of a GaAs/$Al_{0.15}Ga_{0.85}As$ MQW, and the DBRs are several pairs of n-type AlAs/$Al_{0.15}Ga_{0.85}As$ at the bottom and several pairs of p-type AlAs/$Al_{0.15}Ga_{0.85}As$. The proton-implanted regions are used as isolation regions, which define the active region area of the VCSEL. The active area is usually 15 to 20 μm in diameter. The aluminum profile of the structure is also shown in Fig. 10.36.

The threshold current of a proton-implanted isolated VCSEL is usually high ($I_t > 2$ mA). Thus, it is impractical to use such an emitter in dense parallel array applications. Selective oxidation has been used recently for VCSEL isolation, which reduces the threshold current to the sub-milliampere range and with a typical power efficiency larger

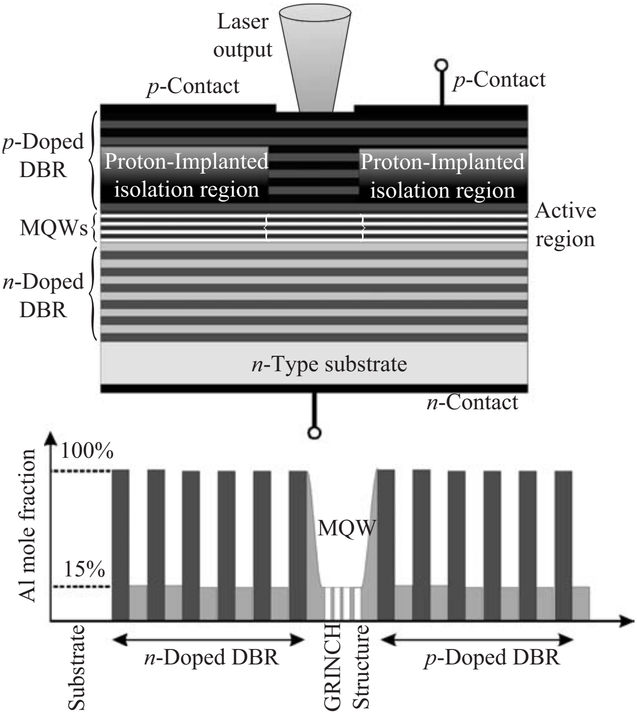


Figure 10.36 A schematic showing the VCSEL device structure whose active region is defined by the proton implantation. The aluminum composition profile of the structure is shown in the lower panel.

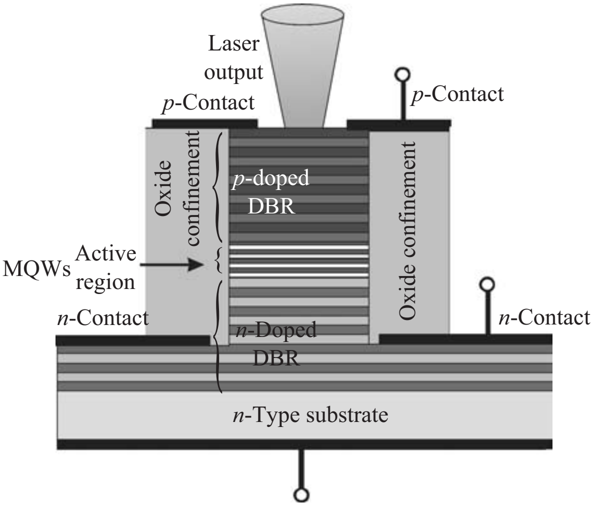


Figure 10.37 A sketch of an oxide-confined VCSEL structure.

than 30 percent. An example of an oxide confinement VCSEL is shown in Fig. 10.37. The active region of the oxide confined VCSEL is usually less than 10 μm in diameter. Because of the small thickness of the active region in the VCSELs, the separation between the longitudinal modes is very large as compared to the separation between the longitudinal modes in edge-emitting lasers.

The refractive index of AlGaAs is ~3.0, while the refractive index of the oxidized layer is on the order of 1.6. The differential change of the refractive index provides an excellent index-guiding optical confinement. Unlike the ion-implanted VCSELs, the index guiding in the selectively oxidized VCSELs produces a monotonic decrease of the threshold currents. For additional discussion on VCSELs, see Cheng and Dutta (2000).

10.4.5 Quantum cascade lasers

The advances in epitaxial growth, in particular the MBE and MOCVD growth methods, allow the layer-by-layer deposition of semiconductor materials with dissimilar bandgaps and lattice constants with high accuracy. The precise control of the dopants and layer thicknesses permits one to engineer the bandgap of complicated structures for specific applications. In many ways, the well-controlled growth conditions can be thought of as wave function engineering. An example of this sophisticated epitaxial growth is the elaborate structures of the quantum cascade lasers. Unlike quantum well laser diodes described in previous sections, the quantum cascade lasers are unipolar devices where the charge carriers are either electrons or holes. Recently, quantum cascade lasers were shown to emit coherent light in the far infrared regions. The basic operational principle of the quantum cascade laser is the downward intersubband transitions where the injected electrons decay from an excited bound state to a low-lying energy level, emitting their energy as photons.

The structures and designs of quantum cascade lasers are numerous. For example, three simple structures are sketched in Fig. 10.38. An interwell transition is shown in Fig. 10.38*a* where the electrons decay under the influence of a forward bias from the ground state in one quantum well to an excited state in an adjacent quantum well emitting a photon in the process. The second structure shown in Fig. 10.38*b* is based on electron resonant tunneling followed by a radiative downward transition. The third structure is comprised of coupled double quantum wells with three bound energy levels. Population inversion is established in the excited state due to the spatial separation of the two states by the barrier layers. The cleaved facets of the structures act as the mirrors.

The lasing threshold gain g_t in a quantum cascade laser can be written as

$$g_t = \frac{\alpha_i + \alpha_m}{\Gamma} \tag{10.141}$$

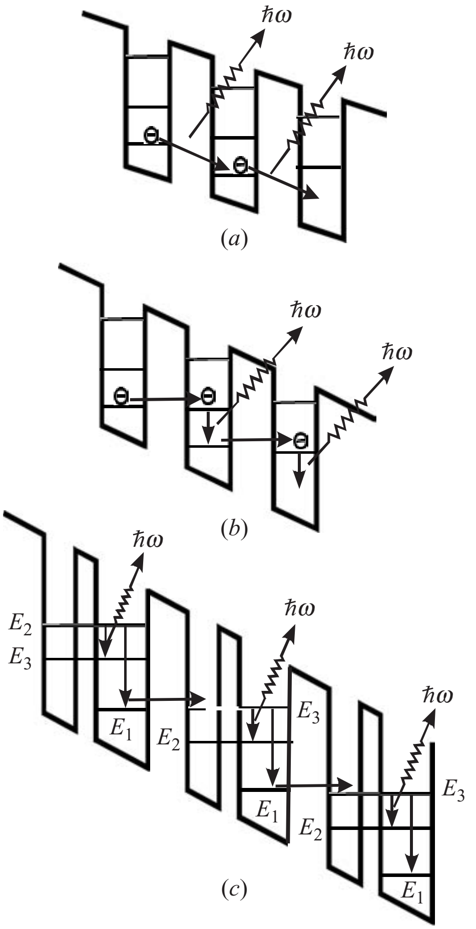


Figure 10.38 Schematics of three different structures of quantum cascade lasers: (*a*) interwell radiative transition, (*b*) intrawell radiative transition, and (*c*) coupled double-well structures with three energy levels.

where Γ is the optical confinement factor, α_m is the mirror loss due to the finite facet reflectivities, and α_i is the internal loss for the optical wave, which results from various absorption mechanisms, such as scattering and free carrier absorption. The confinement factor can be approximated as

$$\Gamma \approx 2\pi^2 \left(n_1^2 - n_2^2\right) \frac{d^2}{\lambda^2} \tag{10.142}$$

where n_1, n_2 = refractive index of active region and cladding region, respectively
d = well width
λ = wavelength

The threshold gain depends on several factors and is related to the threshold current density J_t according to the following relation:

$$g_t = \frac{J_t L T_s \eta_{\mathrm{pi}} \eta_r \eta_{\mathrm{inj}}}{e} \tag{10.143}$$

where L = parameter with dimension of length related to emission wavelength
T_s = parameter with dimension of time that depends on inverse line-width of spontaneous emission
η_{pi} = population inversion efficiency
η_r = radiative efficiency
η_{inj} = injection efficiency

For more details on these parameters, see Helm (2000).

More complicated quantum cascade structures have been reported by several groups (see, for example, Helm 2000). Two typical examples are shown in Fig. 10.39. The first structure, shown in Fig. 10.39*a*, consists of three coupled quantum well active regions with the three energy levels separated by injection regions. The injection region is composed of several *n*-type short-period superlattices or digitally graded alloys, which serve as the collector for the preceding active region and as an emitter for the following active region. The discrete tunneling scheme presented in this figure prevents the formation of electric domains as observed in the continuous sequential tunneling in the superlattice structures. It also eases the rigorous requirement of position-dependent energy level alignment over the whole superlattice structures.

The second quantum cascade laser structure shown in Fig 10.39*b* is basically two coupled quantum wells with three energy levels sandwiched between superlattices that act as distributor Bragg reflectors. This scheme is less sensitive to the defects, interface roughness, and impurity fluctuations. Additionally, the DBR possesses minibands and a minigap as shown in the figure. The minigap prevents the electrons from escaping to the continuum.

The structure of the quantum cascade laser is based on type II superlattices or broken bandgap superlattices. A schematic structure of a type II quantum cascade laser is shown in Fig. 10.40. Several variations of this structure have been reported by many research groups. The structure represents a discrete tunneling scheme where the actual device structure is composed of many active regions separated by an *n*-type doped injection region. These injection regions consist of either graded AlInAsSb quaternary layers or digitally graded InAs/AlSb superlattices.

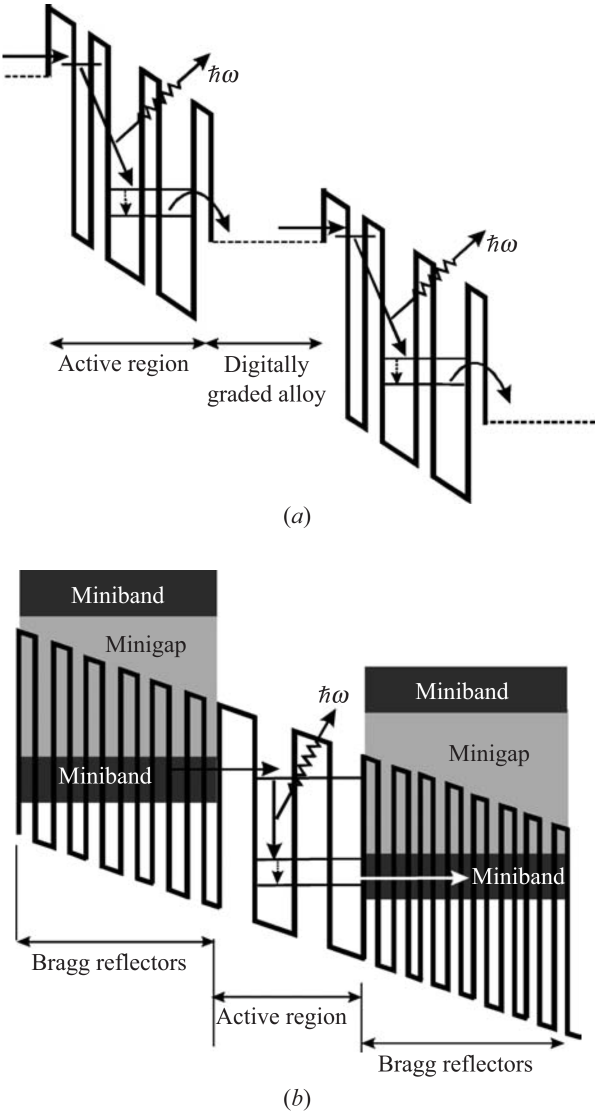


Figure 10.39 (*a*) A schematic presentation of the conduction band energy diagram of a portion of a quantum cascade laser based on intersubband transition. (*b*) The conduction band energy diagram of a portion of a vertical transition quantum cascade laser structure. The dashed arrows indicate electron transitions due to phonon scattering.

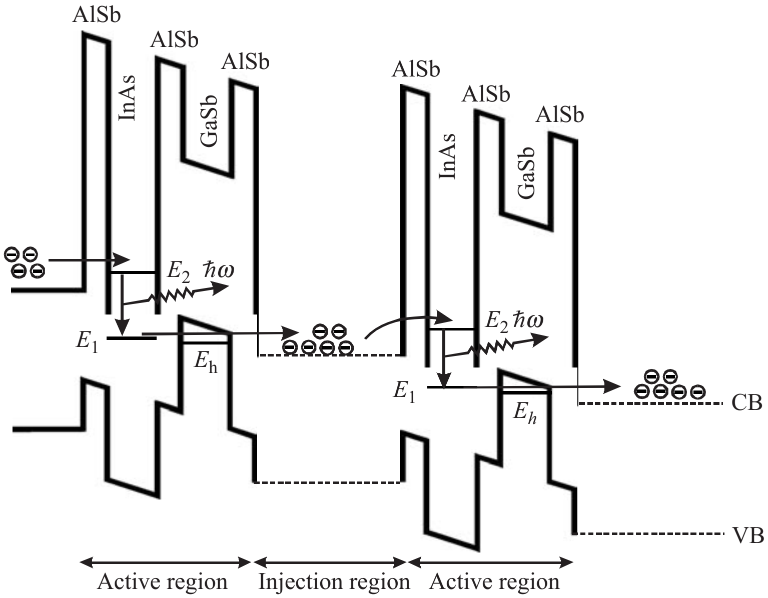


Figure 10.40 A portion of a quantum cascade laser structure based on type II superlattices.

10.4.6 Quantum dot lasers

The search for new and novel structures for quantum wire and dot lasers is motivated by their low-threshold current density, weaker temperature dependence of the threshold current density or temperature-insensitive threshold current, tuning the gain spectrum width and wavelength, and low shift of the laser wavelength with respect to the injection current (low chirp). The quantum well lasers are now used for most commercial applications. The new frontier, however, is the quantum wires and quantum dot lasers where the density of states spans a small energy region. As shown in Fig. 10.41, the density of states and the optical gain of the laser are sketched in four confinement systems. The first is the bulk semiconductor system where the optical gain spectrum width is broad and the density of states is continuous. See Chap. 5 for a detailed discussion on the density of states. The second is a quantum well structure where the density of states is merely a step, and the optical gain is more localized as compared to the bulk system. The third system is a quantum wire where the confinement is along dimensions with one degree of freedom and the density of states is proportional to the inverse of the square root of the energy. This behavior provides a much narrower optical gain. The fourth system is a quantum dot. With this structure the density of states and the optical gain are both spread over

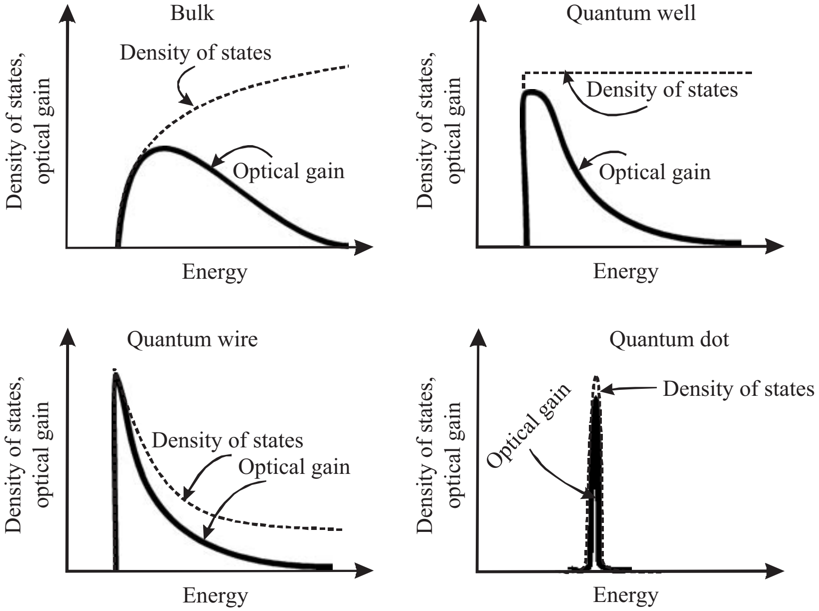


Figure 10.41 Schematic presentations of the density of states and the optical gain spectrum plotted for the active regions of different confinement systems. Bulk semiconductor material is a zero-dimensional confinement system, a quantum well is a one-dimensional confinement system, a quantum wire is a two-dimensional confinement system, and a quantum dot is a three-dimensional confinement system.

a very small region of energy. The quantum dot system thus provides a drastic change in the density of states and optical gain (discrete carrier distribution) which is highly ideal for lasing action with a low threshold current and a high temperature stability.

A typical quantum dot structure used for lasers is shown in Fig. 10.42*a*. The most commonly investigated quantum dot system is the InAs matrixed in GaAs layer, where the GaAs layer acts as the barrier. The cladding layers are added to form the optical cavity or the optical confinement layer. The quantum dots were depicted as semispherical shapes that are randomly distributed in the active region. The band structure diagram is shown in Fig. 10.42*b*. The downward interband radiative transition is shown as the sum of the bandgaps of the quantum dot material, ΔE_n and ΔE_p, where ΔE_n and ΔE_p are the confined energy level positions of the electron in the conduction band and the holes in the valence band, respectively. The band structure shown in Fig. 10.42*b* is sketched for an ideal quantum dot with a cubic shape. In reality, quantum dots are usually pyramidal in shape. The implication is that the band structure of the quantum dots resembles a graded structure where the bandgap decreases along the growth direction.

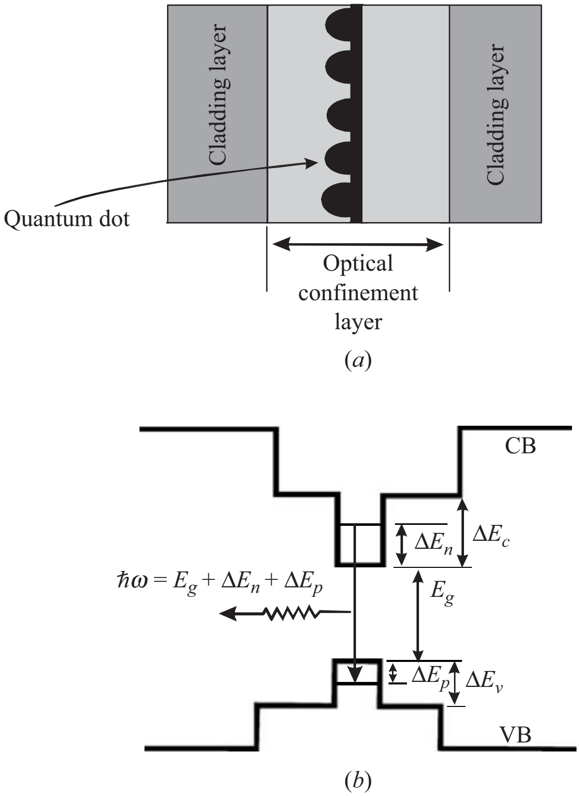


Figure 10.42 (*a*) An illustration of the quantum dot laser active layer defined by the cladding layers. (*b*) A sketch of the energy band diagram showing the downward radiative interband transition.

One drawback of the quantum dot structure is the nonuniformity of the self-assembled quantum dot size. For the ideal case where all quantum dots are identical in size, the optical gain should resemble the density of states in the quantum dots. This means the optical gain spectrum should take the δ-function distribution form. In practice, the self-assembled quantum dots are nonuniform in size. The variation of the dot size produces an inhomogeneous line broadening in the optical gain as shown in Fig. 10.43. In this figure, three quantum dots with different sizes are sketched. The interband transitions are shown as E_1, E_2, and E_3. The variation in these energies causes an inhomogeneous broadening in the optical gain spectrum as shown in the figure.

The inhomogeneous line broadening or dispersion due to the quantum dot size affects the laser characteristic parameters. For example, the maximum gain decreases, the threshold current increases, the internal

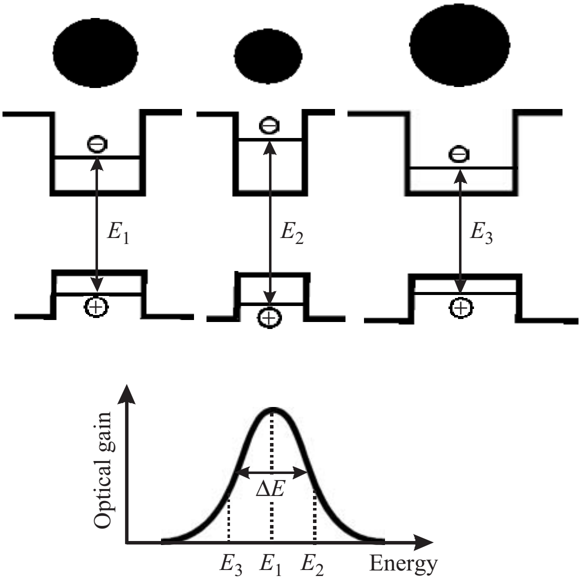


Figure 10.43 Illustration of the inhomogeneous line broadening due to the variation of the quantum dot size.

differential efficiency decreases, the output power decreases, and the threshold current density becomes more sensitive to temperature. The inhomogeneous line broadening ΔE can be expressed as (see Asryan and Luryi 2004).

$$\Delta E = \left(\Delta E_n \frac{\partial \Delta E_n}{\partial a} + \Delta E_p \frac{\partial \Delta E_p}{\partial a} \right) \delta \tag{10.144}$$

where a is the quantum dot mean size and δ is the root mean square of the relative quantum dot size fluctuation. The maximum optical gain is inversely proportion to the ΔE. Thus, the larger the inhomogeneous line broadening, the smaller the maximum optical gain.

Another parameter affected by the variation of the dot size is the chirp or the change in the laser emission wavelength during the current injection. The origin of this shift is related to the coupling of the real and imaginary parts of the complex susceptibility in the laser active region. The variation of the gain due to the current injection causes a variation in the refractive index. The variation of the refractive index modifies the phase of the optical mode in the laser cavity. The coupling strength is defined by the line-width enhancement factor α_c, which is defined as

$$\alpha_c = \frac{\Delta n_r}{\Delta k_r} \tag{10.145}$$

where Δn_r and Δk_r are the changes in the real and imaginary parts of the refractive index, respectively. The change in the imaginary part of the refractive index is related to the change in the net gain, Δg, that occurs during the laser relaxation oscillation, according to the following relation:

$$\Delta g = -\frac{4\pi\nu}{c}\Delta k_r = -\frac{4\pi}{\lambda}\Delta k_r \tag{10.146}$$

where ν is the laser frequency. Substituting Eq. (10.146) into (10.145) yields

$$\alpha_c = -\frac{4\pi}{\lambda}\frac{\Delta n_r}{\Delta g} = \frac{2E}{\hbar c}\frac{\Delta n_r}{\Delta g} \tag{10.147}$$

The change in the gain increases linearly with the injected carrier density while the refractive index decreases linearly with the injected carrier density. Thus, one may rewrite Eq. (10.147) as follows:

$$\alpha_c = -\frac{4\pi}{\lambda}\frac{\partial n_r/\partial N}{\partial g/\partial N} \tag{10.148}$$

where N is the carrier density. For an ideal quantum dot laser, the variation of the refractive index with respect to the carrier density is zero, which means $\alpha_c = 0$ and the laser is chirp-free.

A stack of multiple quantum dots can be used as the active region for a VCSEL as shown in Fig. 10.44. To maintain a low loss, the mirrors need

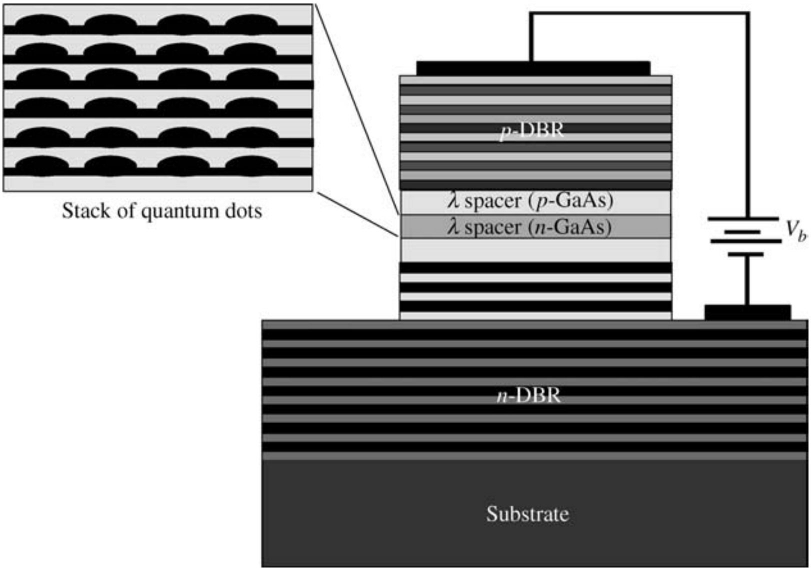


Figure 10.44 A sketch of the multiple-quantum-dot VSCEL with AlO/AlGaAs DBP.

to be highly reflective, which is usually achieved by growing stacks of $\lambda/4$ distributed Bragg reflectors. The mesa diameter is usually on the order of 10 μm with a vertical cavity determined by the total thickness of the multiple-quantum-dot stack. Since the cavity length is very short, the separation between the laser modes (Fabry-Perot modes) is relatively very large. A typical structure of this laser is InGaAs quantum dots and GaAs barriers. The DBRs are usually made of AlGaAs/AlAs. Other DBRs, such as AlGaAs/AlO, are used for the quantum dot VCSELs. The spacer thicknesses are usually grown to match the wavelength of the laser cavity.

The difficulty of achieving multiple-quantum-dot VCSELs comes from the need to have the quantum dots be uniform in size and to have precise epitaxial growth homogeneity and control. For example, if the spacers and the DBR thicknesses do not match the laser mode wavelength, a larger threshold current may be needed to produce stimulated emission.

Summary

This chapter presented detailed discussions on the optoelectronic devices fabricated from quantum wells and dots. As the name implies, optoelectronic devices are those that can either absorb light and then generate an electrical signal or emit light under the influence of an injected electric current. The discussion in this chapter first focused on the infrared quantum detectors where the figures of merit, such as responsivity, noise equivalent power, detectivity, and noise equivalent temperature difference were derived. The basic concepts of the photoconductivity were discussed including the photocurrent and the photoconductive gain. The discussion then focused on light-emitting devices.

The source of noise in infrared detectors was discussed. The most important noise sources are the Johnson noise, the shot noise, the generation-recombination noise, and the $1/f$ noise. Johnson noise is dominant at high frequencies, while $1/f$ noise is dominant at low frequencies.

Examples of quantum detectors were presented. The multiple-quantum-well infrared photodetectors, known as QWIPs, were discussed in more detail. Examples including miniband and embedded miniband quantum structures were presented. In principle, the basic operation of these detectors is based on the electrons that undergo the intersubband transitions in the quantum wells. This transition is excited in the case of the quantum well by photons that are incident at certain angles or p-polarized light. For the intersubband transitions in quantum dot systems, the light is absorbed at any angle or any light polarization due to the fact that the energy dispersion is absent and the selection rules do not exist. This case is valid for an ensemble of quantum dots with high size uniformity.

The photoconductivity measurements were described for single-color and two-color infrared detectors based on multiple quantum wells. The dark current, or the current that is observed under bias voltage in the absence of light illumination, was discussed for the quantum infrared detectors. The two major components of the dark currents are the thermionic and tunneling currents.

Different quantum dot structures used for infrared detectors were discussed with the emphasis on the photoconductivity measurements. It should be pointed out that the quantum dot systems are inherently limited for certain spectral regions. This is because there is a limit on the lower number of monolayers that can be deposited to produce self-assembly quantum dots and there is an upper limit on the number of monolayers that can produce a quantum dot systems with quantum confinement useful for infrared detectors.

Light-emitting diodes (LEDs) based on *pn* junctions and heterojunctions were briefly discussed. The responsivity and the frequency response of the LEDs were presented. These devices are based on the spontaneous emission of light, where the electrons are injected in the structure followed by downward transitions of the electrons from the conduction band to the valence band. As a result of this downward transition, the electrons emit their energy as photons. The important parameters of these devices are the recombination lifetime and the optical gain.

Semiconductor lasers were discussed at the end of the chapter. The laser can be thought of as a waveguide with two mirrors to form a feedback cavity. The discussion covered the basic operational principles of the laser in general. In particular, the spontaneous emission and stimulated emission processes were discussed. The discussion covers the population inversion and the transparency point, which are needed for the lasing action to occur.

Several laser systems were presented. In particular, the edge-emitting lasers, vertical cavity surface-emitting lasers, and quantum cascade lasers were discussed in reasonable detail without going into the full details of deriving their parameters. Finally, quantum dot laser systems were briefly discussed. This system is a fairly new subject, and the full potential of the quantum dot systems has not been fully realized. In the next several years, we should witness the evolution of the quantum dots and their use as laser devices. Thus an update of the subject will be presented in future editions of this textbook.

Problems

10.1 Calculate the responsivity of a photoconductor with an area of 1.0 mm^2 when a blackbody source is used. Assume filters were used such that the limits of the light wavelengths seen by the detectors are 1.0 and 100.0 μm. The measured photocurrent is 0.3 A, and the photon flux is measured as 10^{17} photons $(\mathrm{m}^{-2} \cdot \mathrm{s})$.

10.2 Show that the spectral current responsivity in the frequency domain is given by Eq. (10.35).

10.3 Show that the shot noise in a photodiode detector is reduced to Johnson noise when the modulation frequency is low and the bias voltage is zero.

10.4 Use the spectra that were measured at 77 K in Fig. 10.6 for both the waveguide and Brewster's angle configurations. Calculate the internal quantum efficiency for both spectra assuming that the light made three passes in the waveguide.

10.5 The photocurrent $i_{\mathrm{ph}}(z)$ in a QWIP can be written as $i_{\mathrm{ph}}(z) = (e/h\nu) \alpha T_r L_t(z)\Phi_z$, where $h\nu$ is the photon energy, α is the absorption coefficient, T_r is the transmission tunneling probability, $L_t(z)$ is the transport length given by $L_t(z) = l(1 - e^{-z/l})$. Here l is the mean free path of the electrons. The optical power Φ_z at a location z is given by $\Phi_z = \Phi_o e^{-\alpha z}$ where Φ_o is the power at $z = 0$. Derive an expression for the quantum efficiency. (Hint: Integrate over the detector length to find an expression for the average photocurrent along the length of the detector L. Then use the definition of the current spectral responsivity.)

10.6 Derive Eq. (10.84), and then show that Eq. (10.85) is valid.

10.7 Calculate the number of modes of a GaAs laser that has a cavity length of 500 mm with an emission wavelength of 855 nm. The refractive index of GaAs is 3.5. What are the frequency and wavelength separations of the normal modes? How many longitudinal modes exist for an output gain bandwidth of 10 nm?

10.8 Two consecutive longitudinal modes in the spectrum shown in Fig. 10.33*b* were observed at 12,424.833 cm^{-1} and 12,423.422 cm^{-1}. Use Eq. (10.121) to estimate the laser cavity length. Assume that the refractive index of the quantum well material is 4.0. Compare your results to the cavity length obtained from the interferogram shown in Fig. 10.34.

Appendix

Tables

TABLE A1 Prefixes and Their Symbols for Decimal Multiples and Submultiples of Units

Decimal multiples			Decimal submultiples		
Factor	Name	Symbol	Factor	Name	Symbol
10^{24}	yotta	Y	10^{-24}	yocto	y
10^{21}	zetta	Z	10^{-21}	zepto	z
10^{18}	exa	E	10^{-18}	atto	a
10^{15}	peta	P	10^{-15}	femto	f
10^{12}	tera	T	10^{-12}	pico	p
10^{9}	giga	G	10^{-9}	nano	n
10^{6}	mega	M	10^{-6}	micro	µ
10^{3}	kilo	k	10^{-3}	milli	m
10^{2}	hecto	h	10^{-2}	centi	c
10^{1}	deka	da	10^{-1}	deci	d

TABLE A2 Effective Masses, in the Unit of m_o, in Selected Semiconductors

Semiconductor	Electron effective mass	Heavy hole effective mass	Light hole effective mass
GaAs	0.067	0.45	0.082
AlAs	0.124	0.5	0.22
InP	0.077	0.60	0.012
InAs	0.026	0.41	0.025
InSb	0.014	0.44	0.016
GaSb	0.043	0.33	0.056
AlSb	0.12		
GaP	0.17	0.67	
GaN	0.20	0.80	0.30
InN	0.11	0.50	0.17
AlN	0.27	—	0.25
Si	$m_l = 0.98$ $m_t = 0.19$	0.49	0.16
SiC	0.60		
ZnO	0.28	0.59	
ZnS	0.28	0.49	
CdS	0.14	0.51	
CdTe	0.09	0.40	

Copyright © 2005 by The McGraw-Hill Companies, Inc. Click here for terms of use.

TABLE A3 Bandgap, Lattice Constant, Thermal Conductivity, and Density of Selected Semiconductors

Semiconductor	Bandgap, eV	Lattice constant, nm	Thermal conductivity, W/(cm · °C)	Density, g/cm^3 (at 300 K)
GaAs (ZB, D)	1.424 (300 K) 1.519 (0 K)	0.5653	0.50	5.318
AlAs (ZB, I)	2.153 (300 K) 2.229 (2 K)	0.5660	—	3.717
GaSb (ZB, D)	0.75 (300 K) 0.811 (2 K)	0.609	—	5.63
GaP (ZB, I)	2.270 (300 K) 2.350 (0 K)	0.5451	0.50	4.129
InP (ZB, D)	1.344 (300) 1.424 (2 K)	0.586	—	4.81
InAs (ZB, D)	0.36 (300 K) 0.418 (2 K)	0.6050	—	5.69
InSb (ZB, D)	0.17 (300 K) 0.237 (2 K)	0.647	—	5.80
AlSb (ZB, I)	1.615 (300 K) 1.686 (30 K)	0.61355	—	4.29
Si (Di, I)	1.12 (300 K) 1.17 (2 K)	0.54311	1.48	2.329
Ge (Di, I)	0.66 (300 K)	0.5658	—	5.3234
GaN (W, D)	3.44 (300 K)	$a_o = 0.3189$ $c_o = 0.5185$	2.20	6.095
AlN (W, D)	6.2 (300 K) 6.28 (10 K)	$a_o = 0.3111$ $c_o = 0.4978$	>3.0	3.255
InN (W, D)	0.69–0.9 (300 K)	$a_o = 0.3544$ $c_o = 0.5718$	0.80	6.81
ZnO (W, D)	3.37 (300 K) 3.438 (2 K)	$a_o = 0.32495$ $c_o = 0.52069$	0.60	5.606
ZnS (W, D)	3.741 (300 K) 3.84 (4 K)	$a_o = 0.3811$ $c_o = 0.6234$	—	4.11
SiC (H, I)	4H: 3.26 (300 K)	$a_o = 0.30730$, $c_o = 1.0053$	3.7	
	6H: 3.03 (300 K)	$a_o = 0.30806$, $c_o = 1.51173$	3.8	3.211

NOTE: ZB = zinc-blende, W = wurzite, Di = diamond, H = hexagonal, D = direct, I = indirect.

TABLE A4 Dielectric Constant of Selected Semiconductors

Semiconductor	Dielectric constant
AlP	9.8
AlAs	10.06
AlSb	12.04
GaP	11.1
GaAs	12.5
GaSb	15.7
InP	12.4
InAs	14.6
InSb	17.7
Si	11.9
Ge	16.0
GaN (W)	Average: 10.0
AlN (W)	Average: 9.14
InN	Average: 15.0
ZnO	7.8
ZnS	9.6
CdTe	9.8
CdS	9.4
CdSe	9.3
SiC (6H)	9.66 $\perp$ *c* axis 10.3 $\|\|$ *c* axis

This page intentionally left blank.

Bibliography

Absteiter, G., E. Bauser, A. Fischer, and K. Ploog, *Appl. Phys*. A16 (1978): 345.

Aleiner, I. L., P. W. Brouwer, and L. I. Glazman, "Quantum Effects in Coulomb Blockade," *Physics Report* 358 (2002): 309.

Ambacher, O., *J. Phys. D: Appl. Phys*. 31 (1998): 2653.

Ando, T., A. B. Fowler, and F. Stern, *Rev. Modern Phys.* 54 (1982): 437.

Anselm, A. I., and B. P. Askerov, *Fiz. Tverd Tela (Leningrad)* 3 (1961): 3668 [*Sov. Phys. Solid State*, 3 (1962): 2665].

As, D. J. In *III-Nitride Semiconductors: Growth*, edited by M. O. Manasreh and I. T. Ferguson, New York: Taylor and Francis, 2003, chap. 9.

Ashcroft, N. W., and N. D. Mermin, *Solid State Physics*, New York: Thomson Learning, 1976.

Asryan, L. V., and S. Luryi. In *Semiconductor Nanostructures for Optoelectronic Applications*, Edited by T. Steiner, Boston: Artech House, 2004, chap. 4.

Averin, D. V., and K. K. Likharev, *Mesoscopic Phenomena in Solids*, edited by B. L. Altshuler, P. A. Lee, and R. A. Webb, Amsterdam: Elsevier Science, 1991, p. 173.

Balkanski, M., and R. F. Wallis, *Semiconductor Physics and Applications*, Oxford, England: Oxford University Press, 2000.

Bardeen, J., and W. Shockley, *Phys. Rev*. 80 (1950): 72.

Bastard, G., *Wave Mechanics Applied to Semiconductor Heterostructures*, New York: Halsted Press, 1988.

Bastard, G., J. A. Brum, and R. Ferreira, "Electronics States in Semiconductor Heterostructures," *Solid State Phys.* (Academic, San Diego), 44 (1991): 229.

Beck, W. A., *Appl. Phys. Lett*. 63 (1993): 3589.

Belyaev, L. M., *Ruby and Sapphire*, Translated from Russian by Rubin I. Sapfir. New Delhi, India: Amerind Publishing Co., 1980. Originally published by Moscow: Nauka Publishers, 1974.

BenDaniel, D. J., and C. B. Duke, *Phys. Rev.* B152 (1966): 684.

Bernardini, F., V. Fiorentini, and D. Vanderbilt, *Phys. Rev*. B56, (1997): R10024.

Bimberg, D., M. Grundmann, and N. N. Ledentsov, *Quantum Dot Heterostructures*, New York: Wiley, 1999.

Birman, J. L., *Theory of Crystal Space Groups and Lattice Dynamics*, Berlin: Springer-Verlag, 1984.

Bohm, D., *Quantum Theory*, Englewood Cliffs, NJ: Prentice Hall, 1953.

Brennan, K. F., *The Physics of Semiconductors with Applications to Optoelectronic Devices*, Cambridge, England: Cambridge University Press, 1999.

Brown, G. J., and F. Szmulowicz. In *Long Wavelength Infrared Detectors*, edited by M. Razeghi, New York: Taylor and Francis, 1996.

Cheng, J., and N. K. Dutta, eds., *Vertical Cavity Surface Emitting Lasers: Technology and Applications*, New York: Gordon and Breach (currently Taylor and Francis), 2000, vol. 10.

Cho, A. Y., *Thin Solid Films* 100 (1983): 291.

Choi, K. K. In *Semiconductor Interfaces, Microstructures and Devices*, edited by Z. C. Feng, Philadelphia: IOP, 1993.

Chynoweth, A. G., W. L. Feldmann, and R. A. Logan, "Excess Tunnel Current in Silicon Junctions," *Phys. Rev*. 121 (1961): 684.

Cohen-Tannoudji, C., B. Diu, and F. Laloë, *Quantum Mechanics*, New York: Wiley, 1977.

Conwell, E., and V. F. Weisskopf, "Theory of Impurity Scattering in Semiconductors," *Phys. Rev*. 77 (1950): 388.

Copyright © 2005 by The McGraw-Hill Companies, Inc. Click here for terms of use.

Davies, J. H., *The Physics of Low-Dimensional Semiconductors: An Introduction*, Cambridge, England: Cambridge University Press, 1998.

Davies, J. H., and J. W. Wilkins, "Narrow Electronic Bands in High Electric Field: Static Properties," *Phys. Rev*. B38 (1988): 1667.

Davydov, A. V., W. J. Boettinger, U. R. Kattner, and T. J. Anderson, "Thermodynamic Assessment of the GaN System," *Physics Status Solidi (a)* 188 (2001): 407.

Demassa, T. A., and D. P. Knott, "The Prediction of Tunnel Diode Voltage-Current Characteristics," *Solid State Electronics* 13 (1970): 131.

Dereniak, E. L., and G. D. Boreman, *Infrared Detectors and Systems*, New York: Wiley, 1996.

Dicke, R. H., and J. P. Wittke, *Introduction to Quantum Mechanics*, Reading, Addison-Wesley, PA: 1960.

Dimitrov, H. D., "Low-Frequency Electronic Resistivity of Crystals due to the Scattering from Dipoles," *J. Phys. Chem. Solids* 37 (1976): 825.

Dresselhaus, G., A. F. Kip, and C. Kittel, "Cyclotron Resonance of Electrons and Holes in Silicon and Germanium Crystal," *Phys. Rev*. B98 (1955): 368.

Elliott, R. J., *Phys. Rev.* 108 (1957): 1384.

Erginsoy, C., "Neutral Impurity Scattering in Semiconductors," *Phys. Rev*. 79 (1956): 1013.

Esaki, L., "New Phenomenon in Narrow Germanium p-n Junction," *Phys. Rev*. 109 (1958): 603.

Falicov, L. M., *Group Theory and Its Physical Applications*, Chicago, University of Chicago Press, 1966.

Ferry, D. K., *Quantum Mechanics: An Introduction for Device Physicists and Electrical Engineers*, Bristol, England: IOP, 2001.

Folk, J. A., S. R. Patel, S. F. Godijn, A. G. Huibers, A. M. Cronewett, C. M. Marcus, K. Campman, and A. C. Gossard, *Phys. Rev. Lett.* 76 (1996): 1699.

Fox, M., *Optical Properties of Solids*, Oxford, Oxford University Press, 2001.

Gasiorowicz, S. *Quantum Mechanics*, 3rd ed. New York: Wiley, 2003.

Gaska, R., M. S. Shur, and A. Khan. In *III-V Nitride Semiconductors: Applications and Devices,* edited by E. T. Yu and M. O. Manasreh, New York: Taylor and Francis, (2003), vol. 16.

Giaever, I., and H. R. Zeller, *Phys. Rev. Lett*. 20 (1968): 1504.

Gildenblat, G., B. Gelmont, and S. Vatannia, *J. Appl. Phys*. 77 (1995): 6327.

Gith, J., and W. T. Petusky, *J. Phy. Chem. Solids* 48 (1987): 541.

Glazman, L. I., et al., *JETP Lett*. 48 (1988): 238 [*Pis'ma Zh. Teor. Fiz*. 48 (1988): 218].

Gorter, C. J., *Physica* 17 (1951): 777.

Grundmann, M., O Stier, and D. Bimberg, "InAs/GaAs Pyramidal Quantum Dots: Strain Distribution, Optical Phonons, and Electronic Structure," *Phys. Rev*. B52 (1995): 11969.

Gunn, J. B. "Microwave Oscillations of Current in III-V Semiconductors," *Solid State Commun*. 1 (1963): 88.

Harris, G. L. In *Properties of Silicon Carbide*, edited by G. L. Harris (Bristol, England: IOP, 1995), p. 4.

Harrison, P., *Quantum Wells, Wires and Dots*, New York: Wiley, 2000.

Haynes, J. R., and W. Shockley, "The Mobility and Life of Injected Holes and Electrons in Germanium," *Phys. Rev*. 81 (1951): 835.

Helm, M., ed., *Long Wavelength Infrared Emitters Based on Quantum Wells and Superlattices*, New York: Gordon and Breach, (currently Taylor and Francis), 2000, vol. 6.

Huang, T. -F., and S. J. Harris, Jr. In *III-nitrided semiconductor: growth*, edited by M. O. Manasreh and I. T. Ferguson, New York: Taylor and Francis, 2003, vol. 19, chap. 10.

Hudson, R. D., *Infrared System Engineering*, New York: Wiley, 1969.

Khan, M. A., A. Bhattarai, J. N. Kuznia, and D. T. Olson, *Appl. Phys. Lett.* 63 (1993): 1214.

Khan, M. A., X. Hu, G. Simin, and J. Yang, *Appl. Phys. Lett*. 77 (2000): 1339–1341.

Khan, M. A., X. Hu, G. Simin, A. Lunev, and J. Yang, R. Gaska, and M. S. Shur, *IEEE EDL* 21 (2000): 63.

Kitabayashi, H., T. Waho, and M. Yamamoto, *Appl. Phys. Lett.* 71 (1997): 512.

Kittel, C., *Introduction to Solid State Physics* 7th ed., New York: Wiley, 1996.

Klaassen, F. M., *IEEE Trans. Electron Devices*, ED-18 no. 10, (1971): 887.

Klitzing, K. V., G. Dorda, and M. Pepper, “New Method for High-Accuracy Determination of the Fine-Structure Constant Based on Quantized Hall Resistance,” *Phys. Rev. Lett.* 45 (1980): 494.

Korotkov, A. N., R. H. Chen, and K. K. Likharev, *J. Appl. Phys.* 78 (1995): 2520.

Koster, G. F., *Space Groups and Their Representations in Solid State Physics 5*, New York: Academic, 1957, pp. 173–256.

Kouwenhoven, L. P., et al., *Z. Phys.* B85 (1991): 367.

Lambe, J., and R. C. Jaklevic, *Phys. Rev. Lett.* 22 (1969): 1371.

Laughlin, R. B., “Anomalous Quantum Hall Effect: An Incompressible Quantum Fluid with Fractionally Charged Excitations,” *Phys. Rev. Lett.* 50 (1983): 1395.

Leonard, D., K. Pond, and P. M. Petroff, *Phys. Rev.* B50 (1994): 11687.

Levine, B. F., *J. Appl. Phys.* 74 (1993): R1.

Liu, H. C. In *Long Wavelength Infrared Detectors*, edited by M. Razeghi, New York: Taylor and Francis, 1996, chap. 1.

Liu, L., and J. H. Edgar, “Substrates for Gallium Nitride Epitaxy,” *Materials Science and Engineering* R37 (2002): 61.

Lloyd, J. M., *Thermal Imaging Systems*, New York: Plenum, 1975.

Look, D. C. *Electrical Characterization of GaAs Materials and Devices*, New York, Wiley, 1989.

Luttinger, J. M., “Quantum Theory of Cyclotron Resonance in Semiconductors: General Theory,” *Phys. Rev.* B102 (1956): 1030.

Madhukar, A., E. T. Kim, Z. Chen, J. C. Campbell, and Z. Ye. In *Semiconductor Nanostructures for Optoelectronic Applications*, edited by T. Steiner, Boston: Artech House, 2004, chap. 3.

Manasreh, M. O., and B. C. Covington, “Infrared-Absorption Properties of EL2 in GaAs,” *Phys. Rev.* B36 (1987): 2730–2734.

Manasreh, M. O., and H. H. Jiang, eds., *III-Nitride Semiconductors: Optical Properties I*, New York: Taylor and Francis, 2002, vol. 13.

Manasreh, M. O., and I. T. Ferguson, eds., *III-Nitride Semiconductors: Growth*, New York: Taylor and Francis, 2003.

Manasreh, M. O., ed., *Semiconductor Quantum Wells and Superlattices for Long Wavelength Infrared Detectors*, Boston, MA: Artech House, 1993.

McMelvey, J. P., *Solid State Physics for Engineering and Materials Science*, Malbar, FL: Krieger Publishing Co., 1993.

McQuarrie, D. A., *Statistical Mechanics*, New York: Harper and Row Publishers, 1976.

Mead, C. A., “Scottky Barrier Gate Field-Effect Transistor,” *Proc. IEEE* 54 (1966): 307.

Merzbacher, E., *Quantum Mechanics*, New York: Wiley, 1970.

Mihály, L., and M. C. Martin, *Solid State Physics: Problems and Solutions*, New York: Wiley, 1996.

Millikan, R. A., *Phys. Rev.* 32 (1911): 349.

Mitin, V. V., V. A. Kochelap, and M. A. Stroscio, *Quantum Heterostructures Microelectronics and Optoelectronics*, Cambridge, England: Cambridge University Press, 1999.

Mooradian, A., and G. B. Wright, *Phys. Prev. Lett.* 16 (1966): 999.

Morkoç, Hadis, Aldo Di Carlo, and R. Cingolani, “GaN-Based Modulation Doped FETs and UV Detectors,” *Solid State Electronics* 46 (2002): 157.

Morkoç, H., R. Cingolani, and Bernard Gil, *Materials Research Innovations* 3 (1999): 97.

Mott, N. F., and H. Jones, “*The Theory of the Properties of Metals and Alloys*,” New York: Dover, 1958.

Moustakas, T. D. In *Semiconductors and Semimetals*, New York: Academic, 1999, vol. 57 p. 33.

Neamen, D. A., *Semiconductor Physics and Devices: Basic Principles*, 3rd ed., New York: McGraw-Hill, 2003.

Nelson, R. D., *Opt. Eng.* 16 (1977): 275.

Oda, O., T. Fukui, R. Hirano, M. Muchida, K. Kohiro, H. Kurita, K. Kainosho, S. Asahi, and K. Suzuki, In *InP and Related Compounds*, edited by M. O. Manasreh, New York, Taylor and Francis, 2000, chap. 2.

Ohring, M., *The Materials Science of Thin Films*, New York: Academic, 1992, chap. 4.

Palik, E. D., G. S. Picus, S. Teitlee, and R. F. Wallis, "Infrared Cyclotron Resonance in InSb," *Phys. Rev*. 122 (1961): 475.

Pankove, J. I., "*Optical Processes in Semiconductors*," New York: Dover Publications, Inc., 1971.

Paskova, T., and B. Monemar. In *III-Nitrided Semiconductor: Growth*, edited by M. O. Manasreh and I. T. Ferguson, New York: Taylor and Francis, 2003, vol. 19, chap. 6.

Pattada, B., J. Chen, M. O. Manasreh, S. Guo, D. Gotthold, M. Pophristic, and B. Peres, *J. Appl. Phys*. 93 (2003): 5824.

Pfann, W. G., *Zone Melting*, 2nd ed., New York: Wiley, 1966.

Pfeiffer, L, K. W. West, H. L. Stromer, and K. W. Baldwin, "Electron Mobilities Exceeding 107 cm^2/V·s in Modulation-Doped GaAs," *Appl. Phys. Lett.* 55 (1989): 1888.

Raman, C. V., *Ind. J. Phys*. 2 (1928): 387.

Razeghi, M., *The MOCVD Challenge*, Bristol, England: IOP, 1989, vol. 1.

Razeghi, M., ed., "Long Wavelength Infrared Detectors," New York: Taylor and Francis, 1996, vol. 1.

Reeber, R. R., and K. Wang, *Mat. Res. Soc. Symp*. 622 (2000): T6.35.1.

Robinson, V. N. E., and J. L. Robins, *Thin Solid Films* 20 (1974): 155.

Rogalski, A. *Infrared Detectors*, Amesterdam: Gordon and Breach, 2000.

Roy, D. K., "On the Prediction of Tunnel Diode I-V Characteristics," *Solid State Electronics* 14 (1971): 520.

Schmidt, T. J., and J. J. Song, In *III-Nitride Semiconductors: Optical Properties II*, edited by M. O. Manasreh and H. X. Jiang, New York: Taylor and Francis, 2002, vol. 14 chap. 2.

Siergiej, R. R., R. C. Clarke, S. Siram, A. K. Agarwal, R. J. Bojko, A. W. Morse, V. Balakrishna, M. F. MacMillan, A. A. Burk, Jr., and C. D. Brandt, *Mater. Sci. Eng*. B61–62, (1999): 9–17.

Simin, G., A. Koudymov, H. Fatima, J. Zhang, J. Yang, M. Asif Khan, X. Hu, A. Tarakji, R. Gaska, and M. S. Shur, *IEEE EDL*, 23 (2002): 458–460.

Simin, G., X. Hu, A. Tarakji, J. Zhang, A. Koudymov, S. Saygi, J. Yang, M. Asif Khan, M. S. Shur, and R. Gaska, *Japanese J. Applied Phys.*, 400 (2001): L921–L924.

Singh, J. *Semiconductor Devices: Basic Principles*, New York: Wiley, 2001.

Singh, J. *Electronic and Optoelectronic Properties of Semiconductor Structures*, Cambridge, England: Cambridge University Press, 2003.

Singh, J. *Modern Physics for Engineers* New York: Wiley, 1999.

Smekal, A., *Naturwissensch* 11 (1923): 873.

Smith, R. A., *Semiconductors*, 2nd ed., Cambridge, London: Cambridge University Press, 1978.

Sols, F., M. Macucci, U. Ravaioli, and K. Hess, *J. Appl. Phys*. 66 (1989): 3892.

Stenger, R. T., and K. K. Bajaj, "Optical Properties of Confined Polaronic Excitons in Spherical Ionic Quantum Dots," *Phys. Rev*. B68 (2003): 45313.

Stern, F. *Phys. Rev*. B5 (1972): 2891.

Sze, S. M., *Physics of Semiconductor Devices*, 2nd ed., New York: Wiley, 1981.

Sze, S. M., *Semiconductor Devices Physics and Technology*, 2nd ed., New York: Wiley, 2002.

Szmulowicz, F., *Phys. Rev*. B51 (1995): 1613.

Timp, G., ed., *Nanotechnology*, New York: Springer-Verlag, 1999.

Timp, G., R. E. Howard, and P. M. Mankiewich. In *Nanotechnology*, edited by G. Timp, New York: Springer-Verlag, 1999, chap. 2.

Tsuchiya, M., H. Sakaki, and J. Yashina, "Room Temperature Observation of Differential Negative Resistance in AlAs/GaAs/AlAs Resonant Tunneling Diode," *Jpn. J. Applied Phys*. 24 (1985): L466.

Tsui, D. C., H. L. Störmer, J. C. M. Huang, J. S. Brooks, and M. J. Naughton, "Observation of a Fractional Quantum Number," *Phys. Rev*. B28 (1983): 2274.

van Wees, B. J., H. van Houten, C. W. Beenakkar, J. G. Williamson, L. P. Kouwenhoven, D. van der Marel, and C. T. Foxon, *Phys. Rev. Lett.* 60 (1988): 848.

Venables, J. A., G. D. T. Spiller, and M. Hanbücken, *Rep. Prg. Phys*. 47 (1984): 399.

Vook, R. W., *Int. Metals Rev*. 27 (1982): 209.

Wallis, R. F., “Theory of Cyclotron-Resonance Absorption by Conduction Electrons in Indium Antimonide,” *J. Phys. Chem. Solids* 4 (1958): 101.

Walton, D., T. N. Rhidin, and R. W. Rollins, *J. Chem Phys*. 38 (1963): 2698.

Weast, R. C. and M. J. Astle, ed., *CRC Handbook of Chemistry and Physics*, 60th ed., New York: CRC Press, 1992.

Wooten, F., *Optical Properties of Solids*, New York: Academic Press, 1972.

Wu, W. Y., J. N. Schulman, T. Y. Hsu, and U. Efron, “Effect of Size Nonuniformity on the Absorption Spectrum of a Semiconductor Quantum Dot System,” *Appl. Phys. Lett.* 51 (1987): 710.

Yokoyama, N, K. Imamura, S. Muto, S. Hiyamizu, and H. Nishi, *Jap. J. Appl. Phys*. 24 (1985): L853.

Young, D. L., J. F. Geisz, and T. J. Coutts, *Appl. Phys. Lett*. 82 (2003): 1236.

Yu, E. T. In *III-V Nitride Semiconductors: Applications and Devices*, edited by E. T. Yu and M. O. Manasreh, New York: Taylor and Francis, 2003, vol. 16.

Yu, P. Y., and M. Cardona, *Fundamentals of Semicoductors: Physics and Materials Properties*, 3d. ed., New York Sprinder, 2003.

This page intentionally left blank.

Index

1/f noise, 469, 470–472, 489, 526
2D structures, 377
2DEG channel, 432
  density 272, 447
  in triangular well, 425
2D-to-3D transition, 390
2H SiC, 367
2 × 2 matrix, 110
3C SiC, 368
4H SiC, 367
4 × 4 matrix, 110
6H SiC, 367
$6 \times 6\,\mathbf{k} \cdot \mathbf{p}$ matrix, 119
Absorbance of intersubband transition, 213
Absorption, 501
  coefficient, 198, 502, 504
  coefficient of interband transition in quantum dots, 221
  of phonons, 205
ac current density, 497
Acoustical phonon modes, 249
Active detector area, 466
Active region, 507
Adatom surface density, 382
Adjoint operator, 21
Adsorption of atoms, 359
Airy functions, 60
  Ai(x), 133, 185, 186
  Bi(x), 133
  first derivative, 134
  integral form, 133
  normalized, 134
Al mole fraction, 173
AlInAsSb quaternary layers, 519
Alloy composition, 330
AlN buffer layer, 365
Ambipolar mobility, 291
Amplifiers, 417
Angular frequency, 195
  of longitudinal acoustic wave, 248
  of transverse acoustic wave, 248
Angular momentum, 6, 33, 34
  total, 33, 107
Antistokes phonon scattering, 251
Anyons, 277
Arbitrary potential barrier, 123
Arbitrary spatially varying potential, 72, 73
Arrhenius plot, 384
Arsine, 350
Asymmetric lineshape, 219
Atomic force microscopy, 388
Atomistic nucleation processes, 382–387
Atom
  adsorption of, 359
  Bohr model of, 6
*atto*farads, 148
Auger electron microscopy, 375
Auger processes, 304
Auger recombination, 288, 458
Axial crystal field, 228

B_2O_3, 335
Background limited infrared performance, 463
Background photon flux density, 464
Background photons, 471
Ballistic injection devices, 403
Ballistic transport without scattering, 448
Band bending, 127, 404
Band diagram of GaAs/AlGaAs heterojunction diode laser, 507
Band edge absorption of a direct semiconductor, 225
Band structure calculation methods, 99
Bandgap, 106, 107, 220, 221, 227, 404, 406, 472, 494, 522
  alignment, 143
  diagram of an ideal n-channel MESFET, 414
  energy, 171
  as function of lattice constant, 331
  offset, 127
  of semiconductor, 140
Band-to-band transitions in bulk materials, 191
Barrier height, 482
Barrier width, 45, 46
Base-collector voltage, 435
Base-emitter voltage, 434
Basic HVPE reactions, 353
Beam flux, 358
Beer's law, 196, 199
BenDaniel-Duke boundary conditions, 115
Bessel-Parseval relation, 27
Bipolar transistors, 123
Blackbody
  detectivity, 462
  irradiance, 462
  NEP, 462
  radiation, 2
  responsivity, 461
  spectral intensity, 2
BLIP, 463, 464
Bloch electron, 84
  in weak periodic potential, 89–92
Bloch function, 114
Bloch state
  in conduction band, 201
  in valence band, 201
Bloch wave function, 59, 90, 93
Bloch's theorem, 84, 166
Body-centered cubic structure, 101
Bohr magneton, 181, 237

Copyright © 2005 by The McGraw-Hill Companies, Inc. Click here for terms of use.

Bohr model of the atom, 6
Bohr radius, 226
Bohr-Sommerfeld quantization rule, 8
Bolometer, 459
Boltzmann constant, 143, 148, 157
Boltzmann distribution function, 494, 508
Boltzmann equation, 294
Boltzmann transport equation, 294–298
Boric oxide, 335
Bose factor, 471
Bose-Einstein statistics, 164–166, 205
Bound exciton, 192, 193, 225
Bound ground state, 481
Bound state of a δ-function potential, 66
Boundary conditions, 37, 38, 42, 43, 49, 70
  at turning points, 76
  for δ-function barrier, 125
Bound-to-bound transitions, 210
Bound-to-continuum transitions, 210, 214
Bra, action of a linear operator on, 21
Bra vector, 17
Bragg plane, 90, 91
Bragg reflections, 89
Breakdown voltage, 420
Brewster's angle, 214, 218, 474, 485
Brewster's configuration, 480
Brillouin zone, 92, 96
  boundary, 180, 249
Broadening effect, 213
  of Landau levels, 275
Broken bandgap superlattices, 519
Bubbler, 372
Buffer layer, 366
Bulk semiconductor crystals, 330
Bulk SiC wafers, 347
Bulk viscous flow, 376

C_2 operation, 103
C_3 operation, 102
Canonical ensemble, 156, 157
Canonical transformations, 219
Capacitance, 147, 417
  coupling, 438
  depletion region, 409
  minority charge, 434
Capillarity theory, 379–382
Capture probability, 490–492
Carrier
  -carrier scattering, 304
  concentration, 168
  diffusion in Haynes-Shockley experiment, 293
  effective mass, 252
  freeze-out, 164
  injection, 403
  lifetime, 245
  saturation velocity, 419, 427
Cavity radiator, 2
Characteristic deposition time, 370
Charge, 342
  carrier mobility, 482
  carrier transport in bulk semiconductors, 277
  neutrality condition, 290
  quantization, 149
  transport, 445
Chemical free energy, 380
Chemical free energy per unit volume, 381
Chemical potentials, 440
Cladding region, 518
Classical ballistic regime, 264
Classical dipole source, 458
Classical equation of motion, 275
Classical probability of a simple harmonic oscillator, 64
Classical transverse size effect, 264
Closure relation, 23
Cluster models
  coalescence, 387
  migration, 387
  sintering, 387
Coherence and mesoscopic systems, 318–324
Coherent transport, 318
Cohesive energy, 245
Collective oscillations, 252
Collector current density, 434
Commutator, 20
Complementary circuits, 443
Complex dielectric constant, 197
Components of the angular momentum, 33
Compton scattering, 5
Concentration gradients, 376
  of acceptors, 142
  of donors, 142
Conductance, 440, 443
  as function of gate voltage, 323
  fluctuation, 322
Conduction band
  diagram of quantum cascade laser, 520
  effective mass in direct semiconductor, 106
  offset, 128, 214, 507
  structure, 486
Conductive medium, 194
Conductivity mobility, 279
Conductivity tensor, 276
Confined subbands, total number of, 173
Confined energy levels, 144
  in 1D system, 176
Confinement directions, 171
  of charge carriers, 506
  of longitudinal optical modes, 506
Convection barrier, 342
Conventional optical bandgap, 191
Core level electron, 360
Corrugated substrate, 396
Coulomb blockade, 147–150, 437–451
Coulomb force, 6
Coulomb oscillations, 438, 440
Coulomb staircase, 439
Coupled wells, 78
Critical concentration of clusters, 383
Critical nucleation rate, 383
Critical radius of the film nucleus, 381
Critical transition temperature, 384
Crucible, 341
Crystal
  imperfections, 275
  momentum, 90, 97
  structure, 395
  weight, 338
Crystal puller, 334

Crystalline solid, 85
Crystallographic directions of Si wafers, 348
Cubic crystal, 166
Current
  balance-equation, 467
  density, 38, 143, 237, 410, 509
  generation-recombination noise, 471
  noise, 490
  responsivity, 460, 461
  -voltage relations for a MESFET, 414
Curvature restriction, 446
  of single electron transistor, 447
Cyclotron, properties of
  angular frequency, 273
  frequency, 235
  radius, 270
  resonance, 106, 234–239, 256, 258, 274, 279, 362
  resonance experimental setup, 235
  resonance spectrum, 238

δ-function, 24, 83, 93, 200, 275
  approximation, 87, 322
  barrier, 124, 125
  line shape, 145
  potential, 64–70
Damping constants, 196, 253
Dark current, 474, 481, 483, 486, 490, 491
  in multiple quantum dot infrared detector, 489
DBP, 513, 519, 525
dc-photogenerated current, 463
de Broglie relation, 8, 9, 35
de Broglie wavelength, 8, 261, 318
Debye model of specific heat of solids, 4
Debye temperature, 4
Decaying function, 47
Decaying length, 131
Decaying wave, 72
Decaying wave vector, 123
Defect-assisted tunneling, 141
Deformation potential, 307
Degeneracy, 154–156, 499
  of energy levels, 220
Degenerate *pn*-junction laser, 506
Degenerate *pn* junctions, 505
Degenerate states, 30
Delay time, 49
Density of states, 166–171, 521
  of bulk electron in presence of magnetic field, 181–184
  in bulk semiconductors, 522
    under influence of electric field, 186, 187
  for conduction band, 167
  definition, 166
  function, 162
  as function of energy in superlattices, 181
  in 1D system, 175
  in other systems, 179–188
  in presence of electric field, 184–188
  in quantum dots, 171, 178, 179, 522
  in quantum wells, 171, 171–174, 522
  in quantum well under influence of electric field, 186
  in quantum wire under influence of electric field, 186
  quantum wires, 171, 175–178, 522
  in semiconductor superlattices, 179–181
  in shell, 154
  in superlattices, 180
  for valence band, 167
  in three-dimensional system, 168
  per Landau level, 277
Dephasing effect, 263, 319
Dephasing length, 318
Dephasing of Landau state, 273
Depletion length, 131
Depletion region, 140, 404, 412, 414, 415
  in HEMT, 312
  width, 142
Depletion width, 404, 405, 408, 448
Deposition
  growth rate, 376
  of metallic structures, 388
  temperature, 385
  time, 385
Derivation
  of transport coefficients, 296–304
  of unit step function, 25
Desorption
  energy, 385
  process, 385
Detectivity, 462
Detector noise, 463
Determinant, 54, 96, 249
  of coefficients, 23, 54, 79, 86
Device efficiency, 493
Diagonal matrix element of the Hamiltonian, 101
Diamond structure, 347
Diatomic crystal, 248
Dielectric constant, 59, 130, 197, 227, 290
  function, 253
  susceptibility tensor, 194
  tensor, 194
Differential cross section, 306
Diffraction angle, 371
Diffusion
  coefficient, 263, 496
  constant, 284
  current, 282
  current density, 282–285, 508
  length, 496
  mean distance, 383
  mechanism, 261
  parasitic resistances, 419
  process, 507
Diffusion velocity, average, 434
Digital circuits, 146
Digitally graded alloys, 519, 520
Dipole matrix element, 224
Dipole moment, 197
Dirac notation, 17
Dirac δ-function, 24
Direct bandgap, 286, 198, 241, 227
Direct recombination, 286, 287
Discrete energy, 2
Dislocation densities, 343
Dislocations, 314

Dispersion
  curve, 249
  relation, 195, 247
Displacement
  current, 417
  density, 285
Dissimilar bandgaps, 517
Dissociation of the adsorbed atoms, 359
Distinguishable number of particles, 157
Distinguishable objects, 157
Distributed Bragg reflectors, 513, 515, 525
Distribution function, 156, 157
  definition, 159
  of dopant near the solid–liquid interface, 339
  of noninteracting particles, 160
Dopant
  concentration, 405
  diffusion coefficient, 339
  fluxes at the interfaces, 339
  weight, 338
Doping concentration, 338
Doping density, 483
Doping distribution in the float-zone process, 344
Double heterojunction structure, 432
Double-barrier potential, 135–139
Drain, 406
  current, 412,525
  -source current, 423
  -source voltage, 440
  voltage, 414
Drift
  current, 495
  current density, 278–282, 284
  diffusion method, 295
  velocity, 255, 263, 266, 278, 280, 419, 482
Drude model, 196–198
Dulong and Petit treatment of specific heat of solids, 2
Dye lasers, 458

Edge-emitting laser, 459
  structure, 513
Effective density of states, 508
Effective Hamiltonian, 108
Effective Landé g-factor, 181
Effective mass, 46, 59, 98, 99, 105, 234, 236, 238
  approximation, 167
  of charge carriers, 482
  of electron, 196
  of heavy hole, 111
  of light hole, 111
  mismatch, 116
  Schrödinger equation, 59
  tensor, 99
  theory, 219
Effective number of the thermally excited carriers, 481
Effective polarization, 216
Effective Richardson constant, 410
Effective segregation coefficient, 338–340
Eigenfunctions, 22, 37
Eigenvectors, 19, 23
Eigenvalues, 22, 28, 37
Einstein coefficient, 500
Einstein model of specific heat of solids, 4
Einstein relations, 285, 500
Einstein temperature, 4
EL2 defect, 242
Elastic constant of AlGaN, 429
Elastic scattering, 263, 318
Electric displacement vector, 194
Electrical conductivity, 194, 198, 285, 301, 465
Electrical conductivity in n-type semiconductors, 300–302
Electrical image potential, 129
Electrically neutral solids, 197
Electroluminescence, 240
Electromagnetic waves, 6
Electron
  affinity, 406, 452
  beam lithography, 388
  charging effect, 438
  concentration, 168, 170
  cyclotron resonance (ECR) plasma source, 362
  density, 170, 173
    in 2D system, 174
  diffraction of solid crystal, 5
  diffusion coefficient, 283, 434
  diffusion length, 507
  effective mass, 59, 135, 138, 144, 151, 168, 201, 211, 219, 238, 261, 262, 280, 296, 301, 302, 481, 485, 508
    in InAs, 144
  -electron interaction, 425
  -electron scattering, 277, 317
  exchange, 438
  ground state in the conduction band, 209
  group velocity, 320
  -hole
    exchange interaction, 230
    orbit, 226
    overlap, 224
    pair, 225, 467
    recombination, 288
  -ion interaction, 89
  mobility, 412
    in n-type GaAs, 313
  momentum relaxation time, 316
  -phonon coupling, 240
  -phonon matrix element, 205
  -photon
    coupling, 214, 485
    coupling selection rules, 223
    interaction, 192
  spin degeneracy, 203, 211
  spin quantum number, 237
  thermal velocity, 283
  transmission probability, 445, 446
Electronic devices, 403
Electronic nanostructure devices, 149
Electronic transitions in multiple quantum dots, 217–222
Electronic energy levels, allowed, 83
Electrostatic coupling, 438
Electrostatic potential, 59, 314
  energy, 184
Ellipsometry, 228, 375
Emission of phonons, 205
Emitters, 457

Empirical tunneling current, 151
Empty states, 140
Energy
  quantization, 147
  relaxation time, 313
  of single-electron process, 148
Energy absorption, rate of, 199
Energy band
  allowed, 89
  barrier, 404
  curvature of, 99
  diagram, 408
  eigenvalue, 16
  structure of GaAs, 104
Energy level
  diagram, 118
  in double quantum well structure, 78–80
  of hydrogen atom, 7
  of infinite potential well, 50
  of simple harmonic oscillator, 62
  in triangular quantum well, 185
Envelop
  function approximation, 59, 113
  wave function, amplitude of, 209
  wave function for two bound states, 210
Epitaxial growth, 330
  of self-assembly quantum dots, 388–299
  technique, 313
Equation of motion of masses, 247
Equilibrium
  distribution function, 296
  dopant concentration, 337
Ettinshausen coefficient, 304
Evanescent wave range, 46
Evaporation, 385
Even parity 66
  solution of δ-function potential, 67
Excess carrier, 465
  concentration, 496
Excess current, 141, 142
Excess electron concentration, 245
Excess minority carriers, 292
Excess minority lifetime, 288
Excited state in a quantum well, 116
Exciton, 192, 225–234
  binding energy, 227, 244
    in quantum dots, 233
    in quantum wells, 232
  in bulk semiconductors, 225
  center of mass, 228
  energy levels, 227
  ground-state energy, 232
  lines in GaN, 229
  in quantum dots, 233, 234
  in quantum wells, 231–233
  radius, 227
  wavevector, 228
Extended defects, 239
External differential quantum efficiency, 510
External efficiency, 493
External quantum efficiency, 469, 489, 490
Extraction efficiency, 493

Fabrication of quantum dots, 377–399
Fabry-Perot cavity, 502, 505
Face-centered cubic structure, 103
Facet reflectivities, 518
Faraday configuration, 237
Feedback, 513
Fermi energy, 130, 139
  of bulk GaAs material, 162
  definition, 166
  as function of 1D carrier concentration, 178
  as function of 2D carrier concentration, 174
Fermi golden rule, 201, 304
  level, 144, 162, 171, 294, 404, 406, 407, 440, 445, 481
  sea of electrons, 145
  wavevector, 166, 314, 315, 446
Fermi-Dirac distribution function, 140, 161–164, 294, 300, 321, 494, 502
  of 2D-system, 173
  of unfilled states, 168
Fermi-Dirac occupation function, 211
Fermi-Dirac statistics, 83
Fermi energy, concept of, 167
Ferroelectric materials, 459
Fick's law, 282, 283
Field emission gun, 388
Field-effect devices, 403
Field-effect transistors, 123
Figures of merit, 460–465
Fill factor, 490
Filling factor of quantum dots, 396
Film uniformity, 354
Fine structure constant, 7
Finesse parameter, 513
Finite depth potential well, 51
First Brillouin zone, 103
  center, 102
  extrema of, 103
  L-point of, 458
First derivative
  of depletion region width, 415
  of wave function, 38
First-order correction to the eigenfunction, 33
First-order correction to the eigenvalue, 31
First-order perturbation theory, 30, 305
First-order phase transition, 390
Float-zone
  apparatus, 343
  growth method, 343–345
Floquet, 84
Forbidden energy bands, 89
  frequency band, 249
Forward-bias voltage, 410
Four energy level system, 501
Fourier series, 26
  transform, 10, 25, 27, 35
    in quantum mechanics, 26
    spectrometer, 512
    spectroscopy, 461
    of step function, 64
Fourth-order effective mass theory, 272
Fractional optical absorption, 480
Fractional quantum Hall effect, 277
Franck-Condon shift, 243
Frank-van der Merwe growth mode, 378, 380
Franz-Keldysh effect, 188

Free carrier absorption, 518
Free electron energy, 87
Free electron theory, 85
Free energy, 331, 332, 379–381, 402
Free energy function, 331
Free exciton, 192, 193, 226–228, 245, 256
  binding energy, 227
Free flight time, 264
Free-electron absorption, 198
Free-electron mass, 201
Free-exciton radius, 228
Frenkel exciton, 192
Frequency-dependent dielectric constant, 195
Frequency-dependent diffusion length, 497
Full width at half maximum, 138, 513
Fundamental bandgap of bulk GaAs, 244

Γ-function, 126, 155, 302
  properties, 155
Γ-point in the Brillouin zone, 103
  symmetry, 199
Γ valley, 281
GaAs
  lattice constant, 379
  single crystal boule, 330
  substrate, 347
GaAs, structural directions of, 348
GaAs/AlGaAs
  HBTs, 434
  heterojunction, 127, 312, 433
    diode laser, 510
    LED, 497
    used as a diode laser, 506
  high electron mobility transistor, 237
  interface, 425, 426
  LED, 495
  MODFET, 424
  multiple quantum wells, 113, 114, 173, 243
  quantum wells, 207, 213, 272, 316
  quantum wire, 174
  superlattices, 474, 475
GaN
  band structure, 229
  decomposition, 353
  HEMT, 428
  single crystal, 352
GaN/AlGaN, 423, 418
  heterojunction, 428
    field-effect transistor, 428–433
    interfaces, 429
  HFET, 430–432
  interface, 216, 428
    channel HFET
  multiple quantum wells, 216
  single channel HFET, 431
  superlattices, 429, 433
GaN-based devices, 347
GaN-related electronic devices, 428
GaP, 368
Gas
  constant, 331
  lasers, 458
  source, 358
  -substrate interface, 375
Gate, 406
  capacitance, 447
  cross-sectional area of a MESFET, 413
  current collapse, 428
  leakage current, 428
  metal, 411
  voltage, 439
Gauss's law, 130
Gaussian distribution, 14
General form of the density of states in superlattices, 180
Generation
  of electrical signal, 459
  of electromagnetism, 4
  of excess carriers, 286
  rate, 467
  and recombination, 286–288
  -recombination noise, 469, 489
Germanium *pn*-junction, 123
Gibbs phase rule, 333
Golay cells, 459
Graded bandgap, 127
Graphic solution of a finite potential well, 53
Ground bound state, 206
Ground state energy level of simple harmonic oscillator, 29
  in quantum well, 116
Group velocity, 14, 98
  axis, 182
  of bulk semiconductors, 334–347
  direction, 95
  of GaAs, 350
  of InP, 350
  mode, 378–380, 387, 485
  parameters, 340
  of quantum dots, 377–399
  rate, 369–372
    of HVPE, 352
  of semiconductor thin films, 347–377
Gunn diode, 282, 324, 458
  oscillations, 282
Half-width at half-maximum, 213
Hall coefficient, 267, 269, 302–304
Hall factor, 267
Hall field, 266
Hall mobility, 267, 302
Hall voltage, 267
Hall effect
  conventional, 265–269, 272
  device geometry, 271
  fractional quantum, 277
  quantum, 269–272, 275–277
  resistivity, 271
  sample configuration, 265
Hamiltonian, 28, 29
  of electron in triangular well, 59
  of exciton, 231
  matrix, 110
  matrix elements, 109
  operator, 16
Harmonic oscillator, 3
Haynes-Shockley experiment, 293–294
HBT, 433
  advantages, 434
Heat sink, 511
Heaviside step function, 206, 309

Heavy hole, 111, 112, 207
  effective mass, 227, 485
Heisenberg uncertainty principle, 12, 13, 295
HEMT, 312, 323, 403, 424
Hermite polynomial, 62, 63
Hermitian, 38, 110
Hermitian conjugate, 21
Hertz, 4
HET, 435
Heterojunction, 127, 133, 262
  bipolar transistor, 433–435
  field-effect transistors, 423–427
  field-effect transistors, 424
  interface, 133
  lasers, 506
Hexagonal unit cell, 365
HFETs, 424, 428
High electron mobility transistor, 58, 403, 424
High symmetry points, 227
High-brightness cathode, 388
High-frequency devices, 285
High-frequency dielectric constant, 197, 252
High-index substrate, 395
High-point symmetry, 92
High-power applications, 412
High-power pulsed-laser beam, 355
High-pressure crystal growth, 352
High-pressure environment, 346
High-pressure growth conditions, 362
High-purity silicon boules, 343
High-purity thin films, 411
High-temperature environment, 346
Hole, 167
  band dispersions, 111
  concentration, 168, 170
  distribution function, 168
  effective mass, 119, 120, 188, 222, 227, 411, 485
  generation rate, 289
  -hole interaction, 277
  recombination rate, 289
Homoepitaxial growth, 394
Homogeneous growth, 381
Homojunction, 139
  pn-diode, 405
Hooke's law, 196
Horizontal Bridgman (HB) technique, 329, 341–343
  HVPE reactor, 354
Hot electron, 280
  transistor, 435, 403
Hot hole, 280
Huang-Rhys factor, 243
HVPE, 352
  reactor, 353
Hydride gas, 350
  vapor-phase epitaxy growth of thick GaN layers, 352–355
Hydrogen atom, 7

Ideal diode, 470
  gas equation, 369
  MESFET, 414
  metal-semiconductor junction, 409
Identity, 102

III-nitride
  heterostructures, 375
  materials, 347, 349, 355
  nanostructures, 375
  semiconductor materials, 457
  semiconductors, 216, 217
III-V semiconductor heterojunction laser diodes, 510
Illustration of quantum dot laser, 523
Imaginary part
  of complex susceptibility, 524
  of dielectric constant, 195, 198
  of refractive index, 524
Impact ionization, 281
Imperfections of the host crystal, 262
Impingement flux, 282, 369, 381
Important quantum mechanics postulates, 18
InAs
  bandgap, 209
  lattice constant, 379
  quantum dots, 396, 485
    effect of annealing time, 392
    effect of substrate temperature, 392
    embedded in an InGaAs quantum well, 486
    formation of, 378
  self-assembled quantum dots, 391
InAs/AlGaSb quantum well, 271, 272, 274
InAs/AlSb
  double-barrier resonant tunneling diode, 146
  double-barrier structure, 144
  superlattices, 519
InAs/AlSb/GaSb resonant interband tunneling diode, 147
InAs/AlSbAs, 437
InAs/GaAs
  multiple quantum dot detector, 485
  pyramidal quantum dot, 233
  quantum dots, 242, 486
InAs/InGaAs/GaAs dot-well device, 488
InAs/InGaSb
  superlattices, 208
  type II superlattices, 179
Incident
  blackbody radiant flux, 462
  flux density, 306
  light wavelength, 460
  photon flux density, 460
Incorporation of atoms, 359
Indirect bandgap semiconductors, 205
Indirect recombination, 286
Indirect transition, 205
Indistinguishable particles, 161
Inelastic collisions, 263
Inelastic scattering, 318
Infrared detectors, 399, 457, 459, 460, 463, 464, 471, 472, 486, 487, 489–491, 526, 527
  based on multiple quantum dots, 485–492
Infrared quantum detectors, 459–492
Infrared radiation, 235
Infrared region, 492
InGaAs, 475
  graded quantum well, 488
  quantum dots, 131, 218, 393, 394, 396, 525
InGaAs/InAlAs multiple quantum wells, 240
InGaAsP/InP-based surface emitting LEDs, 515

InGaP/GaAs, 435
Inhomogeneous growth, 396
Inhomogeneous line broadening, 523,524
Injected minority carrier density, 507
Injection
  efficiency, 493, 519
  process, 507
InP single crystal, 336
  growth, 336
In-situ, 355, 357, 359, 361, 363, 374, 375, 377
Integer filling factor, 277
Intensity distribution, 503
Intensity of the electromagnetic wave, 196
Interactions of electromagnetic waves, 193
Interatomic separation, 100, 126
Interband
  matrix element, 115, 201
  transition, 191, 199, 523, 458
  transitions in multiple quantum wells, 223
Interdiffusion, 359
Interface roughness, 314, 316
Interfacial tension
  between film and substrate, 380
  between substrate and vapor, 380
  vapor and film, 380
Interference patterns, 8, 370, 371, 448, 455, 513
Interference phenomenon, 12
Interferogram, 513, 514, 528
  of the edge-emitting laser, 514
Internal losses due to free carrier absorption, 458
Internal quantum efficiency, 469, 479, 480, 509
  of the interband transition, 245
Intersubband transition, 191, 480, 458
  in low-dimensional systems, 210
  in multiple quantum wells, 223, 473
Intervalley scattering, 279
Intraband transition, 191
Intrinsic carrier concentration, 170, 171
Intrinsic Fermi energy level, 171
Intrinsic noise mechanism, 470
Inverse Fourier transform, 25
Inverted vertical HVPE reactor, 355
Inverter, 443, 445
Ionized acceptor concentration, 290
Ionized donor concentration, 290
Irreducible T_d group, 103
Island growth mode, 380
I-V characteristic of the resonant tunneling diode, 144

JFETs, 403
Johnson current noise, 470
Johnson noise, 469
Johnson noise voltage, 470
Joint density of states, 494
Junction field-effect transistor, 403, 420–423

$\mathbf{k} \cdot \mathbf{p}$ theory, 99, 102–113, 218
Kane theory, 99
Ket vector, 17
Kinetic energy, 35
Kinetic model, 392
Kinetic modeling of nucleation, 384
Kinetic processes, 359, 375
Knudsen cells, 358
Kohn-Luttinger parameters, 112–113
KrF excimer laser, 357
Kronig-Penney model, 84, 85–89, 90, 93
k-space, 10, 167, 172, 176, 178, 180

LA phonon mode in GaAs, 251
Lack of dispersion curve in quantum dots, 178
Lagrange method of undetermined multiplier, 158, 162, 164
Landau energy levels, 239
Landau gauge, 238
Landau levels, 269, 272, 274
Landau magnetic length, 277
Landau orbit, 275
Landau quantum number, 181, 239
Landauer formula, 322, 445, 451
Landauer equation, 268
Laplace transformation, 292
Laplacian operator, 9
Laser, 64, 191, 210, 235, 234, 251–253, 355–357, 366, 388, 400, 457–526
  light, 235
  output power, 509
  parameters, 356
  -plume interaction, 356
  -target interaction, 356
  threshold gain, 517
Latent heat of fusion, 335
Lattice, 83
  constant, 246, 395, 517
  -matched heterojunction, 395
  scattering, 279
  vibration, 245–255
Layer growth mode, 380
Layer thickness, 370
Layer-by-layer deposition, 517
Leakage current, 417
Leakage current density, 509
LEC
  growth of compound semiconductors, 336
  single crystal growth, 334, 335
LED, 191, 374, 375, 492–497, 512, 513, 515, 527
  frequency response, 495,496
Lely growth method, 346, 347
Length of cavity, 513
Length of gate, 412, 415
Light group velocity, 509
Light hole, 111–112, 207
  effective mass, 120
Light source, 458
Light-emitting diodes, 191, 492–497
Lightly doped semiconductor, 169
Limitation of resonant tunneling diode, 146
Linear electron density in quantum wire, 176, 177
Linear one-dimensional diatomic lattice model, 247
Linear operator, 16, 17, 20
Lines shape of the optical absorption coefficient in superlattices, 210
Line-width enhancement factor, 524
Liquid encapsulated Czochralski (LEC) technique, 329, 334–340

Liquid encapsulated Kyropoulos (LEK) technique, 329
Liquid phase epitaxy (LPE) technique, 329, 348, 349
  advantages of, 349
  disadvantages of, 349
Liquid-solid interface, 335, 338
LO phonon mode in GaAs, 251
Load-lock chamber, 357
Local oscillators, 139
Localized function, 100
Locking circuits, 139
Long wavelength infrared detectors, 457, 464
Longitudinal acoustic wave, 247
Longitudinal displacement of atoms, 245
Longitudinal elastic constant, 308
Longitudinal modes, 505
Longitudinal optical modes, 503
LOPC modes, 254
Lorentz force, 235, 336
Lorentz model, 196–198, 218, 219
Lorentzian, 138
  line shape, 197
Loss coefficient, 504
Loss factor, 513
Low-diffusion coefficient, 434
Low-dimensional semiconductor structures, 131
Low-dimensional semiconductor systems, 387
Low-dimensional system, 192
Low-field oscillatory conductivity, 273
Low-level carrier injection, 291
Low-power microwave devices, 139, 280
Low-temperature grown AlN, 433
Low-threshold current density, 521
L-valley, 281

Magnetic field, 193, 236, 275
Magnetic induction, 238
Magnetic length, 182, 270
Magnetoresistance, 269
Major scattering mechanisms, 305
Major transport mechanisms in a multiple quantum dot detector, 491
Majority carrier concentration, 287
Majority carriers, 508
Many-body contribution, 443
Many-body effect, 213
Mass action law, 170
Mass flow controllers, 372
Mass flux, 376
Matching conditions, 52
Mathematical tools, 19
Matrix element, 30
Maxwell
  -Boltzmann approximation, 170
Maxwell-Boltzmann distribution function, 296, 300, 498
  -Boltzmann statistics, 157–161
Maxwell's classical electromagnetic theory, 192
Maxwell's equations, 2
MBE, 330, 334, 359, 423, 428, 458, 488, 515, 517
  chamber, 359, 393, 370
  growth, 359
    InAs, 391
    process, 357
    of quantum dots, 377
  sketch, 361
Mean energy of the oscillator, 2
Mean free path, 262, 359
Mean free time, 262
Mean value of the Hamiltonian, 28
Melt-back effect, 349
MESFET, 403
Mesoscopic devices, 263
Mesoscopic regime, 264
Metal-organic chemical vapor deposition technique (MOCVD), 252, 329, 372–377, 311
Metal-organic compounds, 372
Metal-organic sources, 358
Metal-oxide semiconductor field effect transistor, 147, 403
Metal-semiconductor field-effect transistors, 403, 411–420
Metal-semiconductor interface, 403, 404
Metal-semiconductor rectifying contact, 406
Mg-dopant, 357
MgO, 368
Microelectronic devices, 403
Micropipes, 347
Microscopic defects, 239
Microstructures, 332
Miniband, 118, 119, 179–181, 208, 209, 474, 475, 478, 519, 520, 526
  in type II superlattices, 208
Minigap, 520
Minimum detectable signal, 463
Minimum resolvable temperature difference, 465
Minority carriers, 290, 291
  concentration, 287
  injection current, 143
Mirror loss, 518
MKS units, 192
Mobility effective mass, 279
  of electrons, 467
  of holes, 467
  in *n*-type semiconductors, 300–302
MOCVD, 330, 334, 352, 372, 373, 423, 428, 458, 515, 517
  growth of quantum dots, 377
  reactor, 373
MODFET, 312, 323, 403, 424, 426
  based on GaAs/AlGaAs heterojunction, 424
Modified Lely method, 347
Modulation-doped heterojunction interface, 128
Modulation doping, 425
Modulation frequency domain, 468
Modulation-doped field-effect transistor, 403
Mole fraction of ternary compounds, 372
Molecular beam epitaxy, 411
Molecular beam epitaxy (MBE), 218, 313, 329, 357–372
Molecular beam epitaxy growth of III-nitrides, 261–372
Molecules concentration, 359
Molten area, 341
Molten zone, 345

Momentum, 10, 201, 261
  matrix element, 107, 111, 202, 212, 216, 218, 220, 223
  operator, 104, 223
Monatomic gas in a cubic container, 160
Monatomic one-dimensional crystal, 250
Monolayer, 360
Monomer
  capture rate, 385
  surface density, 382
Monte Carlo method, 295
MOSFET, 403
MOSHFETs, 430
MOVPE, 372
Multichannel HFETs, 433
Multiple quantum dot infrared detector, 489
Multiple quantum dot VCSELs, 524
Multiple quantum well
  detectors, 478
  infrared photodetectors, 472–485
Multiple targets, 355

$Na_2B_4O_7$, 337
Nanostructures, 65, 123, 148–150, 153, 193, 239, 240, 243, 375, 388, 457, 472
Narrow bandgap materials, 492
Narrower photoresponse, 474
n-channel, basic operation of, 422
n-channel devices, 420
n-channel MESFET, 411
Nd:YAG, 458
Nearest neighbor interaction, 98
Negative differential resistance, 280
Negative effective mass, 111
NEP, 461, 464
Nernst coefficient, 304
Net recombination rate, 288
NETD, 464
Neutron transmutation doping, 344
NH_3, 351, 353
Nitridation of sapphire, 365
Nitrogen flux of the dc source, 362
Nitrogen plasmas, 362
Noise
  current, 461
  -equivalent bandwidth, 462
  -equivalent irradiance, 463, 464
  equivalent power, 461
  -equivalent temperature difference, 464
  in photodetectors, 469–472
  sources, 469
  voltage, 462
Nondegenerate band, 105
Nondegenerate conditions, 284
Nondegenerate electrons, 311
Nondegenerate energy levels, 228
Nondegenerate semiconductor, 301, 404
Nondispersive media, 13
Nonequilibrium distribution function, 294, 297
Nonequilibrium process, 37, 45
Noninteracting particle distribution function, 156, 157
Noninteracting particle system, 155
Nonmagnetic materials, 194
Nonpolar scattering, 309
Nonradiative recombination lifetime, 245
Nonradiative recombination process, 458
Nonrectifying contact, 403
Nonresonant multiphoton process, 356
Nonsaturation region, 412
Nontrivial solution, 54, 86
Nonuniformity of the self-assembled quantum dots, 523
Nonvanishing matrix element, 212
NOR logic circuit, 437
Normal incident, 222
Normal incident photons, 485
Normalization, 10
  density, 390
Normally off MESFET, 417
Normally on MESFET, 418
n-type collector, 433
n-type emitter, 433
n-type GaAs, 130
n-type quantum wells, 210
n-type QWIP, 483
n-type Schottky diode, 406, 407
n-type semiconductors, 266
n-type Si, 337
Nucleation, 379–387
  parameters of three-dimensions, 386
  parameters of two-dimensions, 386
  rate, 381, 382
  regimes, 386
  site density, 382
Number of longitudinal modes, 504

Occupation number, 156
Occupied energy states, 140
Odd parity, 66
  solution of δ-function potential, 69
Ohmic contacts, 403, 405, 411
One electron approximation in a periodic δ-functions, 93, 94
One-dimensional crystal, 83
One-dimensional electron gas, 323
One-dimensional harmonic oscillator, 239
One-dimensional monatomic lattice, 246
One-dimensional potential profile, 131
One-dimensional system, 37, 171
One-dimensional wave packet, 10, 27
One-electron single-crystal approximation, 83
Optical absorbance, 480
Optical absorption coefficient 196, 204, 205, 207–211, 217, 219, 220, 220, 229, 256, 257, 475, 480, 490
Optical absorption coefficient
  absorption
    of the exciton in bulk semiconductors, 231
    spectra of the intersubband transitions, 474, 475
    spectrum, 478
  confinement factor, 518
  in direct bandgap semiconductors, 199–204
  emission spectra of the element, 6
  gain, 368, 501–505, 521–524, 527
    in bulk semiconductors, 522
    in quantum dot, 522
    in quantum well, 522
    in quantum wire, 522

  in indirect bandgap semiconductors, 205, 206
  of interband transition in quantum wells, 206–208
  of interband transition in type II superlattices, 208–210
  of intersubband transition in multiple quantum wells, 210–216
  of intersubband transitions in GaN/AlGaN multiple quantum wells, 216, 217
  interconnect, 515
  interference method, 370, 371
  irradiant flux, 466
  lithography technique, 388
  loss due to active region, 509
  loss in the active region, 510
  mode distribution in a single quantum well laser, 514
  phonon energy, 309
  phonon modes, 249
  phonon scattering, 309
  power, 495
  properties, 191
  transmission, 375
Optoelectronic devices, 232, 457
Optoelectronics, 64, 239, 329, 347, 367, 395, 428, 437
Orbit radius, 6
Orthogonal properties, 31
Orthogonal wave vectors, 34
Oscillator
  frequency, 2
  strength, 204, 207, 213
Oscillators, 280
Ostwald ripening, 387
Output power of edge-emitting laser, 511
Overlap integral, 79, 101, 207
Overpressure, 342
Oxide-confined VCSELs structures, 516
Oxide confinement, 515

Packets of energy, 4
Parabolic potential well, 61–64
Parallel facets, 503
Paramagnetic detectors, 459
Parasitic capacitance, 427
Parasitic effects, 414
Parasitic resistance, 417
Parity, 66
Particles per quantum state, average, 159
Parseval-Plancherel formula, 27
Patterned quantum dots, 388
Pauli exclusion principle, 90, 164
p-channel JFET, 420
Periodic Bloch function 114
Periodic δ-function potential, 84
Periodic function, 83
Periodic nanostructures, 388
Periodic potential, 83
Periodic potential due to a line of atoms in a crystal, 84
Periodicity of the crystal, 202
Permeability of free-space, 193
Permittivity, 439
  of free space, 6, 130, 193, 252, 285, 290
  of GaAs, 425
  of semiconductor materials, 405, 408
Perturbation, 29–33
Phase coherence transport, 322
Phase diagram, 332
  for GaN, 333
  of single component system, 332
Phase velocity, 14
Phonon, 245–255
  angular frequency, 241
  density of states, 250
  energy, 205, 241
  modes semiconductors, 251
  -plasmon coupled mode, 252
  replicas, 241, 242
  wavevector, 205
Phosphine, 350
Photoconductive gain, 460, 466–468, 471, 483, 490
Photoconductivity, 475
  basic concepts, 465–469
  measurements, 487
  spectrum, 487
Photocurrent, 460, 472, 490, 491
Photodissociation, 356
Photoelectric effect, 4
Photoexcited carriers, 491
Photo-ionization, 356
Photoluminescence, 239–245
  definition, 240
  emission, 225
Photon, 164
  -assisted tunneling, 489
  density, 509
  -electron interaction, 200, 201
  -electron selection rules, 485
  emission, 457
  momentum, 224
  noise, 463
Photonic detectors, 459
Photoreflectance, 375
  spectrum in GaN, 230
Photoresponse, 460, 478
  spectra, 476, 479, 488
Photoresponsivity, 472
Physical parameters
  of GaN, AlN, and InN, 363
  of sapphires, 364
  of SiC, 364
Physical properties
  of GaAs, 368
  of quantum dots, 393
  of Silicon, 368
Piezoelectric coefficients, 429
Piezoelectric potential, 308
  scattering, 308
Pileup process, 282
Pinch-off, 412, 416, 421, 422–426, 427
  current, 416
  voltage, 416, 421, 425
Planar distribution of elements, 360
Planar self-organization of quantum dots, 396
Planck constant, 2
Planck expression of the energy flux, 2
Planck's photon emittance, 464

Plasma frequency, 197, 252
Plasma sources, 362
Plasmon, 252
  damping rate, 255
p-like wave function, 92
pn junction under the influence of forward bias voltage, 493
pn-junction, 148
  diode, 291
  lasers, 505
  LED, 492
  photodiodes, 472
  tunneling diode, 139–143
Point contact diode, 123
Poisson's equation, 290
Polar scattering, 309
Polarity, 366
Polarization
  component, 222
  of electric field, 236
    selection rule, 214
    vector, 194
  light component, 485
Polarized induced charges, 216
Polycrystalline materials, 334, 341, 343
Polycrystalline SiC, 347
Polysilicon rod, 344
Population inversion, 501, 506, 517
  efficiency, 519
Postgrowth annealing, 391, 392
Potential barrier with finite height, 41–49
Potential distribution near the semiconductor interface, 60
Potential energy, 15
  -effect devices, 403
  well with finite-depth, 51–56
  well with infinite depth, 49–51
Poynting vector, 200
Preamplifier noise, 471
Pressure-controlled LEC, 329
Pressure-temperature phase diagram, 333
Primitive unit cell, 167, 175, 202
Probability, 153
  amplitude, 9
  current, 38–40
  density, 38
  function 50
  of obtaining nondegenerate eigenvalue, 18
Propagating wave, 72
Propagation
  constant, 246
  vector of envelop wave function, 209
  vectors, 37, 49, 195
    of unbound particle, 56
Properties of δ-function, 25–26
Pseudopotential methods, 103
PSpice, 421, 423
p-type multiple quantum wells, 223
p-type quantum wells, 210, 216
p-type QWIP, 483, 484
p-type Schottky diode, 406, 407
p-type semiconductors, 266
p-type Si, 337
Pulsed-laser deposition technique, 355–357
Pyramidal-like shape, 131
Pyroelectric detectors, 459
Pyrolytic boron nitride (PBN), 358
Pyrometer, 373, 374

Quantized energy levels in triangular potential well, 77
Quantum box, 264
Quantum cascade laser, 459, 517–521
  based on type II superlattices, 521
  basic operation principle, 517
Quantum conductance, 148, 321, 439
Quantum detectors, 459, 490
Quantum dots, 29, 64, 66, 84, 99, 120, 124, 131, 184, 216, 218–222, 225, 233, 234, 242, 243, 256, 329, 330, 347, 357, 374, 377, 379, 380, 387–400, 402, 437, 451, 459, 472, 489, 490, 522–527
  Absence of wavevector selection rules, 178
  discrete energy levels, 178
  production of, by lithography technique, 377
Quantum efficiency, 459, 463, 479
Quantum Hall effect, 269–272, 275–277
  device geometry, 271
Quantum interference, 450
Quantum lasers, 521–526
Quantum mechanics, 1
Quantum photodetectors, 472
Quantum regime, 264
Quantum resistance, 321, 439
Quantum transmission probability, 445
Quantum transport, 319
Quantum well devices, 264
  edge-emitting lasers, 511–514
  infrared photodetectors, 472
  in presence of electric field, 184
Quantum wire, 171, 175–178, 183, 185–189, 263, 264, 275, 325, 347, 387, 398, 300, 521, 522
Quartz crucibles, 337
Quasi-bulk thick GaN layers, 352
Quasi-continuous eigenvalues, 159
Quasi-continuous energy level, 162
Quasi-Fermi energy level, 494, 495
Quaternary alloys, 457
QWIP, 472, 479, 481, 483–485, 526, 528
  performance, 481

Radiative downward transitions, 517
Radiative efficiency, 519
Radiative recombination lifetime, 245
Raising and lowering operators, 34
Real part of refractive index, 524
Raman
  intensity, 253
  scattering, 251, 252
    spectra, 255
    spectrum of InGaAsN/GaAs single quantum well, 253
  thermal motion of charge carriers, 469
Range of de Broglie wavelength, 262
Rapidly varying Bloch function, 115
Rate constants, 384
Rayleigh-Jeans expression, 2

Ready-to-use wafers, 330
Real part of the dielectric constant, 195, 198
Real part of the refractive index, 195
Real-space transfer devices, 403
Reciprocal space, 96
Recombination
  efficiency, 493
  lifetime, 471, 495, 496, 502
  rate, 286, 287
  time, 461, 467, 468
Recursion formula, 62
Reduced density of states, 206, 207
Reduced effective mass, 220, 494
Reduced-zone representation, 89
Reflected wave packet, 48
  coefficient, 41, 42, 45, 323, 469
  coefficient of unbound particle, 57
  high energy electron diffraction (RHEED), 357
  and transmission of electromagnetic waves, 222
Reflectivity, 490, 504
Refractive index, 195, 371, 503, 524, 525
  of active region, 518
  of AlGaAs, 517
  of GaAs, 405
  of medium, 500
Relaxation time, 296
  approximation, 296
Relaxed heterojunction, 395
Representation of single electron transistor, 441
Resistivity measurements, 269
Resistivity tensor, 276
Resolve sheer stress, 337
Resonant cavity, 502
Resonant interband tunneling diode, 147
Resonant peaks, 136
Resonant tunneling diode, 123, 135, 143–147
Responsivity, 460, 490
  of LED, 495
  of quantum detectors, 460
  of quantum dot infrared photodetector, 487
Reverse saturation current density, 143
RF induction, 355, 373
RF systems, 417
RHEED, 356–358, 360, 361, 375, 377, 389, 390
  diffraction patterns, 360
  dynamic mode, 360
  images, 389, 377
  oscillations, 360
  patterns, 360, 390
  static mode, 360
RHET, 436
Rhombohedral unit cell, 365
Righi-Leduc coefficient, 304
Ripening effect, 387
Ruby, 458
Rydberg, 101
Rydberg constant for the hydrogen atom, 226

S operation, 103
S_4operation, 103
Sapphire nitridation, 363
Saturation
  current, 414, 416
  drift velocity, 482
  mode, 417–419
  region, 412
  reverse current, 470
  velocity, 282
Scalar product, 19
Scanning tunneling microscopy, 388, 393, 396
Scattering
  cross section, 305
  form δ-function potential, 310
    acoustic phonons, 307
    alloy potential, 310
    dipole, 311
    dislocations, 310
    impurities, 262
    ionized impurities, 306
    neutral impurity, 307
    short-range potentials, 310
  by interface roughness, 315
  mechanisms, 296, 302
    in bulk semiconductors, 304–312
  by remote ionized impurity, 314
  time, 198, 274, 279
  in two-dimensional electron gas, 312
Schottky barrier, 408, 448
    diode, 409
    height, 410
    photodiodes, 472
Schottky contacts, 405, 411
Schottky diode, 123, 405, 406–411
Schrödinger equation, 9, 16, 36, 37, 49, 73, 85, 226
  for coupled well, 78
  of δ-function, 65
  for pyramidal potential barrier, 132
  for smooth potential barrier, 126
Second law of thermodynamics, 331
  order correction to the eigenfunction, 33
  order correction to the eigenvalue, 32
  order perturbation, 31, 32
Secular equation, 23
Seebeck coefficient, 304
Seed single crystal, 335
Seeded sublimation growth, 347
Semiconductor laser power output, characteristic behavior of, 510
Segregation coefficient, 338
Selection rules, 222–225
  determined from momentum matrix elements, 218
  for direct transitions, 202
  of electron-photon coupling in quantum wells, 222
Self-assembly quantum dots, 388–399
Semiconductor
  alloys, 330
  growth technologies, 329
  heterojunction lasers, 505–526
  heterojunctions, 29, 153
  interfaces, 58, 60
  lasers, 458, 498–514
Semi-insulating substrates, 337

Separation between fringes, 371
  of charges, 127, 130
  of variables, 15
Sheet carrier concentration, 429
  of ionized impurity, 315
Shift integral, 79
Short range electron-electron scattering, 317
Short-period superlattices, 487, 519
Shot noise, 463, 470, 483
Shubnikov-de Haas effect, 269–275, 279
Si/SiGe, 435
SiC melt, 346
SiC substrates, 366
SiC-based devices, 347
Si-dopant, 357
Signal fluctuation limits, 463
Signal-to-noise ratio, 461, 462
Silicon-based single-electron transistor, 443
Silicon single-crystal boule, 330
Silicon substrate, 366
Simple harmonic oscillator, 28, 36, 61
Simple harmonic oscillator, solution of, 61
Single crystal boules, 347
  crystal growth, 341
  electron energy, 148
    in a periodic potential, 83
    solid, 104
    transistor, 149, 437–451, 454, 455
    transistor as an inverter, 445
    tunneling, 443
  heterojunction structure, 432
  quantum well edge-emitting laser diode, 511
  valley model, 302
Single electron transistor, sketch of 438, 442
Single electron transistor based on Si:Te, sketch of, 444
Singularities, 68
  s-like wave function, 92
Slope efficiency, 510
Small-diameter crystals, 344
Small-perturbation case, 299
Small-signal analysis, 419
Smooth potential barrier, 125, 126
Signal-to-noise ratio (SNR), 462
Solid source, 358
Solid-phase crystallographic structure, 335
Solid-state lasers, 458
Source, 406
Source shutters, 393
Space charge density, 409
Space charge region, 412
Space charge volume density, 409
Spacer, 312
  layer, 425
Specific contact resistance, 405
Specific heat capacity, 3
  of solids, 2
Spectral current density, 483
Spectral current responsivity, 468
Spectral cutoff energy, 483
Spectral radiant input power, 490
Speed of light, 195
Spherical surface of a constant energy, 154
Spin
  degeneracy, 180, 183, 320
    of electrons, 172
  eigenvalues, 181
  -orbit coupling, 107
    Hamiltonian, 107
    interaction, 229
    splitting energy, 272
    valence band, 110
Spin-down wave function, 107
Spin-up wave function, 107
Split channels, 450
Spontaneous emission, 494, 498, 512
  lifetime, 500
  region, 511
Spontaneous polarization, 216
  fields, 429
  of GaN, 429
Stacks of quantum wells, 477
States in 1D system, total number of, 176
Stationary Schrödinger equation in triangular quantum well, 185
Stationary state, 16, 37
Statistical mechanics, 153
Statistical thermodynamics, 153
Steady-state, 286, 296, 299, 468, 507
  continuity equation, 376
Step function, 15, 173, 176
Step potential, 37
Stimulated emission, 498, 501, 503, 512
  region, 511
STM, 438
  images, 391
  images of corrugated GaAs template, 397
Stoichiometric coefficient, 331
Stokes phonon scattering, 251
Strained heterojunction, 395
Stranski-Krastanov growth mode, 218, 378, 380, 387, 486
Strength of a δ-function, 64
Subbands, 173
Sublimation, 347
Submonolayer, 357
Substrate temperature, 360, 375, 391
Supercooled solution, 348
Superlattice barrier, 216
Superlattices, 95–98
Superposition principle, 9
Suppression of the current, 148
Surface carrier depletion depth, 239
Surface construction, 289
Surface density of semiconductor crystals, 370
Surface diffusion
  coefficient, 383
  diffusion energy, 382
Surface emissivity, 373
Surface migration, 359
Surface nonradiative recombination, 495
Surface of substrate, 393
Surface preparation, 363

TA phonon mode, 243
  in GaAs, 251
Taylor series, 97, 241, 288
T_d symmetry, 106

TEM, 389
  images, 396–399
  -insensitive threshold current, 521
  spreading of phases, 319
  of target, 356
Ternary alloys, 457
Tetrahedral point group symmetry, 102
Thermal conductivity, 335
  of GaAs, 337
  of Si, 337
Thermal desorption, 359
Thermal detectors, 459
Thermal energy, 148, 410
Thermal generation rate, 286
  of carriers, 471
Thermal imager systems, 465
Thermal imagers, 464
Thermal kinetic energy, 246, 261
Thermal spreading, 263
Thermal velocity, average, 305
Thermionic emission, 265, 481, 489
Thermionic mechanism, 410
Thermocouples, 459
Thermodynamic analysis, 330
Thermodynamic process, 375
Thermodynamic quantities, 154
THET, 435
Thin tunnel junction, 148
Three-dimensional crystal, 251
Three-dimensional (3D) density of states, 168
Three-dimensional energy level system, 501, 502
Three-dimensional infinite cubic potential well, 154
Three-dimensional island formation, 391, 394
Three-dimensional structures, 377
Three-dimensional system, 166, 171
Three-dimensional wave packet, 10
Threshold
  conditions for lasing, 502
  current, 511
  current density, 509, 519
  gain, 504, 519
Tight-binding method, 99–102
Time evolution of the wave packet, 13
Time-dependent current density, 194
Time-dependent transient density of stable nuclei, 385
TO phonon mode in GaAs, 251
Top of the valence band, 171
Transconductance, 416, 418, 434, 446, 447, 451
Transfer integral, 79
Transfer matrix procedure, 71, 135
Transient time, 419
Transistor gain, 416
Transit time, 264, 451, 461, 468
Transition
  between heavy hole and light hole energy levels, 216
  matrix element, 304
  probability, 313, 502
  from 2D to 3D growth, 378
Transition probability, total, 200, 203
Transmission
  coefficient, 41, 42, 44–46, 71, 123, 124, 323
    through δ-function barrier, 125
    in double-barrier potential, 136
    through heterojunction, 128
    through Schottky barrier, 130
    of unbound particle, 57
  electron microscopy, 388
  in finite double-barrier potential wells, 79
  probability, 44
    coefficient, 450
  spectrum of cyclotron resonance, 237
Transparency point, 502
Transport
  mechanisms, 265
  properties, 144, 261, 262
    of bulk semiconductors, 277
    of photodetectors, 467
  regimes in semiconductor devices, 264
  without scattering, 318
Transverse displacement of atoms, 247
Transverse magnetoresistance, 304
Transverse modes, 247
Transverse momentum, 140
Trapping probability, 490
Trial function, 28, 36
Trial kets, 28
Triangular potential well, 58, 76, 77
Tunnel through a barrier, 404
Tunneling, 37
  through barrier, 265
  carriers, 123
  current, 140
  current density, 405
  effect, 46
  electron transistor, 435–437
  hot electron transistor, 435
  of particle through potential barrier, 47
  phenomenon, 123
  probability, 140, 141, 405
  process, 489
  through pyramidal potential barrier, 131–135
  of states into the bandgap, 187
  transmission coefficient, 320, 482
Turning point of decaying function, 76
Turning points, 72, 75, 124, 137, 150, 482
Two channels, 449
Two-barrier structure, 131
Two-dimensional arrays of VCSELs, 515
Two-dimensional density of states, 173
Two-dimensional electron gas, 58, 271, 315
Two-dimensional electron gas density, 213
Two-dimensional system, 171, 206
Two-energy level system, 499
Two-stage heater, 342
Type I quantum wells, 114, 206, 207
Type II superlattices, 114, 179, 208, 209, 519, 521

Ultrahigh vacuum conditions, 357
Ultrahigh vacuum in the MBE chamber, advantages of, 359
Ultraviolet LEDs, 457
Unbound motion of a particle, 56
Uncertainty principle, 263
Uniform doping distribution, 340

Unipolar devices, 517
Unipolar resonant-tunneling hot electron transistor, 436
Unit cell of sapphire, 365
Unit vector of light polarization, 214
Unperturbed basis states, 32
Unperturbed Hamiltonian, 29, 30
Unwanted impurities, 336

Valence
  band offset, 209, 507
  band parameters, 111
van der Pauw geometry, 267
Vapor pressure, 359
Vapor-phase epitaxy (VPE)
  advantages of, 351
  disadvantages of, 351
  growth of GaN, problems associated with, 351
  method, 349–352
  reactor, 351
  reactor, cross section of, 350
Variation method, 28, 29, 231
Variational parameter, 231
VCSELs, 515, 517, 526
  device structure, 516
Vector
  function, 299
  potential, 201, 238
Velocity
  overshoot, 281
  undershoot, 281
Vertical cavity surface emitting lasers, 459, 515–517
Vertical gradient freezing (VGF) method, 329
Vertical HVPE reactor, 354
Vertical transport, 265
Virtual states in the continuum, 136
Volmer-Weber growth mode, 378, 380
Voltage
  -controlled devices, 403
  spectral responsivity, 460
Volume
  charge density distribution function, 38
  occupied by one state, 167

Wafer
  holder, 375
  surface, 330
Wafering process, 330
Wannier function, 100
Wannier-Mott exciton, 192
Warped spheres, 112
Wave equation, 195
Wave function, 9, 10, 15, 16
  of an infinite potential well, 50
  space, 19
Wave packet, 47, 48
Waveguide, 214, 218, 503
  configuration, 215, 216, 475, 480
Wavelength, 35
  of various spectral regions, 458
Wavelike aspect of matter, 8
Wavelike behavior, 262
Wavelike phenomena, 263
Wavelike properties, 263
Weak periodic potential, 92
Wentzel-Kramers-Brillouin (WKB) method, 123, 128, 137
  approximation, 72–78, 124, 128, 134, 482
  boundary conditions, 137
  connecting formulas, 75
  semiclassical approximation, 80
Wetting layer, 378, 380, 390
Wide bandgap materials, 492
Work function, 5, 404, 407
  for several metals, 406
Wurtzite
  materials, 252
  structure, 216

x-ray diffraction, 289
X-TEM images, 396–399

YAG laser, 251
Young's equation, 380

Zener-Bloch oscillations, 319
Zero degree
  of Airy function, 60
  of freedom, 178
  phonon line (ZPL), 241, 242
Zero-dimensional system, 171
Zinc-blende, 252
  crystal, 107
  GaN, 369
  semiconductors, 106, 308
  structure, 102, 347
ZnO, 228, 352

This page intentionally left blank.

McGraw-Hill's
Digital Engineering Library
A powerful online repository of engineering content from the field's premier publisher

UNPARALLELED CONTENT AND FUNCTIONALITY

DIGITAL ENGINEERING LIBRARY
McGRAW-HILL ENGINEERING ONLINE
Home | About DEL | Subscription Information | Help | Contact Us
My DEL | View Cart | SiteMa
Log In
Log In
Not registered yet?
Search
Search Digital Engineering Library
Go
● Search All Topics
○ Search Dictionary
Advanced Search
Messages
Choose a Topic
Aerospace Engineering
Bio Engineering
Chemical Engineering
Civil Engineering
Communications Engineering
Construction & Arch Engineering
Electrical Engineering
Engineering Math & Science
Environmental Engineering
Industrial Engineering
Materials Science & Engineering
Mechanical Engineering
Headline News
New Content
Cam Design Handbook
Harold A. Rothbart
Highway Engineering Handbook: Building and

Click a topic below to browse subtopics and articles

McGraw-Hill Construction names Harvey Bernstein as vice president of Industry Analytics. McGraw-Hill Professional applies digital object identifiers (DOIs) to Digital Engineering Library content.

Masthead
(7-Jun-2004)
EDITOR-IN-CHIEF Janice L. Tuchman, (212) 904-3251, jan_tuchman@mcgraw-hill.com

Target Date Slides as Electric

- ✓ 4,000 articles derived from 150+ McGraw-Hill titles
- ✓ 12 major engineering disciplines
- ✓ 500+ Topics
- ✓ Easy and effective search or topical browse
- ✓ By the most trusted and respected authors
- ✓ Subscription or pay-per-view option

No other online engineering resource is this AUTHORITATIVE!

Nothing else is this COMPREHENSIVE!

SIGN-UP FOR A FREE 30-DAY TRIAL:

www.digitalengineeringlibrary.com
(CLICK ON "SUBSCRIPTION INFORMATION")