

A D V A N C E D

THEORY
O F
SEMICONDUCTOR
DEVICES

Karl Hess

ADVANCED THEORY OF SEMICONDUCTOR DEVICES

IEEE Press
445 Hoes Lane, P.O. Box 1331
Piscataway, NJ 08855-1331

IEEE Press Editorial Board
Robert J. Herrick, *Editor in Chief*

J. B. Anderson	S. Furui	P. Laplante
P. M. Anderson	A. H. Haddad	M. L. Padgett
M. Eden	S. Kartalopoulos	W. D. Reeve
M. E. El-Hawary	D. Kirk	G. Zobrist

Kenneth Moore, *Director of IEEE Press*
John Griffin, *Acquisition Editor*
Marilyn G. Catis, *Assistant Editor*
Surendra Bhimani, *Production Editor*

IEEE Electron Devices Society, Sponsor
ED-S Liaison to IEEE Press, Kwok Ng

IEEE Solid-State Circuits Society, Sponsor
SSC-S Liaison to IEEE Press, Stuart K. Tewksbury

Composition: William Henstrom
Illustration: Robert F. Mac Farland
Cover design: Sharon Klein, *Sharon Klein Graphic Design*

Books of Related Interest from IEEE Press

NONVOLATILE SEMICONDUCTOR MEMORY TECHNOLOGY:
A Comprehensive Guide to Understanding and Using NVSM Devices
Edited by William Brown and Joe E. Brewer
1998 Hardcover 616 pp ISBN 0-7803-1173-6

SEMICONDUCTOR MEMORIES: Technology, Testing, and Reliability
Ashok K. Sharma
1997 Hardcover 480 pp ISBN 0-7803-1000-4

HIGH-TEMPERATURE ELECTRONICS
Edited by Randall Kirschman
1998 Hardcover 912 pp ISBN 0-7803-3477-9

ADVANCED THEORY OF SEMICONDUCTOR DEVICES

Karl Hess

University of Illinois at Urbana-Champaign

IEEE Electron Devices Society, Sponsor

IEEE Solid-State Circuits Society, Sponsor



The Institute of Electrical and Electronics Engineers, Inc., New York



A JOHN WILEY & SONS, INC., PUBLICATION

New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

© 2000 THE INSTITUTE OF ELECTRICAL AND ELECTRONICS
ENGINEERS, INC. 3 Park Avenue, 17th Floor, New York, NY 10016-5997
All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 and 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4744. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 605 Third Avenue, New York, NY 10158-0012. (212) 850-6011, fax (212) 850-6008, E-mail:
PERMREQ@WILEY.COM.

**For ordering and customer service, call 1-800-CALL-WILEY.
Wiley-IEEE Press ISBN 0-7803-3479-5**

10 9 8 7 6 5 4 3 2

Library of Congress Cataloging-in-Publication Data

Hess, Karl, 1945-

Advanced theory of semiconductor devices / Karl Hess.

p. cm.

Includes bibliographical references (p.).

“IEEE Electron Devices Society, sponsor.”

“IEEE Solid-State Circuits Society, sponsor.”

ISBN 0-7803-3479-5

1. Semiconductors. I. Title.

TK7871.85.H475 1999

621.3815'2--dc21

99-44500

CIP

To the memory of my father
Karl Joseph Hess

CONTENTS

Preface	xiii
Acknowledgments	xv
Chapter 1 A Brief Review of the Basic Equations	1
1.1 The Equations of Classical Mechanics, Application to Lattice Vibrations	2
1.2 The Equations of Quantum Mechanics	9
Chapter 2 The Symmetry of the Crystal Lattice	19
2.1 Crystal Structures of Silicon and GaAs	19
2.2 Elements of Group Theory	22
2.2.1 Point Group	22
2.2.2 Translational Invariance	26
2.3 Bragg Reflection	29
Chapter 3 The Theory of Energy Bands in Crystals	33
3.1 Coupling Atoms	33
3.2 Energy Bands by Fourier Analysis	34
3.3 Equations of Motion in a Crystal	42
3.4 Maxima of Energy Bands—Holes	46
3.5 Summary of Important Band-Structure Parameters	50
3.6 Band Structure of Alloys	50
Chapter 4 Imperfections of Ideal Crystal Structure	57
4.1 Shallow Impurity Levels—Dopants	58
4.2 Deep Impurity Levels	60
4.3 Dislocations, Surfaces, and Interfaces	62

Chapter 5	Equilibrium Statistics for Electrons and Holes	67
5.1	Density of States	67
5.2	Probability of Finding Electrons in a State	73
5.3	Electron Density in the Conduction Band	75
Chapter 6	Self-Consistent Potentials and Dielectric Properties	81
6.1	Screening and the Poisson Equation in One Dimension	82
6.2	Self-Consistent Potentials and the Dielectric Function	83
Chapter 7	Scattering Theory	89
7.1	General Considerations—Drude Theory	89
7.2	Scattering Probability from the Golden Rule	94
7.2.1	Impurity Scattering	94
7.2.2	Phonon Scattering	96
7.2.3	Scattering by a δ -Shaped Potential	102
7.3	Important Scattering Mechanisms in Silicon and Gallium Arsenide	103
Chapter 8	The Boltzmann Transport Equation	109
8.1	Derivation	109
8.2	Solutions of the Boltzmann Equation in the Relaxation Time Approximation	114
8.3	Distribution Function and Current Density	121
8.4	Effect of Temperature Gradients and Gradients of the Band Gap Energy	125
8.5	Ballistic and Quantum Transport	127
8.6	The Monte Carlo Method	129
Chapter 9	Generation-Recombination	135
9.1	Important Matrix Elements	135
9.1.1	Radiative Recombination	135
9.1.2	Auger Recombination	139
9.2	Quasi-Fermi Levels (Imrefs)	139
9.3	Generation-Recombination Rates	140
9.4	Rate Equations	144
Chapter 10	The Heterojunction Barrier	147
10.1	Thermionic Emission of Electrons over Barriers	147

10.2	Free Carrier Depletion of Semiconductor Layers	151
10.3	Connection Rules for the Potential at an Interface	153
10.4	Solution of Poisson's Equation in the Presence of Free Charge Carriers	154
10.4.1	Classical Case	154
10.4.2	Quantum Mechanical Case	157
10.5	Pronounced Effects of Size Quantization and Heterolayer Boundaries	162
Chapter 11	The Device Equations of Shockley and Stratton	167
11.1	The Method of Moments	167
11.2	Moment for the Average Energy and Hot Electrons	170
11.2.1	Steady-State Considerations	171
11.2.2	Velocity Transients and Overshoot	175
11.2.3	Equation of Poisson and Carrier Velocity	176
Chapter 12	Numerical Device Simulations	181
12.1	General Considerations	181
12.2	Numerical Solution of the Shockley Equations	184
12.2.1	Numerical Simulation Beyond the Shockley Equations	188
Chapter 13	Diodes	193
13.1	Schottky Barriers—Ohmic Contacts	194
13.2	The <i>p</i> – <i>n</i> Junction	201
13.2.1	Introduction and Basic Physics	201
13.2.2	Basic Equations for the Diode Current	207
13.2.3	Steady-State Current in Forward Bias	211
13.2.4	AC Carrier Concentrations and Current in Forward Bias	213
13.2.5	Short Diodes	215
13.2.6	Recombination in Depletion Region	216
13.2.7	Extreme Forward Bias	219
13.2.8	Asymmetric Junctions	221
13.2.9	Effects in Reverse Bias	223
13.3	High-Field Effects in Semiconductor Junctions	226
13.3.1	Role of Built-In Fields in Electron Heating and <i>p</i> – <i>n</i> Junction Currents	226

13.3.2	Impact Ionization in <i>p-n</i> Junctions	229
13.3.3	Zener Tunneling	236
13.3.4	Real Space Transfer	240
13.4	Negative Differential Resistance and Semiconductor Diodes	241
Chapter 14	Laser Diodes	247
14.1	Basic Geometry and Equations for Quantum Well Laser Diodes	248
14.2	Equations for Electronic Transport	250
14.3	Coupling of Carriers and Photons	253
14.4	Numerical Solutions of the Equations for Laser Diodes	257
Chapter 15	Transistors	265
15.1	Simple Models	266
15.1.1	Bipolar Transistors	266
15.1.2	Field Effect Transistors	272
15.2	Effects of Reduction in Size, Short Channels	278
15.2.1	Scaling Down Devices	278
15.2.2	Short Gates and Threshold Voltage	279
15.3	Hot Electron Effects	281
15.3.1	Mobility in Small MOSFETs	281
15.3.2	Impact Ionization, Hot Electron Degradation	284
Chapter 16	Future Semiconductor Devices	291
16.1	New Types of Devices	291
16.1.1	Extensions of Conventional Devices	291
16.1.2	Future Devices for Ultrahigh Integration	293
16.2	Challenges in Nanostructure Simulation	295
16.2.1	Nanostructures in Existing Semiconductor Devices	296
16.2.2	Quantum Dots	297
16.2.3	Structural, Atomistic, and Many-Body Effects	297
Appendix A	Tunneling and the Golden Rule	301
Appendix B	The One Band Approximation	305

Contents	xi
Appendix C Temperature Dependence of the Band Structure	307
Appendix D Hall Effect and Magnetoresistance	309
Appendix E The Power Balance Equation	311
Appendix F The Self-Consistent Potential at a Heterojunction	315
Appendix G Schottky Barrier Transport	317
Index	321
About the Author	333

PREFACE

This book evolved from my earlier book of the same title. Chapters have been added (e.g., one on laser diodes); others have been completely rewritten (e.g., the chapter on the Boltzmann equation).

Semiconductor devices are now the substrates of information and computation—the substrates of Internet browsers that sift with great speed through a world of information and represent the information visually to the user, and the substrates of artificial intelligence. They form the basis of all computer chips, of solar cell arrays, and of the newer red lights on cars. They are essential in fiber communications, and laser diodes are among the most sophisticated semiconductor devices. They are truly ubiquitous and can be found in increasing numbers in cars, kitchens and even in electronic door locks. Trillions of the basic semiconductor devices, p - n junction diodes, are fabricated daily, and Moore's law of increasing the integration and reducing the device size every 18 months has been persistently obeyed.

My goal is to present a description of the theoretical concepts underlying device function and to cover device theory from the principles of condensed matter physics and chemistry to the numerical mathematics of device simulation tools, all in a form understandable for anyone who knows advanced calculus and some numerical algorithms important for the solution of the device equations, the Boltzmann equation, and the Schrödinger equation. This goal could not be achieved. Instead I have presented only an overview of some of the most important concepts of selected devices. To obtain a truly broad knowledge of device theory, the reader will need to study additional books that are referenced, particularly the *Solid State Theory* edited by Landsberg, the encyclopedic description of most devices by Sze, and the text on numerical device simulation by Selberherr.

Karl Hess
University of Illinois at Urbana-Champaign

ACKNOWLEDGMENTS

I would like to express my sincere thanks to Wolfgang Fichtner, who has stimulated the revision and invested much time to give advice for improvement. B. G. Streetman, R. Dutton, and M. Lundstrom have given valuable advice during many important stages of the development of this book. Others have contributed to various sections: P. D. Yoder to the section on the density of states and Monte Carlo simulations, J. Bude to the sections on impact ionization through his insight, S. Laux to the chapter on diodes as the major contributor to the new treatment of $p-n$ junctions, Alex Trellakis to the explicit solution for the even part of the distribution function, M. Grupen to the insights presented in the chapter on laser diodes based on his pathbreaking work on laser diode simulation, and L. F. Register to the theory of collision broadening in Monte Carlo simulations and to some of the treatment of the electron phonon interactions.

I thank J. P. Leburton, U. Ravaioli, M. Staedele, F. Oyafuso, B. Klein, F. Register, E. Rosenbaum, and B. Tuttle for reading selected chapters and suggesting improvements, and the students in my classes who have found and corrected many mistakes.

Special thanks go to L. R. Cooper from the Office of Naval Research, to M. Stroscio from the Army Research Office, and to George Lea from the National Science Foundation for their insights regarding the importance of topical areas and for their encouragement. G. J. Iafrate has worked with me on many topics and has influenced my thinking from velocity overshoot to quantum capacitance.

The Beckman Institute of the University of Illinois and its first directors, T. L. Brown and J. Jonas, have provided an ideal environment to cover theoretical expertise of a range of disciplines including basic physics, chemistry, electrical engineering, and numerical mathematics.

William Henstrom has performed above and beyond duty in creating layout and composite work and correcting my feeble attempts in \LaTeX .

Very special thanks go to my wife Sylvia, my daughter Ursula S., and my son Karl H. for their loving support.

Karl Hess
University of Illinois at Urbana-Champaign

CHAPTER 1

A BRIEF REVIEW OF THE RELEVANT BASIC EQUATIONS OF PHYSICS

From a mathematical viewpoint, all equations of physics (both microscopic and macroscopic) are relevant for semiconductor devices. In an absolutely strict mathematical way, we therefore would have to proceed from the fundamentals of quantum field theory and write down the $\approx 10^{23}$ coupled equations for all the atoms in the semiconductor device. Then we would have to solve these equations, including the complicated geometrical boundary conditions. However, the outcome of such an attempt is clear to everyone who has tried to solve only one of the 10^{23} equations.

Any realistic approach oriented toward engineering applications has to proceed differently. Based on the experience and investigations of many excellent scientists in this field, we neglect effects that would only slightly influence the results. In this way many relativistic effects become irrelevant. In my experience, the spin of electrons plays a minor role in the theory of most current semiconductor devices and can be accounted for in a simple way (the correct inclusion of a factor of 2 in some equations).

Most effects of statistics can be understood classically, and we will need only a very limited amount of quantum statistical mechanics. This leaves us essentially with the Hamiltonian equations (classical mechanics), the Schrödinger equation (quantum effects), the Boltzmann equation (statistics), and the Maxwell equations (electromagnetics).

It is clear that the atoms that constitute a solid are coupled, and therefore the equations for the movement of atoms and electrons in a solid are coupled. This still presents a major problem, a many-body problem. We will see, however, that there are powerful methods to decouple the equations and therefore make single particle solutions possible. The many interacting electrons in a solid are then, for example, replaced by single independent electrons moving in a periodic potential. Complex many body effects, such as superconductivity, are then excluded from our treatment, which is justified because of the low electron density in typical semiconductors. We also exclude in our treatment effects of extremely high magnetic fields because these are unimportant for most device applications.

In this way, the fundamental laws of physics are finally reduced to laws of semiconductor devices that are tractable and whose limitations are clearly stated. The following sections are written with the intent to remind the reader of the basic physics underlying device operation and to review some of the physicist's tool kit in solid-state theory.

1.1 THE EQUATIONS OF CLASSICAL MECHANICS, APPLICATION TO LATTICE VIBRATIONS

Hamilton was able to give the laws of mechanics a very elegant and powerful form. He found that these laws can be closely linked to the sum of kinetic and potential energy written as a function of momentumlike (p_i) and spacelike (x_i) coordinates.

This function is now called the *Hamiltonian function* $H(p_i, x_i)$. The laws of mechanics are

$$\frac{dp_i}{dt} = -\frac{\partial H(p_i, x_i)}{\partial x_i} \quad (1.1)$$

and

$$\frac{dx_i}{dt} = \frac{\partial H(p_i, x_i)}{\partial p_i} \quad (1.2)$$

where t is time and $i = 1, 2, 3$. Instead of x_i , we sometimes denote the space coordinates by x, y, z .

Some simple special cases can be solved immediately. The free particle (potential energy = zero) moves according to

$$H = \sum_i p_i^2 / 2m$$

and we have from Eq. (1.1)

$$\frac{dp_i}{dt} = 0; \quad p_i = \text{constant},$$

which is Newton's first law of steady motion without forces.

If we have a potential energy $V(x_1)$ that varies in the x_1 direction, we obtain from Eq. (1.1)

$$\frac{dp_1}{dt} = -\frac{\partial V(x_1)}{\partial x_1} \equiv F_o \quad (1.3)$$

The quantity defined as F_o is the force, and Eq. (1.3) is Newton's second law of mechanics.

A more involved example of the power of Hamilton's equations is given by the derivation of the equations for the vibrations of the atoms (or ions) of the crystal lattice. As we will see, these vibrations are of utmost importance

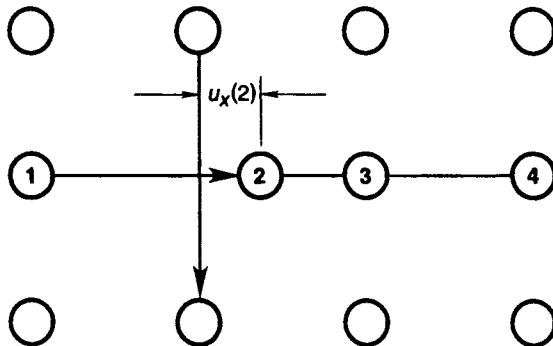


Figure 1.1 Displacement $u_i(r)$ of atoms in a crystal lattice.

in describing electrical resistance. They also give a fine example of how the many-body problem of atomic motion can be reduced to the solution of a single differential equation by using the crystal symmetry (group theory from a mathematics point of view) by a cut-off procedure. Here we cut off the interatomic forces beyond the nearest neighbor interaction. We also introduce below cyclic boundary conditions, which are of great importance and convenience in solid-state problems.

It suffices for this section to define a crystal, and we will be mostly interested in crystalline solids, as a regular array of atoms hooked together by atomic forces. "Regular" means that the distance between the atoms is the same throughout the structure. Many problems involving lattice vibrations can be solved by classical means (i.e., using the Hamiltonian equations) because the atoms that vibrate are very heavy. Then we only have to derive the kinetic and potential energy. Because we would like to describe vibrations (i.e., the displacement of the atoms), we express all quantities in terms of the atomic displacements $u_i(r)$, where $i = x, y, z$ and r is the number (identification) of the atom. It is important to note that r is not equal to the continuous space coordinate r in this chapter, although it has similar significance because it labels the atoms. The displacement of an atom in a set of regularly arranged atoms is shown in Figure 1.1.

We follow the derivations in Landsberg [5] and express the kinetic energy T by

$$T = \frac{1}{2} M \sum_n \dot{u}_i^2(r) \quad \text{where} \quad \dot{u}(r) = \frac{\partial u_i(r)}{\partial t} \quad (1.4)$$

and M is the mass of the atoms (ions).

We assume now that the total potential energy U of the atoms can be expressed in terms of a power series in the displacements,

$$U = U_0 + \sum_n B'_i u_i(r) + \frac{1}{2} \sum_{ij} B''_{ij} u_i(r) u_j(s) + \dots \quad (1.5)$$

where s also numbers the atoms as r does.

The following results rest on this series expansion and truncation, which makes a first principle derivation (involving many body effects) unnecessary.

Equation (1.5) is, of course, a Taylor expansion with

$$B_i^r = \frac{\partial U}{\partial u_i(r)} \quad (1.6)$$

Because the crystal is in equilibrium, that is, at a minimum of the potential energy U , the first derivative vanishes and

$$B_i^r = 0 \quad (1.7)$$

We further have

$$B_{ij}^{rs} = \frac{\partial^2 U}{\partial u_i(r) \partial u_j(s)} = B_{ji}^{sr} \quad (1.8)$$

We now use the fact that the crystal is translationally invariant—that is, we can shift the coordinate system by s atoms (start to count s atoms later), and the crystal is transformed into itself (at least if it is infinite). Therefore,

$$B_{ij}^{rs} = B_{ij}^{(r-s)0} \quad (1.9)$$

Furthermore, a rigid displacement (all u_i equal) of the crystal does not change U and we therefore have, from Eqs. (1.5) and (1.9),

$$\sum_r B_{ij}^{0r} = 0 \quad (1.10)$$

To derive Eq. (1.9), we have assumed an infinite crystal. We also could have introduced so-called periodic or cyclic boundary conditions; that is, continue the crystal by repeating it over and over. In one dimension, this means we consider only rings of atoms (Figure 1.2). This approach amounts to neglecting any surface effects or other effects that are sensitive to the finite extension of crystals.

We can now derive the equations of motion by using Eqs. (1.1) and (1.2) with coordinates $u_i(r)$ instead of x_i :

$$\dot{p}_i(r) = -\frac{\partial H(p_i, u_i)}{\partial u_i} \quad (1.11)$$

and

$$p_i(r) = M \dot{u}_i(r) \quad (1.12)$$

Eq. (1.11) gives

$$\dot{p}_i(r) = -\frac{1}{2} \frac{\partial}{\partial u_i(r)} \left[\sum_{mn} B_{ij}^{mn} u_i(m) u_j(n) \right] \quad (1.13)$$

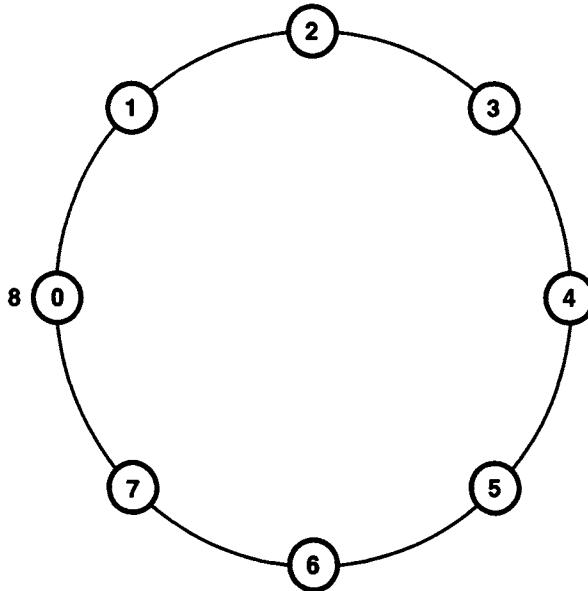


Figure 1.2 A ring of atoms representing cyclic boundary conditions.

Here, also, the indices m, n are used to number the atoms (as r, s above). Therefore

$$\dot{p}_i(r) = - \sum_{s,j} B_{ij}^{rs} u_j(s)$$

and, together with Eq. (1.12), one obtains

$$M\ddot{u}_i(r) + \sum_{s,j} B_{ij}^{rs} u_j(s) = 0 \quad (1.14)$$

Remember that the index s in Eq. (1.14) runs over a large number of atoms; that is, up to about 10^{23} in a typical crystal. The r can also assume any of these numbers. In other words, we have about 10^{23} coupled equations to solve. This situation is very typical for any type of solid-state problem, but by far not as hopeless as it may seem. Powerful methods have been developed to reduce the number of equations and the following treatment is representative. We will assume for simplicity that the crystal is one dimensional and avoid the complicated geometrical arrangement of atoms in a real crystal. (We will learn more about this when we discuss the electrons and their motion in crystals.)

In the three-dimensional case, Eqs. (1.7) through (1.10) are very helpful; they reduce the numbers of parameters. Without going into details, we mention that this reduction of parameters is generally accomplished by group theoretical arguments, and Eq. (1.9) is a direct consequence of the translational invariance (group of translations).

To proceed explicitly with our one-dimensional model, we need to make the drastic assumption that each atom interacts only with its nearest neighbor. (We can use the same method also for second, third ... nearest neighbor interaction, if we proceed numerically and use a high-speed computer.) Our assumption means

$$B^{r,s} \neq 0 \quad \text{only for } s = r \pm 1$$

Notice that we dropped the indices i, j because this is a one-dimensional problem. Without any loss of generality, we may assume $r = 0$. Then we have

$$B^{0s} \neq 0 \quad \text{for } s = \pm 1$$

and

$$B^{0s} = 0 \quad \text{otherwise} \quad (1.15)$$

Furthermore,

$$B^{01} = B^{-10}$$

according to Eq. (1.9) and

$$B^{-10} = B^{0(-1)}$$

according to Eq. (1.8). Therefore,

$$B^{01} = B^{0(-1)}$$

From Eq. (1.10) we obtain

$$B^{00} = -2B^{01} \quad (1.16)$$

It is now customary to denote $B^{01} = B^{0(-1)}$ by $-\alpha$ (α is the constant of the "spring" forces that hold the crystal together) and therefore B^{00} by 2α . The equation of motion, Eq. (1.14), then becomes for any r

$$M\ddot{u}(r) = -2\alpha u(r) + \alpha u(r-1) + \alpha u(r+1) \quad (1.17)$$

Eq. (1.17) leaves us still with 10^{23} coupled differential equations. However, these equations are now in *tridiagonal form*, all with coefficient α . Such a form can be reduced to one equation by skillful substitution. The substitution can be derived from Bloch's theorem, which we will discuss later. It also can be guessed:

$$u(r) = ue^{iqr a} \quad (1.18)$$

Note that the amplitude u is still a function of time. Here a is the distance between atoms (i.e., the lattice constant). Eq. (1.17) becomes

$$M\ddot{u}e^{iqr a} = -2\alpha ue^{iqr a} + \alpha ue^{iqr a} e^{-iqa} + \alpha ue^{iqr a} e^{iqa} \quad (1.19)$$

which gives

$$M\ddot{u} = \alpha u(2 \cos qa - 2)$$

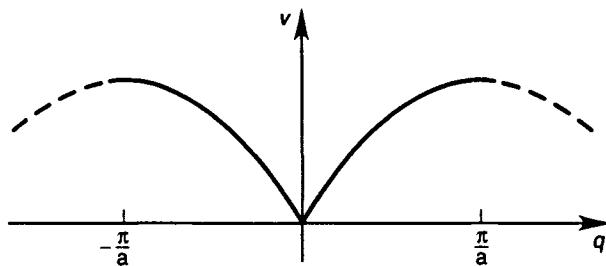


Figure 1.3 Dispersion relation $v(q)$ for lattice vibrations. (Remember, $E = hv$.)

and

$$\ddot{u} = \frac{-4\alpha u}{M} \sin^2\left(\frac{aq}{2}\right)$$

This gives

$$\ddot{u} + v^2 u = 0 \quad (1.20)$$

with

$$v = 2 \left| \sin \frac{aq}{2} \right| \sqrt{\frac{\alpha}{M}} \quad (1.21)$$

This means that the atoms are oscillating in time with frequency v , which is a function of the wave vector q . The function is shown in Figure 1.3.

There are several important points to notice. First, at $q = -\pi/a$ and $q = \pi/a$, the energy has its highest value. For these q , the wavelength $\lambda = 2\pi/q$ has the value $\lambda = 2a$. As can be seen from Figure 1.4, this is the shortest wavelength that we really need to describe the physics of the lattice vibrations. Shorter wavelengths lead only to “wiggles” between the atoms, but the displacements are actually the same. For example, if $q = 3\pi/a$ and $\lambda = 2/3a$, the atoms are displaced in exactly the same way as for $q = \pi/a$. In other words, for any q outside the zone $-\pi/a \leq q \leq \pi/a$, which is called the *Brillouin zone*, we can find a q inside the zone that describes the same displacement, energy, and so

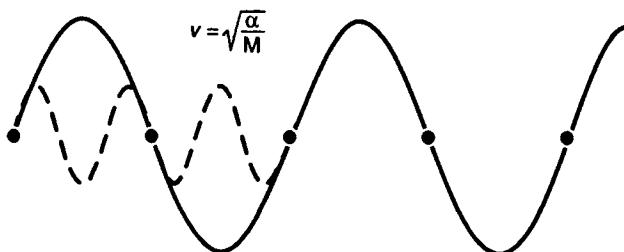


Figure 1.4 Illustration of the shortest possible physical wavelength of lattice vibrations.

on. Notice that in a real crystal the arrangement of atoms is different in different directions. Therefore, the three-dimensional Brillouin zone is usually a complicated geometrical figure (see Chapter 2).

Second, q is not a continuous variable because of the boundary conditions. Consider, for example, the ring of Figure 1.2 with eight atoms and

$$u(0) = u(8)$$

or, in general, for N atoms, we have

$$u(N) = u(0) \quad \text{and} \quad u(N) = u(0)e^{iqNa}$$

Therefore e^{iqNa} equals one, and we conclude that $q = 2\pi l/Na$ where l is an integer. If we restrict q to the first Brillouin zone, we have $-N/2 \leq l \leq N/2$. This means that q assumes only discrete (not continuous) values. However, because of the large number N , it can almost be regarded as continuous.

Third, without emphasizing it, we have developed a microscopic theory of sound propagation in solids. For small wave vectors q (i.e., large λ), we have

$$\sin \frac{qa}{2} \approx \frac{qa}{2}$$

and

$$v = \sqrt{\frac{\alpha}{M}} qa \quad (1.22)$$

Using $\lambda v = v_s$, where v_s is the velocity of sound, we obtain

$$v_s = 2\pi a \sqrt{\frac{\alpha}{M}} \quad (1.23)$$

which is a microscopic description of the sound velocity.

In real crystals additional complications arise from the fact that we can have two or even more different kinds of atoms. These atoms may oscillate as the identical atoms in the above example. There are, however, different modes of oscillation possible. If we think of a chain with two different kinds of atoms, it can happen that one kind of atom (black) oscillates against the other kind (white).

Such an oscillation can take place, and indeed does, at a very high (optical) frequency, and the corresponding lattice vibrations are called *optical phonons*. It is very important to note that in principle all black atoms can oscillate in phase against the white ones. This means that we can have high frequencies (energies) even if the wavelength is very large or the q vector is very small (Figure 1.5).



Figure 1.5 Two different kinds of atoms oscillating against each other. This represents a wave with high energy (frequency) and small wave vector.

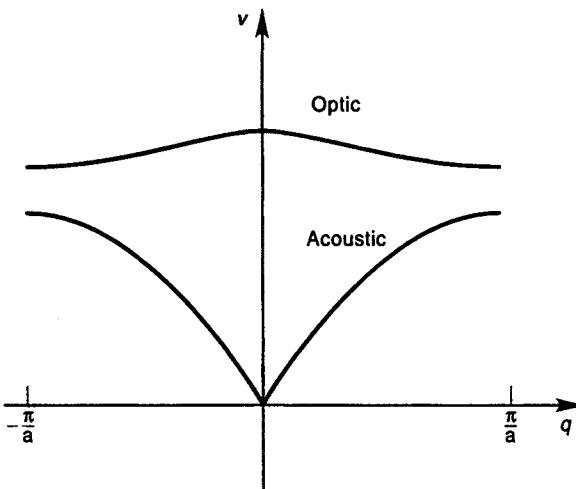


Figure 1.6 Schematic $v(q)$ diagram for acoustic and optic phonons in one dimension.

The energy versus q relation can then have two branches, the acoustic and the optic, as shown in Figure 1.6.

The presence of two different atoms can also cause long-range coulombic forces owing to the different charge on the two atom types (ionic component). The long-range forces cannot be described by simple forces between neighboring atoms, and one calls the phonons *polar optical phonons* if these long-range forces are important.

As mentioned, lattice vibrations are important in various ways. Electrons interact with the crystal lattice exciting (emitting) and absorbing lattice vibrations (the net lost energy is known as Joules heat). The system of electrons by itself is therefore not a Hamiltonian system; that is, one in which energy is conserved. It is only the sum of electrons and lattice vibrations which is Hamiltonian.

The interested reader is encouraged to obtain knowledge of a detailed quantum picture of lattice vibrations (also phonons) and their interactions with electrons as described, for example, by Landsberg [5].

1.2 THE EQUATIONS OF QUANTUM MECHANICS

At the beginning of the twentieth century, scientists realized that nature cannot be strictly divided into waves and particles. They found that light has particle-like properties and cannot always be viewed as a wave, and particles such as electrons revealed definite wave-like behavior under certain circumstances. They are, for example, diffracted by gratings as if they had a wavelength

$$\lambda = \frac{h}{|\mathbf{p}|} \quad (1.24)$$

where $\hbar \equiv h/2\pi \approx 6.58 \times 10^{-16}$ eVs is Planck's constant and \mathbf{p} is the electron momentum.

Schrödinger demonstrated that the mechanics of atoms can be understood as boundary value problems. In his theory, electrons are represented by a wave function $\psi(\mathbf{r})$, which can have real and imaginary parts, and follows an eigenvalue differential equation:

$$\left(-\frac{\hbar^2}{2m} \nabla^2 + V(r) \right) \psi(r) = E\psi(r) \quad (1.25)$$

The part of the left side of Eq. (1.25) that operates on ψ is now called the Hamiltonian operator H . Formally this operator is obtained from the classical Hamiltonian by replacing momentum with the operator $\nabla\hbar/i$ (i = imaginary unit), where

$$\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right)$$

The meaning of the wave function $\psi(\mathbf{r})$ was not clearly understood at the time Schrödinger derived his famous equation. It is now agreed that $|\psi(\mathbf{r})|^2$ is the probability of finding an electron in a volume element $d\mathbf{r}$ at \mathbf{r} . In other words, we have to think of the electron as a point charge with a statistical interpretation of its whereabouts (the wave-like nature). It is usually difficult to get a deeper understanding of this viewpoint of nature; even Einstein had trouble with it. It is, however, a very successful viewpoint that describes exactly all phenomena we are interested in. To obtain a better feeling for the significance of $\psi(\mathbf{r})$, we will solve Eq. (1.25) for several special cases. As in the classical case, the simplest solution is obtained for constant potential. Choosing an appropriate energy scale, we put $V(\mathbf{r}) = 0$ everywhere.

By inspection we can see that the function

$$C \exp(i\mathbf{k} \cdot \mathbf{r}) = C(\cos \mathbf{k} \cdot \mathbf{r} + i \sin \mathbf{k} \cdot \mathbf{r}) \quad (1.26)$$

is a solution of Eq. (1.25) with

$$\frac{\hbar^2 k^2}{2m} = E \quad (1.27)$$

and C a constant.

The significance of the vector \mathbf{k} can be understood from analogies to well-known wave phenomena in optics and from the classical equations. Because E is the kinetic energy, $\hbar\mathbf{k}$ has to be equal to the classical momentum \mathbf{p} to satisfy $E = p^2/2m$. On the other hand, in optics

$$|\mathbf{k}| = 2\pi/\lambda \quad (1.28)$$

which gives, together with Eq. (1.24),

$$\hbar\mathbf{k} = \mathbf{p}$$

which is consistent with the mechanical result.

How can the result of Eq. (1.26) be understood in terms of the statistical interpretation of $\psi(\mathbf{r})$? Apparently

$$|\psi(\mathbf{r})|^2 = |C|^2(\cos^2 \mathbf{k} \cdot \mathbf{r} + \sin^2 \mathbf{k} \cdot \mathbf{r}) = |C|^2$$

This means that the probability of finding the electron at any place is equal to C^2 . If we know that the electron has to be in a certain volume V_{ol} (e.g., of a crystal), then the probability of finding the electron in the crystal must be one. Therefore,

$$\int_{V_{\text{ol}}} |C|^2 d\mathbf{r} = V_{\text{ol}} |C|^2 = 1$$

and

$$|C| = 1/\sqrt{V_{\text{ol}}} \quad (1.29)$$

In other words, the probability of finding an electron with momentum $\hbar\mathbf{k}$ at a certain point \mathbf{r} is the same in the whole volume and equals $1/V_{\text{ol}}$. We will give a more detailed discussion of this somewhat peculiar result in the next section. The unfamiliar reader is referred to an introductory text (e.g., Feynman [3]).

Note that by confining the electron to a volume, we have already contradicted our assumption of constant potential $V(\mathbf{r}) = 0$. (Electrons can only be confined in potential wells.) If, however, the volume is large, our mistake is insignificant for many purposes.

Let us now consider the confinement of an electron in a one-dimensional potential well (although such a thing does not exist in nature). We assume that the potential energy $V(\mathbf{r})$ is zero over the distance $(0, L)$ on the x -axis and infinite at the boundaries 0 and L .

The Schrödinger equation, Eq. (1.25), reads in one dimension (x -direction, $V(x) = 0$)

$$-\frac{\hbar^2}{2m} \frac{\partial^2 \psi(x)}{\partial x^2} = E\psi(x) \quad (1.30)$$

Inspection shows that the function

$$\psi(x) = \sqrt{\frac{2}{L}} \sin \frac{n\pi}{L} x \quad \text{with} \quad n = 1, 2, 3, \dots \quad (1.31)$$

satisfies Eq. (1.30) as well as the boundary conditions. The boundary conditions are, of course, that ψ vanishes outside the walls, since we assumed an infinite impenetrable potential barrier. In the case of a finite potential well, the wave function penetrates into the boundary and the solution is more complicated. If the barrier has a finite width, the electron can even leak out of the well (tunnel). This is a very important quantum phenomenon the reader should be familiar with. We will return to the tunneling effect below.

The wave function, Eq. (1.31), corresponds to energies E (called *eigen energies*)

$$E = \frac{n^2 \pi^2 \hbar^2}{2mL^2} \quad (1.32)$$

Because n is an integer, the electron can assume only certain discrete energies while other energies are not allowed. These discrete energies that can be assumed are called *quantum states* and are characterized by the *quantum number* n . The wave function and corresponding energy are therefore also denoted by ψ_n , E_n .

Think of a violin string vibrating in various modes at higher and lower tones (frequency v), depending on the length L , and consider Einstein's law:

$$E = hv \quad (1.33)$$

If we compare the modes of vibration of the string with the form of the wave function for various n , then we can appreciate the title of Schrödinger's paper, "Quantization as a Boundary Value Problem."

Devices that contain a well and feature quantized energy levels similar to the ones given in Eq. (1.32) do exist. Quantum well lasers typically contain one or more small wells and the well size controls the electron energy. However, in most devices, the wells are not rectangular. In silicon metal oxide semiconductor field effect transistors (MOSFETs), the well is closer to triangular and its shape depends on the electron density (i.e., the charge in the well). We will deal with this charge-dependent well shape in Chapter 10. Here we discuss only well-defined potential problems—cases where the potential is given and fixed. With current high-end workstations, the Schrödinger equation can then be solved numerically for an arbitrary (but given) potential shape. One-dimensional problems can be solved by standard discretization (transforming the differentiations into finite differences) and by solving the resulting matrix equations by standard solvers such as found in EISPACK and LAPACK. For two- and three-dimensional problems, this procedure still leads to a prohibitively large number of equations (which grows with the third power of the number of discretization points). Therefore the discretized mesh must be coarsened even when using the fastest supercomputers. Often, however, one is interested only in relatively small sets of eigen values, for example, the first three [as for $n = 1, 2, 3$ in Eq. (1.31)].

One then can use so-called subspace interaction techniques that only resolve certain intervals of eigenvalues. These techniques are well established for symmetric real matrices as they occur in well-defined potential problems (see, e.g., Golub and Loan [4]). A useful computer code is the RITZED eigenvalue solver by Rutishauser [6].

Frequently, one needs to obtain an explicit solution of Schrödinger's equation for an arbitrary complicated form of the potential, provided only that it is small and represents just a small perturbation to a problem for which the solution is known. This scenario is typical for scattering problems such as an electron propagating in a perfect solid and then encountering a small imperfection and being scattered. Fortunately, for this type of problem there is a powerful method of approximation, perturbation theory, that gives us the solution for arbitrary weak potentials. The method is very general and applies to any kind of equation.

Consider an equation of the form

$$(H_0 + \epsilon H_1)\psi = 0 \quad (1.34)$$

where H_0 and H_1 are differential operators of arbitrary complication and ϵ is a small positive number.

If we know the solution ψ_0 of the equation

$$H_0\psi_0 = 0$$

then we can assume that the solution of Eq. (1.34) has the form $\psi_0 + \epsilon\psi_1$. Inserting this form into Eq. (1.34), we obtain

$$(H_0 + \epsilon H_1)(\psi_0 + \epsilon\psi_1) = H_0\psi_0 + \epsilon H_1\psi_0 + H_0\epsilon\psi_1 + \epsilon^2 H_1\psi_1$$

We now can neglect the term proportional to ϵ^2 (because ϵ is small), and because $H_0\psi_0 = 0$, we have

$$H_1\psi_0 + H_0\psi_1 = 0 \quad (1.35)$$

This equation is now considerably simpler than Eq. (1.34) because ψ_0 is known. Therefore, ψ_1 can be determined easily if H_0 has a simple form no matter how complicated H_1 is. Repeated application of this principle leads to perturbation theory including higher orders ($\epsilon^2, \epsilon^3, \dots$). The derivation is given in many textbooks on quantum mechanics (see Baym [1]). Here we quote only the result that is used at several occasions.

Assume that we know the solutions of a Schrödinger equation:

$$H_0\psi_n = E_n\psi_n \quad n = 1, 2, 3, \dots \quad (1.36)$$

and we would like to know the solutions of

$$(H_0 + H_1)\phi_m = W_m\phi_m \quad \text{with} \quad H_1 \ll H_0 \quad (1.37)$$

Then it is shown in elementary texts on quantum mechanics (Baym [1]), by repeatedly using the method of perturbation theory as outlined, that

$$W_m = E_m + M_{mm} + \sum_{n \neq m} \frac{|M_{mn}|^2}{E_m - E_n} \quad (1.38)$$

with

$$\phi_m = \psi_m + \sum_{n \neq m} \frac{M_{mn}}{E_m - E_n} \psi_n \quad (1.39)$$

and

$$M_{mn} = \int_{V_{\text{ol}}} \psi_n^* H_1 \psi_m d\mathbf{r} \quad (1.40)$$

where $d\mathbf{r}$ stands for $dx dy dz$ (integration over volume V_{ol}) and ψ_n^* is the complex conjugate of ψ_n .

First-order perturbation theory (to order ϵ) amounts to setting $\psi_m = \Phi_m$ and $W_m = E_m + M_{mm}$. The only change then is in the value of the eigen energy by M_{mm} , which can be obtained by the integration in Eq. (1.40); the integrand is known from the solution of Eq. (1.36). This means that the numerical problem

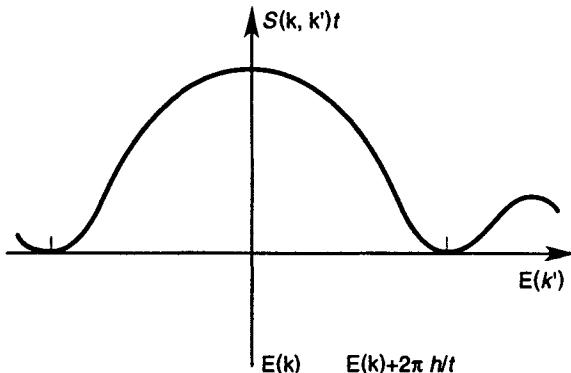


Figure 1.7 Probability of a transition from \mathbf{k} to \mathbf{k}' according to the Golden Rule after the potential has been on for time t .

is reduced to a volume integration (in three dimensions). To obtain solutions to higher order, one also needs to perform the summations in Eqs. (1.38) and (1.39).

The formalism outlined above and the examples given are independent of time, and the electrons are perpetually in appropriate (eigen) states. In many instances, however, we will be interested in the following type of problem: The electron initially is in an eigenstate of H_0 , denoted, for example, by a wave vector \mathbf{k} for the free electron. What is the probability that the electron will be observed in a different eigenstate characterized by the wave vector \mathbf{k}' , after it interacts with a potential $V(\mathbf{r}, t)$? In other words, what is the probability $S(\mathbf{k}, \mathbf{k}')$ per unit time that the interaction causes the system to make a transition from \mathbf{k} to \mathbf{k}' ?

The answer to this question is the famous Golden Rule of Fermi, which is also derived in almost every text on quantum mechanics by so-called time-dependent perturbation theory. The unfamiliar reader is urged to acquire a detailed understanding of the Golden Rule as derived, for example, in the text of Baym [1]. Here we only illustrate its generality and discuss results for important special cases.

1. Assume that a potential $V(\mathbf{r})$ is switched on at time $t = 0$ but is time independent otherwise. One then obtains

$$S(\mathbf{k}, \mathbf{k}') = \left| \int_{V_{\text{ol}}} \psi_{\mathbf{k}}^* V(\mathbf{r}) \psi_{\mathbf{k}'} d\mathbf{r} \right|^2 \cdot \left[\frac{\sin((E(\mathbf{k}') - E(\mathbf{k}))t/2\hbar)}{(E(\mathbf{k}') - E(\mathbf{k}))\sqrt{t/2}} \right]^2 \quad (1.41)$$

The function in brackets deserves special attention and is plotted in Figure 1.7. Notice that as t approaches infinity, the function plotted in Figure 1.7 becomes more and more peaked at its center ($E(\mathbf{k}') = E(\mathbf{k})$). In the limit $t \rightarrow \infty$, the so-called δ -function is approached, which is defined by

$$\lim_{t \rightarrow \infty} \frac{4 \sin^2[(E(\mathbf{k}') - E(\mathbf{k}))t/(2\hbar)]}{(E(\mathbf{k}') - E(\mathbf{k}))^2 t} = \frac{2\pi}{\hbar} \delta(E(\mathbf{k}') - E(\mathbf{k})) \quad (1.42)$$

and can always be understood as a limit of ordinary functions. It does have some remarkable properties, however, and the unfamiliar reader should consult some of the references at the end of this section. A most important property of the δ -function is the following: For any continuous function $f(E')$, we have

$$\int_{-\infty}^{\infty} f(E') \delta(E - E') dE' = f(E) \quad (1.43)$$

2. We assume that the perturbation is harmonic, which means we have a potential of the form

$$V(\mathbf{r}, t) = V(\mathbf{r})(e^{-i\omega t} + e^{i\omega t})$$

For $t \rightarrow \infty$, we obtain the transition probability

$$S(\mathbf{k}, \mathbf{k}') = \frac{2\pi}{\hbar} \left| \int_{V_{ol}} \psi_{\mathbf{k}'}^* V(\mathbf{r}) \psi_{\mathbf{k}} d\mathbf{r} \right|^2 [\delta(E(\mathbf{k}) - E(\mathbf{k}') - \hbar\omega) + \delta(E(\mathbf{k}) - E(\mathbf{k}') + \hbar\omega)] \quad (1.44)$$

It is clear that the δ -function simply takes care of energy conservation. For a constant potential, we have to conserve energy as t increases. For a harmonic perturbation, the system can gain or loose energy corresponding, for example, to the absorption or emission of light.

3. We now turn our attention to the first term in Eq. (1.41), the matrix element, which also plays a vital role in time-independent perturbation theory. The significance of the matrix element is best illustrated by the following special cases of well-defined potential problems (problems in which the potential is given by a certain function of coordinates):

- (a) $V(\mathbf{r}) = \text{constant}$. The matrix element is then

$$\text{constant} \frac{1}{V_{ol}} \int_{V_{ol}} e^{-i\mathbf{k}' \cdot \mathbf{r}} e^{i\mathbf{k} \cdot \mathbf{r}} d\mathbf{r} \quad (1.45)$$

The integration is over the volume V_{ol} of the crystal. In many practical cases, this volume will be much larger than the de Broglie wavelength λ of the electron, which is of the order of 100 Å in typical semiconductor problems. This means that the integral of Eq. (1.45) will be very close to zero, because the cosine and sine functions to which the exponents in Eq. (1.45) are equivalent are positive as often as they are negative in the big volume. There is only one exception: In the case $\mathbf{k}' = \mathbf{k}$, the integral is equal to the volume and the matrix element is equal to constant. Therefore, we can write

$$\text{constant} \frac{1}{V_{ol}} \int_{V_{ol}} e^{-i\mathbf{k}' \cdot \mathbf{r}} e^{i\mathbf{k} \cdot \mathbf{r}} d\mathbf{r} = \text{constant} \delta_{\mathbf{k}', \mathbf{k}} \quad (1.46)$$

where $\delta_{\mathbf{k}', \mathbf{k}} = 1$ for $\mathbf{k} = \mathbf{k}'$ and is zero otherwise. This is known as the *Kronecker delta symbol*. Consequently, the matrix element has

taken care of momentum conservation; the free electron in a constant potential does not change its momentum.

- (b) Second, we consider an arbitrary potential having the following Fourier representation:

$$V(\mathbf{r}) = \sum_{\mathbf{q}} V_{\mathbf{q}} e^{i\mathbf{q} \cdot \mathbf{r}} \quad (1.47)$$

Then the matrix element, which we now denote by $M_{\mathbf{k},\mathbf{k}'}$, becomes

$$\begin{aligned} M_{\mathbf{k},\mathbf{k}'} &= \sum_{\mathbf{q}} \frac{V_{\mathbf{q}}}{V_{\text{ol}}} \int_{V_{\text{ol}}} e^{i(\mathbf{k}-\mathbf{k}'+\mathbf{q}) \cdot \mathbf{r}} \\ &= \sum_{\mathbf{q}} V_{\mathbf{q}} \delta_{\mathbf{k}'-\mathbf{k},\mathbf{q}} \end{aligned} \quad (1.48)$$

How do we interpret this result? If we also allow the potential of Eq. (1.47) to have a time dependence (e.g., as $e^{i\omega t}$), then the potential can be interpreted as that of a wave (e.g., an electromagnetic wave). In this case Eq. (1.49) simply tells us that the wave vectors of all scattering agents (i.e., their momenta) are conserved, because we have

$$\mathbf{k}' - \mathbf{k} = \mathbf{q} \quad (1.49)$$

It is important to notice that Eq. (1.49) is also valid for a static, time-independent potential—that is, even a static potential “supplies” momentum according to its Fourier components—in the same way a wave does. This seems strange at first glance. To see the significance, consider the boundary of a billiard table. This boundary is an impenetrable abrupt potential step whose Fourier decomposition involves all values of \mathbf{q} . Indeed, the boundary can supply any momentum to the ball to make it bounce back. The above two examples show that the Golden Rule essentially takes care of energy and momentum conservation. This is also the reason for its generality and importance. Remember, however, that this is true only for cases when time t at which we observe the scattered particle is long after the potential is switched on. For short times (in practice these are times of the order of 10^{-14} s), the function in Eq. (1.42) cannot be approximated by a δ function, and energy need not be conserved in processes on this short time scale. This is at the heart of the energy time uncertainty relation. To illustrate the great generality of the Golden Rule, one more example is given.

- (c) Consider the “tunneling problem” of Figure 1.8. Although an electric field F is applied in the z -direction, the electron in Figure 1.8 is confined in a small well. Classically, it would stay in the well. However, because the barrier is not infinite, as assumed in Eq. (1.31), the wave function is not zero at the well boundary but penetrates the

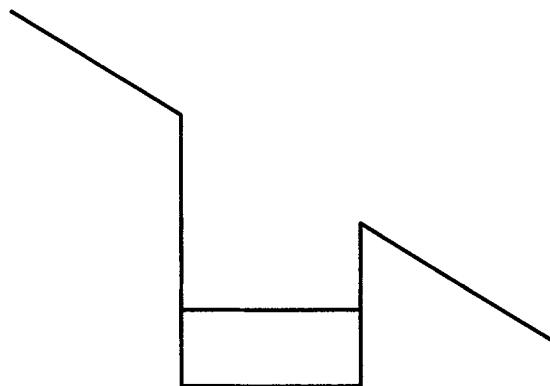


Figure 1.8 Electrons in a potential well plus applied electric field F .

boundary. In other words, there is a finite probability of finding the electron outside the well.

We can calculate the probability per unit time that the electron leaks out if we know the wave function ψ_{in} in the well and ψ_{ou} outside the well, and regard the electric field as a perturbation. This perturbation gives a term (the potential energy) eFz in the Hamiltonian. The Golden Rule tells us that

$$S(w, ou) = \left| \int_{V_{ol}} \psi_{ou}^* eFz \psi_{in} d\mathbf{r} \right| \cdot \frac{2\pi}{\hbar} \delta(E_{ou} - E_{in}) \quad (1.50)$$

In writing down this equation (which was first derived by Oppenheimer), I have swept under the rug the fact that ψ_{ou} and ψ_{in} are the solutions of different Hamiltonians. ψ_{in} is obtained from the solution of the Schrödinger equation of the quantum well and ψ_{ou} is the solution of a free electron in an electric field with

$$H = -\frac{\hbar^2 \nabla^2}{2m} - eFz$$

An exact justification of this procedure is complicated and is discussed in great detail in Duke's treatise of tunneling [2] (see also Appendix A).

We emphasize that the matrix elements represent all that needs to be known to obtain perturbation theory solutions. These matrix elements are given by three dimensional integrals. Alternatively they can be viewed as scalar products in a vector space denoting ψ_n by a vector $|n\rangle$. For those unfamiliar with Dirac's notation the following definition can just be used as a shorthand way of writing the integral:

$$\langle m | H_1 | n \rangle = \int_{V_{ol}} \psi_m^* H_1 \psi_n d\mathbf{r}. \quad (1.51)$$

PROBLEMS

1.1 Solve by perturbation theory (to first order) for y :

$$\epsilon \frac{\partial y \sin(x)}{\partial x} + \frac{\partial y}{\partial x} = y$$

where ϵ is a small positive quantity.

1.2 Calculate the matrix elements for wave functions of the form $\psi = \frac{1}{\sqrt{V_{\text{ol}}}} e^{i\mathbf{k}\cdot\mathbf{r}}$ and

(a) $H_1 \propto e^{i\mathbf{q}\cdot\mathbf{r}}$

(b) $H_1 \propto \delta(\mathbf{r})$

(c) $H_1 \propto |\mathbf{r}|^{-2}$

(d) $H_1 \propto \exp\left\{\frac{-|\mathbf{r}|}{r_0}\right\}$ $r_0 > 0$ Polar coordinates are helpful in parts c and d.

1.3 Consider a one-dimensional crystal lattice with two ions (atoms) repeated in a circular arrangement. The two ions (atoms) are identical, with mass M , but are connected by springs of alternating strength (D_1, D_2).

(a) Derive the equations of motion. (Consider only nearest neighbor interactions, where the force is proportional to the difference in displacements.)

(b) Find and sketch the dispersion relation of the possible vibrational modes. (Assume all displacements are traveling waves with sinusoidal time dependence, that is, $u_i(ra) = \epsilon_i e^{i(qra - \omega t)}$.)

(c) Discuss the form of the dispersion relation and the nature of the modes for $q \ll \pi/a$ and $q = \pi/a$, where q is the wave vector.

(d) Find the velocity of sound (ω/q for $q \rightarrow 0$).

(e) Show that the group velocity $\partial\omega/\partial q$ becomes zero at the Brillouin zone boundary. (This is a general result.)

REFERENCES

- [1] Baym, G. *Lectures on Quantum Mechanics*. New York: Benjamin, 1969, p. 248.
- [2] Duke, C. B. *Tunneling in Solids*. New York: Academic Press, 1969, p. 207.
- [3] Feynman, R. P. *Lectures on Physics*, Vol. III. Reading, MA: Addison-Wesley, 1964, pp. 1.1–4.15.
- [4] Golub, G. H., and Loan, C. F. *Matrix Computation*. John Hopkins Univ. Press: Baltimore, 1989.
- [5] Landsberg, P. T. *Solid State Theory*. New York: Wiley/Interscience, 1969, pp. 327–56.
- [6] Rutishauser, H. “Numerical eigenvalue solver,” *Numerical Mathematics*, vol. 13, 1969, p. 4.

CHAPTER 2

THE SYMMETRY OF THE CRYSTAL LATTICE

This chapter and the next three are a crash course in the solid-state physics underlying semiconductor devices. They can be read as a reminder to the reader of the most important solid-state physics principles that we will need later. They are also written in a way to enable any novice to gain the necessary solid-state knowledge. However, this cannot be accomplished by casual reading, but only by going over the material with a pen.

2.1 CRYSTAL STRUCTURES OF SILICON AND GaAs

In Chapter 1 we discussed the lattice vibrations without defining exactly what a crystal lattice is. Here we give this definition and we see that a crystal is an object of high symmetry. This symmetry can be used to obtain general information about the properties of crystals and also to abbreviate complicated algebra. Full use of the symmetry requires knowledge of group theory. This knowledge is not required for the reader of this book. Nevertheless, an attempt is made here to introduce group theoretical techniques via practical examples.

A crystal consists of a *basis* and a Bravais lattice. The basis can be anything ranging from atoms to giant molecules, such as deoxyribonucleic acid (DNA). The Bravais lattice is a set of points $\{\mathbf{R}_l\}$ that is generated by three non-coplanar translations $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$, which are vectors of three-dimensional space.

$$\mathbf{R}_l = l_1 \mathbf{a}_1 + l_2 \mathbf{a}_2 + l_3 \mathbf{a}_3 \quad (2.1)$$

and the l_i are integers.

According to the properties of this lattice, under reflection, rotation, and so on, one can distinguish 14 types of Bravais lattices. For us, only the cubic types matter (Figure 2.1). The important semiconductors are characterized by a *tetrahedral arrangement* of the nearest neighbor atoms. Their lattice can be viewed as a face-centered cubic lattice with a basis of two atoms. For silicon and

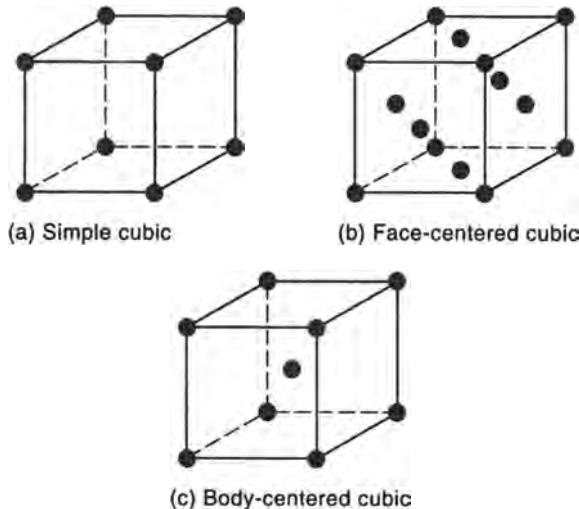


Figure 2.1 The three types of cubic Bravais lattices.

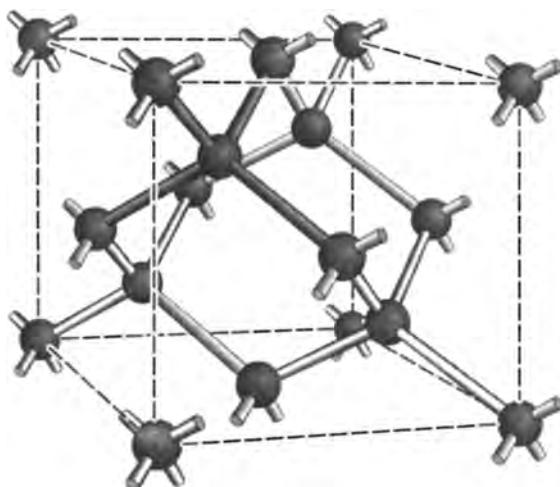


Figure 2.2 Crystal structure of silicon (or GaAs if two kinds of atoms are on the appropriate lattice sites). Notice the tetrahedral arrangement of nearest neighbor atoms and the equivalence to a face-centered cubic lattice (if a two-atom basis is assumed).

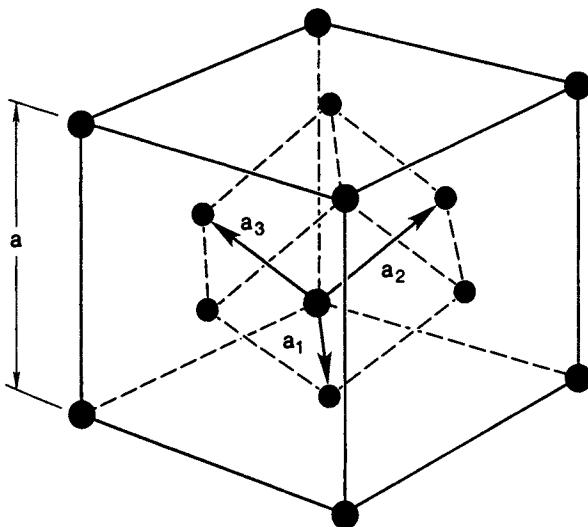


Figure 2.3 Vectors \mathbf{a}_1 , \mathbf{a}_2 , \mathbf{a}_3 generating the face-centered cubic lattice.

germanium, these two atoms are equal; for GaAs and the III-V compounds, the two atoms are different. This is illustrated in Figure 2.2.

Can one view the lattice of silicon under all circumstances as a face-centered cubic crystal with a basis of two atoms? In principle, the symmetry of one face-centered cubic crystal is present. For some effects, however, the existence of the two basis atoms is vital. Consider, for example, lattice vibrations. It is clear that the two basis atoms are connected by different force (spring) constants α than two atoms on the side of the cube (see Figure 2.2). Therefore, optical phonons will exist (see Problem 1.3). Instead of two different kinds of atoms vibrating against each other as in the problem, the two sublattices associated with the two basis atoms can vibrate against each other. By two *sublattices*, we mean that we can also view the silicon crystal as two interconnected face-centered cubic lattices (sublattices), each having one basis atom.

The three translation vectors \mathbf{a}_1 , \mathbf{a}_2 , \mathbf{a}_3 that generate a cubic face-centered lattice are shown in Figure 2.3. It is important to note that these vectors are different from the vectors that generate the simple cubic lattice. If these vectors (along the sides of the cube) were chosen to generate the lattice points \mathbf{R}_l of the face-centered lattice, some points could not be reached with integer values for l_i .

Figures 2.4a and 2.4b are photographs of a gallium-arsenide crystal model in the [110] and [100] directions. They illustrate the anisotropy of a crystal. In other words, lattice waves or electrons traveling in different directions encounter, in general, different patterns (and therefore a different Brillouin zone boundary).

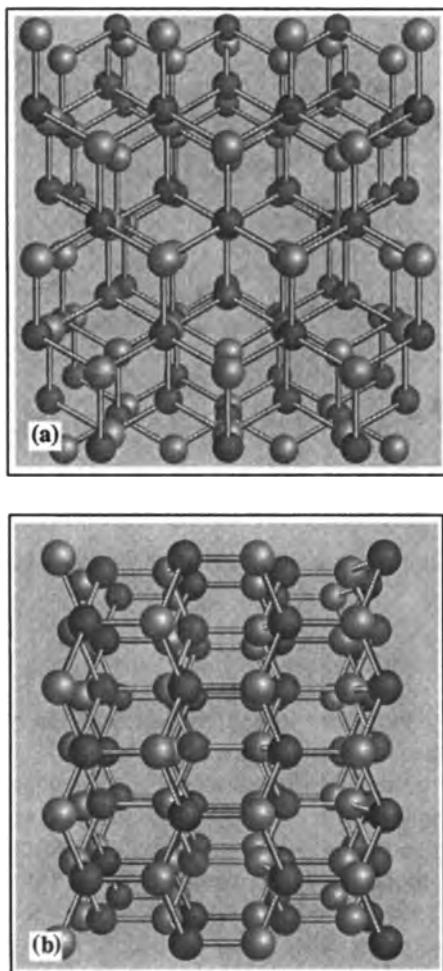


Figure 2.4 Illustrations of a GaAs crystal model in the (a) [111] and (b) less regular crystallographic directions.

2.2 ELEMENTS OF GROUP THEORY

2.2.1 Point Group

From a more mathematical viewpoint, it is important to note that crystal lattices—and with them all their physical properties—are transformed into themselves by certain geometrical operations (rotations, reflections, etc.). The set of all these operations is called the point group of the crystal lattice. Forty-eight such operations (translations are excluded for the moment) transform a cube into itself. The group of these 48 operations is called O_h . Twenty-four of the 48 operations that form the subgroup T_d are shown in Table 2.1. In this table the

Table 2.1 Elements of the Point Group T_d .

$Q_1(x_1x_2x_3)$	$Q_2(x_1\bar{x}_2\bar{x}_3)$	$Q_3(\bar{x}_1x_2\bar{x}_3)$	$Q_4(\bar{x}_1\bar{x}_2x_3)$
$Q_5(x_2x_3x_1)$	$Q_6(\bar{x}_2x_3\bar{x}_1)$	$Q_7(\bar{x}_2\bar{x}_3x_1)$	$Q_8(x_2\bar{x}_3\bar{x}_1)$
$Q_9(x_3x_1x_2)$	$Q_{10}(\bar{x}_3\bar{x}_1x_2)$	$Q_{11}(x_3\bar{x}_1\bar{x}_2)$	$Q_{12}(\bar{x}_3x_1\bar{x}_2)$
$Q_{13}(\bar{x}_1x_3\bar{x}_2)$	$Q_{14}(\bar{x}_1\bar{x}_3x_2)$	$Q_{15}(\bar{x}_3\bar{x}_2x_1)$	$Q_{16}(x_3\bar{x}_2\bar{x}_1)$
$Q_{17}(x_2\bar{x}_1\bar{x}_3)$	$Q_{18}(\bar{x}_2x_1\bar{x}_3)$	$Q_{19}(x_1x_3x_2)$	$Q_{20}(x_1\bar{x}_3\bar{x}_2)$
$Q_{21}(x_3x_2x_1)$	$Q_{22}(\bar{x}_3x_2\bar{x}_1)$	$Q_{23}(x_2x_1x_3)$	$Q_{24}(\bar{x}_2\bar{x}_1x_3)$

Source: After Morgan, D. J., in Landsberg, P. T., ed., *Solid State Theory: Methods and Applications*, Table C.7.1. Copyright 1969 John Wiley & Sons, Ltd. Reprinted by permission.

operation $Q_3(\bar{x}_1, x_2, \bar{x}_3)$ means, for example,

$$Q_3 f(x_1, x_2, x_3) = f(-x_1, x_2, -x_3)$$

for any function f of the coordinates.

Operating on these 24 transformations with the inversion $Q_0 f(\bar{x}_1, \bar{x}_2, \bar{x}_3)$ gives another 24 symmetry operations, which, together with the operations in Table 2.1, form the 48 operations of O_h .

If $f(x_1, x_2, x_3)$ is a physical property of the crystal, and the crystal has the symmetry O_h , then we obtain the value of f at many other points simply by applying the symmetry operations. This can save much computation time. (We explain the use of this for band-structure calculations in Chapter 3.)

The following example also gives a clear illustration of the advantageous use of the symmetry operations. Bardeen, Schrieffer, and Stern realized that electrons can form a two-dimensional gas at the interface between Si and SiO_2 , in a metal–oxide semiconductor (MOS) transistor. Figure 2.5 shows the basic geometry of a MOS transistor (which is described in detail in Chapter 15).

It is known that bulk silicon exhibits an isotropic conductivity σ . The interesting question arose, then, whether the conductivity of the two-dimensional electron sheet is also isotropic in the interface plane and whether the conductivity depends on the crystallographic surface orientation of silicon. To settle these questions, experiments were performed on (100), (110), and (111) surfaces (the reader should be familiar with the Miller indices) by fabricating many transistors on various wafers of these three surface orientations. It was found that on (100) and (111) surfaces the conductivity is isotropic. The (110) surface, however, shows an anisotropic electrical conductivity.

Below we show that this result can be obtained by a straightforward calculation. The current density \mathbf{j} as a function of electric field \mathbf{F} is given by Ohm's law

$$\mathbf{j} = \sigma \mathbf{F} \quad (2.2)$$

In isotropic materials the conductivity σ is a scalar quantity. If we allow for anisotropy, σ becomes a matrix and Eq. (2.2) assumes the form (in two dimen-

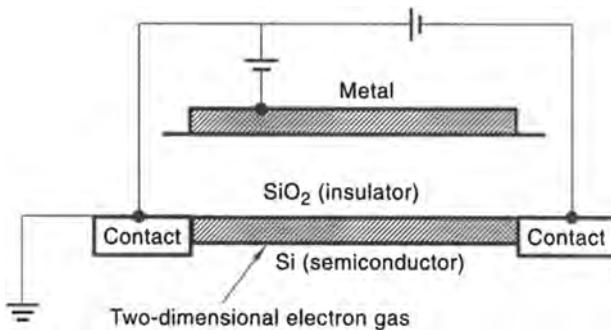


Figure 2.5 MOS transistor with a two-dimensional sheet of electrons at the Si-SiO₂ interface.

sions x, y)

$$\begin{pmatrix} j_x \\ j_y \end{pmatrix} = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{pmatrix} \begin{pmatrix} F_x \\ F_y \end{pmatrix} \quad (2.3)$$

Denoting the conductivity matrix by $\hat{\sigma}$ and the current density and electric field again by their vectors \mathbf{j} and \mathbf{F} , we have

$$\mathbf{j} = \hat{\sigma} \mathbf{F} \quad (2.4)$$

We now apply to Eq. (2.4) one of the symmetry operations Q of a regular square (our system is two dimensional, and the (100) surface has the symmetry of a square instead of a cube). The specific symmetry operation we choose is a rotation by 90°. In the notation of Table 2.1 and in three dimensions, such a rotation would be $Q_0 Q_{14}$.

Because we have given the conductivity in matrix form, we also would like to express the rotation in matrix form. From calculus, we know that a rotation by an angle ϕ can be represented by

$$Q_\phi = \begin{pmatrix} \cos \phi & \sin \phi \\ -\sin \phi & \cos \phi \end{pmatrix}$$

which gives for $\phi = 90^\circ$

$$Q_{90} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

Applying the operation Q_{90} to Eq. (2.4) from the left, we obtain

$$Q_{90} \mathbf{j} = Q_{90} \hat{\sigma} Q_{90}^{-1} Q_{90} \mathbf{F} \quad (2.5)$$

Here we have inserted before the field vector the operation $Q_{90}^{-1} Q_{90}$. Q_{90}^{-1} is the inverse operation of Q_{90} and therefore $Q_{90}^{-1} Q_{90}$ is the identity matrix

$$Q_{90}^{-1} Q_{90} \mathbf{F} = \mathbf{F}$$

From a physical point of view, it is now important to note that for $\phi = 90^\circ$, we have done nothing but turned the (100) surface into itself. Consequently, the current density has to be related to the field in the same way before and after the rotation. Therefore, the conductivity matrix must be the same and

$$Q_{90}\hat{\sigma}Q_{90}^{-1} = \hat{\sigma} \quad (2.6)$$

which reads in matrix form

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{pmatrix} \quad (2.7)$$

Performing the multiplication on the left-hand side of Eq. (2.7), we have

$$\begin{pmatrix} \sigma_{yy} & -\sigma_{yx} \\ -\sigma_{xy} & \sigma_{xx} \end{pmatrix} = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{pmatrix}$$

from which it follows that

$$\sigma_{yy} = \sigma_{xx} \quad (2.8)$$

and

$$-\sigma_{yx} = \sigma_{xy} \quad (2.9)$$

Using, in addition to the 90° rotation, the reflection symmetry $Q_0(\bar{x}_1, x_2)$ one easily obtains

$$\sigma_{xy} = \sigma_{yx} = 0 \quad (2.10)$$

Eq. (2.10) together with Eq. (2.9) gives $\sigma_{xy} = \sigma_{yx} = 0$, which together with Eq. (2.8) proves that \mathbf{j} and \mathbf{F} point in the same direction on a (100) surface; that is, the surface is isotropic. We can do the same proof for a (111) surface that is turned into itself by a rotation of $\phi = 120^\circ$. However, the (110) surface has only a $\phi = 180^\circ$ symmetry and this is not enough to prove that this surface exhibits isotropic behavior. Indeed, experiments show that the (110) surface conductivity is anisotropic.

These results are a special case of a more general rule: Any physical property that can be represented as a matrix of rank r_a is scalar in crystallographic systems that can be transformed into themselves by a number of rotations (around all main axes) larger than r_a .

The rank of a matrix is given by the number of indices. Thus the conductivity is of rank two; a matrix with elements a_{ikem} would be of rank 4. In our example, the number of rotations transforming the surface into itself was four for the (100) surface (90° rotation), three for the (111) surface (120° rotation), but only two for the (110) surface (180° rotation), which is not larger than the rank of the conductivity matrix.

We can use this rule to determine in what form (scalar, vector, matrix) we deal with certain physical properties of semiconductors of interest. The conductivity and similarly the optical absorption, microwave conductivity, and the

like are all matrices of rank two and therefore can be treated in cubic crystals as scalar. We can achieve this simplification without knowing the theory of the conductivity, which is actually developed in Chapter 8, just on the basis of the symmetry properties of the crystal. An even more powerful symmetry, the translational invariance is treated next.

2.2.2 Translational Invariance

We have not yet discussed in detail the other type of symmetry: the symmetry of translations. If we apply a translation \mathbf{R}_l [see Eq. (2.1)] to a crystal, the crystal is transformed into itself (except at the boundaries, which we disregard). We can now argue (as we did for the point group) that any physical property, denoted by $f(\mathbf{r})$, of the crystal has to be the same before and after this translation:

$$f(\mathbf{r} + \mathbf{R}_l) = f(\mathbf{r}) \quad (2.11)$$

where \mathbf{r} is the space coordinate and \mathbf{R}_l is a lattice vector. In other words $f(\mathbf{r})$ is a function that is periodic with respect to all lattice vectors \mathbf{R}_l . The periodicity is a hint that Fourier expansion will be a powerful mathematical tool to treat all those functions $f(\mathbf{r})$. Therefore, it is customary to introduce physical properties of crystals in terms of Fourier series. Because we are dealing with a three-dimensional entity with given periodicity, any physical property or function $f(\mathbf{r})$ is expanded as

$$f(\mathbf{r}) = \sum_{\mathbf{K}_h} A_{\mathbf{K}_h} e^{i\mathbf{K}_h \cdot \mathbf{r}} \quad (2.12)$$

with

$$A_{\mathbf{K}_h} = \frac{1}{\Omega} \int_{\Omega} f(\mathbf{r}) e^{-i\mathbf{K}_h \cdot \mathbf{r}} d\mathbf{r} \quad (2.13)$$

Ω is the basic volume that generates the crystal when repeated over and over by \mathbf{R}_l . The subscript h of \mathbf{K}_h is an integer and labels the vectors \mathbf{K} . The choice of the basic volume is not unique. We could, for example, choose the cell that has the three vectors $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ as boundaries. We can also choose the so-called Wigner-Seitz cell, which is obtained as follows. Connect all nearest neighbor atoms by lines $(\mathbf{a}_1, -\mathbf{a}_1, \mathbf{a}_2, -\mathbf{a}_2, \mathbf{a}_3, -\mathbf{a}_3)$ and cut the connections in half by planes. The geometrical figure enclosed by all these planes is the Wigner-Seitz cell. Note that this cell looks very different for face-centered cubic, body-centered cubic, and simple cubic lattices.

However we choose the cell, the volume is the same and is given by the following product of the three lattice generating vectors

$$\Omega = \mathbf{a}_1 \cdot (\mathbf{a}_2 \times \mathbf{a}_3) \quad (2.14)$$

Using Eq. (2.11), we have

$$f(\mathbf{r} + \mathbf{R}_l) = \sum_{\mathbf{K}_h} A_{\mathbf{K}_h} e^{i\mathbf{K}_h \cdot (\mathbf{r} + \mathbf{R}_l)} = f(\mathbf{r}) \quad (2.15)$$

and therefore,

$$e^{i\mathbf{K}_h \cdot \mathbf{R}_l} = 1 \quad (2.16)$$

From Eq. (2.16), it follows that

$$\mathbf{K}_h \cdot \mathbf{R}_l = 2\pi \text{ (integer)} \quad (2.17)$$

It can be shown (by inspection) that any vector

$$\mathbf{K}_h = h_1 \mathbf{b}_1 + h_2 \mathbf{b}_2 + h_3 \mathbf{b}_3 \quad (2.18)$$

with

$$\mathbf{b}_1 = \frac{2\pi}{\Omega} \mathbf{a}_2 \times \mathbf{a}_3,$$

$$\mathbf{b}_2 = \frac{2\pi}{\Omega} \mathbf{a}_3 \times \mathbf{a}_1,$$

$$\mathbf{b}_3 = \frac{2\pi}{\Omega} \mathbf{a}_1 \times \mathbf{a}_2$$

satisfies Eq. (2.17).

The vectors \mathbf{K}_h are called *reciprocal lattice vectors* (their unit is cm^{-1}). These vectors also generate a lattice, the reciprocal crystal lattice, which is complementary to the crystal lattice. Cubic lattices have reciprocal cubic lattices. However, the reciprocal lattice of a face-centered cubic crystal is body-centered cubic, and vice versa.

We learned in Chapter 1 about the importance of a reciprocal lattice vector, the vector $2\pi/a$ (in one dimension) or any multiple of it. We have seen that the wave vector q of phonons is basically restricted to a zone $-\pi/a \leq q \leq \pi/a$ and assumes the discrete values $2\pi l/Na$, where $-N/2 \leq l \leq N/2$.

For a three-dimensional crystal, the possible values that the wave vector \mathbf{q} can assume are

$$\mathbf{q} = \mathbf{K}_h / N \quad \text{with} \quad 0 \leq |h_1|, |h_2|, |h_3| \leq N/2 \quad (2.19)$$

Here we have assumed that the crystal contains N repetitions of the basis in each of the main directions $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$.

As mentioned, the largest physical values of $|\mathbf{q}|$ define an area called the Brillouin zone (Figure 2.6). This zone is the Wigner-Seitz cell of the reciprocal lattice, as can be seen from Eq. (2.19). The concept of the Brillouin zone is not only important for phonons, but also for electrons. The relevance and significance of the zone concept for electrons can be seen from Bloch's theorem, and the following discussions of Bragg reflection.

Because we like to discuss the consequences of translational invariance for electrons, we need to know the consequences for the wavefunction ψ . If we translate the crystal into itself then it is not ψ that is invariant but $|\psi|^2$, which gives the probability density of finding the electron. This means that ψ can change by a phase factor whose square equals unity as then $|\psi|^2$ is invariant.

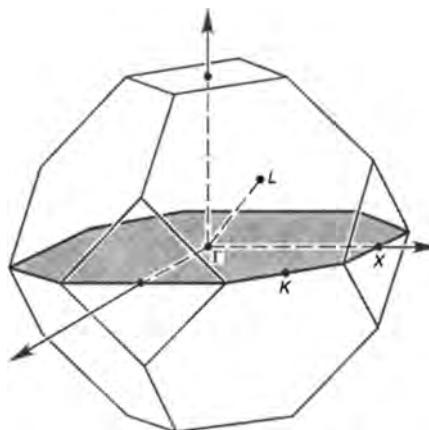


Figure 2.6 Brillouin zone of the face-centered cubic lattice (body-centered cubic reciprocal lattice). Notice that the zone extends to $2\pi/a$ in one direction (X) in contrast to the one-dimensional zone. The labels Γ , X and K denote symmetry points. Γ is the zone center $[0,0,0]$; X the endpoint in $[1,0,0]$ direction; and L and K , the zone endpoints in $[1,1,1]$ and $[1,1,0]$ directions, respectively.

Bloch found that the wave function ψ of an electron in a crystal can be labeled by a wave vector \mathbf{k} (analogous to \mathbf{q} for the phonons) and fulfills the relation

$$\psi(\mathbf{k}, \mathbf{r} + \mathbf{R}_l) = e^{i\mathbf{k}\cdot\mathbf{R}_l} \psi(\mathbf{k}, \mathbf{r}) \quad (2.20)$$

for any lattice vector \mathbf{R}_l .

It can be shown that Eq. (2.20) is equivalent to [see Problem 2.1]

$$\psi(\mathbf{k}, \mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} u_{\mathbf{k}}(\mathbf{r}) \quad (2.21)$$

where $u_{\mathbf{k}}(\mathbf{r})$ is a function periodic with respect to \mathbf{R}_l . That is,

$$u_{\mathbf{k}}(\mathbf{R}_l + \mathbf{r}) = u_{\mathbf{k}}(\mathbf{r}) \quad (2.22)$$

It is important to note that the wave function is unchanged if we replace \mathbf{k} by $\mathbf{k} + \mathbf{K}_h$, where \mathbf{K}_h is a reciprocal lattice vector. This can be seen from Eq. (2.21); $u_{\mathbf{k}}(\mathbf{r})$ can be expanded into the Fourier series given by Eq. (2.15) because it is periodic. The multiplication of this series by $e^{i\mathbf{K}_h\cdot\mathbf{r}}$ just leads to an identical series, which is only reordered in the sequence of reciprocal lattice vectors. Therefore it is clear that the wave vector \mathbf{k} , which for a free electron is proportional to the momentum, must have a different meaning in the crystal. To explore this meaning we will use perturbation theory by regarding the crystal as a perturbation of the free electron behavior. Detailed derivations of the above equations have been given in the literature [1].

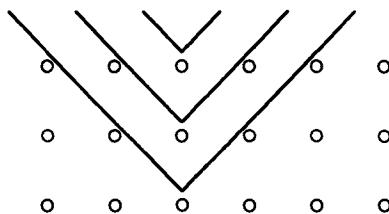


Figure 2.7 Bragg reflection of electrons (waves) by crystal planes.

2.3 BRAGG REFLECTION

To calculate the matrix elements, we Fourier decompose the periodic crystal potential $V(\mathbf{r})$:

$$V(\mathbf{r}) = \sum_h B_{\mathbf{K}_h} e^{i\mathbf{K}_h \cdot \mathbf{r}} \quad (2.23)$$

The matrix element is then

$$\langle \mathbf{k}' | V(\mathbf{r}) | \mathbf{k} \rangle = \sum_h B_{\mathbf{K}_h} \delta_{\mathbf{k}-\mathbf{k}', -\mathbf{K}_h} \quad (2.24)$$

and vanishes except for

$$\mathbf{k} + \mathbf{K}_h = \mathbf{k}' \quad (2.25)$$

The collision of the “light” electron with the “huge” crystal lattice will be elastic (we ignore lattice vibrations), and therefore the energy before and after the collision will be the same, which is equivalent to

$$|\mathbf{k}| = |\mathbf{k}'| \quad (2.26)$$

Squaring Eq. (2.25) and using Eq. (2.26), we have

$$-2\mathbf{k} \cdot \mathbf{K}_h = |\mathbf{K}_h|^2 \quad (2.27)$$

This represents the well-known condition for Bragg reflection. It is also clear that Bragg reflection occurs for \mathbf{k} at the Brillouin zone boundary (solve Eq. (2.27) in one dimension). Bragg reflection simply means that the electrons are reflected by crystal planes, so Eq. (2.25) is valid (Figure 2.7).

We return now to the Bloch theorem [Eq. (2.21)] and regard $u_{\mathbf{k}}(\mathbf{r})$ as a constant. One then recovers the wave function of the free electron, and $\hbar\mathbf{k}$ is the momentum of the free electron. For a free electron, momentum is conserved. In a crystal, \mathbf{k} is not conserved because the crystal itself can contribute vectors \mathbf{K}_h , as can be seen from Eq. (2.25). We now understand that there is not only one \mathbf{k} attributed to the wave function, but all $\mathbf{k} + \mathbf{K}_h$ —or, as stated before, all wave functions with wave vectors $\mathbf{k} + \mathbf{K}_h$ are equivalent. In other words, we can again restrict ourselves to use \mathbf{k} within the Brillouin zone. If \mathbf{k} lies outside the zone, we subtract \mathbf{K}_h until we obtain a value inside the zone.

What happens to the energy of the electrons? Is there a maximum energy at the Brillouin zone boundary as in the case of lattice vibrations? There is not; the electron energy as a function of \mathbf{k} is multiple valued and only limited within certain bands. A helpful analogy is the following.

We compare energy and wave vector with the true kinetic energy and rotational frequency of a spinning wheel in a movie. As the wheel spins faster and faster, the picture of the wheel seems to stop as soon as the frequency of the moving pictures v_0 is the same as the spinning frequency of the wheel v . Increasing v leads to a picture in which the wheel seems to spin in the opposite direction until $v = 2v_0$ and then in forward direction again for $2v_0 \leq v \leq 3v_0$ and so on. If we see only the movie, we do not know which kinetic energy or frequency v the wheel really has unless somebody tells us in which frequency range

$$nv_0 \leq v \leq (n+1)v_0 \quad (2.28)$$

the wheel is spinning.

This situation is analogous in crystals with respect to the energy. For any given \mathbf{k} vector, the energy can be viewed as multiple valued and we need to specify a range, or band, in order to find the energy E from \mathbf{k} . The theory of $E(\mathbf{k})$, band-structure theory, is treated in Chapter 3.

PROBLEMS

- 2.1** Show the equivalence of Eqs. (2.20) and (2.21)

$$\psi(\mathbf{r} + \mathbf{R}_l) = e^{i\mathbf{k}\cdot\mathbf{R}_l} \psi(\mathbf{r})$$

and

$$\psi = e^{i\mathbf{k}\cdot\mathbf{r}} u_{\mathbf{k}}(\mathbf{r})$$

where $u_{\mathbf{k}}(\mathbf{r})$ has the period of the lattice.

- 2.2** Find the surface atomic densities of (100), (110), and (111) planes in a silicon structure with a lattice constant of a . Which plane has the highest density?

- 2.3** (a) Find the reciprocal lattice of the three-dimensional face-centered cubic lattice. Use as lattice vectors (\hat{x} , \hat{y} , \hat{z} are the unit vectors in the respective direction)

$$\begin{aligned} \mathbf{a}_1 &= \frac{a}{2}(\hat{x} + \hat{y}) \\ \mathbf{a}_2 &= \frac{a}{2}(\hat{y} + \hat{z}) \\ \mathbf{a}_3 &= \frac{a}{2}(\hat{z} + \hat{x}) \end{aligned}$$

- (b) Sketch the first Brillouin zone.

- (c) Find the volume of the first Brillouin zone.

(d) Repeat parts a–c for the body-centered cubic lattice with primitive lattice vectors

$$\mathbf{a}_1 = \frac{a}{2}(\hat{\mathbf{x}} + \hat{\mathbf{y}} - \hat{\mathbf{z}})$$

$$\mathbf{a}_2 = \frac{a}{2}(-\hat{\mathbf{x}} + \hat{\mathbf{y}} + \hat{\mathbf{z}})$$

$$\mathbf{a}_3 = \frac{a}{2}(\hat{\mathbf{x}} - \hat{\mathbf{y}} + \hat{\mathbf{z}})$$

REFERENCE

- [1] Madelung, O. *Introduction to Solid State Theory*. New York: Springer-Verlag, 1978, pp. 36–55.

CHAPTER 3

THE THEORY OF ENERGY BANDS IN CRYSTALS

3.1 COUPLING ATOMS

In Chapter 2 we hinted at a band structure for the $E(k)$ relation from rather formal arguments. We now introduce the bands from phenomenological considerations.

Consider a series of quantum wells, as shown in Figures 3.1a through 3.1c. The wells in Figure 3.1a are separated and essentially independent. Each well has, therefore, a series of discrete levels. In Figure 3.1b, the wells are closer together and coupled by the possibility of tunneling. This coupling causes a splitting of the energy levels into N closely spaced levels if we have N coupled quantum wells. The effect is much the same as the phenomenon associated with coupled oscillators (the frequency of the oscillators then splits into a series of frequency maxima) or coupled pendulums in mechanics. We can see that the single levels are, therefore, replaced by *bands of energies*.

If the wells are coupled very closely together, a new phenomenon can happen: It is possible that the bands spread more and overlap. Such an overlap is typical for metals. There is, however, an effect that can split up the bands, even if we put the wells closer and closer together. This happens, in fact, in semiconductors such as diamond and silicon. The effect is known as bonding-antibonding splitting, which is schematically explained in Figures 3.2a and 3.2b. The wave functions plotted in the figures give approximately the same probability for finding an electron in either well. However, the probability of finding an electron between the wells vanishes in Figure 3.2a and is finite in Figure 3.2b. The situation is known from molecules where the electron holds together the positive nuclei of the ions by being in between them. Therefore, a state resembling that in Figure 3.2b is called a bonding state, whereas Figure 3.2a (with no probability of finding an electron in between) represents an antibonding state.

Under certain circumstances, this separation into bonding and antibonding states can lead to an additional splitting of the bands and therefore to the appearance of additional “energy gaps”—regions without states for the electrons. (In

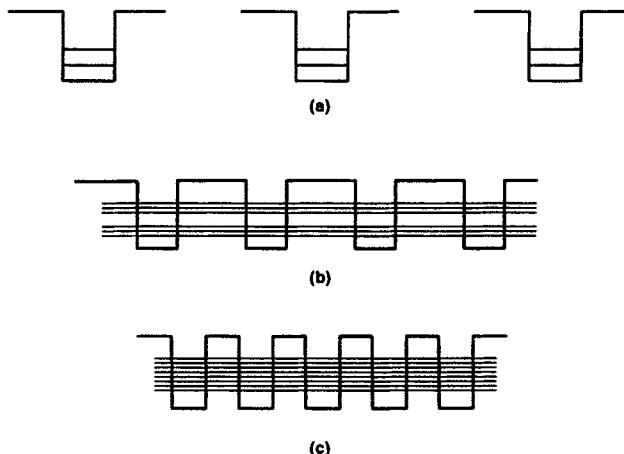


Figure 3.1 Quantum wells coupled together with increasing coupling strength.

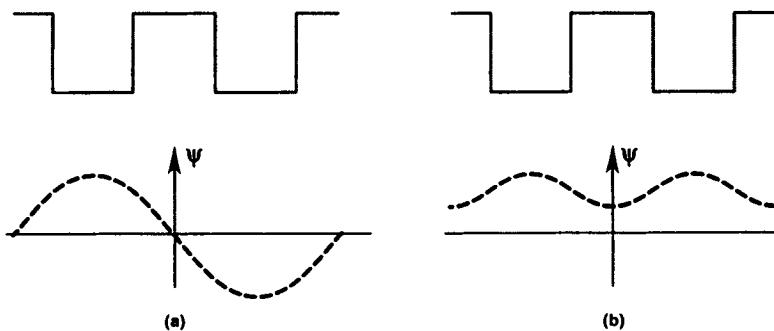


Figure 3.2 Possible forms of wave function for two coupled wells.

the case of diamond and silicon, the splitting follows the formation of so-called sp^3 hybrids, which is discussed in Harrison [5].)

3.2 ENERGY BANDS BY FOURIER ANALYSIS

Although the above discussion establishes the bands by moving the wells (atoms) closer to each other, we can also go about this in another way. We can start with a free electron and introduce the crystal just by using our knowledge about Brillouin zones and restricting the energy function to this zone. We then have to plot the parabola of the free electron $E(\mathbf{k})$ relation, $E = \hbar^2\mathbf{k}^2/2m$, as shown in Figure 3.3.

At first glance, it seems that we have done nothing more than replot a parabola in a very complicated way. However, we will see in the following that the $E(\mathbf{k})$ relation in a crystal is similar to Figure 3.3 except that at the intersections and zone boundaries the function splits, as indicated by the dashed lines, which leads

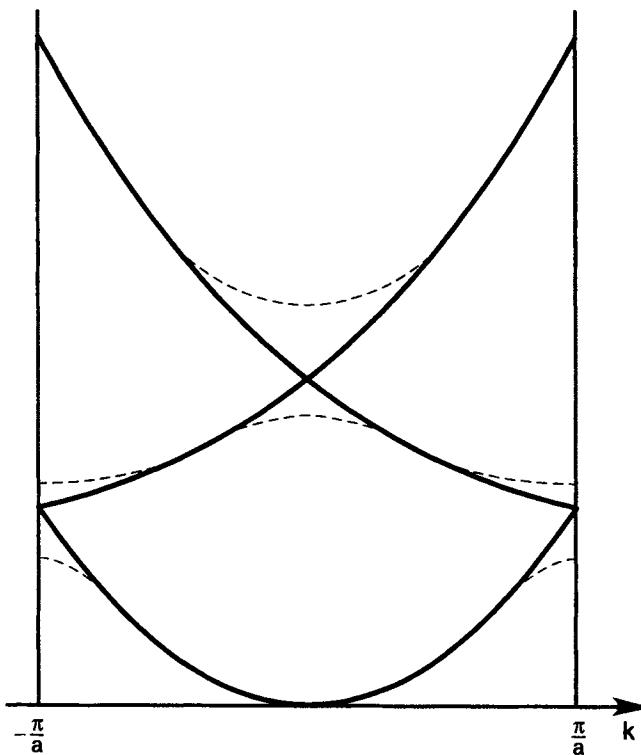


Figure 3.3 The free electron parabola plotted in the Brillouin zone. Notice that the energy now becomes a multiple valued function of \mathbf{k} and we have to label the different branches (numbers 1, 2, 3, ...) to distinguish among them.

to the formation of bands and energy gaps. A rather rigorous theory of the $E(\mathbf{k})$ relation that contains most of these features automatically is described in the following discussion.

This theory is based on direct Fourier analysis of the Schrödinger equation. Bloch's theorem tells us that the wave function can be written in the form

$$\psi(\mathbf{k}, \mathbf{r}) = e^{i\mathbf{k} \cdot \mathbf{r}} u_{\mathbf{k}}(\mathbf{r}) \quad (3.1)$$

where $u_{\mathbf{k}}(\mathbf{r})$ is periodic. Therefore, we can Fourier expand $u_{\mathbf{k}}$ in terms of reciprocal lattice vectors as discussed in Eq. (2.12) to obtain

$$\psi(\mathbf{k}, \mathbf{r}) = e^{i\mathbf{k} \cdot \mathbf{r}} \sum_h A_{\mathbf{K}_h} e^{i\mathbf{K}_h \cdot \mathbf{r}} \quad (3.2)$$

Inserting Eq. (3.2) into the Schrödinger equation [Eq. (1.25)] with a periodic

crystal-potential $V(\mathbf{r})$, we obtain

$$\begin{aligned} \frac{\hbar^2}{2m} \sum_h |\mathbf{k} + \mathbf{K}_h|^2 A_{\mathbf{K}_h} e^{i(\mathbf{k} + \mathbf{K}_h) \cdot \mathbf{r}} + V(\mathbf{r}) \sum_h A_{\mathbf{K}_h} e^{i(\mathbf{k} + \mathbf{K}_h) \cdot \mathbf{r}} \\ = E(\mathbf{k}) \sum_h A_{\mathbf{K}_h} e^{i(\mathbf{k} + \mathbf{K}_h) \cdot \mathbf{r}} \end{aligned} \quad (3.3)$$

The method of solving differential equations by Fourier transformation proceeds now by multiplying Eq. (3.3) by a term, $e^{-i(\mathbf{k} + \mathbf{K}_l) \cdot \mathbf{r}}$, and integrating over the volume V_{ol} of the crystal. Using Eqs. (1.46) and (1.51), we then obtain

$$(E_l^0 - E(\mathbf{k})) A_{\mathbf{K}_l} + \sum_h A_{\mathbf{K}_h} \langle \mathbf{K}_l | V(\mathbf{r}) | \mathbf{K}_h \rangle = 0 \quad (3.4)$$

where we have used the definition

$$\frac{\hbar^2}{2m} |\mathbf{k} + \mathbf{K}_l|^2 \equiv E_l^0 \quad (3.5)$$

Because Eq. (3.4) is homogeneous, a nonzero solution for the $A_{\mathbf{K}_h}$ exists only if the determinant of the coefficients vanishes; that is,

$$\det \left| (E_l^0 - E(\mathbf{k})) \delta_{l,h} ; \langle \mathbf{K}_l | V(\mathbf{r}) | \mathbf{K}_h \rangle \right| = 0 \quad (3.6)$$

where l and h are integers ($l, h = 0, \pm 1, \pm 2, \dots$).

Equation (3.6) is called the secular equation and gives us the possible values of the energy $E(\mathbf{k})$ as a function of wave vector \mathbf{k} . It can only be solved numerically; a large number of Fourier components (\mathbf{K}_h) are usually necessary to give a reasonable description of energy bands in semiconductors. However, we can see some significant features of the solution if we restrict ourselves to a one-dimensional model with two reciprocal lattice vectors 0 ($h = 0$) and $-2\pi/a$ ($h = -1$), which corresponds to one incident and one reflected wave. Then the only matrix element that matters is $\langle 0 | V(\mathbf{r}) | -1 \rangle$, which we denote by M . The secular equation then reads (putting $\langle 0 | V(\mathbf{r}) | 0 \rangle = \langle -1 | V(\mathbf{r}) | -1 \rangle = 0$ by proper choice of the energy scale):

$$\begin{vmatrix} E_0^0 - E(\mathbf{k}) & M \\ M^* & E_{-1}^0 - E(\mathbf{k}) \end{vmatrix} = 0 \quad (3.7)$$

which gives

$$E(\mathbf{k}) = \frac{1}{2}(E_0^0 + E_{-1}^0) \pm \frac{1}{2}\sqrt{(E_0^0 - E_{-1}^0)^2 + 4|M|^2} \quad (3.8)$$

with M^* being the complex conjugate of M . Plotting this $E(\mathbf{k})$ relation gives exactly the first two bands (0 and 1) of Figure 3.3 with the splitting at the Brillouin zone boundary of magnitude $2|M|$. This splitting is called the energy gap E_G .

The student of $E(\mathbf{k})$ relations may wish to pause here and consider the special case $V(\mathbf{r}) = 0$. As can be seen from Eq. (3.6), this leads to the empty lattice

bands $E(\mathbf{k}) = E_l^0$. Note, however, that owing to the complicated reciprocal lattice vectors of the cubic face centered semiconductors, the bands are more complicated than in Figure 3.3. Nice examples are given in [7].

The calculation, as we have presented it, looks rather simple, and the question arises why the band structures of a broad class of semiconductors have been accurately calculated only after 1965. One of the reasons for the problems in band-structure calculations is the fact that it is difficult to calculate the potential $V(\mathbf{r})$. It is clear that $V(\mathbf{r})$, the potential an electron “feels,” is generated not only by the atomic nuclei but also by all the other electrons in the crystal. Therefore, it is not even true that the potential can be written as a locally defined function $V(\mathbf{r})$. However, experience has shown that this is not too bad an approximation for most semiconductors. But how do we determine $V(\mathbf{r})$?

From Eq. (3.6) we can see that it is not even necessary to know $V(\mathbf{r})$. All one needs are the matrix elements—that is, its Fourier components with respect to reciprocal lattice vectors. Cohen and Bergstresser [3] have demonstrated that only a few Fourier components are necessary to obtain relatively accurate $E(\mathbf{k})$ relations if they are chosen wisely. Below we discuss how this choice should be made. We start from the fact that the potential $V(\mathbf{r})$ must be a sum of all the contributions of the atoms that constitute the crystal.

If the atoms are located at \mathbf{R}_l , by introducing a suitable function ω , we can write

$$V(\mathbf{r}) = \sum_l \omega(\mathbf{r} - \mathbf{R}'_l) \quad (3.9)$$

\mathbf{R}'_l is not necessarily a lattice vector. In the case of silicon, we have two atoms in the Wigner-Seitz cell and \mathbf{R}'_l has to reach both. The matrix element can be written as a Fourier coefficient obtained from

$$M_{mn} = \langle \mathbf{K}_n | V | \mathbf{K}_m \rangle = \frac{1}{V_{\text{ol}}} \int_{V_{\text{ol}}} \sum_l \omega(\mathbf{r} - \mathbf{R}'_l) e^{-i\mathbf{K} \cdot \mathbf{r}} d\mathbf{r} \quad (3.10)$$

where we have properly normalized the wave function to the crystal volume, the integration extends over this volume, and $\mathbf{K} = \mathbf{K}_m - \mathbf{K}_n$ (m and n are integer indices as l and h above).

We now rewrite Eq. (3.10) as

$$M_{mn} = \frac{1}{V_{\text{ol}}} \sum_l e^{-i\mathbf{K} \cdot \mathbf{R}'_l} \int_{V_{\text{ol}}} e^{-i\mathbf{K} \cdot (\mathbf{r} - \mathbf{R}'_l)} \omega(\mathbf{r} - \mathbf{R}'_l) d(\mathbf{r} - \mathbf{R}'_l) \quad (3.11)$$

Notice that we can replace $\mathbf{r} - \mathbf{R}'_l$, by \mathbf{r}' in the integral, which then does not depend on the index l . Remember also that we have two atoms (ions) in the Wigner-Seitz cell. To account for this, we label the position of the second atom by $\mathbf{R}_l + \tau$ and position the first atom at \mathbf{R}_l . \mathbf{R}_l is now a lattice vector. (It is also common to position \mathbf{R}_l in between the two atoms and to add and subtract a vector $\tau/2$ to reach the two atomic positions.) Therefore, using Eq. (2.17), we

can rewrite Eq. (3.11) as

$$M_{mn} = \sum_l (1 + e^{-i\mathbf{K}\cdot\tau}) \frac{1}{V_{ol}} \int_{V_{ol}} e^{-i\mathbf{K}\cdot\mathbf{r}'} \omega(\mathbf{r}') d\mathbf{r}' \quad (3.12)$$

where the label l now runs over all Wigner-Seitz cells (not atoms). The function $\omega(\tau)$ is the same in each Wigner-Seitz cell.

We assume also that $\omega(\mathbf{r}')$ vanishes rapidly if we go outside the cell and therefore

$$\int_{V_{ol}} e^{-i\mathbf{K}\cdot\mathbf{r}'} \omega(\mathbf{r}') d\mathbf{r}' = \int_{\Omega} e^{-i\mathbf{K}\cdot\mathbf{r}'} \omega(\mathbf{r}') d\mathbf{r}' \quad (3.13)$$

where Ω is the volume of the Wigner-Seitz cell given by Eq. (2.14). Because the volume of the crystal $V_{ol} = N\Omega$, we finally obtain from Eq. (3.12)

$$M_{mn} = (1 + e^{-i\mathbf{K}\cdot\tau}) \frac{1}{\Omega} \int_{\Omega} e^{-i\mathbf{K}\cdot\mathbf{r}'} \omega(\mathbf{r}') d\mathbf{r}' \quad (3.14)$$

The first term in Eq. (3.14) is the structure factor $S(\mathbf{K})$

$$S(\mathbf{K}) = (1 + e^{-i\mathbf{K}\cdot\tau}) \quad (3.15)$$

whereas the second term is the form factor. The structure factor can easily be calculated from the reciprocal lattice vectors $\mathbf{K}_m - \mathbf{K}_n = \mathbf{K}$.

The reciprocal lattice vectors are the vectors of the cubic body-centered reciprocal lattice (e.g., for silicon) given by

$$\begin{aligned} K_h &= h_1 \mathbf{b}_1 + h_2 \mathbf{b}_2 + h_3 \mathbf{b}_3 \\ &= \frac{2\pi}{a} (-h_1 + h_2 + h_3, h_1 - h_2 + h_3, h_1 + h_2 - h_3) \end{aligned} \quad (3.16)$$

where h_1, h_2, h_3 are integers. It is important to realize that there is a particular form of the combination of integers that enters the vector components of K_h and not all integer combinations are possible. For example, $K_n = \frac{2\pi}{a}(1, 0, 0)$ is not allowed because it would correspond to values of $h_2 = h_3 = 1/2$, which is not integral and therefore not permitted. This is the reason for the complicated form of empty lattice bands for three-dimensional cubic face-centered lattices. A more extensive illustration can be found in the book by Landsberg [7] on page 222.

It remains to determine the form factors. Cohen and Bergstresser [3] assumed that only reciprocal lattice vectors \mathbf{K} of squared magnitude 0, 3, 4, 8, and 11 (in units of $2\pi/a$) contribute and all the other Fourier components vanish. Therefore, the form factor is replaced by a few unknowns. These five unknowns are adjusted to obtain an $E(\mathbf{k})$ relation that fits best the existing experiments (optical absorption, etc.). Band-structure calculations are then reduced to the solution of the system of Eqs. (3.4) and (3.6), which can be achieved with standard numerical routines.

In this way, Cohen and Bergstresser found the $E(\mathbf{k})$ relation of many semiconductors. Their results for Si, GaAs, and other materials are shown in Figure 3.4. The \mathbf{k} vectors are plotted along some of the major directions in the

Brillouin zone (see Figure 2.6) from Γ to X , Γ to L , and so on. Not all bands are shown in the figure—only the most important ones that contribute to electronic conduction. These are the highest bands that are still filled with electrons, the valence band, and the next higher band separated by an energy gap E_G , the conduction bands.

The indices of Γ , L , and X in Figure 3.4 denote the symmetry of the wave functions and their behavior with respect to transformations of the coordinates x , y , z . Γ_1 means, for example, that the wave function at this point has s symmetry (i.e., is spherically symmetric); Γ_{15} means that the wave function has p symmetry (i.e., is cylindrically symmetric around the x , y , z axes); and so on. The method described above is called the empirical pseudo-potential method.

To obtain a more complete view of the band structure, it is customary to plot lines or surfaces of constant energy in k space. Such a plot is shown in Figure 3.5. The label on the curves is the energy of the particular curve in electron volts. It is important to note that for free electrons the lines of equal energy would be circles (disregarding relativistic effects). The Bragg reflection from crystal planes gives rise to the very complicated pattern of Figure 3.5. Only at very low energies in certain bands do the lines approach circles (for silicon, even this is not true).

In actual calculations, it is costly to compute the $E(k)$ relation for all points of the Brillouin zone. However, we know from Chapter 2 that 48 operations transform the cube into itself and the same 48 operations transform the Brillouin zone of a cubic lattice into itself. We, therefore, need to calculate the $E(k)$ relation only in $1/48$ of the Brillouin zone and then apply the symmetry operations of Table 2.1 (replacing $x_1x_2x_3$ with $k_xk_yk_z$) to obtain $E(k)$ everywhere, because

$$QE(\mathbf{k}) = E(\mathbf{k}') = E(\mathbf{k}) \quad (3.17)$$

The $1/48$ part of the zone that can be turned into the full Brillouin zone by the operations Q is given by the conditions

$$0 \leq k_z^* \leq k_y^* \leq k_x^* \leq 1 \quad (3.18)$$

and

$$k_x^* + k_y^* + k_z^* \leq 3/2 \quad (3.19)$$

Here the starred k components are normalized by $2\pi/a$; for example, $k_z^* = k_z(2\pi/a)$. The volume defined by Eqs. (3.21) and (3.22) is called the irreducible wedge, and is shown in Figure 3.6. It should be noted that the translational invariance leads also to an important law for $E(\mathbf{k})$:

$$E(\mathbf{k} + \mathbf{K}_h) = E(\mathbf{k}) \quad (3.20)$$

for the very same reasons which lead to Eq. (3.17).

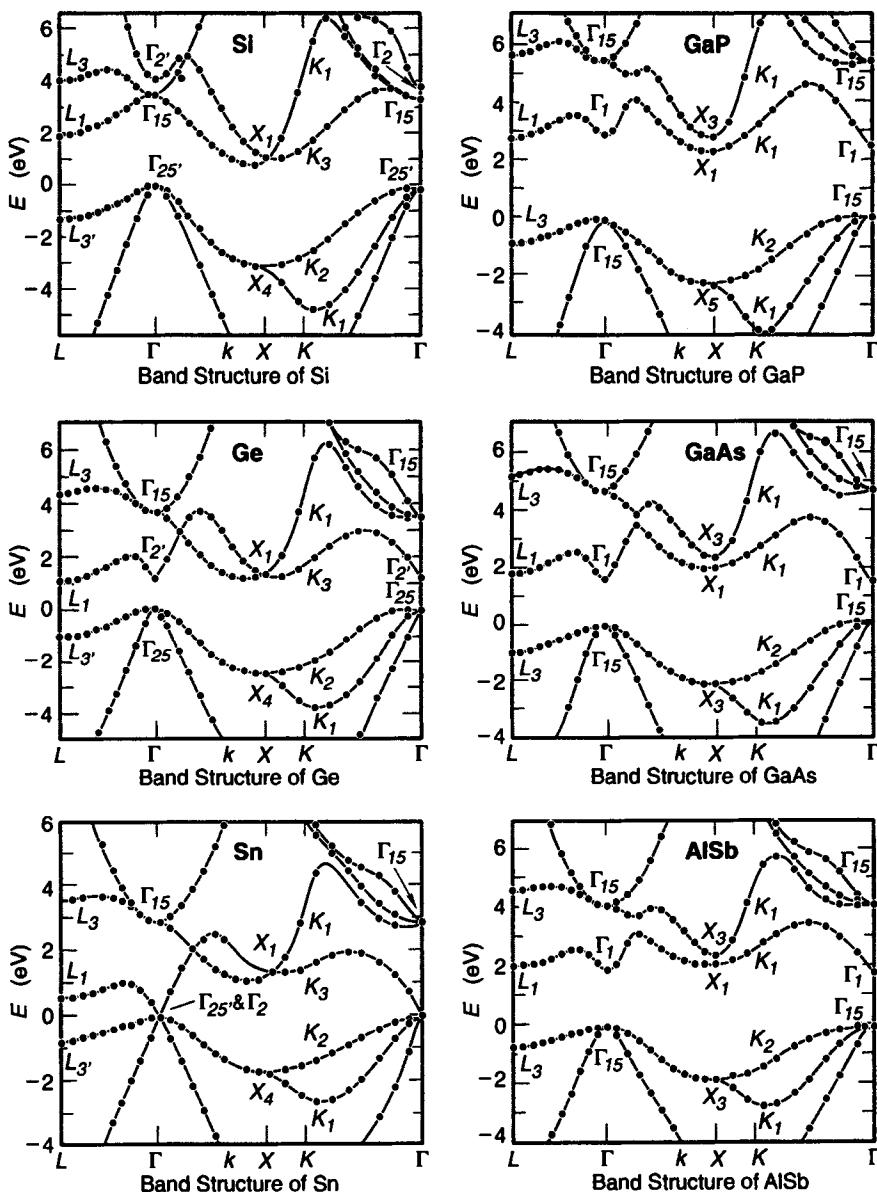


Figure 3.4 Band structure of important semiconductors. [After Cohen and Bergstresser [3].]

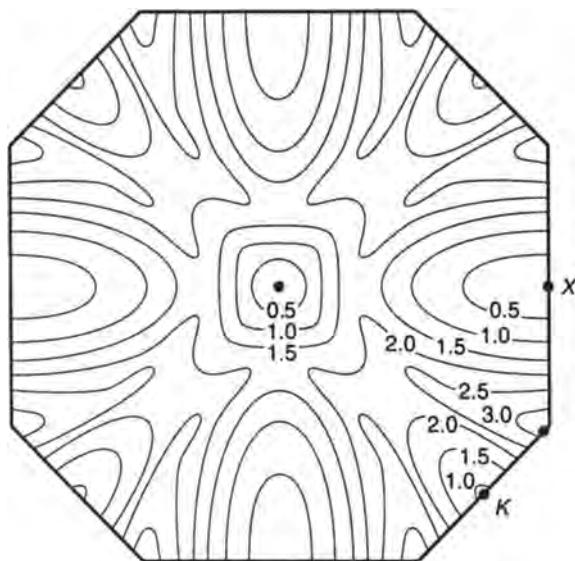


Figure 3.5 Lines of equal energy in k space for a certain cut of the Brillouin zone and the conduction band of GaAs. [After Shichijo and Hess [9], Figure 4.]

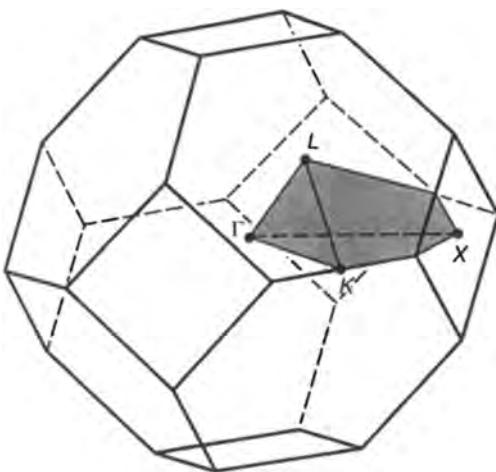


Figure 3.6 Sampling region for the calculation of the band structure. The region is a 1/48 part of the Brillouin zone.

3.3 EQUATIONS OF MOTION IN A CRYSTAL

How does an electron move in such a complicated energy band? The equations of motion are simpler than one would expect. We sketch only the derivation; the full derivation takes considerable space. It is shown in Appendix B that if we restrict ourselves to the consideration of one particular band (one band approximation), the Schrödinger equation can be written as

$$(E(-i\nabla) - eV_{\text{ext}})\psi = E\psi \quad (3.21)$$

if a weak external potential V_{ext} is applied to the crystal.

Here $E(-i\nabla)$ simply means that we take the function $E(\mathbf{k})$ and Taylor-expand it. Then we replace all \mathbf{k} by $-i\nabla$. To give a one-dimensional example, we have

$$E\left(-i\frac{\partial}{\partial x}\right) = E(0) - \frac{\partial E}{\partial k}\Bigg|_{k=0} i\frac{\partial}{\partial x} - \frac{1}{2}\frac{\partial^2 E}{\partial k^2}\Bigg|_{k=0} \frac{\partial^2}{\partial x^2} \quad (3.22)$$

This equation, viewed as a differential equation, is, of course, more difficult to solve than the original Schrödinger equation. The given form, however, is useful for the general considerations that are discussed below because it does not contain the unknown crystal potential.

To proceed, we need to invoke Ehrenfest's theorem, which gives the law of motion for the mean values of coordinates and momenta of a quantum system. It says that Eq. (1.2) is replaced by the following equation for the average velocity component v_i :

$$v_i = \frac{d\langle\psi|x_i|\psi\rangle}{dt} = \frac{1}{\hbar}\langle\psi|\frac{\partial H}{\partial k_i}|\psi\rangle \quad (3.23)$$

If we take the Hamiltonian from Eq. (3.21) and a Bloch wave function for ψ as in Appendix B, we obtain from Eq. (3.23)

$$\frac{\partial}{\partial k_i}H|\psi\rangle = \frac{\partial}{\partial k_i}E(\mathbf{k})|\psi\rangle \quad (3.24)$$

which gives for the vector of the average velocity:

$$\mathbf{v} = \frac{d\mathbf{r}}{dt} = \frac{1}{\hbar}\nabla_{\mathbf{k}}E(\mathbf{k}) \quad (3.25)$$

where

$$\nabla_{\mathbf{k}} = \left(\frac{\partial}{\partial k_x}, \frac{\partial}{\partial k_y}, \frac{\partial}{\partial k_z} \right)$$

The momentum changes only in time if we apply an external electric (or magnetic field), and from the Ehrenfest's theorem one obtains as in Eq. (1.1) with $\mathbf{F} = -\nabla V_{\text{ext}}$

$$\hbar\frac{d\mathbf{k}}{dt} = -e\mathbf{F} \quad (3.26)$$

where \mathbf{F} is the electric field.

Equation (3.26) is identical to the equation for a free electron and is true only because we restrict ourselves to one band. Equation (3.25) tells us that the velocity \mathbf{v} of the electron points always perpendicular to the curves of constant energy in \mathbf{k} space (because it is proportional to the gradient of $E(\mathbf{k})$). Therefore, the motion of an electron in a crystal can be much more complicated than the motion of a free electron.

Why can we describe the conductivity of a metal and semiconductor in simple terms? After understanding the complicated relation between energy and momentum as a consequence of Bragg reflection in the crystal, it is surprising that simple models that regard the electrons as free (and assume a quadratic relation between energy and momentum) have worked so well in the past. In fact, elementary introductions to semiconductor physics and electronics can be given without much knowledge of the band structure at all. The reason is that electrons reside, at least when close to equilibrium (small current densities), close to the minima of the $E(\mathbf{k})$ relation. Then we can expand the $E(\mathbf{k})$ relation in a Taylor series (in one dimension):

$$E(k') = E(k_0) + \frac{\partial E}{\partial k} \Big|_{k=k_0} (k' - k_0) + \frac{1}{2} \frac{\partial^2 E}{\partial k^2} \Big|_{k=k_0} (k' - k_0)^2 + \dots \quad (3.27)$$

where k_0 designates the location of the minimum (maximum) of the $E(k)$ relation. We now can choose the energy scale so that $E(k_0) = 0$. Furthermore, because the first derivative vanishes at an extremum, we may rename $(k' - k_0)$ by k and the number

$$\frac{\partial^2 E}{\partial k^2} \Big|_{k=k_0} \quad \text{by} \quad \frac{\hbar^2}{m^*} \quad (3.28)$$

to obtain

$$E(k) = \frac{\hbar^2 k^2}{2m^*} \quad (3.29)$$

which is the equation for the free electron with the mass replaced by an “effective mass” m^* .

If the surfaces of equal energy in \mathbf{k} spaces are ellipsoidal (as they are in silicon and germanium), one obtains by the same reasoning, using the coordinate system of the main ellipsoidal axes

$$E(\mathbf{k}) = \sum_j \frac{\hbar^2}{2m_j^*} k_j^2 \quad (3.30)$$

In silicon, the minima of the conduction band do have an ellipsoidal shape with the two masses $m_l^* = 0.91m_0$, and $m_t^* = 0.19m_0$ (l stands for the longitudinal and t for the transverse axis of the ellipsoid of revolution). This seems puzzling; we know from group theory that the conductivity of silicon is isotropic. The explanation is that silicon has six equivalent ellipsoids (minima) in the [100], [-100],

[010], [0-10], [001], and [00-1] directions, and the average overall ellipsoids give an isotropic conductivity mass m_σ of

$$\frac{1}{m_\sigma^*} = \frac{1}{3} \left(\frac{1}{m_l^*} + \frac{2}{m_t^*} \right) \quad (3.31)$$

This effective mass treatment close to a minimum can be put in a very general form, because not only functions of numbers but also functions of operators can be written as a Taylor series [Eq. (3.22)]. This general form is the *effective mass theorem*. The effective mass theorem is derived by expanding the one band approximation formula [Eq. (3.21)] into the Taylor series of Eq. (3.22) and truncating after the third term (the term proportional to $\partial^2/\partial x^2$). Weak external potentials V_{ext} can be simply added in as they are in Eq. (3.21). Of course, they must not be strong enough to perturb the band structure itself, which is the case if all their Fourier components (with respect to the reciprocal lattice vectors) are smaller than the Fourier components of the crystal potential energy $V(\mathbf{r})$.

The effective mass theorem can then be stated in the following form. Assume we have a periodic crystal potential energy $V(\mathbf{r})$ and additional (*although weaker in all Fourier components*) external potentials V_{ext} . If one is interested only in the properties near an extremum $E(\mathbf{k}_0)$ of the $E(\mathbf{k})$ relation, then the Schrödinger equation

$$\left(-\frac{\hbar^2}{2m} \nabla^2 + V(\mathbf{r}) - eV_{\text{ext}} \right) \psi = E\psi \quad (3.32)$$

which is equivalent to

$$(E(-i\nabla) - eV_{\text{ext}})\psi = E\psi \quad (3.33)$$

can be replaced by

$$\left(-\sum_j \frac{\hbar^2}{2m_j^*} \frac{\partial^2}{\partial x_j^2} - eV_{\text{ext}} \right) \phi = (E - E(\mathbf{k}_0))\phi \quad (3.34)$$

Here ϕ is the so-called envelope wave function, which contains all the effects of the external potential (as long as it varies slowly) while all the effects of the crystal potential are absorbed in the effective mass m_j . The effective mass theorem then provides a new equation for the envelope wave function in which the crystal potential has disappeared, and all we had to do for this is to replace the mass by an effective mass and count the energy from the minimum $E(\mathbf{k}_0)$. The ellipsoidal form of Eq. (3.34) can be transformed into the familiar isotropic spherical form by suitable coordinate transformations.

Equation (3.34) is basic for the theory of electron devices. It forms the basis for the theory and definition of holes, donor and acceptor states, and a basis to separate the electron energy in the familiar form of kinetic energy above $E(\mathbf{k}_0)$ and potential energy V_{ext} as we do for free electrons. It also permits us to gain understanding of such complex problems as impurity bands or quantum wells in

semiconductors as we show in the three following examples and others in later chapters.

1. Assume V_{ext} is the potential of a single charged impurity in a crystal:

$$V_{\text{ext}} \approx \frac{e}{4\pi\epsilon_0\epsilon r}$$

where ϵ is the dielectric constant, ϵ_0 is the dielectric constant of vacuum, and e is the magnitude of the electronic charge. Then, for the isotropic case, we have a Schrödinger equation

$$\left(-\frac{\hbar^2}{2m^*} \nabla^2 - \frac{e^2}{4\pi\epsilon_0\epsilon r} \right) \phi = E\phi \quad (3.35)$$

where $E(\mathbf{k}_0) = E_c$ is the energy of the band edge of the conduction band. This equation is identical to the equation of the hydrogen atom except for the appearance of the dielectric constant $\epsilon_0\epsilon$ and the effective mass m^* . It follows that such an impurity behaves like a hydrogen atom with the energy measured from E_c .

2. Let V_{ext} be a square well; then close to the band edge the electron will have energy levels in this square well exactly as in free space, but with an effective mass instead of the real mass.
3. Generally, any solved potential problem is solved for crystal electrons close to the band minimum as long as the Fourier components of V_{ext} are smaller than the components of $V(\mathbf{r})$. Assume, for example, that we have a regular lattice of impurities embedded in the host lattice of the semiconductor; that is, V_{ext} is periodic, denoted by $V_{\text{ext}}^{\text{per}}$. Then the Schrödinger equation for isotropic effective mass reads

$$\left(-\frac{\hbar^2}{2m^*} \nabla^2 + V_{\text{ext}}^{\text{per}} \right) \phi = E\phi \quad (3.36)$$

which is the equation for a *band structure*.

Indeed, for large numbers of impurities, one observes an “impurity band” below the conduction band of a semiconductor, as shown in Figure 3.7.

Can we really use the effective mass theorem for an impurity band? One would assume that the Fourier components of the impurity potentials (N_I impurities per cubic centimeter) are small as long as the spacing d between the impurities is much larger than the crystal lattice constant; that is, we have to postulate

$$d = N_I^{-1/3} \gg a \quad (3.37)$$

For typical lattice constants $a \approx 5 \times 10^{-8}$ cm, this means that the impurity concentration is $N_I \leq 10^{19}$ cm⁻³.

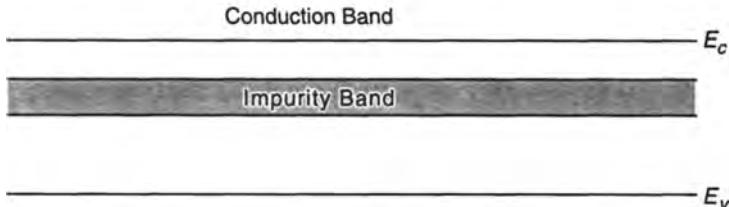


Figure 3.7 Impurity band below conduction band edge in a semiconductor.

For higher densities there is another powerful theorem that gives us again a “method” to “transform away” complications introduced by the crystal structure. This is the virtual crystal approximation, which is discussed later.

3.4 MAXIMA OF ENERGY BANDS—HOLES

Up to now we have mainly looked at the minima of energy bands. Let us look at maxima in the valence band. The new feature is that the effective mass is now *negative*. This brings us to the concept of *holes*.

Assume a valence band as found in the zinc blende (GaAs) and diamond type (Si) and shown in Figure 3.8. First note that there are two $E(\mathbf{k})$ relations present in this band. This means we can have two types of electrons close to the top of the valence band (there is a deeper split-off band caused by spin-orbit interaction, which is not discussed here). We can treat these two types of electrons separately (in Chapter 5 we discuss how to determine their respective numbers). For reasons that will be clarified below, the upper curve is called *heavy hole curve*, while the lower $E(\mathbf{k})$ relation is the *light hole curve*.

We label below all relevant quantities (such as the wave vector) by the subscript el for electrons since we are now introducing holes as additional “quasi-particles.”

Each parabola of Figure 3.8 can be described by

$$E_{\text{el}} = -\frac{\hbar^2 \mathbf{k}_{\text{el}}^2}{2|m^*|} \quad (3.38)$$

Remember that the effective mass at the maximum is negative (second derivative). The electric current density j is obtained by summing overall velocities of filled states \mathbf{v}_k^{el} multiplied by the elementary charge (which is negative for

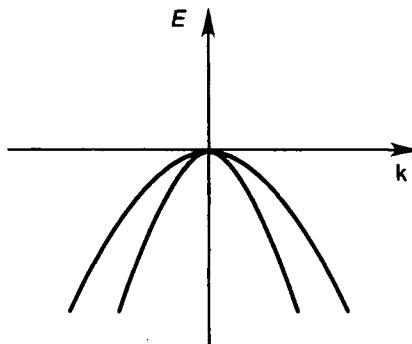


Figure 3.8 Typical form of $E(\mathbf{k})$ relation at the top of the valence band. The deeper split-off band due to the spin-orbit interaction is not shown.

electrons)

$$j = -\frac{e}{V_{\text{ol}}} \sum_{\substack{\mathbf{k}_{\text{el}} \\ (\text{full})}} \mathbf{v}_{\mathbf{k}}^{\text{el}} \quad (3.39)$$

Using the fact that the total wave vector of the electrons in a filled band is zero ($\sum \mathbf{k} = 0$), which follows from the geometrical symmetry of the Brillouin zone, every fundamental lattice type has symmetry under the inversion operation about any lattice point. It follows that the Brillouin zone of the lattice also has inversion symmetry. If the band is filled, all pairs of orbitals \mathbf{k} and $-\mathbf{k}$ are filled, and the total wave vector is zero. Therefore,

$$\sum_{\substack{\mathbf{k}_{\text{el}} \\ (\text{full})}} \mathbf{v}_{\mathbf{k}}^{\text{el}} + \sum_{\substack{\mathbf{k}_{\text{el}} \\ (\text{empty})}} \mathbf{v}_{\mathbf{k}}^{\text{el}} = 0 \quad (3.40)$$

Equations (3.39) and (3.40) give

$$j = +\frac{e}{V_{\text{ol}}} \sum_{\substack{\mathbf{k}_{\text{el}} \\ (\text{empty})}} \mathbf{v}_{\mathbf{k}}^{\text{el}} \quad (3.41)$$

In most practical cases, Eq. (3.41) is much more convenient than Eq. (3.39), because one has fewer empty states than full states. The question arises of whether we can describe the current by a fictitious particle with positive charge. Also, we would prefer particles with positive mass (as we are used to) and we would like to attribute the empty states to these quasi-particles because we prefer to sum over the smaller number of states. Does such a quasi-particle exist? Before we can answer that question, we have to ask ourselves how the empty states evolve in time, because our quasi-particles have to do exactly the same.

The unoccupied levels in a band evolve in time under the influence of applied fields precisely as they would if they were occupied by real electrons. This is so because, given the values of \mathbf{k} and \mathbf{r} at $t = 0$, the semiclassical Eqs. (3.25) and

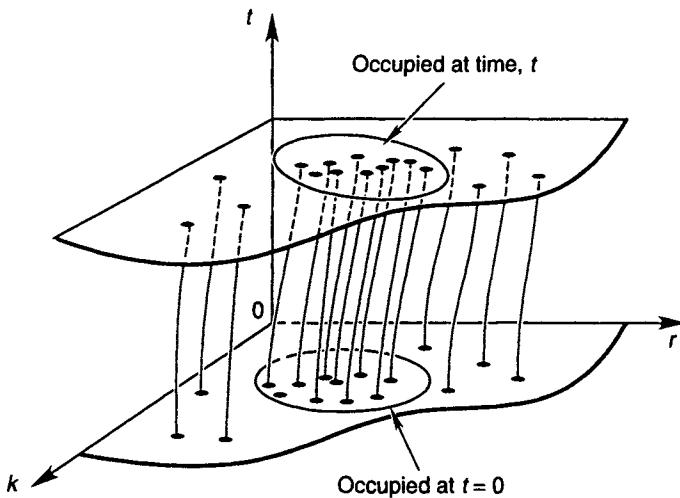


Figure 3.9 Schematic illustration of the time evolution of orbits in semiclassical phase space (here r and k are each indicated by a single coordinate). The occupied region at time t is determined by the orbits that lie in the occupied region at time $t = 0$. [From Solid State Physics by Neil W. Ashcroft and N. David Mermin. Copyright ©1976 by Holt, Rinehart & Winston [1]. Reprinted by permission.]

(3.26), being six first-order equations in six variables, uniquely determine \mathbf{k} and \mathbf{r} at all subsequent (and all prior) times, just as in ordinary classical mechanics.

The position and momentum of a particle at any instant determine the entire orbit in the presence of specified external fields. In Figure 3.9 we indicate schematically the orbits determined by the semiclassical equations, as lines in a seven-dimensional rkt space. Because any point on an orbit uniquely specifies the entire orbit, two distinct orbits can have no points in common. We can, therefore, separate the orbits into occupied and unoccupied orbits (see Figure 3.9) according to whether they contain occupied or unoccupied points at $t = 0$. At any time after $t = 0$, the unoccupied levels will lie on unoccupied orbits, and the occupied levels on occupied orbits. Thus the evolution of both occupied and unoccupied levels is completely determined by the structure of the orbits.

This structure depends only on the form of the semiclassical equations and not on whether an electron happens actually to be following a particular orbit. This means the empty states evolve in the seven-dimensional rkt space according to Eqs. (3.25) and (3.26)

$$\frac{d\hbar\mathbf{k}_{el}}{dt} = -e\mathbf{F} \quad (3.42)$$

(\mathbf{F} is the electric field), as well as Eq. (3.25)

$$\mathbf{v}_k^{el} = \frac{1}{\hbar} \nabla_{k_{el}} E_{el}(\mathbf{k}_{el}) \quad (3.43)$$

Can a quasi-particle, which we will call a “hole,” have positive charge and

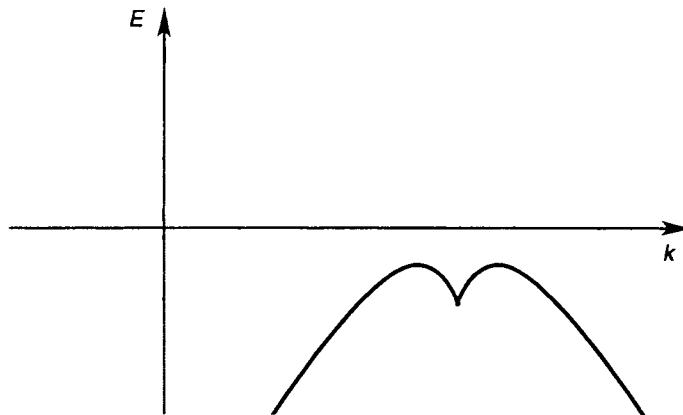


Figure 3.10 Schematic shape of the top of the valence band in tellurium.

positive mass and satisfy Eqs. (3.41), (3.42), and (3.43), and can it be identified with the missing electron(s)? In other words, are the following equations equivalent to Eqs. (3.41), (3.42), and (3.43)?

$$\mathbf{j} = \frac{e}{V_{\text{ol}}} \sum_{\mathbf{k}_h} \mathbf{v}_k^h \quad (3.44)$$

$$\frac{d\hbar\mathbf{k}_h}{dt} = e\mathbf{F} \quad (3.45)$$

$$\mathbf{v}_k^h = \frac{1}{\hbar} \nabla_{\mathbf{k}_h} E_h(\mathbf{k}_h) \quad (3.46)$$

$$E_h = \frac{\hbar^2 \mathbf{k}_h^2}{2|m^*|} \quad (3.47)$$

(h stands for hole).

The equations are equivalent if we have

$$\mathbf{k}_h = -\mathbf{k}_{\text{el}} \quad (3.48)$$

Since then

$$\nabla_{\mathbf{k}_h} = -\nabla_{\mathbf{k}_{\text{el}}} \quad (3.49)$$

and therefore if we measure E_h downwards, that is, $E_h = -E_{\text{el}}$ (or count the mass of the hole positive)

$$\mathbf{v}_k^h = \mathbf{v}_k^{\text{el}} \quad (3.50)$$

and Eq. (3.41) is satisfied and looks like an equation for positive charge. Because the time variable t was not transformed and because of Eq. (3.50), our holes evolve in the $\mathbf{r}t$ space like electrons. In \mathbf{k} space, however, this is not so; there our fictitious particles evolve on the negative side [Eq. (3.48)] or can be seen from the defining equation (3.48).

The effective mass at the top of the valence bands of semiconductors is not always negative. In tellurium, for example, the $E(\mathbf{k})$ relation has the shape shown in Figure 3.10. It is then not entirely appropriate to identify missing electrons by holes.

The effective masses can, in principle, be calculated by the pseudopotential method and Eq. (3.30). They also can be measured by cyclotron resonance and other methods, and are by now very well known for most semiconductors.

3.5 SUMMARY OF IMPORTANT BAND-STRUCTURE PARAMETERS

It is interesting to note the generally complex anisotropic shape of the $E(\mathbf{k})$ relation in the valence band. The effective mass is very anisotropic, and any experiment will measure complicated averages corresponding to the specifics of the experiment. It is, therefore, difficult to give a single effective mass for either the heavy or the light hole band. Nevertheless, we have compiled some approximate values in Table 3.1 (the density of states masses, which we will discuss in a later chapter). Also shown in Table 3.1 are other important properties of semiconductors, such as the energy gap and the dielectric constant.

Figure 3.11 summarizes graphically the main features of conduction and valence bands in the important semiconductors. This finishes our discussion of bulk semiconductor band structures. Before we proceed to the discussion of imperfections, however, we discuss the energy bands of alloys and of junctions in heterojunctions semiconductors.

3.6 BAND STRUCTURE OF ALLOYS

In the discussion of alloys, we are mainly interested in ternary alloys, such as $\text{Al}_x\text{Ga}_{1-x}\text{As}$, where we have added a certain mole fraction x of Al to GaAs. We still have a perfect crystal, but some Ga atoms are randomly replaced by Al atoms (compositional disorder). A schematic sketch of the potential of such an alloy is shown in Figure 3.12.

How do we calculate the band structure in this case? We have seen that we cannot stretch the effective mass theorem too far to account for an “impurity crystal” of similar lattice constant comparable to the host crystal. Such a potential *cannot* be divided into $V(r) + V_{\text{ext}}$ with the Fourier components of $V_{\text{ext}} \ll V(r)$. There is still a method, however, that allows a simple treatment of such a system. This method is called the *virtual crystal approximation*.

The virtual crystal approximation was used by Nordheim to suggest that for a crystal such as $\text{Al}_x\text{Ga}_{1-x}\text{As}$, we can replace the actual potential by an average potential

$$V_{\text{av}} = (1 - x)V_{\text{GaAs}} + xV_{\text{AlAs}} \quad (3.51)$$

where V_{GaAs} and V_{AlAs} are the crystal potentials of GaAs and AlAs, respectively.

Table 3.1 Approximate Values of Material Parameters

	Si	Ge	AlAs	AlSb	GaAs	InP	InAs	InSb
E_G (300K)	1.11	0.65	2.17	1.62	1.439	1.35	0.356	0.180
E_G (4.2K)	1.21	0.74	2.22	1.70	1.52	1.42	0.409	0.235
a (Å) lattice constant	5.431	5.658	5.661	6.138	5.642	5.868	6.058	6.479
ρ (g/cm ³)	2.39	5.323	3.73	4.26	5.36	4.787	5.667	5.775
m_e^*	0.19/0.916	0.082/1.6	0.35	0.39	0.067	0.078	0.023	0.013
m_{h1}^*	0.15	0.042	0.15	0.11	0.087		0.025	0.16
m_{h2}^*	0.52	0.34	0.76	0.5	0.462	0.8	0.41	0.43
Phonon Lo/T ₀ (meV)	51.0/57.4	28.1/33.3	50.1/44.9	42.1/39.5	36.2/33.3	42.8/37.7	30.2/27.1	24.2/22.6
Index of Refraction	3.5	4.1	3.1	3.4	3.4	3.1	3.5	3.9
$\epsilon(\omega \rightarrow \infty)$	11.7	16.0	8.16	10.24	10.92	9.52	11.8	15.7
$\epsilon(\omega \rightarrow 0)$	11.7	16.0	10.06	14.4	12.9	12.35	14.55	17.72

m_{h1}^*, m_{h2}^* denote the effective density of states masses for electrons, light holes, and heavy holes, respectively [see Eq. (5.12)]. The values are given in terms of the free electron mass. $\epsilon(\omega \rightarrow \infty)$ is the dielectric constant for high frequencies ω , and $\epsilon(\omega \rightarrow 0)$ in the static dielectric constant.

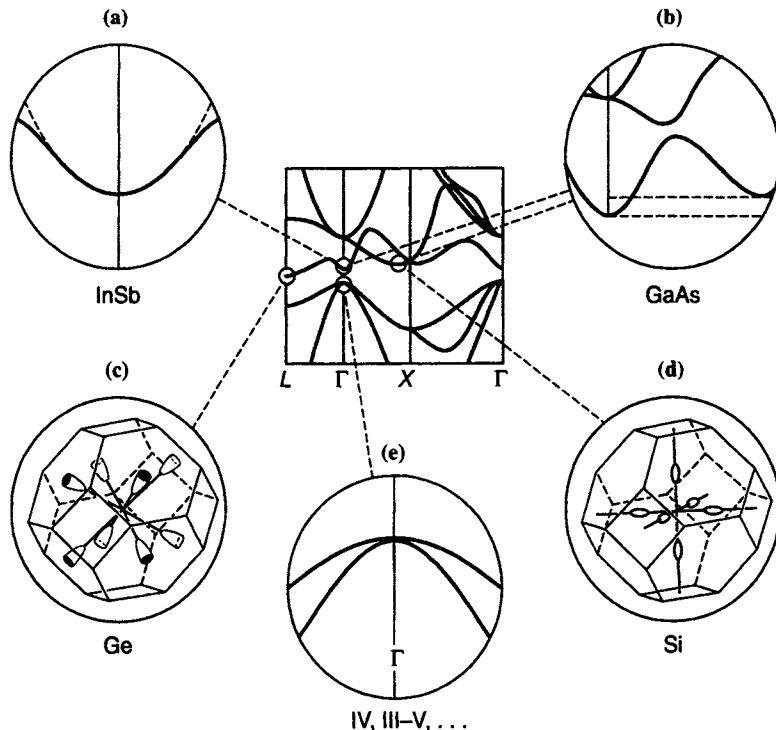


Figure 3.11 Position and form of the conduction band minima (*a*, *b*, *c*, *d*) and valence band maxima (*e*) of important semiconductors. Notice the ellipsoidal form for silicon (*X*) and germanium (*L*) minima. The minima at Γ are spherical. [After Madelung [8], Figure 17.]

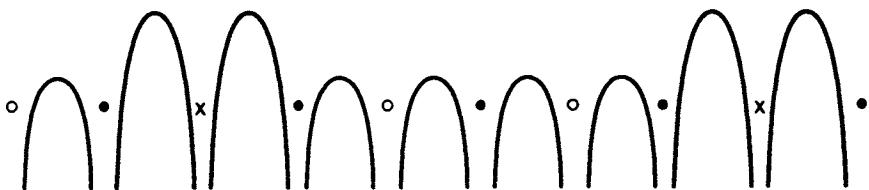


Figure 3.12 Sketch of the atomic potentials of the ternary alloy $\text{Al}_x\text{Ga}_{1-x}\text{As}$.

In the empirical pseudopotential method we can easily include Eq. (3.51) to calculate the necessary form factors if they are known for each compound. In many cases a linear interpolation approximates complex physical quantities well such as the effective mass, the dielectric constant ϵ , and so on. Thus we have

$$\epsilon_{\text{Al}_x\text{Ga}_{1-x}\text{As}} \approx (1 - x)\epsilon_{\text{GaAs}} + x\epsilon_{\text{AlAs}} \quad (3.52)$$

A similar equation can be written for the energy gap between the top of the valence band and the Γ , *X*, and *L* minima. The typical dependencies for this

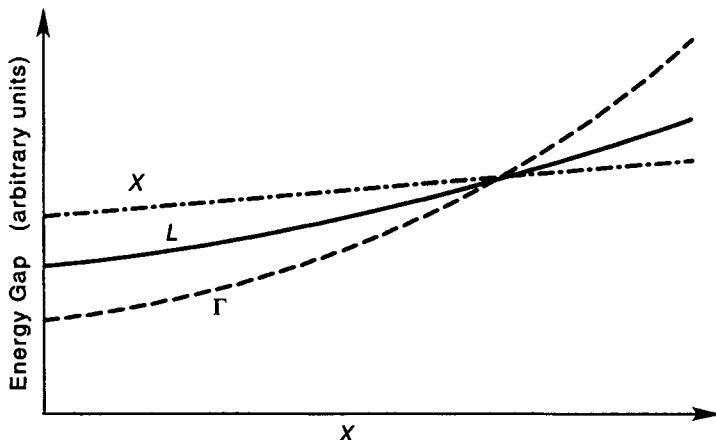


Figure 3.13 Γ , X , L conduction band minima of a typical compound $A_xB_{1-x}C$ as a function of mole fraction x .

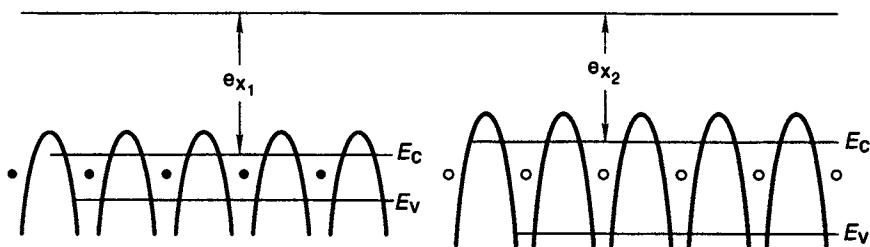


Figure 3.14 Schematic plot of the potential of atoms in two neighboring lattice-matched semiconductors. E_c and E_v denote the conduction and valence band edges, respectively.

type of alloy are shown in Figure 3.13 and do not deviate much from the linear approximation. As shown in the figure, a “disorder bowing” of the energy gap is measured. Methods have been developed to calculate this bowing (Jones and March [6]), but they go beyond the scope of this text.

Alloys are finding increasing use in semiconductor electronics (see Casey and Panish [2] for an overview of many useful material systems). Of special interest are lattice-matched layers (heterolayers) of semiconductors. “Lattice-matched” means that the interatomic distance is the same in two neighboring semiconductor layers, which gives the basic possibility of generating ideal interfaces without defects and distortions. Notice, however, that lattice match is only a necessary condition for perfect interfaces, not a sufficient one.

Figure 3.14 shows schematically the atomic potentials near the junction of two materials. The energies $e\chi_1$ and $e\chi_2$ are necessary to remove an electron from the respective conduction band edge to an infinite point, the vacuum level.

Notice that χ_1 and χ_2 are only loosely connected to quantities that are easy to

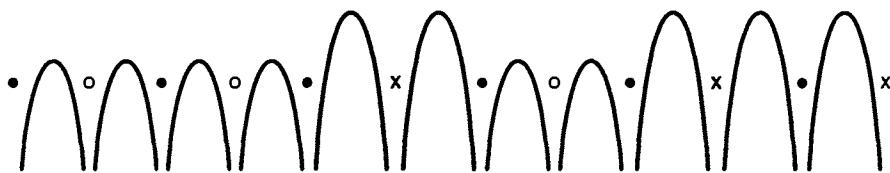


Figure 3.15 Heterointerface between binary and ternary alloys.

measure, such as the *photothreshold*. The reason is that image force and surface reconstruction (see end of chapter) contribute significantly to the photothreshold. Therefore, the absolute values $e\chi_1$ and $e\chi_2$ relative to the vacuum level are not very useful. The relative values, the variation from material to material $e\chi_1 - e\chi_2$, can be expected to be more meaningful, although a calculation is still difficult. Even if there were no surface and image force effects, the quantities $e\chi_1$ and $e\chi_2$ are large and their difference is small, which will give enormous requirements of accuracy to our band-structure calculation to obtain the correct difference (Harrison [5], p. 252). The difference between the conduction (valence) band edges in the two materials is called the band edge discontinuity $\Delta E_c(\Delta E_v)$.

If a ternary alloy is the “atomic” neighbor of a binary compound, we have a natural transition region at the interface (Figure 3.15). It is important to realize that in this case the transition between the two materials can go over a single or many atomic layers.

It is also important to realize that a heterojunction is not the same as a potential step, since the effective mass can be different on the two junction sides. If one uses effective mass theory, one has, therefore, a problem in matching wave functions at the interface since only the true (Bloch) wave functions and their derivatives are continuous, and not the effective mass (quasi-free) wave functions.

One still can assume, however, that the quasi-free electron wave functions (envelope functions) in the two different materials away from the junction satisfy the relation:

$$\left(\begin{array}{c} \phi \\ \partial\phi/\partial z \end{array} \right)_{\text{GaAs}} = \left(\begin{array}{cc} T_{11} & T_{12} \\ T_{21} & T_{22} \end{array} \right) \left(\begin{array}{c} \phi \\ \partial\phi/\partial z \end{array} \right)_{\text{AlGaAs}} \quad (3.53)$$

T_{ij} is called the transition matrix; for the $\text{Al}_x\text{Ga}_{1-x}\text{As}$ and $x \leq 0.3$, it is close to the unit matrix, with $T_{22} = (m_{\text{GaAs}}^*/m_{\text{AlGaAs}}^*)$ and $T_{11} \approx 1$, while $T_{12} = T_{21} = 0$. Of course, to obtain the exact transition matrix, the band structure of the heterojunction needs to be calculated, by the pseudopotential method, for example (Frensley and Kroemer [4]).

PROBLEMS

- 3.1 Calculate the wave functions corresponding to the two solutions of Eq. (3.8) and show that their form corresponds to the bonding-antibonding case.

3.2 The spacing between successive (100) planes in NaCl is 2.82Å; suppose that a given X ray is found to give rise to first-order Bragg reflection at a grazing angle of $11^\circ 57'$. Find the wavelength of the X-ray and the angle at which the second-order Bragg reflection would occur.

3.3 From the Schrödinger equation

$$\left(\frac{-\hbar^2}{2m} \nabla^2 + V \right) \Psi_{\mathbf{k}} = E_{\mathbf{k}} \Psi_{\mathbf{k}}$$

(a) Derive

$$\frac{2m}{\hbar^2} \Psi_{\mathbf{k}} \nabla_{\mathbf{k}} E_{\mathbf{k}} = -2i \nabla \Psi_{\mathbf{k}} - \left[\nabla^2 + \frac{2m}{\hbar^2} (E_{\mathbf{k}} - V) \right] e^{i\mathbf{k}\cdot\mathbf{r}} \nabla_{\mathbf{k}} u_{\mathbf{k}}$$

for the Bloch function, $\Psi_{\mathbf{k}} = e^{i\mathbf{k}\cdot\mathbf{r}} u_{\mathbf{k}}$. (*Hint:* Differentiate with respect to \mathbf{k} .)

(b) Using the result of part a, show that

$$\mathbf{v} = \frac{1}{\hbar} \nabla_{\mathbf{k}} E_{\mathbf{k}}$$

where

$$\mathbf{v} = \frac{\hbar}{im} \int \Psi_{\mathbf{k}} \nabla \Psi_{\mathbf{k}}$$

3.4 Find the energy gap corresponding to the potential $V_0 \cos\left(\frac{2\pi r}{a}\right)$ for a one-dimensional crystal with lattice constant a .

REFERENCES

- [1] Ashcroft, N. W., and Mermin, N. D. *Solid State Physics*. Philadelphia: Saunders, 1976, p. 227.
- [2] Casey, H. C. Jr., and Panish, M. B. *Heterostructure Lasers. Part B. Materials and Operating Characteristics*. New York: Academic Press, 1978, pp. 15–48.
- [3] Cohen, M. L., and Bergstresser, T. K. “Band structures and pseudopotential form factors for fourteen semiconductors of the diamond and zinc-blend structures,” *Physical Review*, vol. 141, 1966, pp. 789–796.
- [4] Frensel, W. R., and Kroemer, H. “Predictions of semiconductor heterojunction discontinuities from bulk band structures,” *Journal of Vacuum Science and Technology*, vol. 13 (X10), 1976, pp. 810–815.
- [5] Harrison, W. A. *Electronic Structure and the Properties of Solids*. San Francisco: Freeman, 1980, pp. 61–80.
- [6] Jones, W., and March, N. *Theoretical Solid State Physics*, vol. 2. New York: Wiley/Interscience, 1973, p. 1093.
- [7] Landsberg, P. T. *Solid State Theory*, New York: Wiley/Interscience, 1969, pp. 305–306.
- [8] Madelung, O. *Introduction to Solid State Theory*. New York: Springer-Verlag, 1978, pp. 62–64.
- [9] Shichijo, H., and Hess, K. “Band structure dependent transport and impact ionization in GaAs,” *Physical Review B*, vol. 23, 1981, pp. 4197–4207.

CHAPTER 4

IMPERFECTIONS OF IDEAL CRYSTAL STRUCTURE

With the exception of the lattice vibrations in Chapter 1, we have discussed up to now only ideal crystals. Real crystals are not periodic with respect to translations, and other symmetry operations owing to imperfection in their structure. Unavoidable imperfections are the phonons. An electron traveling in a crystal is scattered after a typical distance of the order of 100\AA (depending on energy) by phonon emission or absorption. This means that the electron does not really “see” the crystal as a whole but instead only a volume of 10^{-18} cm^{-3} , which contains about 3000 atoms before it is dephased (see Chapter 8) owing to interactions with phonons. In addition to the phonons, there are a number of avoidable imperfections. We distinguish the following:

- Point defects* impurities, vacancies
- Line defects* dislocations
- Areal defects* surfaces, interfaces
- Total defects* glasses, amorphous solids

Adequate control of these imperfections is necessary if a material system is to be used to fabricate electron devices (“device grade material”). As will become clear in the following chapters, point defects need to be controlled on a subpart per million level. Ideally, one wants to control the density of point defects on a level of 10^{14} cm^{-3} , which means having only one imperfection in about 10^8 atoms. High grade semiconductor material should be virtually free of line defects (dislocations), and indeed, silicon can be grown virtually free of dislocations.

As a rule of thumb, one can say that no high-speed or highly integrated semiconductor device should include bare surfaces in its operating parts, because of the detrimental effects of surface states. Also, it is difficult to see how amorphous solids can form the basic electron transport medium for advanced devices. Amorphous solids are rather expected to be useful for the fabrication of necessarily cheap devices with simple functions, such as solar cells.

Imperfections, in particular surfaces, have indeed presented major obstacles for the early development of device ideas. Shockley's ideas to develop a field effect device ran into major obstacles because the electrons that were induced by the field effect and supposed to switch on a current were captured in localized electron states at the surface (surface states or traps) that did not permit nearly free electron propagation as the perfectly periodic solid does. It was Bardeen who identified the surface states as the underlying problem and in the course of this work discovered the point contact transistor together with Brattain.

The point defects, on the other hand, can not only be useful, but some of them—substitutional atoms of various kind (dopants)—indeed form the basis of practically all existing devices. These are therefore described in some detail in the following section.

4.1 SHALLOW IMPURITY LEVELS—DOPANTS

Many early investigations concentrated on point defects that can be classified into interstitial and substitutional impurities and vacancies (Figures 4.1 and 4.2). The most important imperfections in semiconductors (in a positive sense) are the substitutional impurities that can be used as “dopants.” The meaning of the word “dopant” becomes clear from the following. Imagine that we replace in a perfect silicon lattice one silicon atom by arsenic, as shown in Figure 4.2a. Because arsenic has five valence electrons and silicon only has four, one electron is redundant and if it is “donated” to the conduction band, it can move freely in the crystal and contribute to the conductivity of silicon. The conditions that determine that the electron resides in the conduction band are discussed in Chapter 5. Let us now quantitatively evaluate what the arsenic atom introduces in the spectrum of energy levels.

If one electron has propagated away, there remains a charged arsenic ion having a potential energy

$$V_{\text{As}} = -\frac{e^2}{4\pi\epsilon\epsilon_0 r} \times (\text{screening factors}) + V_2 \quad (4.1)$$

The screening factors are typically of the form $e^{-r/L}$ where L is a characteristic length determined by the presence of other charges (see Chapter 6). The potential V_2 is a complicated contribution that depends on the character of the impurity atom. Its size compared to that of the host atoms will determine the rearrangement of the host atoms and therefore result in various contributions, V_2 . For low densities of impurities like arsenic (so that the impurities are independent of each other) we have, roughly, $V_2 \approx 0$ and $L \approx \infty$, so that

$$V_{\text{As}} = -\frac{e^2}{4\pi\epsilon\epsilon_0 r} \quad (4.2)$$

We know, then, from the effective mass theorem that the arsenic behaves like a

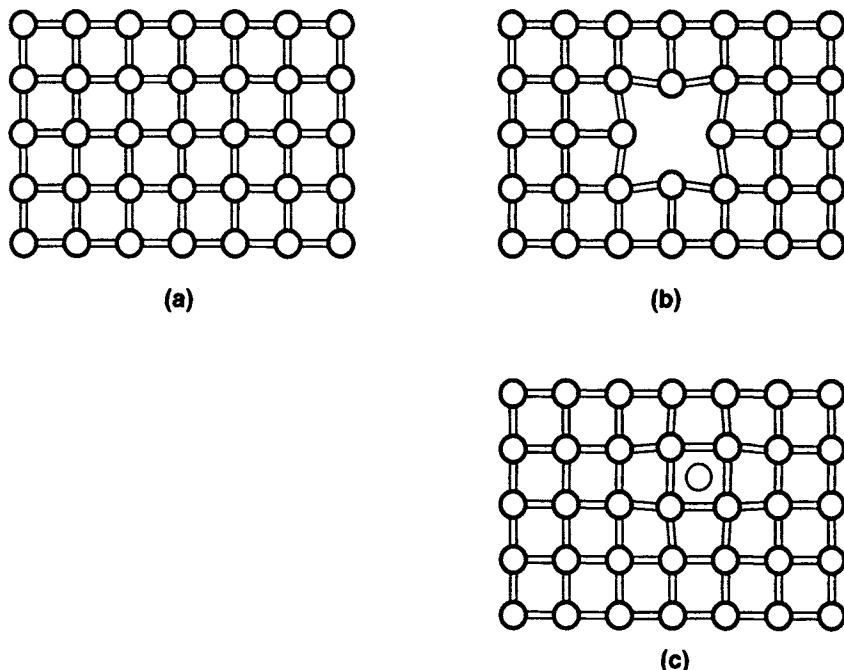


Figure 4.1 (a) Perfect crystal; (b) vacancy; (c) interstitial atom.

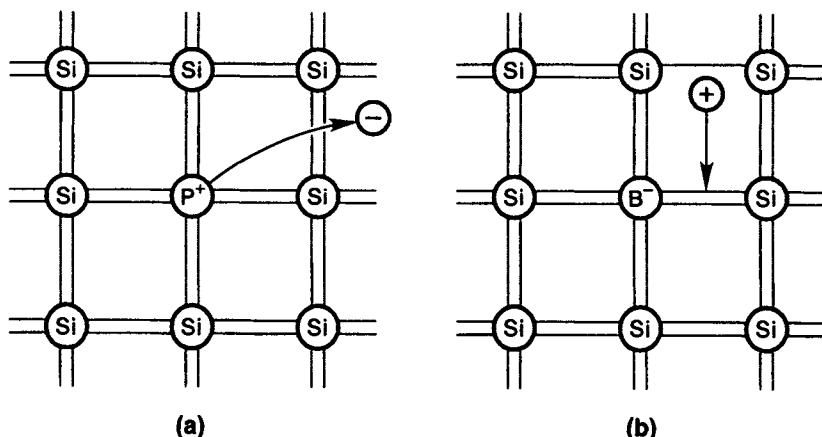


Figure 4.2 (a) Column V substitutional donor impurity, and (b) column III substitutional acceptor impurity in silicon. Notice that, of course, the electrons (holes) indicated in the figure are delocalized in the conduction and valence band, respectively.

hydrogen atom, which gives the energy levels (see any text on quantum mechanics)

$$(E_n - E_c) = -\frac{m^* e^4}{32\pi^2 n^2 \hbar^2 \epsilon^2 \epsilon_0^2} \quad (4.3)$$

and a Bohr radius of

$$a_n = \frac{4\pi n^2 \hbar^2 \epsilon \epsilon_0}{m^* e^2} \quad (4.4)$$

It is important to note that $E_n - E_c$ is reduced (therefore, the name shallow impurity) compared to the free space value (~ 13 eV) by the factor $m^*/\epsilon^2 \sim 10^{-3} m$, while the Bohr radius is increased by a factor $\epsilon/m^* \sim 100/m$, where m is the mass of the free electron.

The increase in the Bohr radius justifies *a posteriori* our choice of the potential given in Eq. (4.2), which uses a dielectric constant and therefore assumes that the electron “sees” the crystal medium and has a large orbit around the positive charge. Indeed, a typical Bohr radius in a III-V compound is of the order of 100 Å.

It is customary to represent impurities by various graphs in real space and \mathbf{k} space. These are shown in Figures 4.3a through 4.3c. The different lengths of the ground state and the excited states in real space and \mathbf{k} space correspond to the uncertainty principle: In the ground state the electron is most localized in real space (short line), which means that the uncertainty of \mathbf{k} is large (long line in \mathbf{k} space graph).

Similar arguments as have been advanced for elements that have one (or more) electrons too many and donate it (donors) can be made for elements that have not enough electrons. These accept electrons (acceptors) in the valence band and create holes. The theory is entirely symmetric to the donor case and is also derived most easily by invoking the effective mass theorem.

Complications of the simple model outlined above arise, for example, from the anisotropy of the silicon conduction band valleys, which is expressed by the term $-\frac{\hbar^2}{2m_j^*} \frac{\partial^2}{\partial x_j^2}$ in the Hamiltonian equation [Eq. (3.34)]. This anisotropy causes

a splitting of the excited states. Another more important complication arises from the overlap of impurity levels at higher doping densities, which leads to impurity bands as has been described in Chapter 3 and will be dealt with in more detail in Chapter 5.

4.2 DEEP IMPURITY LEVELS

Electrons can also be bound closer to the nucleus; for example, if they have multiple charge or if the effective mass is large. Then the above approach, using a dielectric constant, is not justified and the energy levels of the impurities can

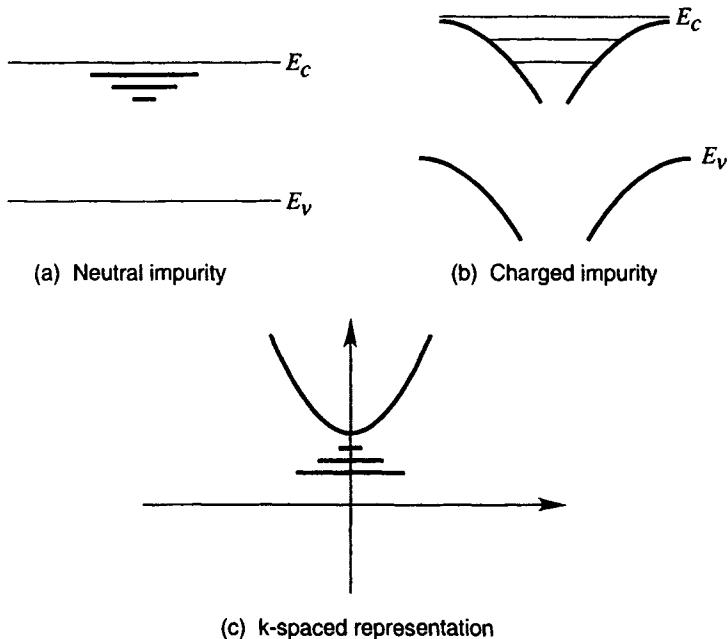


Figure 4.3 Graphs characterizing a donor impurity in a crystal. Parts (a) and (b) are graphs in real space; part (c) is a \mathbf{k} space representation.

be quite different from Eq. (4.3). The energies $E_n - E_c$ are then often larger, and one speaks about deep impurity levels. The theory of deep impurity levels is involved; it is the potential close to the nucleus and the core electrons that matters most and that binds the electrons strongly. The relevant energy scale is then electron volts (instead of milli-electron volts), and minor approximations in a theoretical calculation may shift the energy level from deep in the gap to the conduction band, or vice versa. Any impurity can, in principle, produce both deep and shallow levels—and deep does not necessarily mean that the level lies deep in the energy gap. It only means that the energy levels are connected to the more central part of the impurity potential.

The binding of electrons closer to the nucleus means necessarily that electrons in such state will not as easily be "donated" to the conduction band as from the shallow donors. Clearly thermal activation from an energy level close to the middle of the energy gap is much more unlikely than from a state close to the conduction band edge. The deep levels are therefore not of the same elementary use (donors, acceptors) than the shallow levels. These levels form "electron traps" or electron "recombination centers," as described in Chapter 9. Although they are still important for device operation, they are often not desirable. For example, the capture of one electron in a deep level and the subsequent capture of a hole in the same level results in the destruction of an electron hole pair that can be undesirable because it competes with optical recombination (and therefore

light generation), which makes it a detriment to the operation of optoelectronic devices. Therefore, from a fabrication point of view, it is often desirable to eliminate elements and, generally, defects that form energy levels deep in the energy gap (close to the middle, as will be seen in Chapter 9). From the point of view of theoretical physics, deep levels form a complex problem. In fact, no theory exists that can tell where the energy levels of any substitutional element in any semiconductor really are located. A theory that comes close to being able to predict energy levels in general has been developed by Dow [1], and the interested reader is referred to his treatment.

We can in this book gloss over the theory of deep impurity levels in this way, because their effects can be included phenomenologically into the theory and simulation of semiconductor devices. All that is needed to describe the action of deep levels in devices are capture and emission cross sections, as described in Chapter 9. These can be deduced most reliably from experiments and somewhat less reliably from existing theories.

4.3 DISLOCATIONS, SURFACES, AND INTERFACES

One-dimensional imperfections (dislocations) have only negative aspects with respect to applications (devices), and one tries to avoid them by elaborate crystal-growth techniques. In Si and GaAs, the art has advanced to such a degree that the crystals can be produced dislocation free. Two-dimensional defects, surfaces, and interfaces, are important in semiconductor electronics and are discussed to some extent below.

A simple-minded chemist's picture of a surface is shown in Figure 4.4. Notice that each silicon atom in the bulk of the semiconductor is surrounded by eight valence electrons and one electron is missing at the surface. As a consequence, a "surface state" is created in analogy to an impurity state. Its energy is typically somewhere in the middle of the energy gap. These states can capture electrons and therefore are highly perturbing in the operation of semiconductor devices. *Surfaces are, therefore, avoided in semiconductor electronics and replaced by (more or less ideal) interfaces.*

Actually, this picture of surface states is artificial, and the only reason it is still discussed in this way in almost all texts is its simplicity. Surfaces do not exist in the simple form as implied by the figure. The surface atoms attempt to relax to a state of lower energy and pair with each other. One says the surface reconstructs and patterns are formed, as demonstrated in Figure 4.5, that allow for additional bonding of the surface atoms. When electrons are added at the surface, the reconstruction pattern may change and the electron is trapped at the surface, as also implied by the simple model. A more detailed discussion is given by Harrison [2].

To avoid trapping electrons in device applications, surfaces are replaced by interfaces. In fact, one of the biggest advantages of silicon is that it forms an

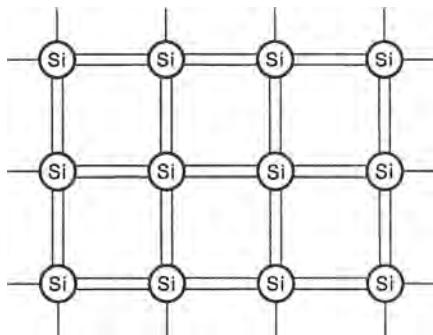


Figure 4.4 Simple-minded explanation of surface states.

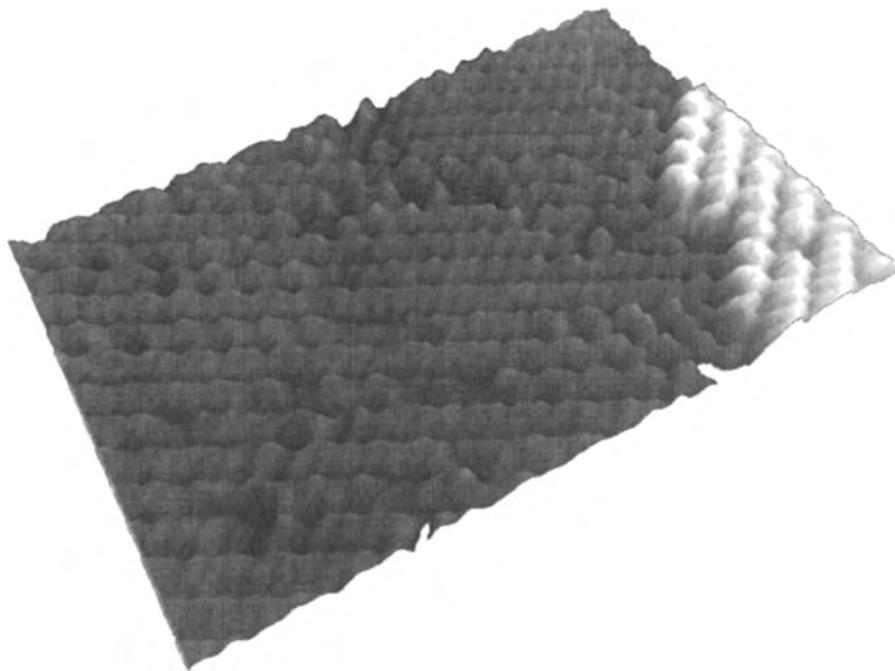


Figure 4.5 Reconstruction pattern of (100) silicon surface. Neighboring atoms pair up to form rows of pairs that can be seen in the figure. As shown, there are domains with different row orientation. [Courtesy of J. W. Lyding.]

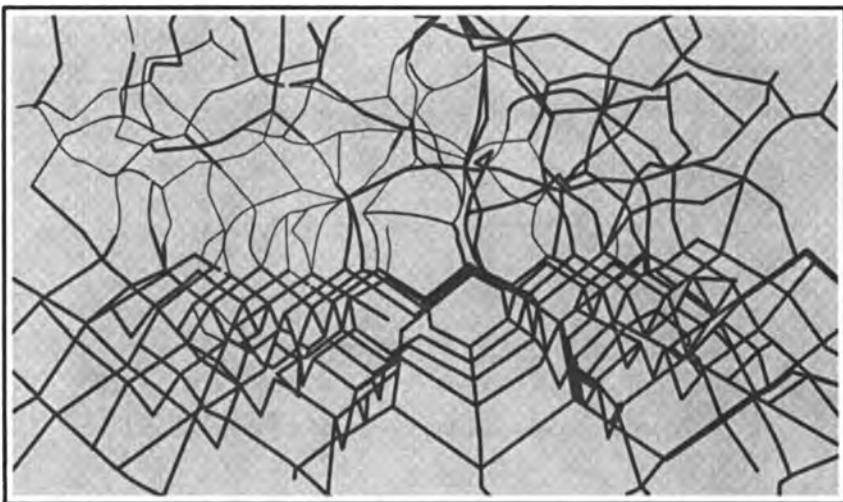


Figure 4.6 Si-SiO₂ interface model. Notice that SiO₂ growing on silicon is not a regular crystal but rather amorphous. However, on short range the SiO₂ atomic arrangement is still almost tetrahedral and the bonding angles are distorted on average only by $\sim \pm 15^\circ$. [After S. T. Pantelides, and M. Long, in *The Physics of SiO₂ and its Interfaces*, S. T. Pantelides, ed., Figure 1. Copyright ©1978 by Pergamon Books Ltd. Reprinted with permission [4].]

almost perfect interface with its dioxide. The topology of the interface Si-SiO₂ is still not exactly known; however, it is believed that amorphous SiO₂ grows on top of Si with only a few “dangling bonds” left and with very little bond angle distortion at the interface. This is shown in Figure 4.6. The leftover dangling bonds can be saturated, for example, with hydrogen, which is made available in the oxidation process. State-of-the-art material has only about 10^{10} cm^{-2} interface traps left (out of 10^{15} cm^{-3}), which is good enough for metal–oxide–silicon transistor applications. Of course, from the viewpoint of the interface, the lattice-matched GaAs-AlAs system is still “infinitely” better.

Let me repeat in passing that lattice match is not the only criterion necessary to achieve ideal interfaces; there are other criteria. For example, if GaAs is grown on top of germanium, the Ga and As atoms have to pair with the correct atom (of the two possible) in the Ge unit cell. They do not always do this because they are usually not “smart.” One, therefore, talks about the “site allocation problem,” which makes the GaAs-Ge interface nonideal even though the materials are lattice-matched. There are other problems of this kind, which have been discussed extensively by Kroemer, [3].

PROBLEMS

- 4.1 Indium arsenide has $E_G = 0.40$ eV; dielectric constant $\epsilon = 15$; electron effective mass $m^* = 0.023m$. Calculate
- the donor ionization energy
 - the radius of the ground state orbit
 - the minimum donor concentration at which appreciable overlap effects between the orbits of adjacent impurity atoms will occur.

REFERENCES

- [1] Dow, J. D. "Localized Perturbations in Semiconductors," in *Highlights of Condensed Matter Theory*, LXXXIX Corso Soc. Italiana di Fisica, Bologna, Italy, 1985.
- [2] Harrison, W. A. *Electronic Structure and the Property of Solids*. San Francisco: Freeman, 1980, p. 229.
- [3] Kroemer, H., Polasko, K. J., and Wright, S. C. "On the (110) orientation as the preferred orientation for the molecular beam epitaxial growth of GaAs on Ge, GaP on Si, and similar zincblend-on-diamond systems," *Applied Physics Letters*, vol. 36, 1980, pp. 763–765.
- [4] Pantelides, S. T., and Long, M., "Continuous-Random-Network Models for the Si-SiO₂ Interface," in *The Physics of SiO₂ and Its Interfaces*, ed. S. T. Pantelides. New York: Pergamon, 1978, pp. 339–343.

CHAPTER 5

EQUILIBRIUM STATISTICS FOR ELECTRONS AND HOLES

Although we discussed the energy band structure (the electronic states) of a semiconductor in detail in the previous chapters, we did not include in our discussion whether these states are actually filled with electrons (two with opposite spin in each state are possible) or not.

In Chapter 4 we have shown, however, that full bands do not contribute to the electronic current. In a defect-free semiconductor at $T = 0$, all bands up to the so-called conduction band are filled. The last filled band is the valence band. As the temperature increases, electrons from the valence band will be excited to the conduction band, and the electrons and holes generated in this way will be able to conduct a current (not as large as in a metal where valence and conduction bands overlap). We also can introduce electrons and holes by doping. To compute the conductivity, we need to know which states are occupied and which are empty. This knowledge is usually acquired by calculating the probability that a state is occupied and by calculating the density of states. The actual carrier occupation is then proportional to the product of these two quantities, which are treated separately below.

5.1 DENSITY OF STATES

Consider a crystal with periodic boundary conditions and N atoms along each of the main coordinate axes. Then, according to Eq. (2.19), the allowed wave vectors are given by

$$\mathbf{k} = \frac{\mathbf{K}_h}{N} = \frac{h_1}{N}\mathbf{b}_1 + \frac{h_2}{N}\mathbf{b}_2 + \frac{h_3}{N}\mathbf{b}_3 \quad (5.1)$$

with

$$0 \leq |h_i| \leq N/2$$

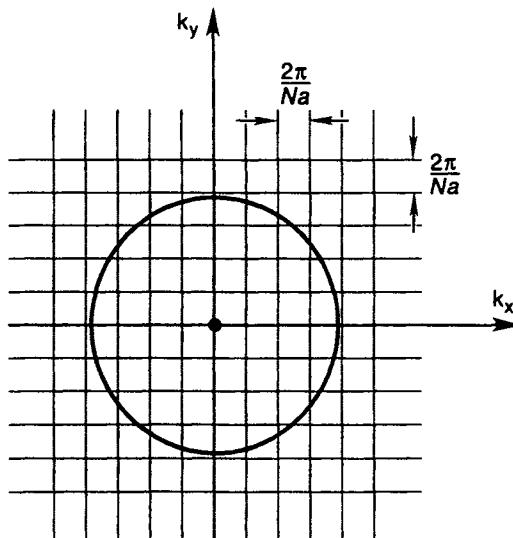


Figure 5.1 Allowed values for the \mathbf{k} vector in two dimensions.

In a simple cubic crystal, \mathbf{b}_1 would be of the form $2\pi/a$ times unit vector (and similarly \mathbf{b}_2 and \mathbf{b}_3), and therefore, the typical form of \mathbf{k} is

$$\mathbf{k} = \frac{2\pi\mathbf{h}_1}{Na} \times \text{unit vector} + \dots \quad (5.2)$$

The components of the allowed \mathbf{k} values form the “lattice” shown in Figure 5.1, where $Na = L$ is the length of the crystal in each direction.

This treatment involving the crystal length and the periodic boundary conditions, may seem somewhat artificial. However, because the end results will not depend on the quantities (N , L , etc.) connected to these artificial conditions, we need not be concerned about it. It is the advantage of periodic boundary conditions that their proper application always gives correct results for bulk properties. Of course, they are invalid when surface properties become important.

We are interested in the number of states at the energy E in the interval $[E, E + dE]$. To calculate this number, we first assume the simple case of *spherical* constant energy surfaces in \mathbf{k} space. The number of allowed \mathbf{k} values in the sphere in Figure 5.1 is then equal to the number of cubes of side length $2\pi/L$ in the sphere. The volume $V_{\mathbf{k}}$ of the sphere is

$$V_{\mathbf{k}} = \frac{4\pi}{3}k^3 = \frac{4\pi}{3} \left[\frac{2m^*(E - E_c)}{\hbar^2} \right]^{3/2} \quad (5.3)$$

The number of states $\bar{N}(E)$ in this volume is

$$\bar{N}(E) = V_{\mathbf{k}} / \left(\frac{2\pi}{L} \right)^3 \quad (5.4)$$

For systems other than simple cubic, one replaces the cubes of reciprocal volume $(\frac{2\pi}{L})^3$ by other unit cells. The outcome is the same.

Because each state can be occupied with two electrons of opposite spin, the number $N(E)$, which finally tells us how many electrons can be accommodated, is twice as large as $\bar{N}(E)$

$$N(E) = 2\bar{N}(E) \quad (5.5)$$

As can be seen from the following algebra, the quantity that is most useful is the number of states per unit sample volume at energy E in the interval $[E, E + dE]$. This quantity is called the density of states $g(E)$, which is given by

$$g(E) = \frac{1}{L^3} \frac{dN(E)}{dE} = \frac{1}{2\pi^2} \left(\frac{2m^*}{\hbar^2} \right)^{3/2} \sqrt{E - E_c} \quad (5.6)$$

To understand the importance of $g(E)$, let us assume for the moment that we know the probability $f(E)$ that an electron occupies a state with energy E . Then we can obtain the density n_c of electrons in the conduction band from

$$n_c = \sum_{\mathbf{k}} f(E) = \int_{E_c}^{\infty} f(E) g(E) dE \quad (5.7)$$

Equation (5.6) can be derived also in a different way: Because the allowed \mathbf{k} values are separated from each other only by very small distances (L is very large), the summation \sum can be replaced by an integration $\int d\mathbf{k}$. Because the number of \mathbf{k} values in the volume $d\mathbf{k} = dk_x dk_y dk_z$ is equal to $d\mathbf{k}/(2\pi/L)^3$, we obtain the number per unit volume (including a factor of 2 for the spin) as

$$2 \left(\frac{1}{2\pi} \right)^3 d\mathbf{k}$$

and we have

$$\frac{1}{V_{\text{ol}}} \sum_{\mathbf{k}} \rightarrow 2 \int \left(\frac{1}{2\pi} \right)^3 d\mathbf{k} \quad (5.8)$$

Using spherical coordinates, we obtain

$$dk_x dk_y dk_z = k^2 dk \sin\theta d\theta d\phi \quad (5.9)$$

It is then easy to show that

$$\frac{2}{(2\pi)^3} \int d\mathbf{k} \rightarrow \int g(E) dE \quad (5.10)$$

For complicated $E(\mathbf{k})$ relations, $g(E)$ takes a form that is more complicated than Eq. (5.6). For example, for one of the lowest conduction band minima in silicon

(one of the six ellipsoids), we have (transforming the ellipsoid to a sphere by a suitable coordinate transformation in \mathbf{k} space)

$$g(E) = \frac{\sqrt{m_l^* m_t^{*2}}}{2\pi^2} \left(\frac{2}{\hbar^2} \right)^{3/2} \sqrt{E - E_c} \quad (5.11)$$

The term $\sqrt{m_l^* m_t^{*2}}$ replaces $m^{*3/2}$ of Eq. (5.6). One therefore defines a density of states mass m_d

$$m_d^* = (m_l^* m_t^{*2})^{1/3} \quad (5.12)$$

which differs from the conductivity mass of Eq. (3.31).

Higher in the conduction band, the proportionality of $g(E)$ to $\sqrt{E - E_c}$ ceases to be true because of the complicated $E(\mathbf{k})$ relation for the conduction band of semiconductors such as silicon or gallium arsenide (Figure 5.2). In this case the density of states cannot be calculated explicitly but is easily calculated by numerical methods. To understand these methods, consider a surface of constant energy E_0 in \mathbf{k} space. This surface is equal to the circle shown in Figure 5.1 only for the simple case of isotropic effective mass. In general this will be a surface (two-dimensional in three-dimensional \mathbf{k} space) of arbitrary complicated shape. We now could calculate the number of states in the volume of this surface similar to Eqs. (5.3) and (5.4). However, we also can directly look at the number of states in the differential volume between the surface E_0 and $E_0 + dE$. The thickness dE expressed in terms of \mathbf{k}_\perp (the unit vector perpendicular to the equal average surface) reads:

$$dE = |\nabla_{\mathbf{k}} E(\mathbf{k})| d\mathbf{k}_\perp \quad (5.13)$$

as known from calculus, and therefore the density of states is given by

$$g(E_0) = \int \frac{ds}{|\nabla_{\mathbf{k}} E(\mathbf{k})|} \quad (5.14)$$

Here ds signifies the surface integral over the surface of equal energy E_0 in \mathbf{k} space. This integral can be evaluated by very efficient numerical methods.

The integration is performed by selecting a random sequence of points \mathbf{k}_j within the first Brillouin zone around which differential contributions to the density of states are calculated. Within a small fixed-size radius R_s of each point \mathbf{k}_j , $E(\mathbf{k})$ is assumed to be a linear function of \mathbf{k} . $E(\mathbf{k}_j)$ and $\nabla E(\mathbf{k}_j)$ are evaluated. If part of a surface $E(\mathbf{k}) = E_0$ lies within the small sphere of radius R_s centered at \mathbf{k}_j , the shortest distance between this surface and \mathbf{k}_j is $R_e = |(E_0 - E(\mathbf{k}_j))/\nabla_{\mathbf{k}} E(\mathbf{k}_j)|$. Therefore, the area of intersection between this surface and the small sphere is $\pi(R_s^2 - R_e^2)$, and its contribution to the density of states is simply $\pi(R_s^2 - R_e^2)/|\nabla_{\mathbf{k}} E(\mathbf{k}_j)|$. This procedure is iterated until the desired accuracy is reached, taking care to normalize $g(E)$ appropriately with the number of points selected.

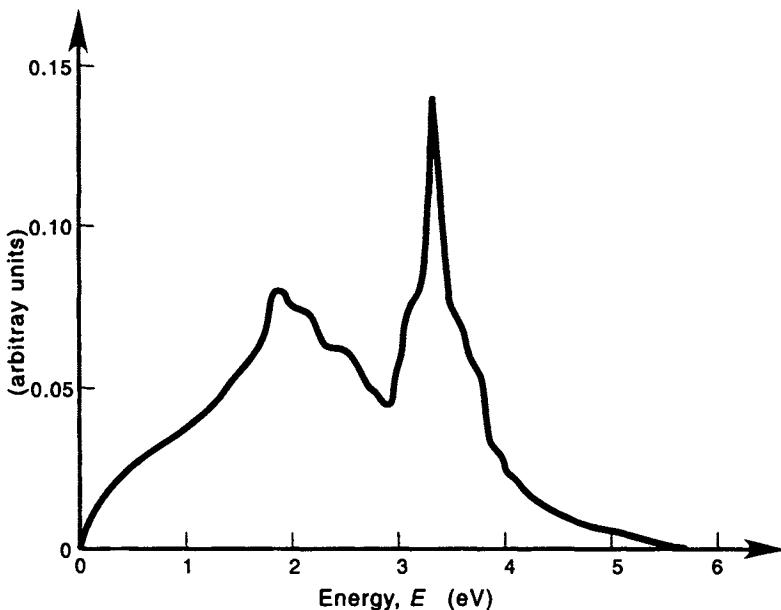


Figure 5.2 The density of states for silicon as calculated from the empirical pseudopotential model (conduction band).

These modifications of the explicit simple treatment of the density of states are necessitated by the complicated form of the energy bands at high energies far above (below) the band edges. Also, at low energies, modifications of the proportionality of $g(E)$ to $\sqrt{E - E_c}$ can be important. These modifications are usually a consequence of the doping. Figure 5.3 shows the impurity-related density of states for various degrees of doping. For light doping (Figure 5.3a), the impurity levels are the single “hydrogen-like” levels. For higher impurity densities (Figure 5.3b), an impurity band develops that merges at very high concentrations with the conduction band (Figure 5.3c). In the latter case, the semiconductor behaves as a metal; that is, it is highly conducting down to the lowest temperatures.

A simple semiquantitative treatment of the impurity “band tail,” shown in Figure 5.3c, has been proposed by Kane [2]. His treatment rests on the assumption of a local conservation of the density of states. He visualizes the impurity potential at high impurity densities as a smoothly varying potential $V_I(r)$, as shown in Figure 5.4.

The density of states at each point r is the density of states of the unperturbed crystal as shown in Figure 5.4. In other words, the conduction band edge is locally shifted and the density of states starts to increase from these shifted energy values. The fact that the conduction band edge is shifted exactly as the potential $V_I(r)$ follows, of course, from the effective mass theorem, which is then the basis of Kane’s theory.

Kane further suggests using the potential $V_I(r)$ as a stochastic variable obey-

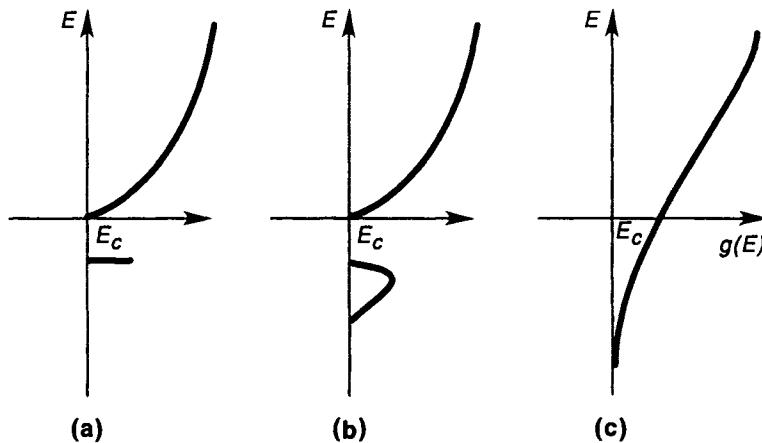


Figure 5.3 Density of states in a semiconductor including impurity levels.

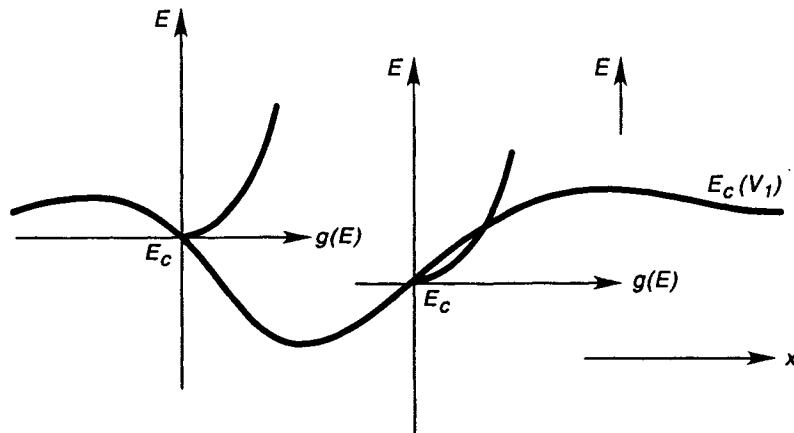


Figure 5.4 The locally conserved density of states in a slowly varying potential of impurities.

ing a Gaussian distribution F

$$F(V_I) = \frac{1}{\sqrt{\pi}} \exp(-e^2 V_I^2 / \eta^2) \quad (5.15)$$

The quantity η has been also estimated by Kane [2]. For our purposes, it is sufficient to regard it as parameter (a typical value for 10^{19} cm^{-3} impurities would be 50 meV). If the unperturbed density of states is denoted by $g(E)$ and the density of states including impurities by $g_1(E)$, we obtain

$$g_I(E) = \int_{-\infty}^{E-E_c} g(E-eV_I) F(eV_I) d(eV_I) \quad (5.16)$$

The equation is general enough to hold also for two-dimensional systems. Remember, however, that Eq. (5.16) is based on the effective mass theorem and will

fail for very close spacing of impurities.

A very different treatment, valid for the “deeper” band tail, has been given by Halperin and Lax [1]. Although this theory does not rely on the simple picture of the effective mass approach, it is valid only in the deeper band tail and is not simple enough for engineering applications. We therefore propose a pragmatic approach for complex simulations related to devices: use Eq. (5.16) and evaluate η from pertinent experimental data for the (restricted) range of doping densities that is of interest.

5.2 PROBABILITY OF FINDING ELECTRONS IN A STATE

Here we describe equilibrium statistics, which are well known and described in many other texts. We therefore list only a few facts that are used in the following and refer the reader for derivation to the excellent treatment by Landsberg [3].

In equilibrium, the probability of finding an electron in a state of energy E is given by the Fermi distribution f , as derived in texts of statistical mechanics

$$f(E) = \frac{1}{e^{(E-E_F)/kT} + 1} \quad (5.17)$$

The parameter E_F is known as the Fermi level (chemical potential) and is determined by the total number of charge carriers.

Figure 5.5 shows $f(E)$ for $T = 0$ and $T \neq 0$. For zero temperature, $f(E) = 1$ below E_F and $f(E) = 0$ above E_F . This means that all the states below E_F will be filled and all above it will be empty if $T = 0$. For higher temperatures, the distribution function exhibits an exponential (Maxwell-Boltzmann) tail.

There are two important approximations to $f(E)$. For low temperatures, $f(E)$ can be approximated by the step function $H(E - E_F)$, as evident from Figure 5.5.

$$f(E)_{T \sim 0} \approx H(E - E_F) \quad (5.18)$$

In this case, the derivative of f is

$$\frac{\partial f(E)}{\partial E} = -\delta(E - E_F) \quad (5.19)$$

A general useful relation is also

$$\frac{\partial f(E)}{\partial E} = -f(E)(1 - f(E))/kT \quad (5.20)$$

If we are only interested in the higher energy tail of $f(E)$, then we can neglect the one compared to the exponent in Eq. (5.17), which gives

$$f(E) \approx e^{(E_F - E)/kT} \quad (5.21)$$

Remember that at finite temperatures E_F is the parameter that characterizes the density of electrons [see Eq. (5.23)] and represents the chemical potential,

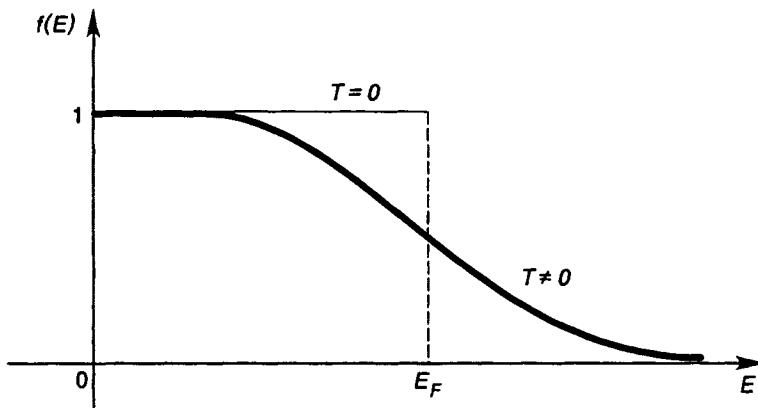


Figure 5.5 Fermi distribution as function of energy.

which is temperature dependent itself. The function of Eq. (5.21) is called the Maxwell-Boltzmann distribution function.

Holes like electrons obey Fermi statistics, but as if their energy were measured “downwards.” This can easily be seen from the fact that the hole distribution f_h equals $1 - f$, where f is the electron distribution.

Equation (5.21) describes the probability of finding an electron in a conduction band state of energy E . Remember that two electrons of opposite spin can occupy such a state. This factor of two is included in our treatment of the density of states. The probability that the energy level of a donor is occupied is more difficult to calculate for various reasons. For one, a donor can usually only accommodate one electron (the one which is then “donated”). There are, however, in addition to the ground state, excited states available for the electron and the various states with and without the electron can have degeneracies (e.g., owing to the fact that two spin states are available to the electron bound by the donor). As a consequence one obtains for the probability $f_{D/A}$ of occupation for a donor/acceptor of energy $E_{D/A}$:

$$f_{D/A} = \frac{1}{1 + \beta_{D/A} \exp(E_{D/A} - E_F)/kT} \quad (5.22)$$

where the factor $\beta_{D/A}$ can often be approximated by $\beta_D = 1/2$ and $\beta_A = 2$ as derived in Landsberg’s [3] treatment. Note that the establishment of such a distribution depends on the communication of the donors with band states and can take considerable time as discussed in Chapter 9.

5.3 ELECTRON DENSITY IN THE CONDUCTION BAND

From Eq. (5.7), by using Eq. (5.22) for the distribution function (which will be justified below), we obtain for the density n in the conduction band

$$n_c = \frac{1}{2\pi^2} \left(\frac{2m^*}{\hbar^2} \right)^{3/2} \int_0^\infty (1 + \exp(E - E_F)/kT)^{-1} \sqrt{E} dE \quad (5.23)$$

Here we have assumed that the conduction band extends toward infinity (the exponent decays rapidly) and we have chosen the conduction band edge as the zero of the energy. Note, also, that an effective mass of a single minimum has been used. Substituting \bar{x} for E/kT and \bar{x}_F for E_F/kT . One is left with a Fermi integral defined by

$$I_{\frac{1}{2}}(\bar{x}_F) = \frac{2}{\sqrt{\pi}} \int_0^\infty \frac{\bar{x}^{1/2}}{1 + \exp(\bar{x} - \bar{x}_F)} d\bar{x} \quad (5.24)$$

which must be evaluated to obtain the carrier concentration n_c in the conduction band. For $\bar{x}_F < -2$, the exponent in the denominator of Eq. (5.24) dominates and, if one seeks accuracy within a few percent only, the one can be neglected. The integral then becomes equal to the Γ function.

$$\int_0^\infty e^{-\bar{x}} \bar{x}^s d\bar{x} = \Gamma(s+1) \quad (5.25)$$

and we obtain

$$n = \frac{1}{4} \left(\frac{2m^* k T}{\pi \hbar^2} \right)^{3/2} e^{E_F/kT} \equiv N_c e^{E_F/kT} \quad (5.26)$$

where N_c is called the effective density of states for the conduction band. Here we have used the following expression for the Γ function:

$$\Gamma(1/2) = \sqrt{(\pi)} \quad (5.27)$$

and

$$\Gamma(s+1) = s\Gamma(s) \quad (5.28)$$

which is valid for s being an integer multiple of $1/2$. For s being integer, we simply have

$$\Gamma(s+1) = s! \quad (5.29)$$

For the density p of holes in the valence band, we obtain in a similar manner

$$p = \frac{1}{4} \left(\frac{2m_h^* k T}{\pi \hbar^2} \right)^{3/2} e^{(-E_G - E_F)/kT} \equiv N_v e^{(-E_G - E_F)/kT} \quad (5.30)$$

where E_G is the band gap energy.

One usually defines an “intrinsic concentration” n_i by the product $n \cdot p$. This product is given by

$$n_i^2 = n \cdot p = \frac{1}{2} \left(\frac{kT}{\pi \hbar^2} \right)^3 (m^* m_h^*)^{3/2} e^{-E_G/kT} \equiv N_c N_v e^{-E_G/kT} \quad (5.31)$$

As can be seen, this product depends only on semiconductor material parameters and not on the Fermi energy. Equation (5.31), therefore, holds also when the semiconductor is doped. Of course, if the doping goes up to levels so that $\bar{x}_F < -2$, the approximations are invalid and so is Eq. (5.31). In this case the integrals $I_{1/2}(\bar{x}_F)$ appear in the final result. Note, however, that even these integrals are correct only if the density of states is proportional to the square root of the energy. For the general case of Eq. (5.16), $I_{1/2}(\bar{x}_F)$ is replaced by a more general integral

$$I_g(\bar{x}) \propto \int_0^\infty \frac{g_1(\bar{x})}{1 + \exp(\bar{x} - \bar{x}_F)} d\bar{x} \quad (5.32)$$

where $g_1(\bar{x})$ can be calculated by (numerical) integration from Eq. (5.16) and represents the energy dependence of the density of states that replaces $\sqrt{\bar{x}}$. To determine the electron and hole concentration in a given band we still need to determine the Fermi level E . The Fermi level E_F can be calculated from the charge neutrality condition. For the pure semiconductor this condition means

$$p = n \quad (5.33)$$

and both n and p (now termed the intrinsic concentrations n_i, p_i) can be calculated from Eq. (5.31).

In the presence of charged donors, of constant density N_D^+ , and charged acceptors, of constant density N_A^- , charge neutrality must be written as

$$N_D^+ + p = N_A^- + n \quad (5.34)$$

To solve Eq. (5.34) for the Fermi level, we need to express N_D^+ and N_A^- as a function of E_F . The total density N_D of donors (or of acceptors N_A) can often be assumed as given, as for example deduced from the method of doping (diffusion, ion implantation). The density of charged donors N_D^+ (or acceptors N_A^-) is then obtained by multiplying the total density with the probability for a donor to be empty or acceptor to be filled, which can be obtained from Eq. (5.22). Equation (5.34) therefore represents, in general, a transcendental equation for E_F . If band tailing is included, the equation may in addition contain numerical double integrations involving Eq. (5.16) and Eq. (5.32). Nevertheless, such equations are easy to solve with current personal computers by use of the Newton method (see, e.g., *Numerical Recipes* [4]).

To remind the reader of the most important special cases, we add here a few analytical considerations:

1. The pure semiconductor with $m^* = m_h^*$. In this case, the equation of charge neutrality, Eq. (5.33), together with Eq. (5.26) and Eq. (5.30), gives

$$E_F = -E_G/2 \quad (5.35)$$

That is, the Fermi energy is in the middle of the energy gap (if $m_h^* \neq m^*$, it will be slightly shifted). This justifies *a posteriori* the use of the Maxwellian distribution for the calculation of n_c because only the tail of the distribution function is within the conduction band. The intrinsic concentration of several semiconductors is shown in Figure 5.6.

2. A semiconductor is doped with N_D donors and $N_D \gg n_i$. Furthermore, we assume that at high temperatures $N_D \approx N_D^+$, while at low temperatures $N_D^+ = 0$ (the electrons “freeze out” back to their parent donors, as shown below). Then

$$n = \frac{n_i^2}{n} + N_D \approx N_D \quad (5.36)$$

for high temperatures, while $n \approx 0$ for low temperatures. Notice, however, that if the temperature becomes very high, the term n_i^2/n in Eq. (5.36) cannot be neglected as the intrinsic concentration rises. This gives the graph n versus T , as shown in Figure 5.7.

3. At low temperatures, the assumption $N_D \approx N_D^+$ does not hold because the thermally excited electrons recombine with the donors largely neutralizing them. N_D^+ is generally given by

$$N_D^+ = N_D(1 - f_D) \quad (5.37)$$

with f_D from Eq. (5.22).

Equation (5.34) then reads in the absence of acceptors

$$N_D(1 - f_D) + p = n \quad (5.38)$$

In the limit $T \rightarrow 0K$, f_D approaches one and all terms in Eq (5.38) approach zero. Physically the electrons “freeze out” from the conduction band and recombine with their “parent” donors.

It is easy to include the ellipsoidal shape of the equal energy surfaces in our calculations. For example, if we consider the silicon conduction band, we have to replace m_d^* by $(m_l^* m_t^{*2})^{1/3}$ and multiply the result by 6 because we have six degenerate ellipsoids located inside the Brillouin zone. Germanium has eight ellipsoids whose center is at the L point of the Brillouin zone. Within the zone there is only one-half of each ellipsoid and we have to multiply the result by a factor of 4 instead of by 6, as in the case of silicon.

Let us finally mention that another complication in the calculation of carrier densities arises from the fact that the energy gap E_G actually depends on temperature. There are different reasons for this temperature dependence, which are

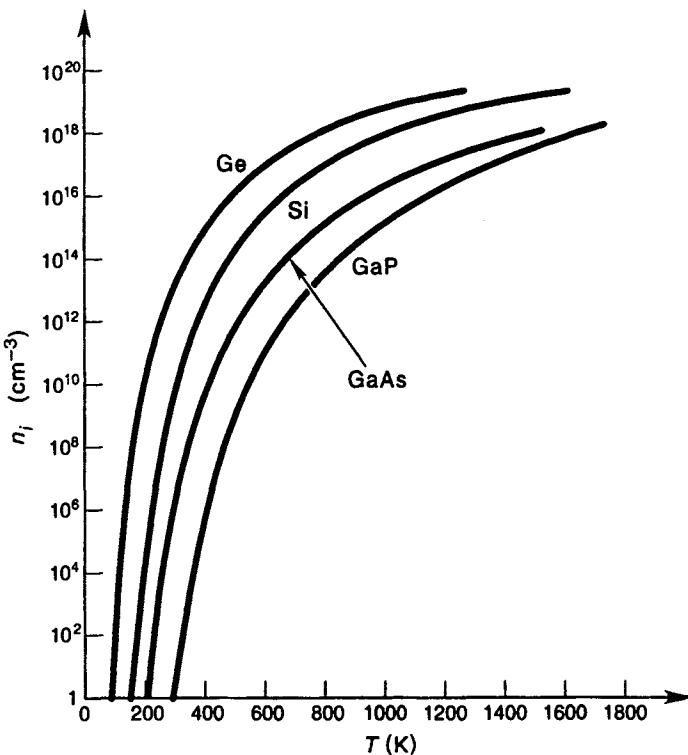


Figure 5.6 The intrinsic carrier concentration, n_i in cm^{-3} , as a function of T for Ge, Si, GaAs, and GaP. [After Thurmond, Figure 14. Reprinted by permission of the publisher, The Electrochemical Society, Inc. [5].]

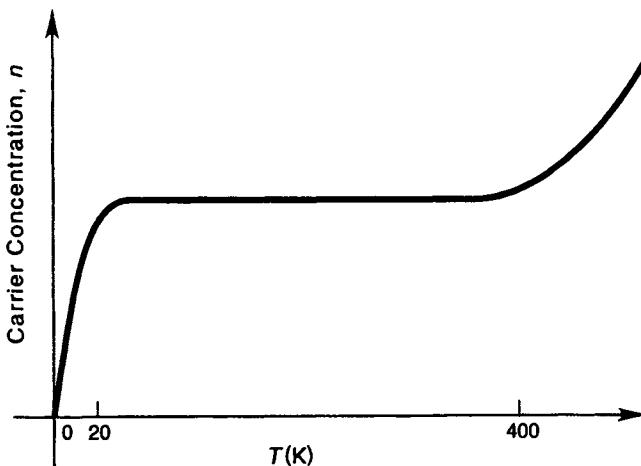


Figure 5.7 n versus T for a doped semiconductor.

described in Appendix C. The band-structure calculation presented in Chapter 1 cannot account for the temperature dependence, because in this calculation the atoms have been assumed to be “frozen” on their exact lattice sites. This functional dependence in temperature is also easy to include in a numerical approach using the equations given in Appendix C.

Of course, the described treatment can only be valid if the external potential V_{ext} , which includes the potential created by the free electrons and holes, themselves, is constant. If for any reason n and p become a function of the space coordinate—for example, because the doping concentrations depend on it—the above treatment is invalid; the external potential depends on the space coordinate, too, and charge neutrality can be violated locally. Globally (over the whole device or chip), the charge is still neutral. In this situation, one needs to solve Poisson’s equation to obtain V_{ext} and one needs a relation between V_{ext} and n and p . Chapter 6 treats this situation using linear response theory. Chapter 13 also gives some explicit considerations of how to include Poisson’s equation for device problems. The full numerical treatment of this problem in presence of electric current forms the heart of device simulation and is described in the final chapters.

PROBLEMS

- 5.1 Find the density of states as a function of energy for $E(k) = C[1 - \cos(|\mathbf{k}|a)]$ in one, two, and three dimensions. Discuss the limiting cases $|\mathbf{k}| \rightarrow 0, \frac{\pi}{a}, \frac{\pi}{2a}$.
- 5.2 Calculate the density of states for a coefficient $u = 50 \text{ meV}$ using Eqs. (5.15) and (5.16) for a two-dimensional system.
- 5.3 Derive the equilibrium electron density N_i for any subband i for a single quantum well assuming two-dimensional considerations apply. Let E_i represent the lowest energy of each subband. Your answer should be valid for all temperatures (see Section 10.4.2).

REFERENCES

- [1] Halperin, B., and Lax, M. “Impurity-band tails in the high-density limit. I. minimum counting methods,” *Physical Review*, vol. 148, 1966, pp. 722–740.
- [2] Kane, E. O. “Thomas-Fermi approach to impure semiconductor band structure,” *Physical Review*, vol. 131, 1963, pp. 79–88.
- [3] Landsberg, P. T., ed. *Solid State Theory*. New York: Wiley/Interscience, 1969, pp. 266–78.
- [4] *Numerical Recipes*, on the Internet at <http://www.nr.com>.
- [5] Thurmond, C. D. “The standard thermodynamic functions for the formation of electrons and holes,” *Journal of the Electrochemical Society*, vol. 122, 1975, pp. 1133–1141.

CHAPTER 6

SELF-CONSISTENT POTENTIALS AND DIELECTRIC PROPERTIES OF SEMICONDUCTORS

As indicated at the end of Chapter 5, the calculation of electron and hole densities becomes far more involved if the donor and acceptor densities are not constant over space. The reason is simply that then, of course, the electron and hole densities also become space dependent and local charge neutrality ceases to be true. If local charge neutrality is violated, then the potentials become themselves dependent on that charge as dictated by the equation of Poisson, which follows from the combination of

$$\nabla \cdot \mathbf{F} = \frac{e}{\epsilon \epsilon_0} (p - n + N_D^+ - N_A^-) \quad (6.1)$$

with

$$\mathbf{F} = -\nabla V(\mathbf{r}) \quad (6.2)$$

where \mathbf{F} is the electric field and $V(\mathbf{r})$ the local potential now a function of space coordinate \mathbf{r} . The solution of Eqs. (6.2) and (6.1) subject to given boundary conditions gives then $V(\mathbf{r})$. Quantum mechanics requires that n and p are calculated from the appropriate wavefunctions if the potential depends on \mathbf{r} and we will return to this below and in Chapter 10. If $V(\mathbf{r})$ varies very slowly then n and p can be related to $V(\mathbf{r})$ also from classical considerations. A typical result is then

$$n = n_0 \exp[-eV(\mathbf{r})/kT] \quad (6.3)$$

with n_0 being the concentration at the point of $V(\mathbf{r}) = 0$. The exponent in Eq. (6.3) represents, of course, a Boltzmann factor, which applies for equilibrium situations. In presence of current flow, Eq. (6.3) is typically replaced by differential equations for current and continuity, which require numerical solutions. As mentioned, this is an important portion of the solution of equations relevant for the simulation of semiconductor devices. The solutions of Eqs. (6.1) through (6.3) result in the so-called self-consistent potential $V(\mathbf{r})$. It is called "self-consistent" because it is a result of both the internal charge distribution $n(\mathbf{r})$ and $p(\mathbf{r})$ subject to the voltages at the boundaries (applied voltages). This

solution can in general only be accomplished numerically. In one dimension, however, it reduces often to one integration, as can be seen from the following arguments.

6.1 SCREENING AND THE POISSON EQUATION IN ONE DIMENSION

The voltage drop at a point z' in one dimension (z -direction) is obtained from Eq. (6.2)

$$V(z') = -zF|_0^{z'} + \int_0^{z'} z \frac{\partial F}{\partial z} dz \quad (6.4)$$

where we assumed $V(0) = 0$. By suitable choice of the zero of the coordinate system one can often achieve that $F(z') \approx 0$ then

$$V(z') \simeq \int_0^{z'} z \frac{\partial F}{\partial z} dz \quad (6.5)$$

and with $\frac{\partial F}{\partial z}$ from Eq. (6.1) we have

$$V(z') \simeq \int_0^{z'} \frac{e}{\epsilon \epsilon_0} (p - n + N_D^+ - N_A^-) z dz \quad (6.6)$$

This is then the promised solution of Poisson's equation by one integration. It is an important tool for fast estimates in semiconductor device theory as we will see momentarily and in Chapter 13.

Consider the case of an overall undoped semiconductor containing a heavily doped contact region at $z' = L_D$. In heavily doped contacts, the electric field is very small, so that $F(L_D) \approx 0$. The contact region will actually spill electrons into the undoped region. The spilling is, of course, dependent on the actual electron concentration n_c of the contact and we write

$$n(z) = n_c \cdot s(z) \quad (6.7)$$

where $s(z)$ is a given function. We can now ask the question what amount of spilled charge causes a voltage equal to kT/e at a certain point in the undoped semiconductor. This point we designate as the zero of the coordinate system. Because the contact is at $z' = L_D$ the length of the spilling region over which a voltage kT/e drops is equal to L_D . Therefore,

$$\frac{kT}{e} \simeq \int_0^{L_D} \frac{en_c}{\epsilon \epsilon_0} s(z) z dz \quad (6.8)$$

and

$$\frac{\epsilon \epsilon_0 kT}{e^2 n_c} \simeq \int_0^{L_D} s(z) z dz$$

The right-hand side has the dimension of length-square and therefore is equal to $L_D M s(M)$ where M is some mean value (determined by the mean value theorem of integral calculus). We now assume $M s(M) \approx L_D^2$. The exact justification follows from the reasoning below and usually can only be obtained numerically. We then obtain

$$L_D^2 \approx \frac{\epsilon \epsilon_0 k T}{e^2 n_c} \quad (6.9)$$

This is the well-known Debye length, L_D , a length very basic in solid-state and semiconductor device theory. The choice of voltage kT/e in Eq. (6.8) is commensurate with the physical intuition that such a thermal voltage will be available for the spilling. In other words, to spill the electrons over from the contact we need energy and this energy is supplied thermally. Any given density of electrons can spill in this way over the length L_D , from contacts or from any other device regions. Therefore, transitions in electron density are gradual and cannot be entirely abrupt. A contact area can then only be defined within the uncertainty of this length. The length itself, as expressed in Eq. (6.9), is of course only approximate and needs to be calculated numerically, particularly in nonequilibrium situations. This can always be achieved by solving the device equations as we will discuss in detail later. The solution of these equations is meanwhile a complete branch of device science and engineering. In this chapter we are not advanced enough to fully appreciate the intricacies of solving Poisson and current as well as continuity equations. However, the issue of a Debye length comes also up on a microscopic level. We can ask ourselves what happens to a positively charged donor in a “bath” of negatively charged electrons, and the answer is that electrons accumulate around the donor and screen its charge within the Debye length. This screening is important for the understanding of scattering of electrons by impurities, and this background is needed in Chapter 7. From the above one asks, how can one derive the potentials and screening length for crystal imperfections of various shapes and forms particularly, because for these small dimensions one needs to know the wavefunction to determine the charge density. The answer is that if one wants to know only linear response close to equilibrium and if one uses perturbation theory and the so-called random phase approximation, then one can come up with a general result. This is the subject of this chapter.

6.2 SELF-CONSISTENT POTENTIALS AND THE DIELECTRIC FUNCTION

Assume then that we put an additional charge on an atom in a crystal. What is the potential of this electronic charge? Some may find an easy answer and, in fact, we have given it in Eq. (4.1). In this equation, however, we introduced without justification a dielectric constant. In this chapter we will derive the dielectric

constant and see that it actually is not a constant but in general is a rather complex function of frequency and wavevector.

The essence of our goal is to solve a many-particle problem. We introduce a charge into a solid, and we would like to know its potential as a function of position. The complication is that this charge exerts forces on all the other charges in the solid and rearranges them. Our applied coulombic potential V_{ap} therefore causes an additional potential V_{ad} owing to the rearrangement. Together the two potentials determine the true potential V_{tr} in the semiconductor.

To develop the theory for this complex situation, we introduce two new techniques. First, we treat the problem self-consistently. In the problems we discussed above, we are faced with the situation that we apply certain potentials (the boundary conditions; potentials at contacts etc.), which then cause a redistribution of charge. This in turn changes the internal potential as dictated by the equation of Poisson the combination of Eqs. (6.1) and (6.2). For the microscopic (atomistic) problems, this means we assume that we know the true impurity potential and denote it by V_{tr} . Then we calculate the redistribution of charge due to this impurity potential and obtain an additional potential V_{ad} . We then use the equation

$$V_{\text{tr}} = V_{\text{ap}} + V_{\text{ad}} \quad (6.10)$$

V_{tr} is thus the self-consistently obtained (true) local potential following the application of V_{ap} and subsequent charge reordering, which gives an additional potential V_{ad} . We actually do not need to split V_{tr} in the two additional components of Eq. (6.10). We can also connect V_{tr} to V_{ap} by some other functional dependence. As we will see below, it is very convenient to deal with the Fourier coefficients of the potential instead of the potential itself because we can solve then Poisson's equation by Fourier expansion techniques. This will be done with a very special trick, called the random-phase approximation: everyone familiar with the calculus of Fourier transformations or expansions knows that in general functions require certain series of terms such as

$$V_{\text{tr}} = \sum_{\mathbf{q}, \omega} V_{\text{tr}}^{\mathbf{q}, \omega} e^{i\mathbf{qr}} e^{i\omega t} \quad (6.11)$$

$V_{\text{tr}}^{\mathbf{q}, \omega}$ are the Fourier coefficients, \mathbf{q} is the wavevector or just a summation index, and ω the frequency. Manipulation of such an expansion and insertion into the appropriate differential equations leads often to products of such sums and thus to mixed terms. The various Fourier coefficients can therefore, in general, not be treated independently. Here, however, this is just what we are going to do by assuming that all mixed terms will give vanishing contributions because of the randomness of the phases of these terms. This is the so-called random-phase approximation. It leads to beautifully simple and complete results. To go beyond this approximation (unnecessary for our purposes) requires major efforts, and there are whole branches of physical science dealing with it (see, e.g., the description of Mahen [1], which shows the intricacies in detail and goes beyond our

simplified description). Using the random phase approximation we can treat all Fourier coefficients independently; that is, we insert for V_{tr} only one coefficient

$$V_{\text{tr}}^{\mathbf{q}, \omega} e^{i\mathbf{q} \cdot \mathbf{r}} e^{i\omega t} \quad (6.12)$$

instead of the sum, do the same for all other qualities involved, then calculate the Fourier coefficients and perform the sum at the end to arrive at the potential V_{tr} .

The functional relation of V_{tr} and V_{ap} is given by the following equation (which can be regarded as the definition of $\epsilon(\mathbf{q}, \omega)$):

$$V_{\text{tr}}^{\mathbf{q}, \omega} = \frac{V_{\text{ap}}^{\mathbf{q}, \omega}}{\epsilon(\mathbf{q}, \omega)} \quad (6.13)$$

$\epsilon(\mathbf{q}, \omega)$ represents, of course, the dielectric constant which is a function of \mathbf{q} and frequency ω .

The wave function of the electron is calculated by perturbation theory. Because we have only one Fourier component, only the matrix elements involving the given \mathbf{q} are finite and we can write the perturbed wave function as [see Eq. (1.39)]

$$\Phi_{\mathbf{k}}^m = \Psi_{\mathbf{k}}^m + b_{\mathbf{k}+\mathbf{q}}^l \quad (6.14)$$

with

$$b_{\mathbf{k}+\mathbf{q}} = \frac{\langle \mathbf{k} + \mathbf{q} | eV_{\text{tr}} | \mathbf{k} \rangle}{E(\mathbf{k})_l - E(\mathbf{k} + \mathbf{q})_m + \hbar\omega} = \frac{eV_{\text{tr}}^{\mathbf{q}, \omega}}{E(\mathbf{k})_l - E(\mathbf{k} + \mathbf{q})_m + \hbar\omega} \quad (6.15)$$

Here we have used the fact that the matrix element of V_{tr} is equal to its Fourier coefficient as shown in Eq. (6.14). This is only correct within the effective mass approximation. The transition from \mathbf{k} to $\mathbf{k} + \mathbf{q}$ can also involve two energy bands, for example, conduction and valence band. (The index m assumes then two values). Then one also has to calculate the matrix element using conduction and valence band wavefunctions and also the approximate $E(\mathbf{k})$ relations for initial (e.g., conduction band) and final (e.g., valence band) state.

For the present purpose we can ignore these difficulties; we will concentrate mostly on one band when evaluating the dielectric constant. The perturbed wave function allows us to calculate the new charge distribution ρ_{new} , which is given by

$$\rho_{\text{new}} = e \sum_{\mathbf{k}, m} |\Phi_{\mathbf{k}}^m|^2 \quad (6.16)$$

Because the “old” charge distribution is the electron distribution in the unperturbed crystal, that is, $\sum_{\mathbf{k}} e/V_{\text{ol}}$, for plane waves we obtain for the change in the charge $\delta\rho$

$$\delta\rho = e \sum_{\mathbf{k}, m} \left(|\Phi_{\mathbf{k}}^m|^2 - \frac{1}{V_{\text{ol}}} \right) \quad (6.17)$$

Using Eq. (6.14) and neglecting higher-order terms (in involving squares of the b coefficients) one gets for each band (index l):

$$\delta\rho = \sum_{\mathbf{k}, l} (b_{\mathbf{k}+\mathbf{q}} e^{i\mathbf{q}\cdot\mathbf{r}} + b_{\mathbf{k}+\mathbf{q}}^* e^{-i\mathbf{q}\cdot\mathbf{r}}) \quad (6.18)$$

Actually, only full states can contribute to a change $\delta\rho$. Therefore, we have to multiply Eq. (6.17) by the probability that a state is occupied ($f^l(\mathbf{k})$), which gives (involving a transformation $\mathbf{k} \rightarrow \mathbf{k} + \mathbf{q}$ in part of the sum)

$$\delta\rho = \frac{e^2 V_{\text{tr}}^{\mathbf{q}, \omega}}{V_{\text{ol}}} \sum_{\mathbf{k}, m, l} \frac{f^l(\mathbf{k}) + f^m(\mathbf{k} + \mathbf{q})}{E(\mathbf{k})_l - E(\mathbf{k} + \mathbf{q})_m + \hbar\omega} + \begin{array}{l} \text{complex} \\ \text{conjugate} \end{array} \quad (6.19)$$

The reader is encouraged to derive Eq. (6.19) in all details.

The next step in this calculation is to determine V_{ad} . This can be done from the Poisson equation:

$$\nabla^2 V_{\text{ad}} = -\frac{1}{\epsilon_0} \delta\rho \quad (6.20)$$

Here ϵ_0 is the dielectric constant of free space.

Fourier transforming Eq. (6.20), we have

$$q^2 V_{\text{ad}}^{\mathbf{q}, \omega} = e^2 / \epsilon_0 V_{\text{tr}}^{\mathbf{q}, \omega} \sum_{\mathbf{k}, m, l} \frac{f^l(\mathbf{k}) + f^m(\mathbf{k} + \mathbf{q})}{E(\mathbf{k})_l - E(\mathbf{k} + \mathbf{q})_m + \hbar\omega} \quad (6.21)$$

We are mostly interested in the static dielectric constant ($\omega = 0$), which then becomes using Eq. (6.13):

$$\epsilon(\mathbf{q}, 0) = 1 + \frac{e^2}{\epsilon_0 q^2} \sum_{\mathbf{k}, m, l} \frac{f^l(\mathbf{k}) + f^m(\mathbf{k} + \mathbf{q})}{E(\mathbf{k})_l - E(\mathbf{k} + \mathbf{q})_m} \quad (6.22)$$

A special case that we can easily evaluate is that of an electron gas at a conduction band minimum in the effective mass approximation. We then have $l = m$ in Eq. (6.22), and we can perform the sum over \mathbf{k} (with the help of a computer for even more complicated $E(\mathbf{k})$ relations). Here we only will evaluate the expression for $\mathbf{q} \rightarrow 0$. To do this, we expand

$$E(\mathbf{k} + \mathbf{q}) = E(\mathbf{k}) + \mathbf{q} \cdot \nabla_{\mathbf{k}} E(\mathbf{k}) \quad (6.23)$$

and

$$f(E(\mathbf{k} + \mathbf{q})) = f(E(\mathbf{k})) + \frac{\partial f}{\partial E} \mathbf{q} \cdot \nabla_{\mathbf{k}} E(\mathbf{k}) \quad (6.24)$$

which gives

$$\begin{aligned} \epsilon(\mathbf{q}, 0) &= 1 + \frac{e^2}{q^2 \epsilon_0} \frac{2}{8\pi^3} \int \frac{\mathbf{q} \cdot \nabla E(\mathbf{k})}{\mathbf{q} \cdot \nabla E(\mathbf{k})} \left(-\frac{\partial f}{\partial E} \right) g(E) dE \quad (6.25) \\ &= 1 + \frac{e^2}{q^2 \epsilon_0} \int \left(-\frac{\partial f}{\partial E} \right) g(E) dE \end{aligned}$$

If we define

$$\frac{e^2}{\epsilon_0} \int \left(-\frac{\partial f}{\partial E} \right) g(E) dE \equiv q_s^2 \quad (6.26)$$

we have

$$\epsilon(\mathbf{q}, 0) = 1 + q_s^2 / q^2 \quad (6.27)$$

q_s^2 can easily be evaluated in special cases. For example, for f being a step function and consequently $\frac{\partial f}{\partial E} = -\delta(E - E_F)$, we have

$$q_s^2 = \frac{e^2}{\epsilon_0} g(E_F) \quad (6.28)$$

$L_s = 2\pi/q_s$ is called the *Thomas-Fermi screening length*. For f being a Maxwell-Boltzmann distribution, one obtains

$$q_D^2 = \frac{e^2 n}{\epsilon_0 k T} \quad (6.29)$$

and $L_D = 2\pi/q_D$ is the Debye screening length which was discussed at the outset of this chapter.

The quantities q_s and q_D are termed “screening wave vector,” which becomes clear from the definition of the dielectric function: Any Fourier component V_{ap}^q with $q \gg q_s$ will result in a very small “true” Fourier component V_{tr}^q . That is, the spatial change of the true potential corresponding to this or smaller wave vectors will be negligible (screened out by the other charge carriers). On the other hand, for $q \gg q_s$ the medium (electron gas) will not have an appreciable influence as $\epsilon(q, 0) = 1$.

The form of the potential in real space is given by

$$V_{tr} \propto \frac{1}{r} e^{-r/L_s} \quad (6.30)$$

This is obtained by “back” Fourier transformation.

Eq. (6.27) does not completely describe the dielectric function of a semiconductor; we have not taken into account the electrons in the valence band. A treatment of both bands introduces the energy gap E_G in the energy denominator of Eq. (6.22). However, because the sum over all \mathbf{k} is taken in Eq. (6.22), it is the energy difference at the place of the highest density of states (jointly of conduction and valence bands) that matters. The density of states is usually very high at the energy somewhat above the X valleys. Therefore, in these cases, the most important energy difference is that around X ; that is, E_G^X .

A rather tedious calculation executed by Penn [2] gives the following result for the static dielectric constant (now denoted by ϵ_{Penn}):

$$\epsilon_{Penn} = 1 + \left(\frac{\hbar \omega_p}{E_G^X} \right)^2 \quad (6.31)$$

with

$$\omega_p^2 = \frac{ne^2}{\epsilon_0 m}$$

where m is the free electron mass, n the total electron density in the valence band, and ω_p the plasma frequency (Ziman [3]). Penn [2] neglected the conduction band electrons in the derivation. If we combine Eqs. 6.31 and 6.27, we obtain

$$\epsilon(\mathbf{q}, 0) \approx \left[1 + \left(\frac{\mathbf{q}_s}{\mathbf{q}} \right)^2 \frac{1}{\epsilon_{\text{Penn}}} \right] \epsilon_{\text{Penn}} \quad (6.32)$$

which is sufficient to describe ϵ of a semiconductor in many practical cases and can be used to calculate the true potential in a semiconductor if the applied potential is given.

Instead of ϵ_{Penn} , one also can use the dielectric constants given in Table 3.1. Notice that the dielectric constant depends in principle also on the frequency ω , which we have neglected in our treatment. In the table, limiting values for $\omega \rightarrow 0$ and $\omega \rightarrow \infty$ (static and optic dielectric constants) are given. The difference arises from contributions of the crystal lattice in polar materials, such as GaAs [3].

PROBLEMS

- 6.1 Let a free electron gas in a positive background be perturbed by a potential

$$\delta U(\mathbf{r}, t) = U_q e^{i\mathbf{q} \cdot \mathbf{r}} e^{i\omega t} e^{i\alpha} + \text{cc}$$

Show that the charge density fluctuations $\delta\rho(\mathbf{r}, t)$ are given by

$$\delta\rho(\mathbf{r}, t) = e \sum_{\mathbf{k}} \frac{f_l(\mathbf{k}) + f_m(\mathbf{k} + \mathbf{q})}{E(\mathbf{k}) - E(\mathbf{k} + \mathbf{q}) + \hbar\omega - i\hbar\alpha} e^{i\mathbf{q} \cdot \mathbf{r}} e^{i\omega t} e^{i\alpha} U_q + \text{cc}$$

where $\text{cc} = \text{complex conjugate}$.

REFERENCES

- [1] Mahen, G. D. *Many Particle Physics*, New York: Plenum Press, 1981.
- [2] Penn, D. R. "Wave-number-dependant dielectric function of semiconductors," *Physical Review*, vol. 128, 1962, pp. 2093–2097.
- [3] Ziman, J. M. *Principles of the Theory of Solids*, 2nd ed. Cambridge: Cambridge Press, 1972, pp. 146–60.

CHAPTER 7

SCATTERING THEORY

7.1 GENERAL CONSIDERATIONS—DRUDE THEORY

A precise knowledge of the motion and scattering of electrons is necessary to understand the conductivity of a solid. One would think that scattering of electrons by impurities impedes only their motion. However, this is not true. A totally perfect solid can exhibit a totally unexpected behavior, which can be seen from the following example. The band structure of a one-dimensional (and, similarly, of a simple cubic three-dimensional) crystal is given by

$$E(k) = E_0(1 + \cos(k \cdot a + \pi))$$

Let's assume that we have one electron in the conduction band and a full valence band. The velocity of the electron is [Eq. (3.25)]:

$$v = \frac{1}{\hbar} \frac{dE}{dk} = -\frac{E_0}{\hbar} a \sin(k \cdot a + \pi)$$

The k value develops as [Eq. (3.26)]:

$$k = -eFt/\hbar + k_0$$

and we assume that the electron starts at $k + 0 = 0$ for $t = 0$.

Therefore, we have

$$v = \frac{E_0}{\hbar} a \sin(eFat/\hbar - \pi)$$

Because the current density $j = nev$, we see that a dc electric field F gives rise to an ac current of angular frequency $\omega = eFa/\hbar$. Thus the pure semiconductor would be an excellent high-frequency generator.

The angular frequency increases with lattice constant a , and, indeed, Esaki and Tsu [4] proposed to produce very pure superlattices having rather large a . They hoped to obtain unlimited frequencies in this way. Their work stimulated the technology, and we can produce very pure superlattices (lattice-matched

layers of semiconductors, e.g., GaAs-AlAs). The Esaki-Tsu oscillator, however, has never become a practical device. Also, for most applications of semiconductors a "normal" conductance is the rule, and we will see that we have a normal conductance only if scattering centers are present; that is, only if the semiconductor is imperfect. By far the most important scattering agents affecting the conductivity are the lattice vibrations (phonons), since they take away energy from the electrons and do not let k or $E(k)$ grow sufficiently for the electrons to approach the Brillouin zone boundary and exhibit an oscillatory velocity.

There are other scattering mechanisms, which are elastic but still influence the electronic conduction substantially. Among these are scattering by charged impurities, neutral impurities, and surfaces. The lattice vibrations themselves are scattered by such impurities and even by the different atomic isotopes of which the crystal is composed. (These scattering processes are the reason for the slow propagation of heat, which ideally could propagate with the velocity of sound.)

Here we concentrate on the scattering of electrons. Imagine an electron propagating in a crystal and colliding with various scattering agents. A consequence of all these collisions is a quasi-Brownian motion of the electron. As mentioned, the collisions also force the electron to stay energetically close to the conduction band edge. Feynman introduced an idealized graphical representation of collisions and the corresponding terms of perturbation theory that also gives additional information in the form of the k vector labels, as shown in Figure 7.1. In this way also very complicated collision processes can be neatly represented.

The whole theory of electric conduction would be very simple if it were not for the following complications:

1. The scattering depends on k and q .
2. More than one electron is present at the same time and we have to calculate the statistics of electron propagation.
3. The band structure enters the equations of motion.

We will deal with these complications later.

At this point we consider Drude's model of conduction, which circumvents these complications by simplifying assumptions. Drude assumed that all electrons move with a velocity v and used the equation of motion

$$\hbar k = mv = F_0 \quad (F_0 = \text{force}) \quad (7.1)$$

To include the band structure one uses m^* instead of m . Because this equation leads to a continuous acceleration, Drude suggested that the scattering agents act like a friction force F_f :

$$F_f \simeq mv/\tau \quad (7.2)$$

where τ is the time constant (relaxation time) of the friction force. (The meaning will be explained immediately.) The total equation of motion is, then, according

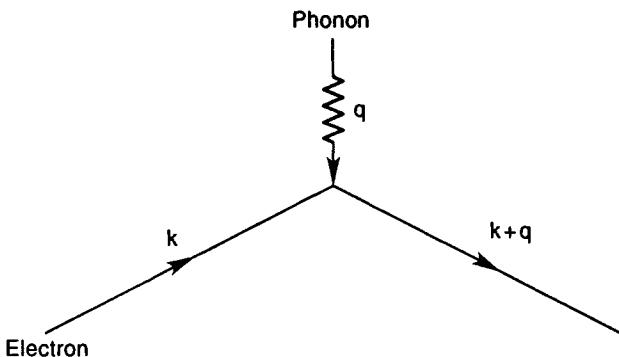


Figure 7.1 Feynman graph of collision.

to Drude

$$m\dot{v} = \mathbf{F}_0 - \mathbf{F}_f = \mathbf{F}_0 - mv/\tau \quad (7.3)$$

If the applied force \mathbf{F}_0 is set equal to zero, then

$$m\dot{v} = mv/\tau \quad (7.4)$$

and

$$v \propto \exp(-t/\tau) \quad (7.5)$$

Equation (7.5) shows that τ is the time constant in which v decays to zero. Thus the assumption of a friction force and a single velocity circumvents difficulties (1) and (2). It is only justified by its success: The Drude theory gives a qualitative explanation of almost all low-field, low-frequency conduction phenomena. The current density j is calculated by multiplying the velocity by the density of electrons:

$$\mathbf{j} = en\mathbf{v} \quad (7.6)$$

Inserting for the applied force \mathbf{F}_0 , the force of an electric field \mathbf{F} , and a magnetic field \mathbf{B} , one has

$$m\dot{v} = -e(\mathbf{F} + \mathbf{v} \times \mathbf{B}) - mv/\tau \quad (7.7)$$

This equation will now be discussed and solved for three special cases:

1. \mathbf{F} is independent of time, and $\mathbf{B} = 0$.

These assumptions imply a steady-state solution, that is, $\mathbf{v} = 0$ and $mv = e\mathbf{F}/\tau$, which indeed gives a dc current (no Bloch oscillations):

$$\mathbf{j} = -en\mathbf{v} = \frac{e^2 \tau n}{m} \mathbf{F}$$

We therefore conclude from the definition, Eq. (2.3), that the conductivity σ is scalar and given by

$$\sigma = en\mu \quad (7.8)$$

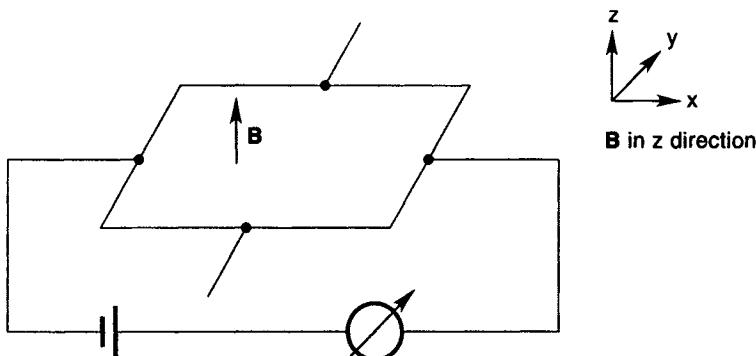


Figure 7.2 Schematic sample geometry for Hall measurements.

with

$$\mu = e\tau/m \quad (7.9)$$

μ is called the mobility and has the dimension cm^2/Vs . Multiplied by the electric field μ gives the negative velocity of the electrons.

2. \mathbf{F} is a low-frequency electric field and $\mathbf{B} = 0$.

$$\mathbf{F} \equiv \mathbf{F}_{ac}e^{i\omega t} + \text{cc}$$

(where cc stands for complex conjugate). To solve Eq. (7.7), we try $\mathbf{v} = \mathbf{v}_0e^{i\omega t} + \text{cc}$, which gives

$$mi\omega v_0e^{i\omega t} = -e\mathbf{F}_{ac}e^{i\omega t} - mv_0e^{i\omega t}/\tau$$

and a similar equation for the complex conjugate. It follows that

$$-\mathbf{v}_0 = \frac{e\mathbf{F}_{ac}}{im\omega + m/\tau} = \frac{\mu\mathbf{F}_{ac}}{1 + i\omega\tau} \quad (7.10)$$

That is, the velocity is equal to the dc velocity divided by $1 + i\omega\tau$. This means that at frequencies comparable to $1/\tau$ (which is usually above 10 GHz), the semiconductor ceases to be a resistor only and represents also an inductive delay. The imaginary part can also be viewed as a contribution of the free electrons to the dielectric constant (Maxwell's equations!). An interesting fact to note is that the real part becomes inversely proportional to τ for very high frequencies, which means that the conductivity is zero in absence of scattering.

3. \mathbf{F} is independent of time; $\mathbf{B} \neq 0$, and $\mathbf{B} = (0, 0, B_z)$.

If we try a solution as in example (1), then we see that the component of the velocity in the y -direction is not equal to zero, even if the applied field points only in the x -direction. The reason, of course, is the magnetic field, which deflects electrons in the y -direction ($\mathbf{v} \times \mathbf{B}$). Because we do not have any current channels in the y -direction, all that happens is

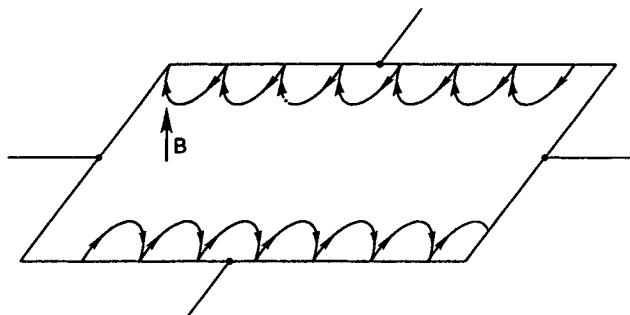


Figure 7.3 Skipping orbits of electrons at the boundary of a Hall sample. Notice that skipping at a given boundary and for a given magnetic field \mathbf{B} is possible only in one direction.

that electrons will accumulate at the sample boundaries and build up an electric field \mathbf{F}_y , which opposes further flow of electrons, as shown in Figure 7.2. It follows that we must admit an \mathbf{F}_y component of the field and obtain from

$$-e(\mathbf{F} + \mathbf{v} \times \mathbf{B}) = m\mathbf{v}/\tau$$

and

$$-e[(F_x, F_y, 0) - (0, -v_x B_z, 0)] = m(v_x, 0, 0)/\tau$$

the solution

$$-v_x = \mu F_x \quad (7.11)$$

and

$$F_y = v_x B_z \quad (7.12)$$

Because F_x , F_y , and B_z can be measured, the results allow a determination of the mobility μ and the velocity v_x . Together with the measurement of the current density $j_x = -env_x$, the electron concentration n can also be obtained. This effect (the occurrence of F_y in a magnetic field) is called the Hall effect. It is frequently used to determine μ and n .

The exact formula for the Hall field (without Drude's approximation) is almost identical to Eqs. (7.11) and (7.12) except for a statistical factor (averaging!), which is discussed in Appendix D. Some values for this factor are discussed in the review of Rode [14], and are typically close to one.

Interesting effects happen in the limit of high magnetic fields. The electrons in the center of the sample then move in cyclotron orbits. At the boundaries, however, the electrons are reflected and therefore move in skipping orbits, as illustrated in Figure 7.3.

These skipping electrons can proceed only in one direction at a given side and in the opposite on the other. Which direction the electron

propagates depends, of course, on the applied electric field. However, if the electron encounters a scattering event that would force it to propagate into the opposite direction, it cannot proceed on the skipping orbits. As a consequence, scattering cannot randomize the velocity and the conductivity becomes extremely large—for practical purposes infinite. These ranges of large conductivity go hand in hand with a steplike change of the Hall voltage which is called the Quantum Hall Effect. The interested reader is referred to the discussion of Datta [3].

In the following, improvements of the Drude theory are discussed. Most of the improvements are based on a careful reassessment of the relaxation time, which is replaced by an energy-dependent function or by an integral scattering operator (see the Boltzmann equation in Chapter 8). A central role in this theory is the scattering probability per unit time $S(\mathbf{k}, \mathbf{k}')$, which can be calculated from the Golden Rule. The remainder of this chapter deals with calculations of $S(\mathbf{k}, \mathbf{k}')$ and its phenomenological connection to the relaxation time and conductivity. The rigorous derivation of the conductivity from $S(\mathbf{k}, \mathbf{k}')$ is discussed in Chapter 8.

7.2 SCATTERING PROBABILITY FROM THE GOLDEN RULE

7.2.1 Impurity Scattering

We perform the calculation of $S(\mathbf{k}, \mathbf{k}')$ for ionized impurity scattering and acoustic phonon scattering in detail and discuss the result for the other important mechanisms for which the derivation is totally analogous.

To obtain $S(\mathbf{k}, \mathbf{k}')$ we need to calculate the matrix element $M_{\mathbf{kk}'}$ which, in the effective mass approximation, is

$$M_{\mathbf{kk}'} = \frac{1}{V_{\text{ol}}} \int_{V_{\text{ol}}} d\mathbf{r} H' e^{i(\mathbf{k}-\mathbf{k}')}$$

For impurity scattering, the perturbing part of the Hamiltonian H' is the true potential energy eV_{tr} for the impurity

$$M_{\mathbf{kk}'} = \frac{1}{V_{\text{ol}}} \int_{V_{\text{ol}}} d\mathbf{r} eV_{\text{tr}} e^{i(\mathbf{k}-\mathbf{k}')}$$

Using $\mathbf{k} - \mathbf{k}' \equiv \mathbf{q}$, we obtain the result that the matrix element is equal to the Fourier transformation of the true potential:

$$M_{\mathbf{kk}'} = \frac{1}{V_{\text{ol}}} eV_{\text{tr}}^{\mathbf{q}} \quad (7.13)$$

The Fourier component $V_{\text{tr}}^{\mathbf{q}}$ is obtained from the applied potential by Eq. (6.13), and V_{ap} can be obtained from the Poisson equation for a point charge (located at

$\mathbf{r} = 0$):

$$\nabla^2 V_{\text{ap}} = -\frac{e}{\epsilon_0} \delta(\mathbf{r}) \quad (7.14)$$

Because we need to know $V_{\text{ap}}^{\mathbf{q}}$, we solve the Poisson equation by Fourier transformation, which gives

$$V_{\text{ap}}^{\mathbf{q}} = -\frac{e}{\epsilon_0 q^2} \quad (7.15)$$

Using the dielectric constant for small \mathbf{q} as derived before, we have

$$V_{\text{tr}}^{\mathbf{q}} = \frac{e}{\epsilon_0 \epsilon (q^2 + q_s^2/\epsilon)} \quad (7.16)$$

and $M_{\mathbf{k}\mathbf{k}'}$

$$M_{\mathbf{k}\mathbf{k}'} = \frac{e^2}{V_{\text{ol}} \epsilon_0 \epsilon (q^2 + q_s^2/\epsilon)} \quad (7.17)$$

and therefore

$$S(\mathbf{k}, \mathbf{k}') = \frac{2\pi}{\hbar} \frac{e^4}{V_{\text{ol}}^2 \epsilon_0^2 \epsilon^2 (q^2 + q_s^2/\epsilon)} \delta(E(\mathbf{k}) - E(\mathbf{k}')) \quad (7.18)$$

It is important to note that scattering by impurities is elastic since the impurity is so much heavier than the electron. Therefore,

$$\begin{aligned} q^2 &= |\mathbf{k} - \mathbf{k}'|^2 = k^2 + k'^2 - 2kk' \cos \theta \\ &= 2k^2(1 - \cos \theta) \\ &= 4k^2 \sin^2 \theta / 2 \end{aligned} \quad (7.19)$$

where θ is the angle between \mathbf{k} and \mathbf{k}' . This simplifies the calculation of $S(\mathbf{k}, \mathbf{k}')$ and of the total scattering rate defined by

$$\frac{1}{\tau_{\text{tot}}} = \sum_{\mathbf{k}'} S(\mathbf{k}, \mathbf{k}') \quad (7.20)$$

The summation over \mathbf{k}' in Eq. (7.20) is conveniently transformed into an integration by choosing a polar coordinate system in which \mathbf{k} points in the k_z -direction. This does not restrict generality and makes the angle θ between \mathbf{k} and \mathbf{k}' equal to the polar angle. Therefore, using an isotropic effective mass, the properties of the δ function ($\delta(a(\mathbf{k} - \mathbf{k}')) \rightarrow \frac{1}{a} \delta(\mathbf{k} - \mathbf{k}')$) and $w = \sin \theta / 2$:

$$\frac{1}{\tau_{\text{tot}}^I} = \frac{16\pi^2 m^* e^4 k}{\hbar^3 V_{\text{ol}} \epsilon^2 \epsilon_0^2} \int_0^1 \frac{w dw}{(2k^2 w^2 + q_s^2/\epsilon)^2} \quad (7.21)$$

where the superscript I of τ_{tot}^I stands for impurity. The integration is then easily performed. (We will calculate below total rates for other scattering mechanisms). The rate given in Eq. (7.21) is for a single impurity (also singly charged). For

\bar{N}_I independent impurities, the rate has to be multiplied by \bar{N}_I , which introduces the impurity density $\bar{N}_I/V_{\text{ol}} = N_I$ as a multiplicative factor. Much research has been devoted to impurity scattering. Approximations beyond the Golden Rule, impurity correlations, and other higher-order effects have been treated. A rather complete reference list has been given by Kuchar et al. [11]. The above treatment is equivalent to that of Brooks and Herring, and is only strictly valid if there are enough screening electrons available. This is the case in inversion layers. Other cases—such as a semiconductor with few free-charge carriers and a high density of donors and acceptors—call for a different treatment.

The total scattering rate represents the probability per unit time that an electron with wave vector \mathbf{k} is scattered into any possible state labeled by the wave vector \mathbf{k}' .

As can be seen, $1/\tau_{\text{tot}}$ is still a function of \mathbf{k} . The electrons are distributed in \mathbf{k} space. The distribution, however, is not necessarily the equilibrium distribution as discussed in Chapter 5 because the application of an electric field will change this distribution function. Therefore, up to this point we are not able to calculate the average value of τ_{tot} which we will denote by $\langle \tau_{\text{tot}} \rangle$. How this average has to be taken will be shown in Chapter 8 by solving the Boltzmann equation.

For the time being, we can argue generally that the average electron energy is given after Boltzmann by $(3/2)kT$, and we therefore should replace the wave vector by an average value corresponding to this energy, that is, by $\mathbf{k} = \sqrt{3kTm^*}/\hbar$ for an isotropic effective mass m^* . This average total scattering time is closely related and approximately equal to Drude's time constant τ . The exact averaging procedure is given by Eq. (8.46). If the screening is weak, the scattering is not randomizing and additional care must be taken to arrive at a relevant time constant which is then the momentum relaxation time or the inverse of the momentum scattering rate.

7.2.2 Phonon Scattering

Even more important than impurity scattering is scattering by the lattice vibrations, the phonons. Phonon scattering provides the mechanism for the energy loss. Impurity scattering is elastic, as we have seen before, and the energy of the electrons would diverge in a constant electric field. The mobility of semiconductors is also often dominated by phonon scattering, at least around room temperature.

Two new aspects are introduced by phonon scattering compared to impurity scattering. First, phonons can be absorbed and emitted, thereby changing the phonon wave function in addition to the electron wave function. The most elegant way to account for this fact is second quantization (Ziman [18]). Here we just state the result: We can treat phonon scattering in the usual way by using Eq. (1.44) and multiplying the result by certain factors. These factors arise from the phonon wave function. It is clear that a phonon can be absorbed only if at least one phonon is present. Therefore, the result for phonon absorption as ob-

tained by Eq. (1.44) has to be multiplied by the average number of phonons that are present in the mode q , just as we multiplied the impurity scattering rate by the number (density) of impurities.

The appropriate multiplier (see Ziman [18]) is the phonon occupation number N_q :

$$N_q = \frac{1}{e^{\hbar\omega_q/kT} - 1} \quad (7.22)$$

In Eq. (7.22) $\hbar\omega_q$ is the phonon energy that for acoustic phonons is given by Eq. (1.21), and is approximately independent of q for optical phonons.

The factor multiplying the phonon emission rate is $(N_q + 1)$. The emission, therefore, is composed of two contributions. One is independent of the phonon occupation number and is termed *spontaneous emission* (just as in the case of light). The second term is proportional to N_q and is called *stimulated emission* because more phonons are emitted if more “stimulating” phonons are present. Again, this is the same as in the case of light where Einstein derived the concept of stimulated emission. The similarity with light is, of course, no accident. Phonons are “bosons” (particles with integer spin) in contrast to electrons (or holes), which are fermions (half-numbered spin). Equation (7.22) is, therefore, also called the Bose distribution.

Although Eq. (7.22) is formally quite similar to Eq. (5.17), the Fermi distribution, the two functions behave numerically differently. The physical reason is that only two electrons with opposite spin fit into a given energy level, whereas the phonons are the merrier the closer and the more of them there are.

An important approximation to Eq. (7.22) is obtained for $\hbar\omega \ll kT$, which gives

$$N_q \approx \frac{1}{1 + \frac{\hbar\omega_q}{kT} - 1} = \frac{kT}{\hbar\omega_q} \quad (7.23)$$

This approximation is also known as equipartition for historical reasons.

The second difference between phonon scattering and impurity scattering is in the scattering potential. Phonons distort the crystal lattice and can create in semiconductors essentially three kinds of potential energy changes: the deformation potential, the piezoelectric potential, and the polar optical potential. We will treat in detail only the deformation potential, which was introduced by Bardeen and Shockley. The basic idea is illustrated in Figure 7.4, which shows the conduction and valence bands as a function of the interatomic distance d . This graph can be calculated from our pseudopotential theory by varying the lattice constant.

If the lattice is displaced by u [see Eq. (1.18)], the energy of the conduction (or valence) band will change by

$$\Delta E_c = E_c(a) - E_c\left(a + \frac{du}{dx}a\right) \quad (7.24)$$

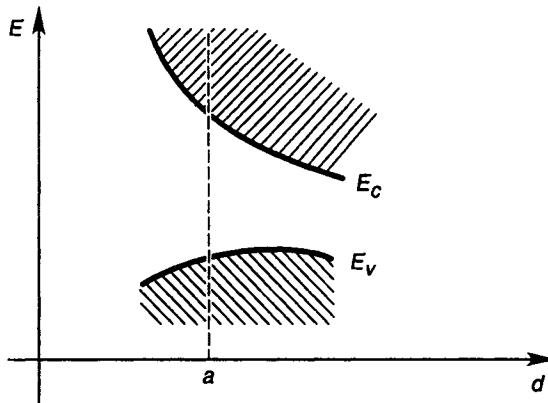


Figure 7.4 Conduction and valence bands as function of interatomic distance. The actual lattice constant of the material is denoted by a .

where a is the lattice constant assumed to be small compared to the phonon wavelength. This gives, by Taylor expansion

$$\Delta E_c = (dE_c/dx)a \quad (7.25)$$

Note that we have implicitly assumed that the lattice displacement u has the same effect as expanding or compressing the whole crystal. This means our concept will actually work only if the phonon wavelength spans many lattice constants. As long as this is true, the effective mass theorem, which we are going to use, also applies.

In three dimensions the change ΔE_c is proportional to the volume change, which is then given in terms of the lattice displacement u by

$$\frac{\Delta V_{\text{ol}}}{V_{\text{ol}}} = \nabla \cdot \mathbf{u}(\mathbf{r}) \quad (7.26)$$

which is basic to Bardeen's deformation potential theory. Therefore,

$$\Delta E_c = V_{\text{ol}} \frac{dE_c}{dV_{\text{ol}}} \nabla \cdot \mathbf{u}(\mathbf{r}) \quad (7.27)$$

The constant $V_{\text{ol}} \frac{dE_c}{dV_{\text{ol}}}$ is usually denoted by Z_A if acoustic phonons are involved, and is called the acoustic deformation potential. Therefore, the perturbing Hamiltonian becomes

$$H_A = \Delta E_c = Z_A \nabla \cdot \mathbf{u}(\mathbf{r}) \quad (7.28)$$

For holes we have a change in the valence band edge ΔE_c , and the deformation potential constant is different. Rigorously, Z_A has to be replaced by a matrix because of the anisotropy of crystals and can even depend on the wavevector [17].

A typical value for the deformation potential is $|Z_A| \approx 8$ eV for semiconductor conduction bands and a somewhat smaller value for valence bands [9].

Because we have already derived \mathbf{u} (see Chapter 1), we can therefore derive the matrix elements. Using Eq. (1.18) in vector form and Eq. (1.49) we find

$$M_{\mathbf{k},\mathbf{k}'} = \pm Z_A i \mathbf{q} \cdot \mathbf{u} \delta_{\mathbf{k}'-\mathbf{k},\pm\mathbf{q}} V_{\text{ol}} \left(N_q + \frac{1}{2} \pm \frac{1}{2} \right)^{1/2} \quad (7.29)$$

Here we have chosen a plus sign for the \mathbf{q} vector in the case of phonon absorption and a minus sign for emission. This choice is made for convenience and has no additional meaning; \mathbf{q} is a vector that can point in any direction. Notice that transversal waves (wave vector \mathbf{q} perpendicular to displacement \mathbf{u}) have a vanishing matrix element (the dot product is zero) and therefore do not scatter the electrons. It is only the longitudinal (\mathbf{q} parallel \mathbf{u}) phonons that contribute to scattering within the approximations that we have used. The factor N_q [lower sign in Eq. (7.29)] is for absorption, while for emission a factor $(N_q + 1)$ needs to be implemented, as discussed above. We do not derive here the value of the displacement \mathbf{u} and refer the interested reader to references [6] and [12].

Noting that \mathbf{k} is changed to $\mathbf{k} \pm \mathbf{q}$ by the scattering process and inserting the value for \mathbf{u} , we can write the matrix element:

$$|\langle \mathbf{k} \pm \mathbf{q} | H'_A | \mathbf{k} \rangle|^2 = \frac{Z_A^2 \hbar \omega_{\mathbf{k}}}{2V_{\text{ol}} \rho v_s^2} \left(N_q + \frac{1}{2} \pm \frac{1}{2} \right) \quad (7.30)$$

Here V_{ol} is the volume of the crystal, v_s the velocity of sound (Chapter 1), ρ the mass density (which is 2.3 g/cm^3 for silicon). The minus sign refers to phonon absorption and the plus sign to phonon emission. Notice that the approximation of Eq. 7.23 gives a matrix element approximately independent of the phonon wave vector \mathbf{q} .

For optical phonons and deformation potential interaction one obtains a similar matrix element. Now, however, one needs to consider the displacement of the two face-centered sublattices against each other. The perturbing energy is then given by $D\mathbf{u}_{\mathbf{q}}$ with D defined in Eq. (7.34) and $\mathbf{u}_{\mathbf{q}}$ being the longitudinal displacement of one subband with respect to the other [7].

$$|\langle \mathbf{k} \pm \mathbf{q} | H'_0 | \mathbf{k} \rangle|^2 = \frac{Z_0^2 \hbar \omega_0}{2V_{\text{ol}} \rho v_s^2} \left(N_q + \frac{1}{2} \pm \frac{1}{2} \right) \quad (7.31)$$

where the index 0 has replaced the indices \mathbf{q} and A of Eq. (7.29). Because $\hbar \omega_0$, the optical phonon energy, is approximately independent of the wave vector \mathbf{q} , N_q is constant and Eq. (7.31) does not depend on \mathbf{q} .

Although the optical phonon matrix element is almost identical to the acoustic one, its origin is quite different; in the case of optical phonons, the two sublattices of the crystal structure vibrate against each other. This is different from the volume change in the case of acoustic phonons. As a consequence, the constant Z_0 and optical phonon scattering are sensitive to the symmetry of the

particular range of the band structure in which the electron is scattered. It turns out—and this is very important to remember—that if the electron is scattered close to the Γ minimum (the same is true for X minima) and if the wave function has spherical symmetry, the matrix element vanishes, that is, optical deformation potential scattering is “forbidden” (Harrison [7]). This is the case for the conduction band minimum of GaAs and is a reason for the high electron mobility in GaAs.

Electrons can also be scattered by polar optical phonons (see Chapter 1). This mechanism is important in the Γ minimum of GaAs and also in InP because of the lack of deformation potential scattering. The polar scattering arises from the polarities of the two different atoms in III-V compounds, as illustrated in Figure 7.5. The displacement illustrated in the figure results in a macroscopic field that can scatter the electrons. Because the potential is of long range, the squared matrix element contains a coulombic factor l/q^2 (Madelung [13]), which means that scattering preferably takes place at small angles ($\mathbf{k} \approx \mathbf{k}'$). The polar crystal also demonstrates very clearly that the electron is not a single entity anymore but has to be viewed together with the crystal lattice. This new entity is called (in the case of a crystal with polar component) a polaron, which is illustrated in Figure 7.6.

This lattice distortion leads to a “renormalization” of the bare electron effective mass as calculated from the band structure of the undistorted crystal. In GaAs, InP, InSb, and so on, this renormalization is rather small (see Appendix C).

In heterostructures, more complicated forms of the polar interaction are possible. For example, at the Si-SiO₂ interface, electrons residing in the silicon can still interact with polar modes (remote polar phonon scattering), as described by Hess and Vogl [10]. The calculation of the total phonon scattering rates from the matrix elements requires still more algebra. Optical deformation potential scattering is relatively easy to calculate. Using Eqs. (7.31) and (7.20), and considering phonon absorption only, we have

$$\frac{1}{\tau_{\text{tot}}^{\text{opd}}} = \frac{\pi Z_0^2 \hbar \omega_0 N_q}{\hbar V_{\text{ol}} \rho v_s^2} \sum_{\mathbf{k}'} \delta(E(\mathbf{k}) - E(\mathbf{k}') + \hbar\omega) \quad (7.32)$$

and only the δ function argument depends on \mathbf{k}' .

The summation over \mathbf{k}' can be converted into an energy integration according to Eq. (5.8). However, care must be taken with the spin factor. Because the electron does not change its spin when interacting with phonons, the initial spin state fixes the final one. Therefore, instead of the density of states with two possibilities for the spin, we now have only one:

$$\begin{aligned} \sum_{\mathbf{k}'} \delta(E(\mathbf{k}) - E(\mathbf{k}') + \hbar\omega) &= \frac{V_{\text{ol}}}{2} \int \delta(E - E + \hbar\omega) g(E') dE' \\ &= \frac{V_{\text{ol}} g(E + \hbar\omega_0)}{2} \end{aligned} \quad (7.33)$$

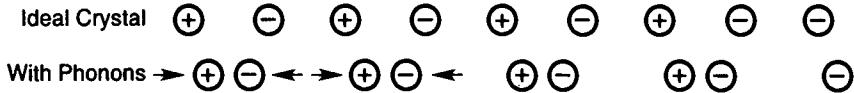


Figure 7.5 A phonon displaces the two sublattices of, for example, GaAs, against each other. For $\mathbf{q} \approx 0$, all negative atoms are displaced toward (or away from) the positive atoms.

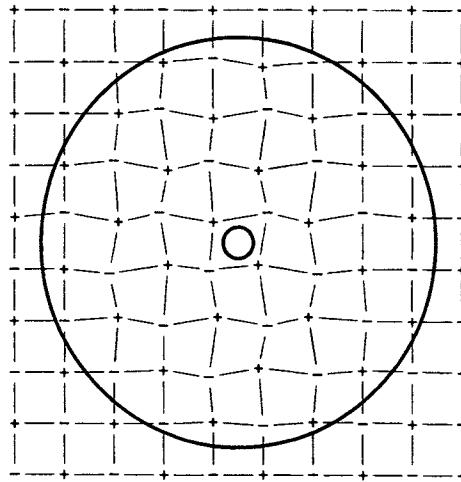


Figure 7.6 Electron plus crystal distortion forming a “polaron.” [After Madelung [13], Figure 59.]

where g is the density of states given by, for example, Eq. (5.6).

In the current literature one frequently finds a coupling constant D defined by

$$D^2 = Z_0^2 \omega_0^2 / v_s^2 \quad (7.34)$$

The total scattering rate for optical deformation potential scattering becomes then

$$\frac{1}{\tau_{\text{tot}}^{\text{opd}}} = \frac{D^2 \sqrt{m_i^* m_t^*}}{\sqrt{2\pi\hbar^3\rho\omega_0}} \left[N_q \sqrt{E + \hbar\omega_0} + (N_q + 1) \sqrt{E - \hbar\omega_0} \right] \quad (7.35)$$

This equation includes absorption and, for $E \geq \hbar\omega_0$, emission of optical phonons as well, for the case that the density of states is given by Eq. (5.11). For low temperatures $N_q = 0$, and the absorption vanishes. It is important to note, however, that spontaneous emission is still possible.

The possible scattering rate $1/\tau_{\text{dot}}^{\text{ac}}$ for acoustic phonon scattering is not as easy to compute because the phonon energy $\hbar\omega_q$ depends on $q = |\mathbf{k} - \mathbf{k}'|$. The summation (or integration) over the δ function then presents a more elaborate algebraic problem, which has been reviewed in detail by Conwell [2]. The result

is (in the equipartition approximation)

$$\frac{l}{\tau_{\text{tot}}^{\text{ac}}} = \frac{\sqrt{2}}{\pi} \frac{Z_A^2 \sqrt{m_l^* m_t^* kT}}{\rho \hbar^4 v_s^2} \sqrt{E} \quad (7.36)$$

For polar optical phonon scattering, Conwell [2] gives the result

$$\frac{l}{\tau_{\text{tot}}^{\text{po}}} = \frac{2e\bar{F}}{\sqrt{2m^* E}} \left[n_q \sinh^{-1} \sqrt{\frac{E}{\hbar\omega_0}} + (N_q + 1) \sinh^{-1} \sqrt{\frac{E - \hbar\omega_0}{\hbar\omega_0}} \right] \quad (7.37)$$

where \bar{F} is a polar coupling field (~ 6000 V/cm for GaAs) and m^* is the effective mass [assumed isotropic for the calculation of Eq. (7.37)].

In contrast to deformation potential scattering, polar optical scattering becomes weaker with higher energy (i.e., the scattering rate decreases). This is because the matrix element is proportional to $l/|\mathbf{k} - \mathbf{k}'|$, which masks the proportionality to the final density of states.

7.2.3 Scattering by a δ -Shaped Potential

As a further example, we calculate the scattering by N_b independent potentials, which have the form of a δ function:

$$V_{\text{ext}} = eV_0 \delta(\mathbf{r}) \quad (7.38)$$

The matrix element is then

$$M_{\mathbf{k}, \mathbf{k}'} = \frac{e}{V_{\text{ol}}} \int_{V_{\text{ol}}} e^{-i\mathbf{k} \cdot \mathbf{r}} V_0 \delta(\mathbf{r}) e^{i\mathbf{k}' \cdot \mathbf{r}} d\mathbf{r}$$

$$M_{\mathbf{k}, \mathbf{k}'} = eV_0 / V_{\text{ol}}$$

which means that $M_{\mathbf{k}, \mathbf{k}'}$ is independent of \mathbf{k} and \mathbf{k}' (a δ function can provide all momenta $\mathbf{q} = \mathbf{k} - \mathbf{k}'$).

Therefore the total scattering rate becomes

$$\sum_{\mathbf{k}'} S(\mathbf{k}, \mathbf{k}') = \frac{2\pi}{\hbar} e^2 \frac{V_0^2}{V_{\text{ol}}^2} \sum_{\mathbf{k}'} \delta(E(\mathbf{k}') - E(\mathbf{k})) \quad (7.39)$$

because we assume that the scattering process is elastic.

Transforming the summation into an integration over energy we get

$$\sum_{\mathbf{k}'} S(\mathbf{k}, \mathbf{k}') = \frac{\pi}{\hbar} e^2 \frac{V_0^2}{V_{\text{ol}}} \int_{-\infty}^{\infty} g(E') \delta(E - E') dE \quad (7.40)$$

$$= \frac{\pi}{\hbar} e^2 V_0^2 g(E) / V_{\text{ol}} \quad (7.41)$$

Equation (7.41) shows that the total scattering rate is proportional to the final density of states. This is a very important result. Of course, it is rigorously true only if the matrix element is independent of \mathbf{k} and \mathbf{k}' . However, because

this proportionality arises from the sum over all possible final states $\sum_{\mathbf{k}'}$ and because the matrix element is in many cases only weakly dependent on \mathbf{k} and \mathbf{k}' , the proportionality to the density of final states often holds approximately. For phonon emission we would obtain a proportionality to $g(E - \hbar\omega)$ and for absorption $g(E + \hbar\omega)$. Notice again that Eq. (7.40) has been divided by a factor of 2 (π/\hbar instead of $2\pi/\hbar$), because in a scattering transition the spin usually does not change and the final spin state is therefore fixed. The δ function potential is a reasonable model for any potential of very short range. The corresponding mobility is

$$\mu = \frac{|e|\tau}{m^*} \approx \frac{V_{\text{ol}}|e|\hbar}{N_b m^* 2\pi e^2 V_0^2 g(3/2kT)} \quad (7.42)$$

where N_b/V_{ol} is the density of scattering centers and $\tau = \langle \tau_{\text{tot}} \rangle$.

7.3 IMPORTANT SCATTERING MECHANISMS IN SILICON AND GALLIUM ARSENIDE

We now discuss the important scattering mechanisms for electrons and holes in silicon and gallium arsenide. First we consider electrons in silicon. The silicon conduction band has six equivalent ellipsoidal minima close to the Brillouin zone boundary at the connection line from Γ to X . Within each minimum we have acoustic phonon scattering, ionized impurity scattering, and perhaps some additional scattering mechanisms such as scattering by neutral impurities. At room temperature, ionized impurity scattering starts to be important if the impurity density exceeds $\sim 10^{17} \text{ cm}^{-3}$. Acoustic phonons are always significant in silicon at room temperature. Optical deformation potential scattering is negligible within the particular minimum because D vanishes for reasons of symmetry. However, scattering between the minima is significant and the corresponding value of D can be large.

It is clear from the conservation of energy and crystal momentum (i.e., including reciprocal lattice vectors) that scattering between the minima on a given axis (e.g., [-100], [100]) is different from scattering between minima on different axis (e.g., [-100], [010]) with respect to phonon energies and momenta. The former process is called g scattering, and the latter is called f scattering. A good approximation for these scattering processes is to use Eq. (7.35) with different coupling constants D and optical phonon energies for the particular processes.

Table 7.1 lists these constants and additional material parameters for silicon. At higher energies a second conduction band becomes important in silicon, which introduces further complications. The total phonon scattering rate, including all scattering in between minima (intervalley) and the two conduction bands (interband) for acoustic as well as optical deformation potential scattering, has been calculated by Tang and Hess [16] and Higman [1] and is shown in Figure 7.7. The literature is actually full of a variety of values for the deformation

Table 7.1 Material Parameters for Silicon

Bulk Material Parameters		
Lattice constant	5.431	Å
Density	2.329	g/cm ³
Dielectric constant	11.7	
Sound velocity	9.04×10 ³	cm/s
X Valley		
Effective masses		
Transverse	0.19	<i>m</i>
Longitudinal	0.916	<i>m</i>
Nonparabolicity	0.5	eV ⁻¹
Acoustic deformation potential	9.5	eV
L Valley		
Effective masses		
Transverse	0.12	<i>m</i>
Longitudinal	1.59	<i>m</i>
Phonons		Deformation
Temperature		Potential <i>D</i>
(K)		(eV/cm)
X-X Intervalley Scattering		
220	3×10 ⁷	<i>f</i>
550	2×10 ⁸	<i>f</i>
685	2×10 ⁸	<i>f</i>
140	5×10 ⁷	<i>g</i>
215	8×10 ⁷	<i>g</i>
720	1.1×10 ⁹	<i>g</i>
X-L Intervalley Scattering		
672	2×10 ⁸	<i>g</i>
634	2×10 ⁸	<i>g</i>
480	2×10 ⁸	<i>g</i>
197	2×10 ⁸	<i>g</i>

potentials. Recent work points toward the validity of a scattering rate, as shown in Figure 7.7, even though different sets of deformation potentials can be used in order to obtain similar total rates. A discussion from more basic principles has been given by Yoder [17].

The scattering mechanisms in the conduction band of gallium arsenide are very different from silicon. Because of the spherical symmetry of the wavefunction at Γ , optical deformation potential scattering is zero (D vanishes because of the symmetry) in the lowest GaAs minimum.

The effective mass in this minimum is very small $m^* = 0.067m_0$. Therefore, acoustic phonon scattering also contributes very little. These two facts are the very reason why GaAs exhibits high electron mobility compared to silicon (at room temperature about a factor of 5 higher than silicon). This higher mobility

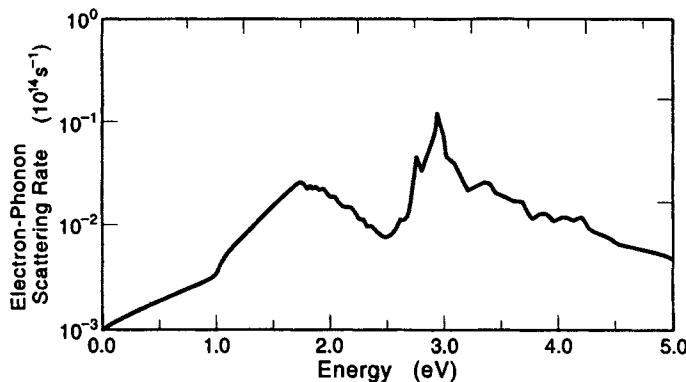


Figure 7.7 The total phonon scattering rate of silicon corresponding to the transport parameters of Table 7.1. [After Higman and Hess [1]].

offers advantages from the viewpoint of high-speed semiconductor devices.

The main scattering mechanisms in GaAs are polar optical scattering and impurity scattering if the density of impurities is significant ($\geq 10^{17} \text{ cm}^{-3}$ at

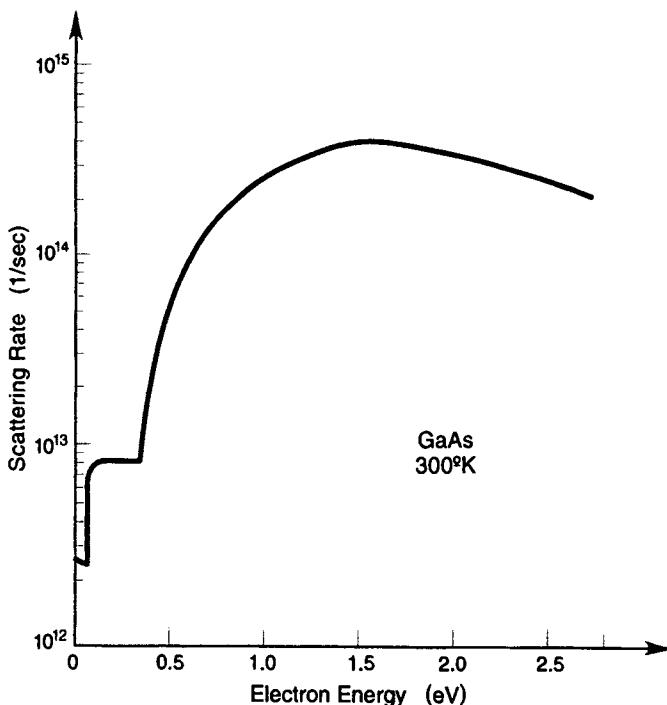


Figure 7.8 Approximate phonon scattering rate of gallium arsenide using the material parameter of Table 7.2. [After Shichijo and Hess [15], Figure 7.]

Table 7.2 Material Parameters for Gallium Arsenide

Bulk Material Parameters		
Lattice constant	5.642	Å
Density	5.36	g/cm ³
Electron affinity	4.07	eV
Piezoelectric constant	0.16	C/m ²
LO phonon energy	0.036	eV
Longitudinal sound velocity	5.24×10^5	cm/s
Optical dielectric constant	10.92	
Static dielectric constant	12.90	
Valley-Dependant Material Parameters		
	$\Gamma[000]$	$L[111]$
Effective mass (m^*/m_0)	0.067	0.222
Nonparabolicity (eV ⁻¹)	0.610	0.461
Energy band gap relative to valence band (eV)	1.439 (0)	1.769 (0.33)
Acoustic deformation potential (eV)	7.0	9.2
Optical deformation potential (eV/cm)	0	3×10^3
Optical phonon energy (eV)	—	0.0343
Number of equivalent valleys	1	4
Intervalley deformation potential D (eV/cm)		
Γ	0	1×10^9
L	1×10^9	1×10^9
X	1×10^9	5×10^8
Intervalley phonon energy (eV)		
Γ	0	0.0278
L	0.0278	0.0290
X	0.0299	0.0293
	$X[100]$	

room temperature). Higher in the conduction band additional minima of ellipsoidal form appear, as shown in Figure 3.4. At an energy of ~ 0.33 eV above the Γ minimum there are “germaniumlike” minima at L , and at ~ 0.5 eV above Γ , silicon-type minima are located close to the X point. Not much is known about the details of scattering in these minima. One can assume, however, that the scattering in the X minima is very similar to scattering in silicon, and scattering in the L minima is similar to that in germanium. There is, of course, also scattering between all these minima such as $X-L$, $\Gamma-X$, and so on. A list of coupling constants that describe well a large set of experimental results (Shichijo and Hess [15]) is given in Table 7.2, and the total scattering rate is shown in Figure 7.8.

The maxima of the valence bands of both Si and GaAs are at Γ and the hole wave function does not have spherical symmetry. Therefore optical deformation potential scattering is important for holes in both GaAs and Si. Also important are acoustic phonon scattering (especially for the heavy holes) and impurity scattering. For a detailed discussion, refer to the review of Rode [14].

The number and variety of scattering mechanisms discussed above for bulk semiconductors are enlarged by the presence of interfaces. Interfaces usually

necessitate dealing with interface roughness scattering (Ferry et al. [5]) and with a host of “remote” scattering mechanisms. Remote here means that the scattering agent and the electrons are in different types of semiconductors. Examples are the already mentioned remote phonon scattering and remote impurity scattering. Remote impurity scattering is an important mechanism in modulation doped AlGaAs-GaAs layers (or other lattice-matched III-V compound systems). Typically the AlGaAs is doped with donors and the electrons move to the GaAs (which has the smaller band gap), where, because of their high density compared to the unintentional GaAs “background doping,” screening is effective and the remote impurities contribute only little to the scattering rate [8]. Consequently, very low total scattering rates and high mobilities can be achieved in these materials, as described in some detail in Chapter 10.

PROBLEMS

- 7.1 Calculate the scattering probability $S(\mathbf{k}, \mathbf{k}')$ for a scattering potential $U(\mathbf{r}) = U_0 \frac{e^{-|\mathbf{r}|/\lambda}}{|\mathbf{r}|^2}$ (use spherical coordinates relative to $\mathbf{q} = \mathbf{k} - \mathbf{k}'$).
- 7.2 Show that random numbers x with a given distribution $f(x)$ in an interval $a \leq x \leq b$ can be obtained (in principle) starting with random numbers r uniformly distributed in the interval $(0, 1)$. Assume that $f(x)$ is normalized on (a, b) .
- [Hint: Let $F(x') = \int_a^{x'} f(x) dx$, and use the fact that $dF = f(x) dx$.] Check your result for $f(x) = \text{constant} = 1/(b-a)$.

REFERENCES

- [1] Abramo, A., et al. “A comparison of numerical solutions of the Boltzmann transport equation for high-energy electron transport in silicon,” *IEEE Transactions on Electron Devices*, vol. 41, 1994, pp. 1646–54.
- [2] Conwell, E. M. “High Field Transport in Semiconductors,” in *Solid State Physics*, ed. F. Seitz, et al. New York: Academic, 1967
- [3] Datta, S. *Electronic Transport in Mesoscopic Systems*, New York: Cambridge Univ. Press, 1995.
- [4] Esaki, L., and Tsu, R. “Superlattice and negative differential conductivity in semiconductors,” *IBM Journal of Research and Development*, vol. 14, 1970, p. 61.
- [5] Ferry, D. K., Hess, K., and Vogl, P. “Physics and Modeling of Submicron Insulated-Gate Field-Effect Transistors 11,” in *VLSI Electronics 2*, ed. N. Einspruch. New York: Academic, 1981, p. 72.
- [6] Ferry, D. K. *Semiconductors*, New York: Macmillan, 1991.
- [7] Harrison, W. A. “Scattering of electrons by lattice vibrations in nonpolar crystals,” *Physical Review*, vol. 104, 1956, pp. 1281–1290.
- [8] Hess, K. “Impurity and phonon scattering in layered structures,” *Applied Physics Letters*, vol. 35, 1979, pp. 484–486.
- [9] Hess, K., and Dow, J. D. “Deformation potentials of bulk semiconductors,” *Solid State Communications*, vol. 40, 1981, p. 371.

- [10] Hess, K., and Vogl, P. "Remote polar phonon scattering in silicon inversion layers," *Solid State Communications*, vol. 30, 1979, pp. 807–809.
- [11] Kuchar, F., Fantner, E., and Hess, K. "Ionized impurity scattering in semiconductors," *Journal of Physics C: Solid State Physics*, vol. 9, 1976, pp. 3165–3171.
- [12] Landsberg, P. T., ed., *Solid State Theory*, New York: Wiley/Interscience, 1969.
- [13] Madelung, O. *Introduction to Solid State Theory*, ed. M. Cardona, et al. New York: Springer Verlag, 1978, pp. 183–187.
- [14] Rode, D. L. "Low-Field Electron Transport," in *Semiconductors and Semimetals*, ed. R. K. Willardson and A. C. Beer, vol. 10. New York: Academic, 1975, pp. 1–89.
- [15] Shichijo, H., and Hess, K. "Band structure dependant transport and impact ionization in GaAs," *Physical Review B*, vol. 23, 1981, pp. 4197–4207.
- [16] Tang, J. Y., and Hess, K. "Impact ionizaton of electrons in silicon," *Journal of Applied Physics*, vol. 54, 1983, pp. 5139–5143.
- [17] Yoder, P. D. PhD Thesis, University of Illinois, Urbana 1993.
- [18] Ziman, J. M. *Elements of Advanced Quantum Theory*, Cambridge: Cambridge Univ. Press, 1969.

CHAPTER 8

THE BOLTZMANN TRANSPORT EQUATION AND ITS APPROXIMATE SOLUTIONS

8.1 DERIVATION

We have derived in the previous chapters a model for the motion and scattering of single electrons in the conduction band of a semiconductor. Ultimately we aim at calculating the electrical current in a device. To achieve this we need to know the distribution of electrons in real space and their velocity. Then we can calculate the current from Eq. (3.48), using Eq. (5.8) to obtain j , the current density:

$$j = -\frac{e}{4\pi^3} \int \mathbf{v} f d\mathbf{k} \quad (8.1)$$

The distribution function appears because we sum only overfilled states as we did for the carrier concentration in Eq. (5.7). However, f is now not equal to the equilibrium Fermi distribution function, because no current can flow in equilibrium. Indeed Eq. (8.1) gives zero current if the Fermi (or Maxwell-Boltzmann) distribution function is inserted; f is then an even function of \mathbf{k} while \mathbf{v} is odd [see Eq. (3.25)]. Therefore we need a nonequilibrium distribution function for f to calculate the current, that is, the distribution function in presence of space-dependent electric fields as they occur in devices. In general this distribution function is a function of \mathbf{r}, \mathbf{v} , and time t [i.e., $f(\mathbf{r}, \mathbf{v}, t)$] because the electric fields can vary in time and space and the velocity distribution of electrons is changed by the strength of the electric fields.

However, the functional dependence on \mathbf{r} and \mathbf{v} gives pause! From a quantum mechanical point of view, the simultaneous specification of space and velocity (momentum) coordinate violates the uncertainty principle. We have made a point throughout that the electrons (and holes) are in extended states and “see” much of the crystal structure. Thus we might think that what we need to do is solve the Schrödinger equation (or equations) for the many-electron system and then calculate the current from the wave function as prescribed in quantum mechanics. However, we know already from Chapter 7 that a quantum approach

leads to an oscillating current or, if we fix the velocity of the electron to certain states in the Brillouin zone, we will obtain a persistent current. In other words, we obtain everything else but the finite dc current proportional to the electric field that Ohms law dictates. What is the reason for this apparent failure of a quantum theory for the electrons? We have explained that reason at the beginning of Chapter 7, introducing friction to obtain the Drude theory. In other words, we need to solve many body Schrödinger equations not only for the electron system but also for the system of lattice vibrations (phonons). This represents in general a very complicated problem that is beyond the scope of this book. The interested reader is referred to the theory of Caldeira and Leggett [1].

A rough understanding of the situation can be obtained by considering the famous two-slit electron interference experiment as described in elementary quantum mechanics texts (e.g., the Feynman lectures). An electron having two slits to pass through creates an interference pattern on a screen that is different from the smooth classical probability curve. However, if we watch the electron (e.g., by light absorption or emission) and then can determine exactly through which slit it went, the pattern on the screen is the classical one. There are, of course, fine details in the transition from quantum interference to the classical probability of detection; but it is clear that by observing the electron (letting it interact with light) we will obtain the classical limit. Similarly the electron is “observed” owing to its interactions with the lattice vibrations (i.e., owing to its generation of Joules heat). The emission of phonons, which can have typically a broad range of energies and wavevectors \mathbf{q} , leads to a “dephasing,” or a loss of coherence of the electron wave function, which in turn leads to a more classical appearance of the electrons. Not all quantum mechanics is “lost,” however; the interactions with the lattice vibrations in semiconductors can be viewed as perturbation, as described previously by the Golden Rule.

Therefore, one can view the electron as propagating unperturbed by phonons over a distance, the inelastic mean free path, which varies in semiconductors typically between 10^{-5} cm and 3×10^{-7} cm, depending on scattering mechanism and electron energy. This distance includes many lattice constants and thus the electron “sees” the crystal structure as a phase coherent quantum mechanical entity with Eqs. (3.25) and (3.26) describing the motion and accelerations. After this unperturbed motion and acceleration, the electron is scattered according to the rules described in Chapter 7, suffers dephasing, and thus appears as a classical electron on a geometric scale that is much larger than the distance between the dephasing scatterings (the inelastic mean free path). Therefore, on that geometric scale, we can regard the electron as a classical object that just follows special equations of motion and scattering rules derived from the underlying quantum mechanics of the crystal. We have thus accounted for the quantum interference effects caused by the crystal and need not be concerned about other quantum interference on a larger scale because of the dephasing. Can we then construct a theory for the distribution function f as Boltzmann did? Not quite yet; we still must be sure that we do not violate the Uncertainty Principle, which

can be stated in terms of the wavevector. If the electron wavevector is known with an uncertainty of Δk we cannot know its position more accurately than Δx determined from the equation $\Delta x \approx 1/\Delta k$. From the above consideration we have already determined that geometric feature sizes need to be larger than the path between dephasing scattering to uphold a semiclassical approach. Therefore we cannot classically describe feature sizes and locate the electron much below a length scale of 10^{-5} cm; for electrons in silicon with an energy corresponding to kT at room temperature, it is more like 10^{-6} cm. Thus $\Delta x \approx 10^{-6}$ – 10^{-5} cm and $\Delta k \approx 10^5$ – 10^6 cm $^{-1}$. Because the Brillouin zone boundary is close to 10^8 cm $^{-1}$, this accuracy in determining \mathbf{k} is often sufficient. Under these conditions, we therefore can define a distribution function $f(\mathbf{k}, \mathbf{r}, t)$ which represents the probability of finding an electron (hole) having a wavevector between \mathbf{k} and $\mathbf{k} + d\mathbf{k}$ and a space coordinate between \mathbf{r} and $\mathbf{r} + d\mathbf{r}$ given the uncertainties dictated by quantum mechanics.

The inexperienced reader may still feel uneasy and prefer to see a derivation of the Boltzmann equation from a rigorous quantum transport equation. This, however, is a cumbersome task and involves all the considerations, restrictions, and discussion that we have given above. The mathematics involved is also excruciatingly difficult and tends to detract from the real issues involved.

The classical derivations of the Boltzmann transport equation (BTE) can proceed in various ways; for example, by using the equation of Liouville. Here we chose a less rigorous derivation that is, however, valid under the conditions discussed above and permits us to automatically include the underlying quantum mechanics of the crystal within the one band approximation and the Golden Rule for scattering, including scattering processes that do not conserve the particle number (electron–hole generation-recombination), a conservation law assumed in the Liouville equation. Consider the rate of change of particles in phase space [seven-dimensional $\mathbf{k}, \mathbf{r}, t$, space that can be treated by considering a cube at \mathbf{r} in real space (Figure 8.1) and at \mathbf{k} in \mathbf{k} -space (Figure 8.2)].

We first calculate how many electrons arrive from the left, enter the cube in the left $dydz$ plane, and how many leave at the corresponding plane to the right in a time period dt . Because the x -direction travel distance of electrons with velocity \mathbf{v} is $v_x dt$ we have

$$\text{incoming: } f(\mathbf{k}, \mathbf{r}, t) d\mathbf{k} dy dz v_x dt \quad (8.2)$$

$$\text{outgoing: } f(\mathbf{k}, \mathbf{r} + dx/y/z, t) d\mathbf{k} dy dz v_x dt \quad (8.3)$$

and the net particle gain, therefore, is

$$\begin{aligned} & -v_x [f(\mathbf{k}, \mathbf{r} + dx/y/z, t) - f(\mathbf{k}, \mathbf{r}, t)] dy dz d\mathbf{k} dt \\ &= -v_x \frac{\partial f}{\partial x} dx dy dz d\mathbf{k} dt \\ &= -\mathbf{v} \cdot \nabla f d\mathbf{k} dr dt \quad \text{in three dimensions} \end{aligned} \quad (8.4)$$

The velocity \mathbf{v} is given by Eq. (3.25).

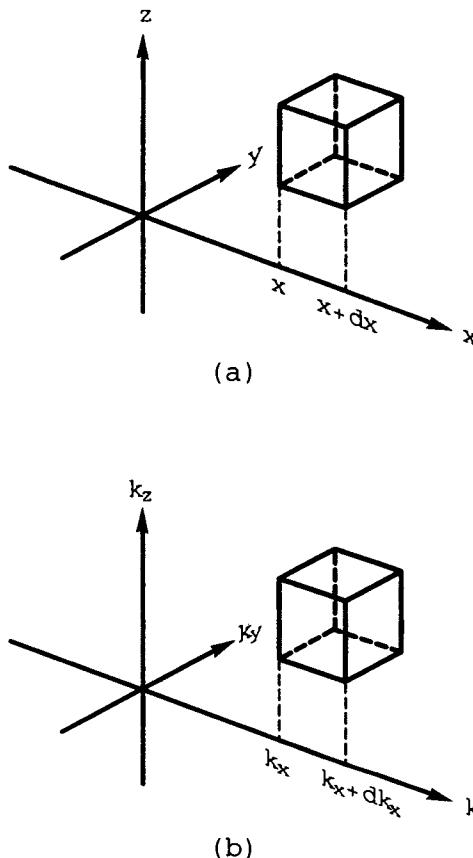


Figure 8.1 Cube in real space used for calculating particle balance in the Boltzmann equation.

In an analogous manner, we obtain the change of the number of electrons at \mathbf{k} in \mathbf{k} -space because of accelerations, replacing dx by dk_x and so on in Figure 8.1 and replacing $\frac{dx}{dt} = v_x$ as used in Eq. (8.2) by $\frac{dk_x}{dt}$. Then we obtain

$$-\frac{d\mathbf{k}}{dt} \cdot \nabla_{\mathbf{k}} f d\mathbf{k} d\mathbf{r} dt \quad (8.5)$$

where according to Eq. (3.26) $\frac{d\mathbf{k}}{dt} = -e\mathbf{F}$ and \mathbf{F} is the electric field.

There is still another possibility to change the number of electrons with wavevector \mathbf{k} at \mathbf{r} . The electrons can be scattered and change their wavevector from \mathbf{k} to \mathbf{k}' at a given point \mathbf{r} in space. Figure 8.2 shows the two infinitesimal volumes in \mathbf{k} space to illustrate the scattering events. The outgoing (out of state \mathbf{k}) electrons are

$$\text{out} = - \sum_{\mathbf{k}'} S(\mathbf{k}, \mathbf{k}') f(\mathbf{k}, \mathbf{r}, t) d\mathbf{k} d\mathbf{r} dt \quad (8.6)$$

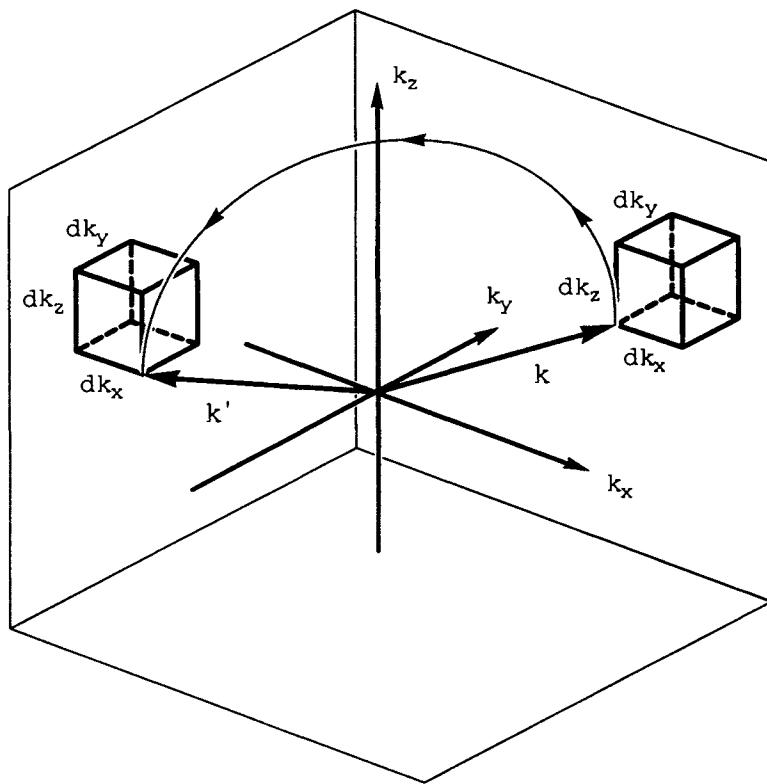


Figure 8.2 Cubes in \mathbf{k} space used for calculating particle balance in the Boltzmann equation.

The factor $f(\mathbf{k}, \mathbf{r}, t)$ is necessary because an electron has to be in the \mathbf{k} state to be scattered out. In degenerate systems (Fermi statistics), an additional factor $1 - f(\mathbf{k}', \mathbf{r}, t)$ arises from the Pauli principle. The incoming (into the \mathbf{k} state) electrons are

$$\text{in} = \sum_{\mathbf{k}'} S(\mathbf{k}', \mathbf{k}) f(\mathbf{k}', \mathbf{r}, t) d\mathbf{k} d\mathbf{r} dt \quad (8.7)$$

Again, the Pauli principle will call for a factor $1 - f(\mathbf{k}, \mathbf{r}, t)$. The Boltzmann equation is obtained by balancing the particle numbers and the change in f given by the net change of incoming and outgoing particles. Therefore, we have

$$\begin{aligned} \frac{\partial f(\mathbf{k}, \mathbf{r}, t)}{\partial t} &= -\mathbf{v} \cdot \nabla f(\mathbf{k}, \mathbf{r}, t) - \frac{1}{\hbar} \mathbf{F}_0 \cdot \nabla_{\mathbf{k}} f(\mathbf{k}, \mathbf{r}, t) \\ &\quad + \sum_{\mathbf{k}'} \{f(\mathbf{k}', \mathbf{r}, t) S(\mathbf{k}', \mathbf{k}) - f(\mathbf{k}, \mathbf{r}, t) S(\mathbf{k}, \mathbf{k}')\} \end{aligned} \quad (8.8)$$

where \mathbf{F}_0 is the force ($-e\mathbf{F}$ for an electric field \mathbf{F}).

If we include the factors arising from the Pauli principle as discussed above

and replace \mathbf{v} as dictated by Eq. (3.25), we arrive at

$$\begin{aligned} \frac{\partial f(\mathbf{k}, \mathbf{r}, t)}{\partial t} = & -\frac{1}{\hbar} \nabla_{\mathbf{k}} E(\mathbf{k}) \cdot \nabla f(\mathbf{k}, \mathbf{r}, t) - \frac{1}{\hbar} \mathbf{F}_0 \cdot \nabla_{\mathbf{k}} f(\mathbf{k}, \mathbf{r}, t) \\ & + \sum_{\mathbf{k}'} \{f(\mathbf{k}', \mathbf{r}, t)(1 - f(\mathbf{k}, \mathbf{r}, t))S(\mathbf{k}', \mathbf{k}) \\ & - f(\mathbf{k}, \mathbf{r}, t)(1 - (f(\mathbf{k}', \mathbf{r}, t))S(\mathbf{k}, \mathbf{k}')\} \quad (8.9) \end{aligned}$$

This equation is extremely difficult to solve because it contains the large sum over all \mathbf{k}' . The sum can be transformed in an integral as prescribed by Eq. (5.8), but then our equation is an integro-differential equation. Our goal is to find solutions to Eq. (8.9) and this can indeed be achieved using the Monte Carlo method, which is discussed in a later section. However, to proceed explicitly, we continue with simplifying Eq. (8.8). We derive from the integro-differential equation a differential equation by involving what one calls the relaxation time approximation. Equation (8.8) is still different from the original Boltzmann equation in that it contains the band structure of the solid through Eq. (3.25). We will, however, often loosely call it the Boltzmann equation.

8.2 SOLUTIONS OF THE BOLTZMANN EQUATION IN THE RELAXATION TIME APPROXIMATION

If the collisions are elastic, the complicated collision sum can be replaced by a relaxation time. However, the most important collisions in semiconductors are the inelastic collisions with lattice vibrations. Fortunately, elasticity is not a necessary condition for simplifying the collision sum (integral). One can show simplification of Eq. (8.8) and even Eq. (8.9) under very general conditions. In fact one needs only certain symmetry properties of the distribution function f and of the scattering rate $S(\mathbf{k}, \mathbf{k}')$. These are very often satisfied and make the relaxation time approximation an excellent tool, even an exact tool. The reason the relaxation time approximation is often blamed for disagreement between experiment and theory is that additional approximations are often made that cannot be justified.

We proceed assuming that the distribution function can be portioned into a function f_0 even in \mathbf{k} and a function f_1 which is odd in \mathbf{k} which means

$$f = f_0 + f_1 \quad \text{with} \quad \begin{cases} f_0(\mathbf{k}, \mathbf{r}, t) = f_0(-\mathbf{k}, \mathbf{r}, t) \\ f_1(\mathbf{k}, \mathbf{r}, t) = -f_1(-\mathbf{k}, \mathbf{r}, t) \end{cases} \quad (8.10)$$

This partition is always possible ($f(\mathbf{k}) - f(-\mathbf{k}) = 2f_1$, $f(\mathbf{k}) + f(-\mathbf{k}) = 2f_0$). To achieve simplification of the integral we assume the sufficient condition:

$$S(k_i, k'_i) = S(k_i, -k'_i) = S(-k_i, k'_i) = S(-k_i, -k'_i) \quad (8.11)$$

Here $i = x, y, z$. In other words, the transition probability is even in all wavevector components.

Equation (8.11) is exact for optical deformation potential scattering and a good approximation for acoustic phonon scattering. Polar optical scattering cannot be treated this way because the matrix element is inversely proportional to $|\mathbf{k} - \mathbf{k}'|$. The precise treatment of polar optical scattering therefore necessitates the solution of Eq. (8.8) without the simplifications that we will sketch in the following.

With the above assumptions, the collision term in Eq. (8.8) reduces to

$$\sum_{\mathbf{k}'} \{ [f_0(\mathbf{k}', \mathbf{r}, t)S(\mathbf{k}', \mathbf{k}) - f_0(\mathbf{k}, \mathbf{r}, t)S(\mathbf{k}, \mathbf{k}')] - f_1(\mathbf{k}, \mathbf{r}, t)S(\mathbf{k}, \mathbf{k}') \} \quad (8.12)$$

The term containing $f_1(\mathbf{k}', \mathbf{r}, t)$ vanishes because f_1 is odd and S is even in \mathbf{k}' . In equilibrium the terms containing f_0 cancel each other; f_0 is the Maxwell-Boltzmann (or Fermi) distribution and we have detailed balance—that is, equal numbers of particles are scattered in and out of each \mathbf{k} -space element. Under nonequilibrium conditions, this ceases to be true, and f_0 differs often significantly from the Maxwell-Boltzmann distribution. Then we denote the terms containing f_0 symbolically by

$$\frac{\partial f_0}{\partial t} |_{\text{coll}}$$

and the Boltzmann equation can be written as

$$\begin{aligned} \frac{\partial f(\mathbf{k}, \mathbf{r}, t)}{\partial t} = & -\mathbf{v} \cdot \nabla f(\mathbf{k}, \mathbf{r}, t) - \frac{1}{\hbar} \mathbf{F}_0 \cdot \nabla_{\mathbf{k}} f(\mathbf{k}, \mathbf{r}, t) \\ & + \frac{\partial f_0}{\partial t} |_{\text{coll}} - f_1(\mathbf{k}, \mathbf{r}, t) \sum_{\mathbf{k}'} S(\mathbf{k}', \mathbf{k}) \end{aligned} \quad (8.13)$$

As shown below, for the case of small electric fields \mathbf{F} , the term $\frac{\partial f_0}{\partial t} |_{\text{coll}}$ vanishes (as F^2) and f_0 approaches the Maxwell-Boltzmann distribution. Then the Boltzmann equation reduces to a partial differential equation of first order. This simplification was our goal when partitioning f into an even and odd part and permits us to transform the BTE into a differential equation. The sum $\sum_{\mathbf{k}}$ is now present only as a factor equal to the total rate $1/\tau_{\text{tot}}$ according to our previous definition.

If Eq. (8.11) does not hold, the scattering is not randomizing; that is, the electron can be scattered in certain preferable directions. Both ionized impurity scattering and the polar optical interaction scatter mainly in the forward direction because the matrix element is inversely proportional to $|\mathbf{k} - \mathbf{k}'|$ and becomes largest for $|\mathbf{k} - \mathbf{k}'| \rightarrow 0$. A relaxation time cannot then be defined, although it may still be a good approximation for certain energy ranges.

For nonrandomizing elastic scattering processes and isotropic effective mass, one still can derive the relaxation time form of the Boltzmann equation except that $1/\tau_{\text{tot}}(\mathbf{k})$ must be replaced by $1/\tau_m(\mathbf{k})$ with

$$\frac{1}{\tau_m(\mathbf{k})} = \frac{V_{\text{ol}}}{8\pi^3} \int d\mathbf{k}' S(\mathbf{k}, \mathbf{k}') (1 - \cos \theta) \quad (8.14)$$

which replaces Eq. (7.20) and differs by the factor $(1 - \cos \theta)$, where θ is the angle between \mathbf{k} and \mathbf{k}' . $\frac{1}{\tau_m}$ is the momentum scattering rate (and τ_m is the momentum relaxation time). Note that scattering events in the direction of the original momentum are not counted because $(1 - \cos \theta) = 0$. Only if the momentum direction changes do we have contributions to $1/\tau_m$. It seems natural that only these processes should be counted, and indeed, for elastic scattering, nothing happens at all if \mathbf{k} does not change direction. This can be an extremely important correction, because impurity scattering, in the range of weak screening occurs mainly in forward direction ($\mathbf{k} \approx \mathbf{k}'$). Then $1/\tau_{\text{tot}}$ can be extremely large compared to $1/\tau_m$. For inelastic processes the difference is not as significant. Imagine an electron at an energy equal to the optical phonon energy and emitting a phonon. This electron loses all its energy in the scattering process even if \mathbf{k} and \mathbf{k}' are parallel. Therefore the wavevector magnitude changes drastically and $|\mathbf{k} - \mathbf{k}'|$, which appears in the denominator of the matrix element, stays finite.

From this discussion, it is clear that the choice of the relaxation time for f_1 in the Boltzmann equation requires care. An exact definition of the relaxation time that rests on the properties of $S(\mathbf{k}, \mathbf{k}')$ may not be possible at all. However, for many important cases, such as scattering for electrons in silicon, the relaxation time approximation is well justified, except if we consider anisotropies, which are relatively unimportant. We will discuss later the Monte Carlo method, which is capable of obtaining solutions of the Boltzmann equation even if the relaxation time approximation fails. Note, however, that the Monte Carlo method is mainly important because of its other capabilities, for example, the exact inclusion of band-structure effects. The errors from the relaxation time approximation are usually too small to justify the use of such a complex method. As we will see immediately, the relaxation time approximation permits us to solve the Boltzmann equation fairly easily and gives us physical insight of what a general solution would look like.

A straightforward solution of Eq. (8.13) can be obtained for time- and space-independent distribution functions in the limiting case of small fields. Then the BTE reads in the relaxation time approximation

$$-\frac{\mathbf{F}_0}{\hbar} \cdot \nabla_{\mathbf{k}} f(\mathbf{k}, \mathbf{r}, t) = \frac{f_1}{\tau_{\text{tot}}} \quad (8.15)$$

Because $f = f_0 + f_1$ with the assumption that $f_1 \ll f_0$ and that f_0 approaches the Maxwell-Boltzmann distribution for small forces \mathbf{F}_0 , we obtain

$$f_1 = -\tau_{\text{tot}} \frac{\mathbf{F}_0}{\hbar} \cdot \nabla_{\mathbf{k}} e^{(E_F - E)/kT} \quad (8.16)$$

For higher electric fields, $\frac{\partial f_0}{\partial t}|_{\text{coll}}$ does not vanish, and the solution of the BTE is more involved. Assuming again steady state ($\partial f / \partial t = 0$) and spatial homogeneity ($\nabla f = 0$), the BTE reads

$$\frac{\mathbf{F}_0}{\hbar} \cdot (\nabla_{\mathbf{k}} f_0(\mathbf{k}) + \nabla_{\mathbf{k}} f_1(\mathbf{k})) = -\frac{f_1}{\tau_{\text{tot}}(\mathbf{k})} + \frac{\partial f_0}{\partial t}|_{\text{coll}} \quad (8.17)$$

This equation contains even and odd functions: f_1 and $\nabla_{\mathbf{k}} f_0$ are odd; $\nabla_{\mathbf{k}} f_1$ and $\frac{\partial f_0}{\partial t}|_{\text{coll}}$ are even. The even and odd functions must cancel separately. We therefore obtain two equations:

$$\frac{\mathbf{F}_0}{\hbar} \cdot \nabla_{\mathbf{k}} f_0(\mathbf{k}) = -\frac{f_1}{\tau_{\text{tot}}(\mathbf{k})} \quad (8.18)$$

and

$$\frac{\mathbf{F}_0}{\hbar} \cdot \nabla_{\mathbf{k}} f_1(\mathbf{k}) = +\frac{\partial f_0}{\partial t}|_{\text{coll}} \quad (8.19)$$

Spatial inhomogeneity and time dependence can easily be included into these equations, but this complicates the algebraic steps shown below. We therefore omitted these terms. Notice that Eq. (8.18) gives the desired function f_1 if we know f_0 . This fact is used later when we develop the equation of moments. We will then assume that f_0 is known and it has a certain functional form. It is this assumption of the form of f_0 that leads to various approximations. The relaxation time approximation inherent in Eq. (8.18) is often a good approximation, particularly in nonpolar semiconductors, except that τ_{tot} may have to be replaced by τ_m . f_0 can, of course, be calculated from Eq. (8.19). This is shown in the following equation using assumptions that simplify the algebra sufficiently to arrive at an explicit solution.

We consider for the collision operator for optical deformation potential scattering only. From Eq. (7.31), we have

$$S(\mathbf{k}, \mathbf{k}') = \bar{C} \left(N_q + \frac{1}{2} \pm \frac{1}{2} \right) \delta(E - E' \pm \hbar\omega_0) \quad (8.20)$$

where \bar{C} is a constant that can be deduced from Eq. (7.31). Therefore,

$$\begin{aligned} -\frac{\partial f_0}{\partial t}|_{\text{coll}}^{\text{opt}} &= \int d\mathbf{k}' \bar{C} [f_0(E)(N_q + 1)\delta(E - E' + \hbar\omega_0) \\ &\quad + f_0(E)N_q\delta(E - E' - \hbar\omega_0) - f_0(E')N_q\delta(E - E' - \hbar\omega_0) \\ &\quad - f_0(E')(N_q + 1)\delta(E - E' + \hbar\omega_0)]. \end{aligned} \quad (8.21)$$

The integration over \mathbf{k}' is best converted into an integration over energy E' by introducing the density of states $g(E')/2$. The factor 1/2 arises because scattering by phonons does not change the spin. Performing the integration over the δ -function gives

$$\begin{aligned} -\frac{\partial f_0}{\partial t}|_{\text{coll}}^{\text{opt}} &= \frac{\bar{C}}{2} [g(E - \hbar\omega_0)f_0(E)(N_q + 1) + g(E + \hbar\omega_0)f_0(E)N_q \\ &\quad - f_0(E + \hbar\omega_0)g(E + \hbar\omega_0)(N_q + 1) - g(E + \hbar\omega_0)f_0(E - \hbar\omega_0)N_q] \end{aligned} \quad (8.22)$$

To simplify Eq. (8.22) we need to deal with the terms containing $f(E \pm \hbar\omega_0)$. These terms, arising from the scattering integral for in- and out-scattering, are cumbersome; from standard calculus we know only how to solve differential

equations for functions of independent variables. But now we have an equation containing $f(E)$ and $f(E \pm \hbar\omega_0)$. For energies E much higher than the phonon energy we can find a remedy to this problem. Normalizing the energy to kT and defining $\tilde{x} = E/kT$ as well as $\tilde{\omega} = \hbar\omega_0/kT$, we can Taylor-expand both $f(E)$ and $g(E)$ up to second order in terms of the normalized quantities \tilde{x} and $\tilde{\omega}$ as

$$f(\tilde{x} \pm \tilde{\omega}) = f(\tilde{x}) \pm \tilde{\omega} \frac{\partial f(\tilde{x})}{\partial \tilde{x}} + \frac{\tilde{\omega}^2}{2} \frac{\partial^2 f(\tilde{x})}{\partial \tilde{x}^2} \quad (8.23)$$

Inserting these expansions into Eq. (8.22) and neglecting terms of order $\tilde{\omega}^3$ or higher, we then obtain

$$\begin{aligned} \frac{\partial f_0(\tilde{x})}{\partial t} \Big|_{\text{coll}}^{\text{opt}} &= \frac{\tilde{C}\tilde{\omega}}{2} \left[2 \frac{\partial g(\tilde{x})}{\partial \tilde{x}} f_0(\tilde{x}) + g(\tilde{x}) \frac{\partial f_0(\tilde{x})}{\partial \tilde{x}} \right] \\ &\quad + \frac{\tilde{C}\tilde{\omega}^2}{2} \left[(2N_q + 1) \frac{\partial g(\tilde{x})}{\partial \tilde{x}} \frac{\partial f_0(\tilde{x})}{\partial \tilde{x}} + \frac{1}{2} (2N_q + 1) g(\tilde{x}) \frac{\partial^2 f_0(\tilde{x})}{\partial \tilde{x}^2} \right] \end{aligned} \quad (8.24)$$

The density of states $g(E)$ is in general a complicated function and can only be obtained numerically, for example, by the procedures described in Chapter 5. However, in three dimensions $g(E)$ is proportional to \sqrt{E} for small E ; therefore, let us assume $g(\tilde{x}) = \tilde{g}\sqrt{\tilde{x}}$ with \tilde{g} being a constant. Note also that we have to put $g(\tilde{x}) = 0$ for $\tilde{x} \leq 0$ if we chose the conduction band edge as the zero of energy. Additionally, for small phonon energies $\tilde{\omega} \ll 1$ we can expand the occupation number N_q as

$$N_q = \frac{1}{e^{\tilde{\omega}} - 1} \approx \frac{1}{\tilde{\omega}} - \frac{1}{2} \quad (8.25)$$

Using these approximations, we may simplify the collision operator as

$$\begin{aligned} \frac{\partial f_0(\tilde{x})}{\partial t} \Big|_{\text{coll}}^{\text{opt}} &= \frac{\tilde{C}\tilde{g}\tilde{\omega}}{2} \left[\frac{1}{\sqrt{\tilde{x}}} f_0(\tilde{x}) + \sqrt{\tilde{x}} \frac{\partial f_0(\tilde{x})}{\partial \tilde{x}} \right] \\ &\quad + \frac{\tilde{C}\tilde{g}\tilde{\omega}}{2} \left[\frac{1}{\sqrt{\tilde{x}}} \frac{\partial f_0(\tilde{x})}{\partial \tilde{x}} + \sqrt{\tilde{x}} \frac{\partial^2 f_0(\tilde{x})}{\partial \tilde{x}^2} \right] \end{aligned} \quad (8.26)$$

$$= \frac{\tilde{C}\tilde{g}\tilde{\omega}}{2} \frac{1}{\sqrt{\tilde{x}}} \frac{\partial}{\partial \tilde{x}} \left[\tilde{x} f_0(\tilde{x}) + \tilde{x} \frac{\partial f_0(\tilde{x})}{\partial \tilde{x}} \right] \quad (8.27)$$

To proceed with the solution of Eq. (8.18) and Eq. (8.19), we first insert Eq. (8.18) into Eq. (8.19) and obtain

$$\frac{1}{\hbar^2} \mathbf{F}_0 \cdot \nabla_{\mathbf{k}} [\tau_{\text{tot}}(\tilde{x}) \mathbf{F}_0 \cdot \nabla_{\mathbf{k}} f_0(\tilde{x})] = - \frac{\partial f_0(\tilde{x})}{\partial t} \Big|_{\text{coll}}^{\text{opt}} \quad (8.28)$$

At this point it is useful to consider how the relaxation time $\tau_{\text{tot}}(\tilde{x})$ depends on the energy \tilde{x} . Because a simple phase-space argument indicates an inverse square root energy dependence for stationary δ -shaped scatterers, let us assume

that $\tau_{\text{tot}}(\bar{x}) = \tau_0 \bar{x}^{-1/2}$. Furthermore, the gradient operator for isotropic functions can be rewritten using the chain rule as

$$\nabla_{\mathbf{k}} = \frac{\mathbf{k}}{|\mathbf{k}|} \frac{\partial}{\partial |\mathbf{k}|} = \frac{\mathbf{k}}{|\mathbf{k}|} \frac{\partial \bar{x}}{\partial |\mathbf{k}|} \frac{\partial}{\partial \bar{x}} = \frac{\hbar^2}{m^* k_B T} \mathbf{k} \frac{\partial}{\partial \bar{x}} \quad (8.29)$$

Using this result repeatedly on (8.28), we get the following sequence of equations

$$\frac{\tau_0}{m^* k_B T} \mathbf{F}_0 \cdot \nabla_{\mathbf{k}} \left[\mathbf{F}_0 \cdot \mathbf{k} \frac{1}{\sqrt{\bar{x}}} \frac{\partial f_0(\bar{x})}{\partial \bar{x}} \right] = - \frac{\partial f_0(\bar{x})}{\partial t} |_{\text{coll}}^{\text{opt}} \quad (8.30)$$

$$\frac{\tau_0 \mathbf{F}_0^2}{m^* k_B T} \frac{1}{\sqrt{\bar{x}}} \frac{\partial f_0(\bar{x})}{\partial \bar{x}} + \frac{\mathbf{F}_0 \cdot \mathbf{k}}{m^* k_B T} \mathbf{F}_0 \cdot \nabla_{\mathbf{k}} \left[\frac{1}{\sqrt{\bar{x}}} \frac{\partial f_0(\bar{x})}{\partial \bar{x}} \right] = \quad (8.31)$$

$$\frac{\tau_0 \mathbf{F}_0^2}{m^* k_B T} \frac{1}{\sqrt{\bar{x}}} \frac{\partial f_0(\bar{x})}{\partial \bar{x}} + \frac{(\mathbf{F}_0 \cdot \mathbf{k})^2}{m^* k_B T} \frac{2\bar{x}}{\mathbf{k}^2} \frac{\partial}{\partial \bar{x}} \left[\frac{1}{\sqrt{\bar{x}}} \frac{\partial f_0(\bar{x})}{\partial \bar{x}} \right] = \quad (8.32)$$

The vector product $(\mathbf{F}_0 \cdot \mathbf{k})^2 / \mathbf{k}^2$ is still an obstacle for an explicit solution and is usually replaced by its average value [2]. In Section 8.3 we show that this average is equal to $\mathbf{F}_0^2 / 3$. This gives

$$\frac{2\mathbf{F}_0^2 \tau_0}{3m^* k_B T} \left[\frac{1}{\sqrt{\bar{x}}} \frac{\partial f_0(\bar{x})}{\partial \bar{x}} + \sqrt{\bar{x}} \frac{\partial^2 f_0(\bar{x})}{\partial \bar{x}^2} \right] = - \frac{\partial f_0(\bar{x})}{\partial t} |_{\text{coll}}^{\text{opt}} \quad (8.33)$$

and after extracting the square root and inserting Eq. (8.27), we get

$$\frac{2\mathbf{F}_0^2 \tau_0}{3m^* k_B T} \frac{1}{\sqrt{\bar{x}}} \frac{\partial}{\partial \bar{x}} \left[\bar{x} \frac{\partial f_0(\bar{x})}{\partial \bar{x}} \right] = - \frac{\bar{C} \bar{g} \bar{\omega}}{2} \frac{1}{\sqrt{\bar{x}}} \frac{\partial}{\partial \bar{x}} \left[\bar{x} f_0(\bar{x}) + \bar{x} \frac{\partial f_0(\bar{x})}{\partial \bar{x}} \right] \quad (8.34)$$

We now replace the force \mathbf{F}_0 with $-e\mathbf{F}$, where \mathbf{F} is the electric field, and collect all constants into a single factor \mathbf{F}^2 / F_c^2 , with F_c called the critical field. After multiplying the equation with $\sqrt{\bar{x}}$ only total derivatives remain and we may integrate over \bar{x} . Because $f_0(\bar{x})$ vanishes as energy \bar{x} approaches infinity, the first integration constant is zero, and we obtain

$$\frac{\mathbf{F}^2}{F_c^2} \frac{\partial f_0(\bar{x})}{\partial \bar{x}} = -f_0(\bar{x}) - \frac{\partial f_0(\bar{x})}{\partial \bar{x}} \quad (8.35)$$

which has the solution

$$f_0(\bar{x}) = C \exp \left(-\bar{x} / \left[1 + \frac{\mathbf{F}^2}{F_c^2} \right] \right) \quad (8.36)$$

It is important to realize that f_0 still looks like a Maxwell-Boltzmann distribution. The only difference is now that the temperature has to be replaced by a temperature T_c of the electrons (charge carriers), which in in general can be different from the temperature of the crystal lattice that enters, for example, N_q .

We will denote from now on the lattice temperature or equilibrium temperature of the electrons (holes) by T_L . Then

$$T_c \approx T_L \left(1 + \frac{F^2}{F_c^2} \right), \quad (8.37)$$

and T_c can be considerably higher than the lattice temperature.

We will later derive an equation like this from energy balance considerations by letting the average energy be equal to $\frac{3}{2}kT_c$. In this way, that is, as a measure for average energy, a carrier temperature always can be defined even if the distribution function is not of Maxwell-Boltzmann type as in Eq. (8.36). Indeed, had we used the complicated real form of the density of states instead of our simple approximation and an energy-dependent τ_0 , we would not have arrived at the simple form of Eq. (8.36), or the Maxwell-Boltzmann type. Nevertheless, in this more complicated case an average energy can be defined and also denoted by $\frac{3}{2}kT_c$, where T_c represents a fictitious temperature. Then, we could again arrive at a formula similar to Eq. (8.37). The critical field F_c , the field at which the carrier temperature T_c is twice the lattice temperature T_L is around 10^4 V/cm for electrons in silicon.

In a transistor, such fields are common and the electron temperature can be many times the lattice temperature. Why is it then that the transistor still may feel cool when touched or investigated with a temperature sensor? The reason is that one feels the lattice temperature only. The electrons cannot leave the crystal because there is a work function (see Chapter 10) to be overcome. This work function is typically several electron volts corresponding to thousands of degrees Kelvin for T_c . For extremely high fields ($F \simeq 10^6$ V/cm), electrons are heated to such high temperatures. However, Eq. (8.37) is then not reliable because at such high energies the density of states becomes extremely nonlinear. Then one needs to use a Monte Carlo approach to calculate distribution function and average energy.

The constant C in Eq. (8.35) can be determined from the boundary condition of constant carrier concentration (if indeed it is constant and generation-recombination processes are negligible):

$$n = \int_0^\infty f_0 g(E) dE$$

For $g(\bar{x}) = \bar{g}\sqrt{\bar{x}}$ we have

$$n = \bar{g} \int_0^\infty \sqrt{\bar{x}} C e^{-E/kT_c} dE$$

which gives

$$C = \frac{nT_L}{\bar{g}kT_c^2} \Gamma(3/2)$$

We can then write Eq. (8.35) exactly in the form of a Maxwell-Boltzmann distribution and therefore define

$$C = \exp(E_{\text{QF}}/kT_c) \quad (8.38)$$

E_{QF} is different from the Fermi level (we do not have equilibrium) and is called the quasi-Fermi level (the term imref, which is Fermi backwards, is sometimes used).

Equation (8.38) gives

$$E_{\text{QF}} = kT_c \ln \left(\frac{\Gamma(3/2)nT_L}{\bar{g}kT_c} \right)^2 \quad (8.39)$$

The major conclusion of this section is then the following: If electric fields are applied to semiconductors, the even part of the distribution function may still be Maxwell-Boltzmann-like. However, temperature has to be replaced by a carrier temperature T_c and the quantity corresponding to the Fermi energy becomes also a function of T_c as can be seen from Eq. (8.39).

Later in this chapter, we will see that even under less stringent approximations T_c assumes the form

$$T_c = T_L + S(F) \quad (8.40)$$

where the function $S(F)$ vanishes at least as F^2 for $F \rightarrow 0$ and T_L is the temperature of the crystal lattice; that is, the temperature without applied field.

As emphasized above, we have to distinguish from now on between electron and lattice temperatures and between Fermi level and quasi-Fermi level, because the two are not necessarily the same. The reason is simply that electrons gain energy from the field and can lose it only if their temperature is raised above the equilibrium value. The concept of quasi-Fermi levels can be generalized to variable carrier concentration applied in Section 8.4 and is in Chapter 9 for the treatment of generation recombination.

8.3 DISTRIBUTION FUNCTION AND CURRENT DENSITY

Having calculated the distribution function, we can obtain the current density from the definition, Eq. (3.48). The integral over $\mathbf{v}f_0$ vanishes since f_0 is even, and we have

$$\mathbf{j} = -\frac{e}{4\pi^3} \int \mathbf{v}f_1 d\mathbf{k} \quad (8.41)$$

From Eq. (8.18), \mathbf{j} becomes

$$\mathbf{j} = -\frac{e^2}{4\pi^3} \int \tau_{\text{tot}}(\mathbf{k}) \frac{\partial f_0}{\partial E} \mathbf{v}(\mathbf{v} \cdot \mathbf{F}) d\mathbf{k} \quad (8.42)$$

In components, Eq. (8.42) reads

$$\sigma_{li} = -\frac{e^2}{4\pi^3} \int \tau_{\text{tot}}(\mathbf{k}) \frac{\partial f_0}{\partial E} v_l v_i d\mathbf{k} \quad (8.43)$$

where σ_{li} is a component of the conductivity matrix as defined by Eq. (2.3).

If $\tau(\mathbf{k})$ depends only on the absolute value of \mathbf{k} (or the energy) and if the band structure is isotropic, σ_{li} becomes diagonal, that is, $\sigma_{li} = \sigma \delta_{l,i}$.

From the definition of the mobility ($\sigma = en\mu$), one obtains

$$en\mu = -\frac{e^2}{4\pi^3} \int \tau_{\text{tot}}(\mathbf{k}) \frac{\partial f_0}{\partial E} v_i^2 d\mathbf{k} \quad (8.44)$$

where $i = x, y, \text{ or } z$. Equation (5.7) then gives

$$\mu = -\frac{e}{4\pi^3} \frac{\int \tau_{\text{tot}}(\mathbf{k}) \frac{\partial f_0}{\partial E} v_i^2 d\mathbf{k}}{\frac{1}{4\pi^3} \int f_0 d\mathbf{k}} \quad (8.45)$$

Assuming that f_0 is a Maxwellian distribution at carrier temperature T_c and writing a specific velocity component v_j in polar coordinates ($v_z = v \cos \theta$), we obtain in the effective mass approximation

$$\mu = \frac{e \langle \tau_{\text{tot}} \rangle}{m^*} = \frac{\frac{2}{3} e \int_0^\infty \tau_{\text{tot}}(\mathbf{k}) e^{-E/kT_c} \frac{E}{kT_c m^*} g(E) dE}{\int_0^\infty e^{-E/kT_c} g(E) dE} \quad (8.46)$$

Here we have made use of the facts that in the numerator $\int_0^\pi \cos^2 \theta \sin \theta d\theta = 2/3$ (in the denominator $\int_0^\pi \sin \theta d\theta = 2$) and $m^* v^2 / 2 = E$.

For the average value $\langle \tau_{\text{tot}} \rangle$ of τ_{tot} one therefore obtains the equation

$$\langle \tau_{\text{tot}} \rangle = \frac{\int_0^\infty \tau_{\text{tot}}(\mathbf{k}) e^{-x} \bar{x}^{3/2} d\bar{x}}{\frac{3}{2} \int_0^\infty e^{-x} \bar{x}^{1/2} d\bar{x}} \quad (8.47)$$

which is equivalent to

$$\langle \tau_{\text{tot}} \rangle = \frac{1}{\Gamma(5/2)} \int_0^\infty \tau_{\text{tot}}(\mathbf{k}) e^{-x} \bar{x}^{3/2} d\bar{x} \quad (8.48)$$

where $\bar{x} = E/kT_c$.

In some cases $\tau_{\text{tot}}(\mathbf{k})$ follows a power law in energy:

$$\tau_{\text{tot}}(\mathbf{k}) = \tau_0 \bar{x}^p \quad (8.49)$$

Then

$$\langle \tau_{\text{tot}} \rangle = \frac{\tau_0 \Gamma(p + 5/2)}{\Gamma(5/2)} \quad (8.50)$$

In the following, we discuss the temperature-dependent mobility for several scattering mechanisms. It is important to realize that τ_{tot} may depend on the carrier temperature and on the lattice temperature at the same time. The carrier

temperature dependence is introduced by the energy dependence of τ_{tot} , and the lattice temperature dependence comes from the phonon occupation number N_q .

For acoustic phonon scattering we have, from Eq. (7.36):

$$1/\tau_{\text{ac}} \propto \sqrt{E}kT_L \quad (8.51)$$

Therefore,

$$\tau_{\text{ac}} \propto \frac{(\bar{x})^{-1/2}}{kT_L \sqrt{kT_c}} \quad (8.52)$$

For optical phonons no simple expression can be found, as can be seen from Eqs. (8.22) and (7.35). At very high electron temperatures, however, optical phonons behave similarly to acoustic phonons [see Eq. (7.35)] and

$$\tau_{\text{op}} \propto \frac{(\bar{x}^{-1/2})}{\sqrt{kT_c}} \quad (8.53)$$

The lattice temperature dependence is still complicated because of the exponents in N_q .

For the case $T_c = T_L$, τ_{op} can be approximated by $\tau_{\text{op}} \propto T_L^p$, where p is close to ~ 2 in silicon and GaAs.

Impurity scattering does, of course, not depend on the lattice temperature (N_q is not involved). There are two limiting cases of impurity scattering, depending on the strength of screening. For very strong screening, we have

$$1/\tau_{\text{imp}} \propto \sqrt{E} \quad (8.54)$$

This square root of energy comes from the density of states, while the matrix element is independent of energy because the screened potential is of short range (see the matrix element of the δ function potential, and Eq. (7.18) with large q_s).

For weak screening (small q_s), on the other hand,

$$\tau_{\text{imp}} \propto E^{3/2} \quad (8.55)$$

This proportionality is only approximate and can be verified from Eq. (7.21) for weak screening. Ionized impurity scattering can therefore be proportional to both $T_c^{-1/2}$ and to $T^{3/2}$, depending on screening.

If there is more than one scattering process of importance, the scattering rate $S(\mathbf{k}, \mathbf{k}')$ is given by the sum of the various scattering processes as long as these processes are independent. This is usually a good approximation (in case of dependencies, higher-order perturbation theory is necessary to calculate $S(\mathbf{k}, \mathbf{k}')$ which goes beyond the scope of this book). Because the total scattering rate is just the sum of $S(\mathbf{k}, \mathbf{k}')$ over final states \mathbf{k}' , these sums add as long as the scattering mechanisms are independent. This in turn means that the inverse of the total scattering times for the various mechanisms adds

$$\frac{1}{\tau_{\text{tot}}^{\text{all}}(\mathbf{k})} = \frac{1}{\tau_{\text{tot}}^I(\mathbf{k})} + \frac{1}{\tau_{\text{tot}}^{ph}(\mathbf{k})_{\text{all}}} + \dots \quad (8.56)$$

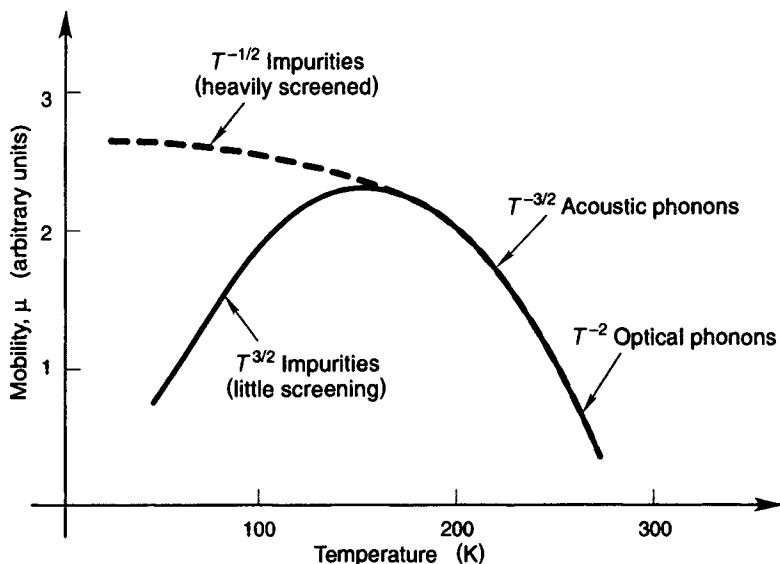


Figure 8.3 Typical temperature dependence of the mobility in semiconductors.

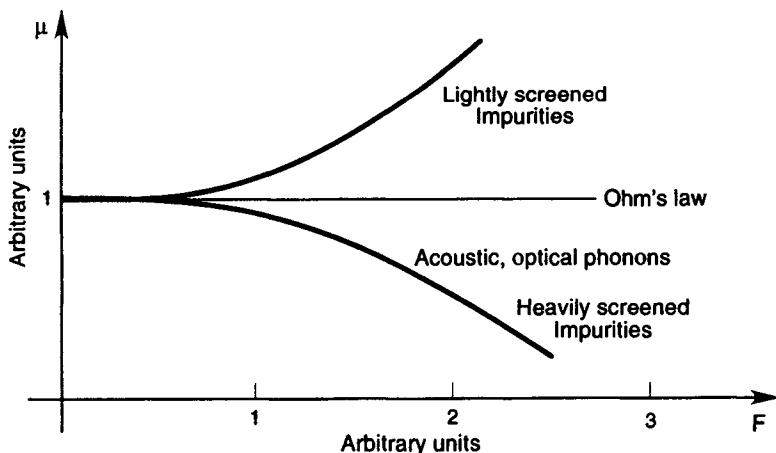


Figure 8.4 Field dependences of the mobility for various scattering mechanisms.

where the superscripts indicate the scattering mechanisms, I for impurity, ph for phonon scattering etc. $\tau_{\text{tot}}^{\text{all}}$ can then be inserted into the integral of Eq. (8.47) to calculate the average relaxation time and from this the mobility. Because this integral is over τ_{tot} and not over the inverse, the average $\langle \tau_{\text{tot}}^{\text{all}} \rangle$ is not equal to the sum of the averages $\langle \tau_{\text{tot}}^{\text{ph}} + \tau_{\text{tot}}^{\text{I}} + \dots \rangle$. Therefore the total mobility μ^{all} is not equal to the sum of mobilities owing to the various scattering mechanism's $\mu^{\text{ph}}, \mu^{\text{I}},$ and so on. However, as long as the various relaxation times have a weak

functional dependence on energy, which is often the case, one can approximate the mobility by

$$\frac{1}{\mu^{\text{all}}} \approx \frac{1}{\mu^{\text{I}}} + \frac{1}{\mu^{\text{ph}}} + \dots \quad (8.57)$$

Equation (8.57) is called the Mathiessen rule. These considerations lead to the typical temperature dependence of the mobility in semiconductors as shown in Figure 8.3. The field dependence follows from the proportionalities to T_c as outlined above. T_c increases with the electric field resulting in the schematic field dependences shown in Figure 8.4.

8.4 EFFECT OF TEMPERATURE GRADIENTS AND GRADIENTS OF THE BAND GAP ENERGY

Gradients in the temperature of charge carriers can easily be treated in the relaxation time approximation if the even part f_0 of the nonequilibrium energy distribution is known; for example, for small forces, f_0 will be close to the equilibrium distribution; that is, either a Fermi or a Maxwell-Boltzmann distribution. Whenever the forces drive f_0 far away from equilibrium, a carrier temperature is not well defined or f_0 becomes difficult to calculate. This can be seen from the procedure outlined in Section 8.2 starting from Eq. (8.22), which now has to include a term arising from $\mathbf{v} \cdot \nabla f$, and thus reads

$$\mathbf{v} \cdot \nabla f_0(\mathbf{k}, \mathbf{r}, t) + \frac{F_0}{\hbar} \nabla_{\mathbf{k}} f_0(\mathbf{k}, \mathbf{r}, t) = -\frac{f_1(\mathbf{k}, \mathbf{r}, t)}{\tau_{\text{tot}}(\mathbf{k})} \quad (8.58)$$

The procedure to solve this equation corresponding to Eq. (8.19) now becomes cumbersome, and it is probably easier to use a Monte Carlo method as described in Section 8.6 to determine f_0 or to use the method of moments that is explained in Chapter 11. Let us assume that we have obtained f_0 by some means. f_1 can then be calculated from Eq. (8.58) and from f_1 the current is easily calculated. To proceed explicitly assume that f_0 is a Fermi (or Maxwell-Boltzmann) distribution with quasi-Fermi level E_{QF} and carrier temperature T_c as defined in Section 8.2, both of them now being dependent on the space coordinate. We obtain

$$\nabla f_0 = \frac{\partial f_0}{\partial E_{\text{QF}}} \nabla E_{\text{QF}} + \frac{\partial f_0}{\partial T_c} \nabla T_c \quad (8.59)$$

This gives f_1 and integrations over \mathbf{k} -space as performed in Section 8.3 give the current density.

These investigations can, of course, be performed explicitly only for a Maxwell-Boltzmann distribution and simple functional dependences (such as power laws) on energy for τ_{tot} . A world of homework problems can be concocted here to solve for various power laws of τ_{tot} and we refer the reader to more introductory texts, which give explicit formulae. For us it is sufficient that the current

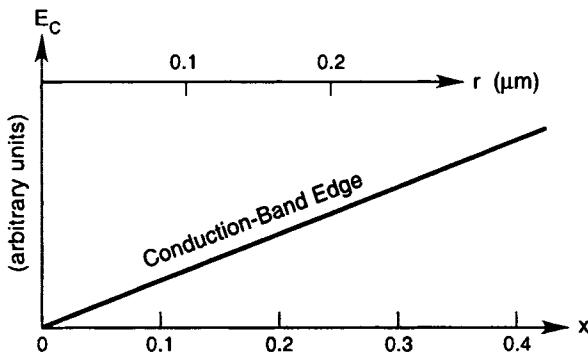


Figure 8.5 Space-dependent band edge. x denotes the Al mole fraction, not the space coordinate. The space coordinate is denoted by r .

can be obtained in general by integration of Eq. (8.58) using Eq. (8.59). Without performing this integration, however, one can deduce certain important proportionalities for the current. The ∇ term gives proportionalities to ∇T_c and ∇E_{QF} as just explained. We now take a closer look at the force term, the term proportional to F_0 in Eq. (8.58). For homogeneous semiconductor material and using the effective mass theorem, the band edge E_c represents the zero of the kinetic energy of electrons (E_v for holes) and follows the external electrostatic potential as discussed in context with Eq. (3.33). The force F_0 is therefore given by

$$F_0 = -e\nabla V_{\text{ext}} \quad (8.60)$$

Because e is, in our notation, the absolute value of the elementary charge, the force would show a positive sign for holes. This means that an external electrostatic potential accelerates holes and electrons in opposite directions. We can now come to an important generalization. As discussed at the end of Chapter 3, it is possible to grow semiconductor alloys with variable composition. For example, one can vary the Al content in $\text{Al}_x\text{Ga}_{1-x}\text{As}$. Then one might get a space-dependent band edge as illustrated in Figure 8.5. Such space-dependent conduction and valence band edges give rise to an additional force term proportional to $-\nabla E_c$ or ∇E_v (for holes) that can now point in the same direction [the different sign for electrons and holes before the gradient follows from Eqs. (3.41) and (3.44)]. Notice that this possibility of forcing electrons and holes to accelerate in the same direction is very important for device applications. In lasers one wishes to collect electrons and holes at the same location to have optimal recombination of them. This can be done using graded or abrupt heterostructures such as shown in Figure 8.5 as suggested originally by Kroemer.

The force on our electron in inhomogeneous material is then

$$F_0 = -\nabla(E_c - eV_{\text{ext}}) \quad (8.61)$$

Collecting all terms, the current must then be proportional to ∇E_{QF} , ∇T_c , and $\nabla(E_c - eV_{\text{ext}})$. In calculating the current one can show that under very general

conditions the term proportional to $\nabla(E_c - eV_{\text{ext}})$ can also be written as proportional to ∇E_{QF} so that the current contains only terms proportional to ∇E_{QF} and ∇T_c . We would like to emphasize that all these discussions are based on the effective mass theorem and become incorrect if the band structure varies rapidly over short distances; corrections are also necessary if the effective mass becomes strongly dependent on alloy composition.

8.5 BALLISTIC AND QUANTUM TRANSPORT

Equation (8.18) represents the key to the understanding of semiclassical transport in semiconductors. As we saw in the derivation, it implies that the electron can be seen as a semiclassical particle; that is, the phase coherence of its wave function is destroyed by inelastic scattering processes. Neglecting these inelastic processes entirely and assuming phase coherence of the wavefunction leads, for example, to an alternating current as demonstrated at the beginning of Chapter 7. There is, however, another possibility of ballistic transport without the Bloch oscillation of the current. This type of transport occurs if short regions of material (shorter than the inelastic mean free path) exhibiting virtually scattering-free transport (also called ballistic regions) that are coupled to other regions, which by design exhibit high scattering rates and semiclassical transport (called reservoirs or contacts), the external voltages also need to be much smaller than the width of the conduction band. An example for such a structure is given by a combination of two highly doped semiconductor regions that sandwich semiconductor material that is not doped or only lightly doped. In the reservoirs, we can then define a distribution function f as before. Assume now that electrons leave these reservoirs only at a relatively low rate compared to the rate of their interaction with one another and with the lattice vibrations in the reservoir. We will then have a distribution function close to the equilibrium distribution; it will have the shape of a Fermi distribution or a Maxwell-Boltzmann distribution in most of the reservoir up to close to its border with the ballistic region. Because we permit a current flow we are, of course, out of equilibrium, and we can define only a quasi-Fermi level for the distribution function and not an equilibrium Fermi level. If we apply a certain voltage eV_{ext} to one reservoir, then the quasi-Fermi level in this reservoir differs just by eV_{ext} . The physical content of this statement becomes clear if we consider two metals separated by an insulator and apply voltage to them. The difference in the metal quasi-Fermi levels must be equal to just this voltage as follows from the definition of the quasi-Fermi level and energy conservation. We now permit coupling between the two reservoirs and let a current flow. To simplify our considerations, we assume for the time being a one-dimensional system with reservoirs as schematically shown in Figure 8.6. Electrons are released from the left contact, which is grounded while voltage V_{ext} is applied to the right contact. In the ballistic region, electrons are transmitted with probability of one if no quantum reflections occur, or in the case

of structure in the material (owing to the existence of a heterojunction) the transmission is with a probability $T_{LR}(E)$. $T_{LR}(E)$ can, of course, be calculated from the Schrödinger equation. The absolute value of the current density j_{LR} out of the left contact going to the right is then

$$j_{LR} = \frac{e}{\pi} \int_0^\infty T_{LR}(E) v_z f_0 dk_z \quad (8.62)$$

with f_0 being given by a Fermi distribution with left-side quasi-Fermi level E_{QF}^L . We can now use the even part of the distribution function in calculating a current because we are only interested in the charge carriers that are leaving the contact from left to right. The inverse current is smaller because of the different quasi-Fermi level of the right contact and therefore a net current exists. Naturally, in the ballistic region the odd part of the distribution function f_1 is large. It is caused by and can be calculated from the two even parts in the contacts emitting ballistically toward opposite directions. We will return to this type of current in Chapter 10 discussing Bethe's thermionic emission theory.

The current flowing from the right contact to the left can be calculated in a similar way, except, that now we have to insert a quasi-Fermi level E_{QF}^R and transmission coefficient $T_{RL}(E)$ for the right side. Because in equilibrium, the two currents j_{RL} and j_{LR} must be equal and cancel each other, it is also easy to see that $T_{RL}(E) = T_{LR}(E)$, and we will name it just $T(E)$ from here on. To estimate the current, we assume that the effective mass approximation is valid and $m^* v_x = \hbar k_x$. Further, we set $T(E) = 1$ neglecting for the moment any classical or quantum reflections. Then we have

$$j_{LR} = \frac{e\hbar}{\pi m^*} \int_0^\infty k_z \frac{1}{\exp\left(\left(\frac{\hbar^2 k_z^2}{2m^*} - E_{QF}^L\right)/kT_L\right) + 1} dk_z \quad (8.63)$$

The integral is easily performed by introducing a new integration variable equal to k_z^2 and one obtains

$$j_{LR} = \frac{2e}{h} kT_L \ln(1 + e^{E_{QF}^L/kT_L}) \quad (8.64)$$

The absolute value of the total current is given by

$$j = j_{RL} - j_{LR} \quad (8.65)$$

Using now the fact that the difference in quasi-Fermi levels is equal to eV_{ext} and Taylor expanding the logarithm for small external voltages one obtains

$$j = \frac{2e^2}{h} \frac{V_{ext}}{e^{-E_{QF}^L/kT} + 1} \quad (8.66)$$

which gives in the limit of large E_{QF}^L/kT_L a conductance G of

$$G = \frac{2e^2}{h} \quad (8.67)$$

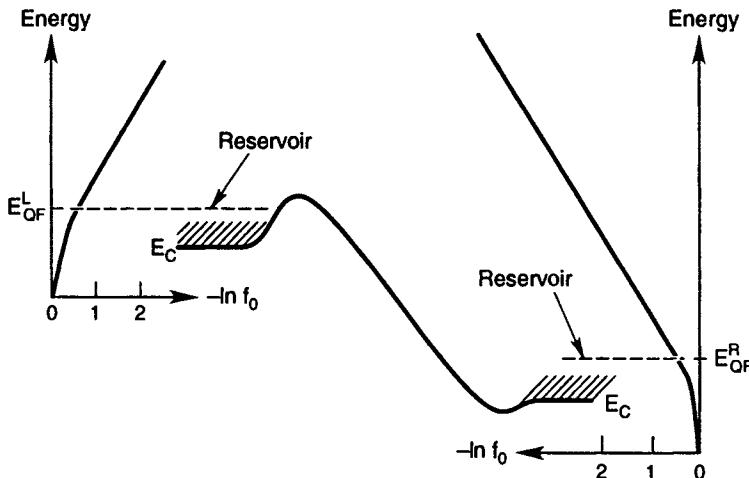


Figure 8.6 Schematic of two reservoirs sandwiching the ballistic region.

This is the famous quantum conductance; its inverse the quantum resistance is close to $13 \text{ k}\Omega$.

One might now ask how we could obtain a quantum conductance, a formula which contains only \hbar and e , from an entirely plausible, virtually classical formula for the current. We did, of course, use a Fermi distribution; however, had we used a Maxwell-Boltzmann distribution, the result would have been the same except the one in the denominator of Eq. (8.66) would disappear. The reason for obtaining a quantum result is that we have counted the quantum states correctly when using Eq. (8.62); this equation is based on the replacement of the sum over all \mathbf{k} -states $\sum_{\mathbf{k}}$ by the integral $\frac{2L}{2\pi} \int dk_x$ as prescribed by the procedure derived in Chapter 5, for one dimension. The correct state counting is sufficient, as so often, to give the correct quantum result. Equation (8.67) is a special case for the general quantum conductance of small structures that was derived by Landauer and Büttiker (see ref. [3]). The ballistic conductance is of great importance in semiconductor device theory. In three dimensions, one derives in the same way as the so-called Richardson current or Bethe's thermionic emission formula, which is described in Chapter 10.

8.6 THE MONTE CARLO METHOD

The Monte Carlo method, as introduced for semiconductor transport by Kurosawa, defines a numerical approach to solve the Boltzmann equation [5]. The basic mathematics involved is a method of integration using random numbers. Consider a dart-board of unit area with any complicated geometric figure, such as your favorite person, on it. Now throw darts at the board randomly and count the number of hits. This number divided by the total number of darts in the

board gives the area of the person (since the board has unit area). One can play this game less violently and more generally by using random number generators and n -dimensional geometric figures instead of boards. As shown in numerous publications (see, e.g., D. K. Ferry's Semiconductors, p. 354 [4]), the solution of the Boltzmann equation can be written in the form of integrals which involves the path along the semiclassical electron trajectory. The integrals can, of course, be calculated using the Monte Carlo method of integration. One can show that this is equivalent to following the path of an electron and its scattering if one considers steady state and homogeneous material, because one can then invoke ergodicity; that is, the equivalence of the averaging of one electron's properties over time with the average over an ensemble of many. In other words, to compute the electron distribution function or the electron's average property, one can proceed in the following way. We calculate the probability that the electron is not scattered in a certain time interval and let the electron move in this time interval according to Eqs. (3.25) and (3.26) or any other appropriate equations of motion. Then the electron is scattered to a different point in \mathbf{k} -space according to its correct probability and the process is iterated and the electrons properties are accumulated.

Current computers are able to handle such a project. It turns out, however, that one does not even need to store all events in memory. One can accumulate averages by using certain "estimators," which are described below. We still have to let the electron propagate and be scattered by chance.

In such a calculation, we need to know the probability P that the electron is not scattered in a time interval $[t_1, t_2]$. We know that the total scattering rate per unit time is $\sum_{\mathbf{k}} S(\mathbf{k}, \mathbf{k}')$ and therefore can relate the probability $P(t)$ to the probability $P(t + dt)$. Because P is the probability that the electron is not scattered, we have

$$P(t + dt) = P(t) \left(1 - dt \sum_{\mathbf{k}'} S(\mathbf{k}, \mathbf{k}') \right) \quad (8.68)$$

where $dt \sum_{\mathbf{k}'} S(\mathbf{k}, \mathbf{k}')$ is the probability of an electron being scattered in the time interval dt .

From the definition of differentiation, Eq. (8.68) is equivalent to

$$\frac{dP}{dt} = -P \sum_{\mathbf{k}'} S(\mathbf{k}, \mathbf{k}') \quad (8.69)$$

Solving this differential equation for P in a time interval $[t_1, t_2]$, we have

$$P = \exp \left\{ - \int_{t_1}^{t_2} \sum_{\mathbf{k}'} S(\mathbf{k}, \mathbf{k}') \right\} \quad (8.70)$$

$\sum_{\mathbf{k}'} S(\mathbf{k}, \mathbf{k}')$ is generally a function of time; k can be a function of time, as is seen from the equations of motion, Eqs. (3.25) and (3.26).

We can now design the following numerical scheme to calculate all the transport effects for the electrons (writing $1/\tau_{\text{tot}}$ for $\sum_{\mathbf{k}'} S(\mathbf{k}, \mathbf{k}')$):

1. We calculate $\tau_{\text{tot}}(\mathbf{k})$ for all important scattering mechanisms. If these mechanisms are independent, then the probabilities for an electron being scattered add, which means that

$$\frac{1}{\tau_{\text{tot}}^{\text{all}}(\mathbf{k})} = \frac{1}{\tau_{\text{tot}}^I(\mathbf{k})} + \frac{1}{\tau_{\text{tot}}^{ph}(\mathbf{k})} + \dots \quad (8.71)$$

where the superscript I denotes impurity and ph phonon scattering.

2. We calculate the time that an electron is not scattered by solving the integral equation, (8.70). This is accomplished in the following way: Because P is a probability, it will assume random numbers between 0 and 1. We can therefore replace P by such a random number r_d equi-distributed between 0 and 1, to obtain

$$r_d = \exp \left[- \int_{t_1}^{t_2} \frac{dt}{\tau_{\text{tot}}^{\text{all}}(t)} \right] \quad (8.72)$$

This equation is now solved for t_2 (t_1 is the time the electron starts).

3. Having calculated t_2 , we let a particular electron accelerate freely between t_1 and t_2 and then the electron is scattered to a different \mathbf{k}' vector. Of course, we have to choose the correct \mathbf{k}' vector according to the correct probability given by the dependence of $1/\tau_{\text{tot}}$ on \mathbf{k}' . If the matrix element does not depend on \mathbf{k}' , any random \mathbf{k}' that conserves energy will do. Otherwise we have to pick \mathbf{k}' with the correct probability distribution (see Problem 8.2).
4. We calculate again a free acceleration time period $[t_2, t_3]$ using a new r_d and

$$r_d = \exp \left[- \int_{t_2}^{t_3} \frac{dt}{\tau_{\text{tot}}^{\text{all}}(t)} \right] \quad (8.73)$$

by solving Eq. (8.73) for t_3 . Again, we let the electron accelerate freely in this time period and then scatter, and we continue this whole process for many time periods, typically $\sim 10^5$. At each instance we calculate the important physical quantities velocity, energy, and so on.

As mentioned before, we do not need to store all the information at all times. Instead, we can accumulate information by using certain estimators that sum and average velocities, energies, and other important physical quantities.

An estimator for the velocity v is, for example, given by

$$v = \frac{\mathbf{r}_{\text{initial}} - \mathbf{r}_{\text{final}}}{\text{total time}} \quad (8.74)$$

Note, however, that this is a time average whose validity depends on the ergodicity of the process, which means that we consider only homogeneous, steady-state

situations. In case these conditions are not met, it is not sufficient to follow a simple electron over a long time. Instead an ensemble of many electrons must be considered (typically 10^5 electrons). Transients and deviations from steady-state and spatial variations are important in devices. Therefore devices require an ensemble Monte Carlo simulation. We will describe in later chapters such Monte Carlo tools. Here we just remark that the Monte Carlo method is ideal to solve the Boltzmann equation, including band-structure effects in the equations of motion and in the scattering integrals [7].

The Monte Carlo method also offers a natural opportunity for generalization. If we have, for example, a quantum barrier that requires tunneling and if this barrier is thin enough (so that the tunneling times and the change in distance is immaterial), we can regard this barrier as a scattering center. When the semiclassical electron impinges on the barrier we can calculate the correct transmission and reflection probability, and, as described for the scattering process, continue the electron after “scattering” (which is now tunneling) either reflected or transmitted through the barrier. In principle this is equivalent to the so-called Bardeen Transfer Hamiltonian formalism described in Appendix A. The Monte Carlo method solves in this way a much more complicated equation than the Boltzmann equation and includes more quantum mechanics. The nice feature is that we do not even need to know the actual governing integro-differential equation that we are solving and that we can include increasingly interesting physics. Note, however, that this cannot be pushed beyond the ranges of validity of the Bardeen Transfer Hamiltonian picture.

Another quantum effect that can be included in the Monte Carlo method is collision broadening (CB). In the presence of scattering, the lifetime of a carrier in the any state \mathbf{k} is finite and, thus, its energy is uncertain. CB is a well-recognized effect exhibited, for example, in the nonzero line widths of the absorption spectra of atoms and molecules. In semiconductors, the CB of high-energy states owing to optical phonon scattering can even exceed the energy of the optical phonons themselves. At low energies, CB may provide the only mechanism by which a process may occur, such as absorption or emission of photons of energy less than the semiconductor band gap.

The Monte Carlo method can be generalized to allow treatment of CB effects if the assumption of energy conservation during scattering with respect to the “bare” (as opposed to augmented or “dressed” by the coupling) carrier and phonon energies, as used in Golden Rule calculations, is relaxed. However, this relaxation should be performed in such a way that, despite the CB, the *total* energy of the *coupled* carrier–phonon multi-particle system *is* conserved through any and all scattering events. An approach recently developed by L. F. Register [6] to modeling CB within the Monte Carlo method that satisfies both of these requirements is to select the \mathbf{k} states sequentially such that

$$E(\mathbf{k}_l) = E(\mathbf{k}_{l-1}) \mp \hbar\omega_{l,l-1} + \Delta E(\mathbf{k}_l) - \Delta E(\mathbf{k}_{l-1}) \quad (8.75)$$

where the $E(\mathbf{k})$ are the bare carrier energies, the $\pm \hbar\omega_{l,l-1}$ are the bare energies of

the emitted or absorbed phonons, respectively, and the $\Delta E(\mathbf{k}_l)$ are the CB contributions. The index l corresponds to the time t_l at which scattering into the state \mathbf{k}_l occurs. There exists a one-to-one correspondence between these latter CB contributions and the carrier-phonon coupling Hamiltonian in the multi-particle Schrödinger equation of the coupled system. Thus, while CB is exhibited with respect to the bare carrier and phonon energies, the total energy of the coupled system, carrier plus phonon *plus coupling*, is conserved through each scattering event. Note that after a long sequence of L scattering events,

$$E(\mathbf{k}_L) = E(\mathbf{k}_0) + \left(\sum_{l=1}^L \mp \hbar\omega_{l,l-1} \right) + \Delta E(\mathbf{k}_L) - \Delta E(\mathbf{k}_0), \quad (8.76)$$

so that there is no accumulation of the CB contributions, which avoids some difficulties associated with earlier approaches. The CB of scattering is illustrated in Figure 8.7. The degree of CB in each state \mathbf{k}_l is inversely proportional to the carrier lifetime $\tau_{\text{tot}}^{\text{all}}(\mathbf{k}_l)$. Basic scattering theory suggests the distribution function for sampling the values of $\Delta E(\mathbf{k}_l)$ should be approximately Lorentzian, $[\hbar/(2\tau_{\text{tot}}^{\text{all}}(\mathbf{k}_l)\pi)]/[\Delta E^2(\mathbf{k}_l) + \hbar^2/(2\tau_{\text{tot}}^{\text{all}}(\mathbf{k}_l))^2]$, which replaces the δ -function of energy conservation in the Golden Rule. Note that the lifetimes depend also on the CB itself:

$$\tau_{\text{tot}}^{\text{all}}(\mathbf{k}_l) = \tau_{\text{tot}}^{\text{all}}(\mathbf{k}_l, E(\mathbf{k}_l) - \Delta E(\mathbf{k}_l)). \quad (8.77)$$

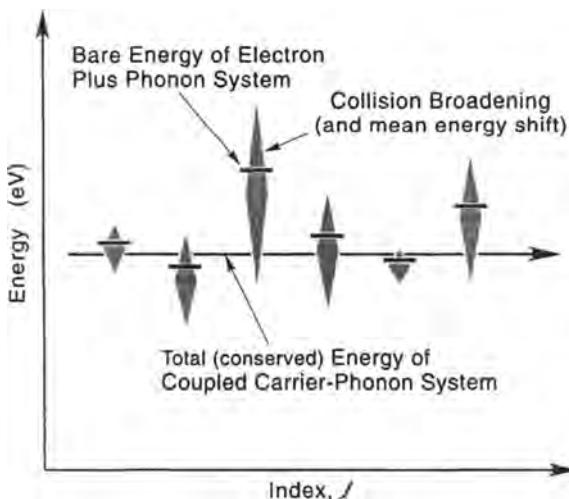


Figure 8.7 The coupled carrier-phonon system of fixed total energy, which includes the bare carrier energy, kinetic and potential, the bare phonon energy, and the carrier-phonon coupling energy, may pass through broadened bare carrier-phonon states of varying nominal energy. The small shifts in the mean energies of the shaded areas from the bare energies correspond to the so-called “self-energy” corrections also produced by the coupling.

These lifetimes, of course, also depend on the energy of the phonon system through the phonon occupation numbers.

PROBLEMS

- 8.1** Solve the following equation, which is close to but not identical to the Boltzmann equation for constant relaxation time τ_{tot} in a time-independent, uniform electric field F in one dimension:

$$\frac{-eF}{m^*} \frac{\partial f(v)}{\partial v} = -\frac{(f(v) - f_{\text{MB}}(v))}{\tau_{\text{tot}}}$$

(v = velocity) where $f_{\text{MB}}(v) = n \left(\frac{2\pi kT}{m^*} \right)^{-1/2} \exp \left(-\frac{m^* v^2}{2kT} \right)$ is the equilibrium distribution function (in this case, Maxwell-Boltzmann). (*Hint:* Integrate from v to ∞ ; your answer will contain an error function.) Partition the solution into even and odd part and show that the even part is not equal to the equilibrium distribution f_{MB} .

- 8.2** Using the effective mass theorem and Eqs. (3.30) and (8.43), derive the conductivity mass, Eq. (3.31), by considering a two-dimensional system with four equivalent constant energy ellipses along the [1,1], [-1,1], [1,-1], and [-1,-1] directions. Each ellipse is characterized in the coordinate system of the main axes by two masses: m_1^* = longitudinal mass, m_t^* = transverse mass.

REFERENCES

- [1] Caldeira, A. O., and Leggett, A. J. "Quantum Tunneling in Dissipative Systems," *Annals of Physics* (NY), vol. 149, 1983, p. 374.
- [2] Conwell, E. M. "High Field Transport in Semiconductors," in *Solid State Physics*, ed. F. Seitz, D. Turnbull, and H. Ehrenreich, Supplement 9. New York: Academic, 1967.
- [3] Datta, S. *Electronic transport in mesoscopic systems*, Cambridge: Cambridge Univ. Press, 1995.
- [4] Ferry, D. K. *Semiconductors*, New York: Macmillan, 1991.
- [5] Price, P. J. "Monte Carlo Calculation of Electron Transport in Solids," in *Semiconductors and Semimetals*, ed. R. K. Willardson and A. C. Beer, vol. 14. New York: Academic, 1979, pp. 249–305.
- [6] Register, L. F., and Hess, K. in the proceedings of the Fourth International Symposium on New Phenomena in Mesoscopic Structures, Kauai, Hawaii, December 7–11, 1998.
- [7] Shichijo, H., and Hess, K. *Physical Review B*, vol. 23, 1981, pp. 4197–4207

CHAPTER 9

GENERATION- RECOMBINATION

Generation-recombination (GR) processes are scattering processes similar to those described in Chapter 7 and can usually be calculated in a similar fashion; that is, by the use of the Golden Rule [3]. The reason another chapter is devoted to these processes is that the number of particles is not conserved in a particular band in the GR process. Typically, an electron from the conduction band will recombine with a hole in the valence band or an electron–hole pair will be generated. The involvement of two bands renders the one-band approximation invalid and also makes a simplistic use of the effective mass theorem impossible. We will, therefore, not be able to calculate explicitly the matrix elements in this chapter, but rather will point out their proportionalities to quantities such as carrier concentration and density of states.

9.1 IMPORTANT MATRIX ELEMENTS

In principle, we can numerically calculate all matrix elements by using wave functions computed, for example, by the method of Chapter 3. Notice that GR gives a richness to the theory of transport in semiconductors (not present in the case of metals), which also forms the basis of the theory of p – n junctions (see Chapter 13). This change is due to interactions with photons, phonons, and the like, and is treated separately below. We describe mainly recombination processes; the generation processes are the exact inverse.

9.1.1 Radiative Recombination

Electrons can lose their energy by emitting a photon, as shown in Figure 9.1. The photon can be emitted by a band-to-band transition or by a transition involving impurity states or exciton states. An exciton is a hydrogen atom-like entity where the nucleus is replaced by a hole; that is, an electron–hole pair are bound to each other. If the effective mass of the hole is large compared to the electron, the

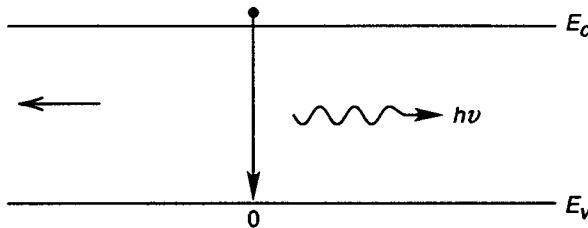


Figure 9.1 Schematic diagram for radiative recombination.

theory of the exciton proceeds like the theory of impurity levels. If the electron and hole masses are comparable, a more complicated approach is in order.

We treat here only band-to-band transitions; band-to-impurity transitions can be treated in an analogous fashion. The transition probability is again obtained from the Golden Rule. The only problem is the derivation of the matrix element $M_{c,v}$ for the electron-photon interaction. Because this derivation is given in many textbooks (see, e.g., Casey and Panish [2], a); we just repeat the result.

The so-called momentum matrix element is given by

$$M_{c,v} = \int_{V_{ol}} \psi_v^* \hat{p} \psi_c d\mathbf{r} \quad (9.1)$$

where ψ_v and ψ_c are the valence and conduction band wave functions, respectively, and \hat{p} is the momentum operator:

$$\hat{p} = \frac{\hbar}{i} \nabla \quad (9.2)$$

We now write the wave function using Bloch's theorem

$$\psi_c = e^{i\mathbf{k}_c \cdot \mathbf{r}} u_c(\mathbf{r}) \quad (9.3)$$

and

$$\psi_v = e^{i\mathbf{k}_v \cdot \mathbf{r}} u_v(\mathbf{r}) \quad (9.4)$$

where \mathbf{k}_c and \mathbf{k}_v are the wave vectors of electrons in the conduction and valence bands, respectively, and $u_{c,v}$ are the periodic parts of the wave functions.

Using Eqs. (9.2), (9.3), and (9.4) in Eq. (9.1), and differentiating, we get

$$\begin{aligned} M_{c,v} &= \frac{\hbar}{i} \int_{V_{ol}} e^{-i\mathbf{k}_v \cdot \mathbf{r}} u_v^*(\mathbf{r}) \nabla e^{i\mathbf{k}_c \cdot \mathbf{r}} u_c(\mathbf{r}) d\mathbf{r} \\ &= \frac{\hbar}{i} \left[\int_{V_{ol}} e^{i(\mathbf{k}_c - \mathbf{k}_v) \cdot \mathbf{r}} u_v^*(\mathbf{r}) \nabla u_c(\mathbf{r}) d\mathbf{r} \right. \\ &\quad \left. + \int_{V_{ol}} e^{-i\mathbf{k}_v \cdot \mathbf{r}} u_v^*(\mathbf{r}) u_c(\mathbf{r}) \nabla e^{i\mathbf{k}_c \cdot \mathbf{r}} d\mathbf{r} \right] \end{aligned} \quad (9.5)$$

Because $u_c(\mathbf{r})$ and $u_v(\mathbf{r})$ are orthogonal and the exponential components vary slowly, the second integral is usually neglected. The first integral can be evaluated in the following way.

The product $u_v^*(\mathbf{r})\nabla u_c(\mathbf{r})$ is the same for each unit cell and is multiplied by a slowly varying function that can be taken out of the integral. We then have

$$\begin{aligned} M_{c,v} &= \frac{\hbar}{i} \sum_{\text{all cells}} e^{i(\mathbf{k}_c - \mathbf{k}_v) \cdot \mathbf{r}} \int_{\Omega} u_v^*(\mathbf{r}) \nabla u_c(\mathbf{r}) d\mathbf{r} \\ &= \frac{\hbar}{i} \int_{V_{\text{ol}}} e^{i(\mathbf{k}_c - \mathbf{k}_v) \cdot \mathbf{r}} d\mathbf{r} \cdot \frac{1}{\Omega} \int_{\Omega} u_v^*(\mathbf{r}) \nabla u_c(\mathbf{r}) d\mathbf{r} \end{aligned} \quad (9.6)$$

And using the definition

$$\langle v | \hat{\mathbf{p}} | c \rangle \equiv \frac{1}{\Omega} \int_{\Omega} u_v^*(\mathbf{r}) \nabla u_c(\mathbf{r}) d\mathbf{r}$$

we have

$$M_{c,v} = \langle v | \hat{\mathbf{p}} | c \rangle \int_{V_{\text{ol}}} e^{i(\mathbf{k}_c - \mathbf{k}_v) \cdot \mathbf{r}} d\mathbf{r} \quad (9.7)$$

We can see that $M_{c,v}$ is then finite only if

$$\mathbf{k}_c - \mathbf{k}_v = 0 \quad (9.8)$$

This means geometrically that only vertical transitions can happen in \mathbf{k} space, as shown in Figure 9.2. This, of course, is important in optical devices (lasers, LEDs, and so on). Transitions that involve changes in \mathbf{k} cannot happen in first order. However, second-order transitions involving a photon and a phonon are possible. The photon is then responsible for the energy change (its momentum is very small, which is the reason for the vertical transition), and the phonon supplies most of the momentum and contributes only little to the energy change. Detailed calculations and discussion have been given by Bebb and Williams [1].

The treatment discussed above is very useful and shows immediately how the matrix element has to be modified if external potentials are present. If we have, for example, a quantum well heterolayer structure, then the envelope (effective mass) wave function is replaced by [see Eq. (1.31)]

$$e^{i(\mathbf{k}_c - \mathbf{k}_v) \cdot \mathbf{r}} \rightarrow e^{i(\mathbf{k}_{c\parallel} - \mathbf{k}_{v\parallel}) \cdot \mathbf{r}} \sin \mathbf{k}_{c\perp} z \sin \mathbf{k}_{v\perp} z \quad (9.9)$$

where \parallel indicates that the vector is parallel to the layers and \perp indicates the direction perpendicular to the layers.

For infinite square wells of width L , we have $\mathbf{k}_{c\perp} = \mathbf{k}_{v\perp} = n\pi/L$, where n is an integer. We see then that the matrix element is only finite if the transition occurs between levels with the same index n ; in other words, we have a selection rule $\Delta n = 0$ (see Figure 9.3). The term $\langle v | \hat{\mathbf{p}} | c \rangle$ is more difficult to evaluate; approximations have been published (see, e.g., Casey and Panish [2], b).

The emission of phonons in band-to-band transitions is a rather rare event; the phonon energy is typically much smaller than the energy of the band gap.

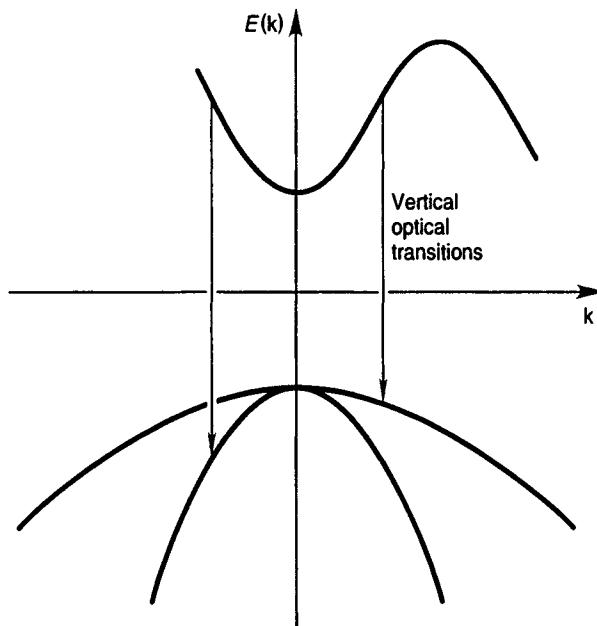


Figure 9.2 Possible transitions in \mathbf{k} space caused by photon emission.

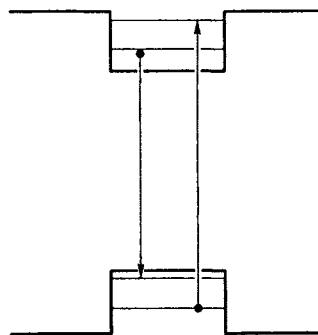


Figure 9.3 Optical transition in a quantum well.

However, if impurity states are involved, the emission of phonons becomes easier for two reasons: First, the energy that must be overcome is smaller since the electron can recombine in cascades, as shown in Figure 9.4. Second, the electron-phonon coupling is enhanced by the impurity since the momentum conservation is relaxed, as the impurity potential can “supply” any necessary component \mathbf{q} . The matrix element for electron capture by an impurity has been reviewed in detail by Ridley [4] and will not be derived here. The matrix element for optical transitions involving impurities has been treated by Casey and Panish [2] (c).

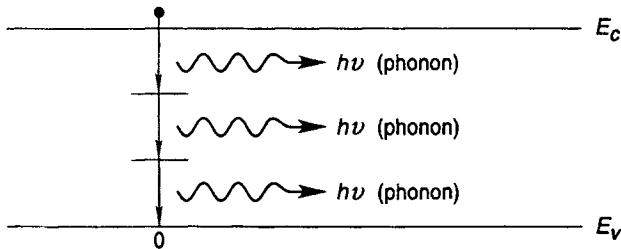


Figure 9.4 Recombination involving phonons and impurities.

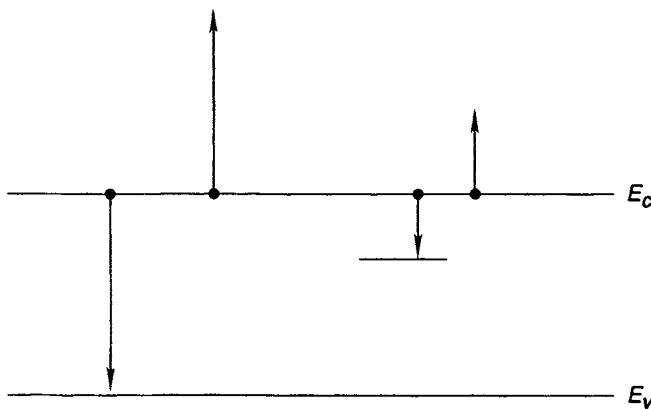


Figure 9.5 Auger recombination of an electron.

9.1.2 Auger Recombination

This recombination process involves more than one electron but no radiation. As shown in Figure 9.5, an electron can transfer its energy to another electron and recombine with a hole. The exact inverse of this process is called impact ionization and is described in Chapter 13. Here it is sufficient to understand that the Auger process is only important if the electron density is high (one needs electron-electron interaction) and therefore is only important in heavily doped semiconductors as the rate increases with the square of the carrier concentration. The interested reader is referred to Ridley's treatment [4]. Before we proceed with GR mechanisms, an important concept must be explained: *quasi-Fermi energy* (sometimes also quasi-Fermi level).

9.2 QUASI-FERMI LEVELS (IMREFS)

We have discussed quasi-Fermi energy (imref) in connection with high field solutions of the Boltzmann equation. If the equilibrium is disturbed optically or by electrical injection of charge carriers (as we will see when developing the theory of $p-n$ junctions), the imref concept also applies. This is especially easy to see

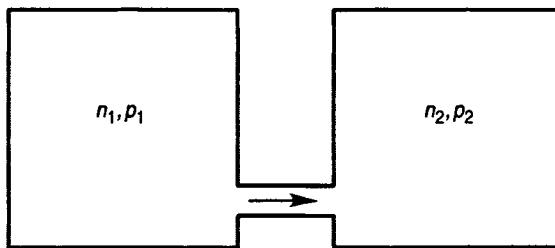


Figure 9.6 Hydrodynamic analogy for the explanation of imrefs.

for the case of optical excitation.

Assume that we radiate light onto a semiconductor and generate a large number of electron-hole pairs. The recombination of an electron with a hole over the band gap by light emission takes place typically in 10^{-9} s. The electron-electron interaction, hole-hole interaction, as well as electron- (hole-) phonon interactions have typically a much shorter time constant (of the order of 10^{-13} s). Therefore the electrons and holes will be in equilibrium with each other in a time of the order of several 10^{-13} s. Then electrons and holes obey Fermi-like distributions separately.

To understand this in more detail, consider the following analogy. In Figure 9.6, two gas containers connected with a very thin tube are shown. The total particle density, which in equilibrium is a measure of the Fermi energy, is different in the two containers ($n_1 \neq n_2$) and so is the pressure ($p_1 \neq p_2$). The system is not in equilibrium because there is a steady stream from one container to the other. However, if the tube is just small enough the particles in each container separately will be in equilibrium with themselves; that is, the even part of the distribution function in each container will have the Fermi shape:

$$f_0(E) = \frac{1}{e^{(E-E_{\text{QF}})} + 1} \quad (9.10)$$

9.3 GENERATION-RECOMBINATION RATES

Using the distribution function of Eq. (9.10) we can calculate the transition rates for most of the GR processes. We know from the treatment of the Boltzmann equation that we have to determine the transition rate by using Eq. (7.39). To calculate this term properly, we must include the $(l - f)$ terms required by the Pauli principle if the distribution function cannot be approximated by a Maxwell-Boltzmann distribution. This is frequently the case for GR problems because

they involve two bands. Therefore,

$$\left. \frac{\partial f_0}{\partial t} \right|_{\text{coll}} = \frac{V_{\text{ol}}}{8\pi^3} \int d\mathbf{k}' \left[S(\mathbf{k}, \mathbf{k}') f_0(\mathbf{k}) \left(1 - f_0(\mathbf{k}') \right) - S(\mathbf{k}', \mathbf{k}) f_0(\mathbf{k}') \left(1 - f_0(\mathbf{k}) \right) \right] \quad (9.11)$$

To proceed with the calculation, we assume that all functions are functions of energy instead of the \mathbf{k} vector, which is usually a good approximation, and we introduce the density of states in the calculation. Generation and recombination are treated separately. That is, we separate the two terms in Eq. (9.11) and take the difference at the end of the calculation. Because we have to deal mainly with two bands, we apply again the indices c, v to energy and \mathbf{k} vector. The edges of the conduction and valence bands are now denoted by E_c^0 and E_v^0 , respectively.

Suppose, then, that a conduction band state of energy E_c is full, a valence band state of energy E_v is empty, and the transition probability per unit time is $S(E_c, E_v)$. The density of full states at energy E_c between E_c and $E_c + dE$ is given by

$$f_c(E_c) g_c(E_c) dE_c \quad (9.12)$$

Here f_c is the Fermi distribution in the conduction band with quasi-Fermi energy E_{QF}^c (we assume that E_{QF}^c is well defined). $g_c(E_c)$ is the density of states in the conduction band. The density of empty valence band states around E_v is given by

$$[1 - f_v(E_v)] g_v(E_v) dE_v \quad (9.13)$$

The meaning of the symbols is the same as above, only now they are applied to the valence band instead of the conduction band. The recombination term of Eq. (9.11) is then

$$\int_{-\infty}^{E_v^0} dE_v f_c(E_c) (1 - f_v(E_v)) g_v(E_v) S(E_c, E_v) \quad (9.14)$$

Notice that, again, as in the case of phonon scattering, the spin of the electron may not change in a particular transition, depending on the mechanism. As described in connection with Eq. (7.33), the final density of states has to be replaced by one-half of its actual value. Here, however, we do not need to be concerned about this factor because it will cancel out in the final results.

In the following discussion we will need the total rate of recombination R , which is obtained by integrating Eq. (9.14) over all conduction band energies (now including the factor of two for the spin). This total rate can be brought into the following form: The product of electron density in the conduction band n and hole density in the valence band p is given by [see derivation of Eq. (5.31)]:

$$n \cdot p = \int_{E_c^0}^{\infty} dE_c \int_{-\infty}^{E_v^0} dE_v f_c(E_c) (1 - f_v(E_v)) g_c(E_c) g_v(E_v) \quad (9.15)$$

It therefore follows that

$$R = n \cdot p \langle S(E_c, E_v) \rangle$$

where $S(E_c, E_v)$ is the average defined by

$$\langle S(E_c, E_v) \rangle$$

$$= \frac{\int_{E_c^0}^{\infty} dE_c \int_{-\infty}^{E_v^0} dE_v f_c(E_c)(1 - f_v(E_v))g_c(E_c)g_v(E_v)S(E_c, E_v)}{\int_{E_c^0}^{\infty} dE_c \int_{-\infty}^{E_v^0} dE_v f_c(E_c)(1 - f_v(E_v))g_c(E_c)g_v(E_v)} \quad (9.16)$$

Inserting for f_c and f_v and assuming a Boltzmann distribution, we can see that the quasi-Fermi levels cancel in Eq. (9.16) and $\langle S(E_c, E_v) \rangle$ is a number independent of the density of electrons and holes (except for Auger processes, where $S(E_c, E_v)$ depends on E_{QF}). This is very fortunate and helps give simple rate equations.

Let us now consider the inverse process the generation of an electron hole pair by the same mechanism. To obtain the total generation rate G we only have to exchange E_c and E_v in Eqs. (9.14)–(9.16), which gives

$$G = \int_{E_c^0}^{\infty} dE_c \int_{-\infty}^{E_v^0} dE_v (1 - f_c(E_c))f_v(E_v)g_c(E_c)g_v(E_v)S(E_c, E_v) \quad (9.17)$$

In many cases, one can neglect f_c versus 1 and approximate $f_v(E_v) \approx 1$. (If this does not hold, the generation undergoes a “Moss-Burstein” shift to higher energies.) With the above approximations, the generation rate is given by

$$G = \int_{E_c^0}^{\infty} dE_c \int_{-\infty}^{E_v^0} dE_v g_c(E_c)g_v(E_v)S(E_c, E_v) \quad (9.18)$$

which is a constant independent of the carrier densities as long as $S(E_c, E_v)$ is not the probability of an Auger process the latter depending on the density. G for equilibrium can be determined from the requirement of detailed balance:

$$G = R$$

Therefore,

$$G = n_i^2 \langle S(E_c, E_v) \rangle \quad (9.19)$$

where n_i is the intrinsic carrier concentration. This equation is also used for situations close to equilibrium but not exactly in equilibrium, which is to be justified in each case. Auger processes can be treated in a similar manner; however, additional dependences on n and p are then introduced.

Some of the most important recombination processes involve a trap for electrons or holes (impurity level), and we therefore have to consider these processes (Shockley-Hall-Read). Assume for simplicity that we have a total density N_{TT}

of identical noninteracting trapping centers with levels at energy E_T in the band gap. Let f_T be the probability of finding a trap full and n_T the number of full traps. We then have

$$n_T = N_{TT} f_T \quad (9.20)$$

f_T is a Fermi distribution in equilibrium but may be very different far away from equilibrium. In some instances the introduction of a quasi-Fermi energy may be appropriate. However, traps usually do not “communicate” with each other, and therefore long-time constants will be needed to establish a Fermi distribution in some instances. Especially for the case of electron injection into traps, f_T may be very different from Fermi-like distributions.

Using Eq. (9.20) and the above definition, we have as first term of the collision operator

$$\sum_{\text{Traps}} f_c(E_c)(1 - f_T)S(E_c, E_T) = N_{TT} f_c(E_c)(1 - f_T)S(E_c, E_T) \quad (9.21)$$

We will need later the sum over all possible recombination events, labeled the recombination term R :

$$R = N_{TT} \int_{E_c^0}^{\infty} dE_c f_c(E_c)(1 - f_T)g_c(E_c)S(E_c, E_T)$$

As before, we can rewrite this as

$$R = N_{TT}(1 - f_T)n \langle S(E_c, E_T) \rangle \quad (9.22)$$

with the average $\langle S(E_c, E_T) \rangle$ being given by

$$c_n \equiv \langle S(E_c, E_T) \rangle = \frac{\int_{E_c^0}^{\infty} dE_c^0 f_c(E_c)g_c(E_c, E_T)S(E_c, E_T)}{\int_{E_c^0}^{\infty} dE_c^0 f_c(E_c)g_c(E_c, E_T)} \quad (9.23)$$

The recombination rate is then rewritten as

$$R = (N_{TT} - n_T)nc_n \quad (9.24)$$

To calculate the generation process we have to exchange E_T and E_c in Eq. (9.21), which gives

$$G = N_{TT} \int_{E_c^0}^{\infty} dE_c f_T(E_T)(1 - f_c(E_c))S(E_T, E_c)g_c(E_c) \quad (9.25)$$

Using a Maxwell-Boltzmann distribution for f_c and assuming $f_c \ll 1$, we arrive similarly as in the above calculation for R at

$$G = n_T e_n \quad (9.26)$$

where e_n is independent of the trap and electron densities. Note, however, that e_n does depend on the energy of the trap as

$$e_n \propto \exp E_T/kT \quad (9.27)$$

because the electrons need to be activated to get out of the trap (escape) (Ridley). Here we measure E_T from the conduction band edge (i.e., it is negative).

The capture coefficient c_n is often independent of E_T ; the probability of falling into a trap (or a deep well) usually does not depend on the depth of the trap (well). There are some exceptions to this rule, however, which have been reviewed by Ridley [4].

9.4 RATE EQUATIONS

The net rate GR of electron loss (or gain) becomes

$$-GR = c_n n (N_{TT} - n_T) - n_T e_n = R(n) - G(n) \quad (9.28)$$

The subscript n in the rates stands for electrons. We can derive similar rates for hole capture and emission at a trapping center. The total change in the density of full traps is then

$$\frac{\partial n_T}{\partial t} = R(n) - G(n) - R(p) + G(p) \quad (9.29)$$

where again n and p stand for electrons and holes, respectively. If we consider both electron and hole recombination, we can have a situation where GR is equal to zero while the system, in fact, is not in equilibrium. There can be a steady capture of electrons and holes in the traps, which results in the destruction of electron-hole pairs. Of course, to make the flow steady we have to replenish the electron-hole pairs, for example, by shining light onto the semiconductor. In this case ($\partial n_T / \partial t = 0$), which is very important for the steady-state theory of $p-n$ junctions, we have

$$R(n) - G(n) = R(p) - G(p) \quad (9.30)$$

Equation (9.30) gives

$$c_n n (N_{TT} - n_T) - n_T e_n = c_p p n_T - e_p (N_{TT} - n_T) \quad (9.31)$$

The rates for the holes differ from the rates for the electrons because a capture of a hole means that the trap is full while an electron capture means that the trap is empty. Therefore, the factor of $c_p p$ is n_T , while the factor of $c_n n$ is $(N_{TT} - n_T)$. We can deduce from Eq. (9.31) n_T , which is

$$n_T = \frac{N_{TT}(c_n n + e_p)}{c_n n + e_n + c_p p + e_p} \quad (9.32)$$

We can use this expression for n_T to calculate the net electron recombination in steady-state U_s :

$$U_s = R(n) - G(n) = \frac{c_n c_p N_{TT} n \cdot p - e_n e_p N_{TT}}{c_n n + e_n + c_p p + e_p} \quad (9.33)$$

The total GR term for electrons is then the generation by light (or in other instances by an electric field) GR_{light} minus U_s :

$$GR = GR_{\text{light}} - U_s$$

Close to equilibrium U_s takes a simple form. Because in equilibrium $U_s = 0$, we have

$$c_n c_p n \cdot p = e_n e_p$$

It is customary to assume that the rates e_n , e_p , c_n , and c_p do not depend on the perturbation and retain their equilibrium value. This assumption is not always valid (see Eq. (9.27) and Figure 13.19) and should be carefully examined before it is used. Nevertheless, we use it in the following to simplify U_s . This is appropriate for not very high electric fields in the junction (see Figure 13.19). Therefore, using Eq. (5.31), we have

$$\frac{e_n e_p}{c_n c_p} = n_i^2$$

and

$$U_s = c_n c_p N_{\text{TT}} \frac{n \cdot p - n_i^2}{c_n n + e_n + c_p p + e_p} \quad (9.34)$$

Denoting the equilibrium electron density by n_0 and hole density by p_0 and approximating in the denominator $n \approx n_0$, $p \approx p_0$ (but not in the numerator, which would vanish), we obtain

$$U_s = \bar{c}(n \cdot p - n_i^2) \quad (9.35)$$

where \bar{c} is a constant given by

$$\bar{c} = c_n c_p N_{\text{TT}} \frac{1}{c_n n + e_n + c_p p + e_p} \quad (9.36)$$

If we generate a small perturbation of the electron density with $n = n_0 + \delta n$ and $\delta n \ll n_0$, then

$$n \cdot p_0 = (n = n_0 + \delta n)p_0 = n_0 p_0 p_0 \delta n$$

and

$$U_s = \bar{c} p_0 \delta n = \frac{\delta n}{\tau_n} \quad (9.37)$$

τ_n is called the lifetime of electrons. This time constant is in its nature quite different from the phonon scattering times because the electrons vanish from the conduction band. There also exists a lifetime for holes that is derived in an analogous fashion.

PROBLEMS

- 9.1** Show that in a semiconductor with only one type of acceptor trapping center the minimum possible electron lifetime is

$$\tau_{no} = \frac{1}{c_n N_{TT}}$$

Hint: Assume here that $\delta n \simeq n$. (Likewise, $\tau_{po} = 1/c_p N_{TT}$.)

- 9.2** Consider the same situation as in Problem 9.1 and assume a disturbed electron and hole population as $n_0 + \delta n$ and $p_0 + \delta p$. If $\delta n = \delta p$ is very small and the trap concentration is also small, show that n and p return to equilibrium with a lifetime

$$\tau = \frac{\tau_{p0}(n_0 + n_1) + \tau_{n0}(p_0 + p_1)}{n_0 + p_0}$$

where $n_1 = e_n/c_n$ and $p_1 = e_p/c_p$ (the rates are equilibrium rates).

- 9.3** Using the expression in Problem 9.2, try to find out whether or not a trap close to the conduction band edge (shallow) is a more efficient recombination center than one at the middle of the gap (deep).

REFERENCES

- [1] Bebb, H., and Williams, E. in *Semiconductors and Semimetals*, vol. 8, ed. R. K. Willardson and A. C. Beer. New York: Academic, 1972, pp. 243–253.
- [2] Casey, H. C. Jr., and Panish, M. B. *Heterostructure Lasers. Part A. Fundamental Principles*. New York: Academic, 1978, (a) pp. 122–27; (b) p. 146; (c) pp. 144–150.
- [3] Landsberg, P. T. *Solid State Theory*, New York: Wiley/Interscience, 1969, pp. 279–294.
- [4] Ridley, B. K. *Quantum Processes in Semiconductors*, Oxford: Clarendon, 1982, pp. 251–263.

CHAPTER 10

THE HETEROJUNCTION BARRIER AND RELATED TRANSPORT PROBLEMS

Up to now, we have considered mainly the equations of motion—the scattering and the statistics of electrons in the conduction and valence bands of semiconductors—that have led us to device equations and to the understanding of some transport effects that are important in semiconductor devices. In all cases, we had assumed homogeneous doping and one type of semiconductor throughout. Devices are usually based on doping inhomogeneities (homojunctions) or the combination of more than one semiconductor (heterojunctions). Obviously the Poisson equation will play an important role, as we already have seen, for the case of space charge limited current. There are, however, also several new transport effects that we need to know to understand what makes devices tick. This chapter is intended to give the transition between bulk (homogeneous) transport and electronic transport in devices. We describe here the transport of electrons close to the interface of two different neighboring semiconductors. Two of the most important devices, the metal–oxide–silicon transistor and the Schottky barrier diode, work on the basis of the principles that are developed below.

10.1 THERMIONIC EMISSION OF ELECTRONS OVER BARRIERS

Consider two different neighboring semiconductors with a conduction band discontinuity ΔE_c such as that of $\text{Al}_x\text{Ga}_{1-x}\text{As}$ – GaAs , which has been described in Chapter 3. We assume that the bands are in a fixed position, as shown in Figure 10.1, and we attempt to calculate the current from GaAs to AlGaAs and vice versa. The calculation can easily be performed by using an approximation introduced by Bethe.

Bethe assumed that the distribution function is equal to the Fermi distribution (or Maxwell-Boltzmann distribution) having a quasi-Fermi level E_{QF} , which is different in the two materials but constant on each side. This means that a strong electron–electron interaction must be present, which causes the distribution function to look Fermi-like independent of the fact that electrons can be lost

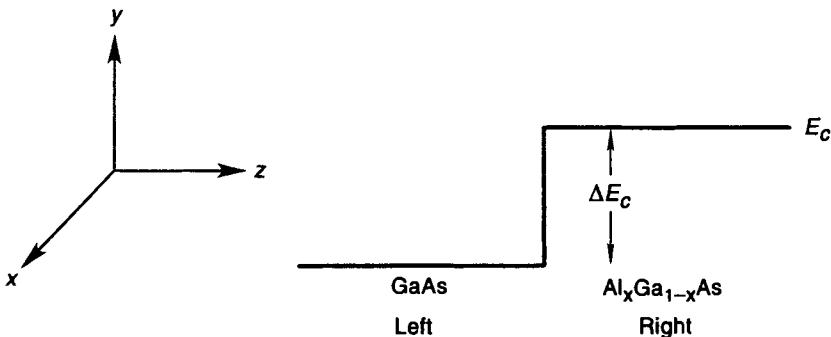


Figure 10.1 The conduction band edge at a heterojunction interface.

at a high rate to the neighboring material. In a Schottky barrier diode this loss of electrons to the neighboring material can be so substantial that it is hard to believe that the electron distribution still stays Fermi- or Maxwell-like. Exact calculations, however, necessitate an ensemble Monte Carlo method. The following explicit treatment is instructive and probably correct within a factor of 2 or so for most practical cases. Classically, all the electrons with a velocity component in the positive z -direction (perpendicular to the interface) large enough to overcome the band-edge discontinuity ΔE_c will propagate into the $\text{Al}_x\text{Ga}_{1-x}\text{As}$ and all the electrons in the $\text{Al}_x\text{Ga}_{1-x}\text{As}$ having a component in the negative z -direction will propagate to the GaAs. The current density from the left (GaAs) to the right ($\text{Al}_x\text{Ga}_{1-x}\text{As}$) is then

$$J_{LR} = \frac{e}{4\pi^3} \int_{k_x k_y} dk_x dk_y \int_{k_z > k_{z0}} dk_z v_z f_0(k) \quad (10.1)$$

where k_{z0} is the minimum \mathbf{k} vector component necessary to overcome the barrier. Classically, we can calculate k_{z0} from

$$\frac{m^*(v_{z0})^2}{2} = \Delta E_c \quad \text{and} \quad m^* v_z = \hbar k_z \quad (10.2)$$

Here v_{z0} is the minimum velocity necessary to overcome ΔE_c . There are several quantum corrections (the quantum-transmission coefficient, deviations from the effective mass approximation, and tunneling, which will be discussed later). Assuming, then, for $f_0(k)$ a Maxwellian distribution at a quasi-Fermi level, we may rewrite the current as

$$J_{LR} = \left(\frac{m^*}{\hbar} \right)^3 \frac{e}{4\pi^3} e^{(E_{QF}^L - E_c^L)/kT} \int_{-\infty}^{\infty} dv_x dv_y \int_{v_z}^{\infty} dv_z v_z \exp \left[-\frac{m^*}{2kT} (v_x^2 + v_y^2 + v_z^2) \right] \quad (10.3)$$

Here E_{QF}^L is the quasi-Fermi level on the left side measured from the left-side conduction band edge. It is easy to perform the x - y integration in cylindri-

cal coordinates. Transforming $dv_x dv_y$ to $v dv d\phi$ ($v = \sqrt{v_x^2 + v_y^2}$) and denoting $(v_x^2 + v_y^2) (m^*/2)$ by \bar{E} , one obtains by using $d\bar{E} = m^* v dv$

$$\int_{-\infty}^{\infty} dv_x dv_y \exp\left(-\frac{m^*}{2kT}(v_x^2 + v_y^2)\right) = \frac{2\pi}{m^*} kT \quad (10.4)$$

From (10.4) we reduce (10.3) to

$$j_{LR} = \frac{e}{2\pi^2} \frac{m^{*2}}{\hbar^3} kT e^{(E_{QF}^L - E_c^L)/kT} \int_{v_z 0}^{\infty} v_z e^{-m^* v_z^2 / 2kT} dv_z \quad (10.5)$$

which gives

$$j_{LR} = A^* T^2 e^{(E_{QF}^L - E_c^L - |\Delta E_c|)/kT} \quad (10.6)$$

j_{LR} as given by Eq. (10.6) is called the thermionic emission current, and $A^* \equiv (em^* k^2) / (2\pi^2 \hbar)$ is the effective Richardson constant, which for the free electron ($m^* = m$) is 120 amp/(cm²K²), and T is the temperature of the carriers. If the electrons are heated to a temperature T_c different from the lattice temperature T_L , then T_c has to appear in the exponent.

The current from the $\text{Al}_x\text{Ga}_{1-x}\text{As}$ toward the GaAs, j_{RL} is given by

$$j_{RL} = A^* T^2 e^{(E_{QF}^R - E_c^R)/kT} \quad (10.7)$$

where E_{QF}^R is the quasi-Fermi level in the $\text{Al}_x\text{Ga}_{1-x}\text{As}$ measured from the conduction band edge in the $\text{Al}_x\text{Ga}_{1-x}\text{As}$.

Assume now that the $\text{Al}_x\text{Ga}_{1-x}\text{As}$ is doped with a density N_D of donors and the GaAs is undoped. The quasi-Fermi levels then have a different distance from the conduction band edge of the two materials and consequently $j_{LR} \neq j_{RL}$, and a net current can flow. This, of course, cannot be true in equilibrium. We, therefore, have to conclude that the diagram in Figure 11.1 changes shape as equilibrium is approached. The reason is easy to understand. If we start out at a certain time $t = 0$ with the circumstance of Figure 11.1, then, of course, electrons start to flow from the $\text{Al}_x\text{Ga}_{1-x}\text{As}$ (doped) to the GaAs (undoped) where their potential energy is lowest. We can estimate roughly the carrier density as a function of time in the $\text{Al}_x\text{Ga}_{1-x}\text{As}$ close to the interface by using the equation of continuity, Eq. (11.4), generation-recombination neglected:

$$e \frac{\partial n}{\partial t} = \frac{\partial j}{\partial z} \quad (10.8)$$

and by assuming that current flows from $\text{Al}_x\text{Ga}_{1-x}\text{As}$ to GaAs only. We also assume that this thermionic emission current flows only within a certain region, which is of the order of the mean free path L_m of the electrons. (Beyond this region the average velocity is much smaller because of collisions.) Then, using $\partial j / \partial z \approx j / L_m$, we have

$$e \frac{\partial n(t)}{\partial t} \approx -A^* T^2 e^{E_F^R(t)/kT} / L_m \quad (10.9)$$

Because from Eq. (5.23), by using a time-dependent quasi-Fermi level

$$n(t) = \int_0^\infty e^{(E_F^R(t)-E)/kT} g(E) / dE$$

we obtain

$$e^{E_F^R(t)/kT} = n(t) 4 \left(\frac{\hbar^2}{2m^* \pi kT} \right)^{3/2} \equiv n(t) C_1 \quad (10.10)$$

Equations (10.9) and (10.10) give

$$\frac{\partial n(t)}{\partial t} = -\frac{A^* T^2}{eL_m} C_1 n(t)$$

and

$$n(t) = n_c \exp \left[-\left(\frac{A^* T^2 C_1}{eL_m} t \right) \right] \quad (10.11)$$

where n_c is the concentration in the $\text{Al}_x\text{Ga}_{1-x}\text{As}$ at $t = 0$. The time constant $eL_m/A^* T^2 C_1$ is, near room temperature, of the order of picoseconds; that is, the $\text{Al}_x\text{Ga}_{1-x}\text{As}$ loses initially its electrons in picoseconds, and positively charged donors remain behind.

On the $\text{Al}_x\text{Ga}_{1-x}\text{As}$ side, these positively charged donors give rise to a potential barrier, which slows down the electron transfer to the GaAs and finally prevents electrons from leaving $\text{Al}_x\text{Ga}_{1-x}\text{As}$ at a higher rate than electrons returning from the GaAs. Then equilibrium has been reached; the corresponding equilibrium band diagram is shown in Figure 10.2.

The dynamics of electrons being emitted out of the GaAs can be treated in an analogous fashion by including the band edge discontinuity into the exponent of the current density. The time constants for emission into the $\text{Al}_x\text{Ga}_{1-x}\text{As}$ can then be considerably longer, depending on the magnitude of the band edge continuity. If the emission out of the GaAs is enhanced owing to heating of the electron gas by electrical fields parallel to the interface, one speaks of real space transfer. In steady state, the in and out currents are balanced and no net current flows perpendicular to the barrier. However, this does not mean that the system is in equilibrium. There can be current flow parallel to the interface. The inverse condition is, of course, also possible.

The theory of transport over a heterobarrier as shown in Figure 10.2 is basic to the theory of semiconductor devices. A rigorous solution to the heterobarrier problem can in general be obtained only by ensemble Monte Carlo simulations combined with solutions of the Poisson equation. Because of the inhomogeneous charge distribution, Poisson's equation plays a significant role. We show below two steps toward an approximate solution. First we treat the right-hand side of Figure 10.2, which is almost depleted of mobile electrons. Next we discuss the left-hand side where electrons accumulate, and the matching boundary conditions that connect the two sides.

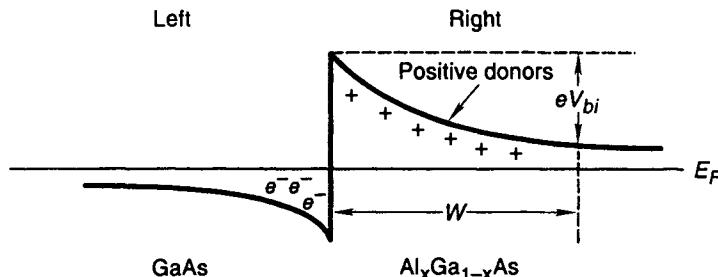


Figure 10.2 As Figure 11.1, but with band bending owing to a charge distribution included.

10.2 FREE CARRIER DEPLETION OF SEMICONDUCTOR LAYERS

As can be seen from Figure 10.2, the $\text{Al}_x\text{Ga}_{1-x}\text{As}$ depletes a distance $\approx W$. In other words, the electron concentration is considerably smaller in this region than it is far away to the right-hand side. There is a “fuzzy” region where the depletion region goes over into the normal semiconductor. This transition occurs within the Debye length. We will return to this point below. The calculation of W and the potential drop V_{bi} (Figure 10.2) necessitates, of course, the complete solution of the Poisson equation (also at the left side of Figure 10.2). However, an important relation between W and V_{bi} can be derived without this knowledge, and this relation will be the subject of this section. The Poisson equation, Eq. (6.20), reads in one dimension (including the fixed charge of positive donors N_D^+ and negative acceptors N_A^- , as well as electrons of density n and holes of density p):

$$\frac{\partial F}{\partial z} = -\frac{e}{\epsilon \epsilon_0} (p - n + N_D^+ - N_A^-) \quad (10.12)$$

with F being the electric field.

We now solve the Poisson equation by using a very general trick that is based on two steps: (1) partial integration and (2) the depletion approximation.

It is convenient to choose the zero of the coordinate system at the interface. We have then

$$V_{bi} = - \int_0^W F dz$$

and

$$V_{bi} = -zF|_0^W + \int_0^W z \frac{\partial F}{\partial z} dz \quad (10.13)$$

The depletion approximation can be stated in the following way: The semiconductor is depleted (i.e., free of mobile electrons or holes) over a distance W at the end of which (at $z = W$) the electric field is zero. The merit of this approximation can be seen directly from Figure 10.2. One assumes that the electric field is screened out by the free carriers after the width W . However, to screen a field we need a Debye or Thomas-Fermi length or the general length $2\pi/q_s$

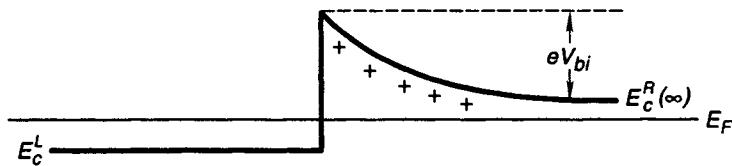


Figure 10.3 Band structure close to a band edge discontinuity. The band bending at the undepleted (left) side is neglected. However, it is important to note that negative charge must accumulate in the GaAs because the semiconductor as a whole must be neutral.

as given in Eq. (6.26). Within this length there is a transition region containing mobile charge and a finite electric field. We therefore can expect the depletion approximation to work only if

$$W \gg 2\pi/q\epsilon \quad (10.14)$$

Then we obtain from Eq. (10.13)

$$V_{bi} \approx \int_0^W z \frac{\partial F}{\partial t} dz \quad (10.15)$$

This equation is very useful for estimates of depletion voltages and can be regarded as one of the basic tools to calculate explicitly the properties of semiconductor devices. Neglecting n and p in the depleted region and assuming that only charged donors are present, we have

$$V_{bi} \approx \int_0^W z \frac{e}{\epsilon \epsilon_0} N_D^+ dz \quad (10.16)$$

For a constant donor concentration N_D^+ this gives

$$V_{bi} \approx \frac{1}{2} \frac{W^2 e}{\epsilon \epsilon_0} N_D^+ \quad (10.17)$$

This equation relates V_{bi} to W but does not give us the value of either. As outlined above, to obtain both we need to solve the Poisson equation on the left side and connect the two solutions. This will be done later, although it is pretty complicated; we cannot use the depletion approximation on the other side. For the time being we approximate the solution on the left side by assuming that the GaAs is much heavier doped than the $\text{Al}_x\text{Ga}_{1-x}\text{As}$, and therefore the bands do not distort substantially in the GaAs.

Within this approximation, we write (Figure 10.3)

$$|eV_{bi}| \approx \Delta E_c + E_c^L - E_c^R(\infty) \quad (10.18)$$

For j_{RL} one obtains [see Eq. (10.6)]

$$j_{RL} = A^* T^2 \exp(E_F^R - E_c^R(\infty) - |eV_{bi}|)/kT \quad (10.19)$$

Of course in equilibrium $E_F^R = E_F^L$ and $j_{LR} = j_{RL}$, that is, no net current is flowing.

For W one obtains

$$W \approx \left[\frac{2\epsilon\epsilon_0(\Delta E_c + E_c^L - E_c^R(\infty))}{e^2 N_D^+} \right]^{1/2} \quad (10.20)$$

For rough estimates, the term $E_F^R - E_c^R(\infty)$ can be neglected.

10.3 CONNECTION RULES FOR THE POTENTIAL AT AN INTERFACE

Up to now we have treated mainly the right side of the junction and assumed that the other side (GaAs) is heavily doped and exhibits a flat conduction band edge (Figure 10.3). We are now going to solve the Poisson equation for the GaAs side.

The connection rules for the solutions of the Poisson equation at the interface of two dielectrics are that the potential ϕ is continuous and the normal field (in absence of interface charge) changes as the ratio of the dielectric constant does. The parallel field is also continuous. To write down the equations, we need to remember that we solve Poisson's equation for the additional charges in the crystal; that is, for donors, acceptors, electrons, and holes, while the band structure (the $E(k)$ relation) of the crystal is predetermined. Also predetermined are the band edge discontinuities between one crystal and another. The external potential just shifts these energy bands rigidly as long as it is not too strong [see the effective mass theorem and Eqs. (3.32) and (3.34)]. It is often convenient to introduce a vacuum reference energy, which is the electron energy (at rest) outside the semiconductor (metal), and to measure the band edge energies from there. We will do this in several instances.

Given these facts, we can connect the left and right sides of the interface potential ϕ_i

$$\phi_i^R - \phi_i^L = \Delta E_c \quad (10.21)$$

and

$$\epsilon_L \frac{\partial \phi_i^L}{\partial z} = \epsilon_R \frac{\partial \phi_i^R}{\partial z} \quad (10.22)$$

In a homogeneous material the potential connects, of course, continuously and

$$\phi_i^L = \phi_i^R \quad (10.23)$$

This condition must also be used in cases when ϕ is calculated from Poisson's equation including only the charge of dopants and free carriers. The band edge discontinuity involves the crystal atoms and does not appear in such a solution. It has to be added subsequently if necessary. Notice that Eq. (10.23) applies only in the absence of interface dipoles. To proceed with the solution of Poisson's

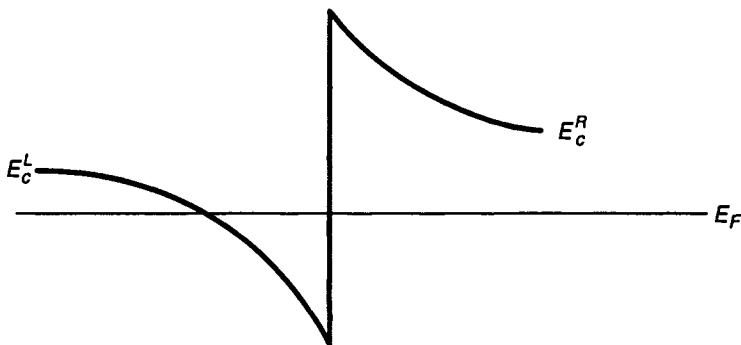


Figure 10.4 Conduction band edge for the case of electron accumulation at the heterojunction.

equation on the left-hand side, we can therefore assume ϕ_i^L and $\partial\phi_i^L/\partial z$ as given boundary values and perform the connection later.

10.4 SOLUTION OF POISSON'S EQUATION IN THE PRESENCE OF MOBILE (FREE) CHARGE CARRIERS

10.4.1 Classical Case

We discuss now the solution of Poisson's equation at the left side of the heterojunction. (The index L, for left, is used only in cases of possible confusion.) Figure 10.4 shows a blowup of a possible form of the conduction band edge at the GaAs (left) side of the junction. To proceed with the calculation of the potential ϕ from Poisson's equation, we need to express the carrier concentration as a function of ϕ . Because n follows the classical Boltzmann's law—to convince yourself solve Eq. (11.10) for $j = 0$ —we have

$$n = n_c(-\infty) \exp(e\phi/kT_c) \quad (10.24)$$

Here $n_c(-\infty)$ is the equilibrium electron concentration far away to the left from the junction. Equation (10.34) is derived in the Problems section.

Denoting $e\phi/kT_c$ by $\bar{\phi}$, Poisson's equation reads (in the absence of holes or acceptors):

$$\frac{\partial^2 \bar{\phi}}{\partial z^2} = -\frac{e^2}{\epsilon \epsilon_0 k T_c} (N_D^+ - n_c(-\infty) e^{\bar{\phi}}) \quad (10.25)$$

If $n_c(-\infty) \approx N_D^+$, as is the case for constant doping, we have

$$\frac{\partial^2 \bar{\phi}}{\partial z^2} = +\frac{e^2 n_c(-\infty)}{\epsilon \epsilon_0 k T_c} (e^{\bar{\phi}} - 1) \quad (10.26)$$

There is no explicit solution of this differential equation and we therefore cannot get $\bar{\phi}$ explicitly. We can, however, obtain the total concentration of excess charge

Q_{tot} at the left side of the junction (GaAs) as a function of interface potential ϕ_i and field F_i . This information is sufficient for some approximate considerations in the theory of devices. A special technique, worthwhile to remember, is still necessary to calculate Q_{tot} as a function of ϕ_i and F_i . We multiply Eq. (10.26) by the normalized electric field $F = -\partial\bar{\phi}/\partial z$ and integrate. The left-hand side of Eq. (10.26) gives

$$-\int \frac{\partial^2 \bar{\phi}}{\partial z^2} \frac{\partial \bar{\phi}}{\partial z} dz = \frac{1}{2} \left[\frac{\partial \bar{\phi}}{\partial z} \right]^2 \quad (10.27)$$

The right-hand side becomes

$$\begin{aligned} -\frac{e^2 n_c(-\infty)}{\epsilon \epsilon_0 k T_c} \int (e^{\bar{\phi}} - 1) \frac{\partial \bar{\phi}}{\partial z} dz &= -\frac{e n_c(-\infty)}{\epsilon \epsilon_0 k T_c} \int (e^{\bar{\phi}} - 1) d\bar{\phi} \\ &= -\frac{e^2 n_c(-\infty)}{\epsilon \epsilon_0 k T_c} (e^{\bar{\phi}} - \bar{\phi} + \text{constant}) \end{aligned} \quad (10.28)$$

Therefore we have the relation

$$\bar{F}^2 = \frac{2e^2 n_c(-\infty)}{\epsilon \epsilon_0 k T_c} (e^{\bar{\phi}} - \bar{\phi} + \text{constant})$$

The constant has to be equal to -1 if we choose $\bar{\phi} = 0$ and $\bar{F} = 0$ for $z = -\infty$. One then obtains

$$\bar{F}^2 = \frac{2e^2 n_c(-\infty)}{\epsilon \epsilon_0 k T_c} (e^{\bar{\phi}} - \bar{\phi} - 1) \quad (10.29)$$

Integrating Eq. (10.25) between $-\infty$ and the location $z = 0$ (the interface), one has

$$\epsilon \epsilon_0 F_i = -Q_{\text{tot}} \quad (10.30)$$

where Q_{tot} is the total excess charge (per unit area) between $z = -\infty$ and 0 .

Equation (10.29) with $\phi = \phi_i$ and Eq. (10.30) can be used together with the connection rules to obtain the complete solution of Poisson's equation. Note, however, that we obtained only the overall charge (integral) and not the charge as a function of z . We will return to this point after the quantum mechanical solution, as we will discuss the consequences of these solutions for semiconductor devices. For large values of $\bar{\phi}$, the exponent dominates Eq. (10.29), and with the help of Eq. (10.30), one obtains $Q_{\text{tot}} \propto e^{\phi_i/2}$.

The combination of Eqs. (10.29) and (10.30) gives us Q_{tot} as a function of the interface potential ϕ_i , which is schematically shown in Figure 10.5. This figure demonstrates that, with increasing interface potential, the net carrier concentration increases exponentially on the GaAs side. In other words, electrons accumulate according to Boltzmann's law in the GaAs. They are located mostly at the interface and one speaks of an accumulation layer.

If the GaAs were doped with acceptors at a density N_A , the derivation above would proceed in almost precisely the same way. However, because we then

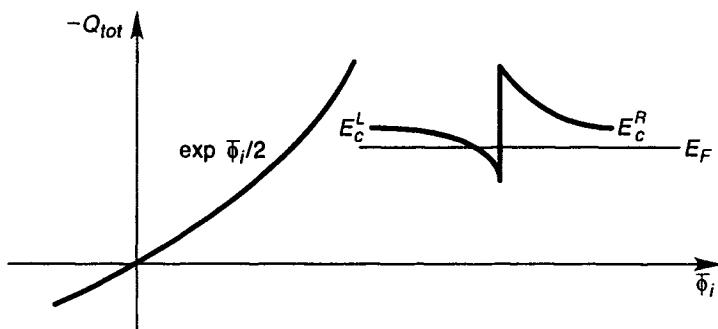


Figure 10.5 Schematic dependence of total charge Q_{tot} on the interface potential. The inset shows a typical band-structure diagram for accumulation.

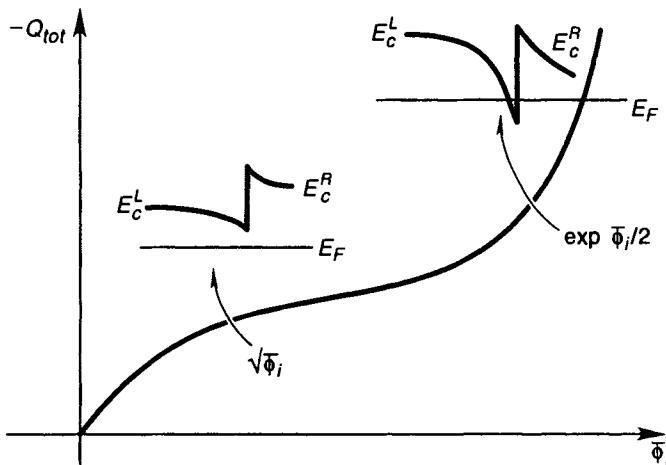


Figure 10.6 Schematic dependence of total charge Q_{tot} on the interface potential. The insets show typical band diagrams for depletion and inversion, respectively.

would have to permit the presence of holes, electrons initially accumulating in the GaAs would recombine with holes, and for small ϕ_i a depletion layer would form before electrons are further accumulated. Because the depletion width W changes as $\sqrt{\phi}$ [see Eq. (10.20)], Q_{tot} changes initially only as $\sqrt{\phi}$. At higher values of ϕ electrons start to accumulate in the *p*-type semiconductor, and one says that an inversion layer forms and again $Q_{\text{tot}} \approx \exp \phi / 2$. This is shown in Figure 10.6.

The onset of strong inversion in Figure 10.6 is not defined rigorously. However, inspection of Eqs. (10.29) and (10.30) shows that it occurs approximately at an interface potential

$$\bar{\phi}_s = 2kT_c \ln(N_A/n_i)/e \quad (10.31)$$

where n_i is the intrinsic concentration at $-\infty$, that is, in bulk GaAs.

10.4.2 Quantum Mechanical Case

The potential well in Figure 10.4 can be very narrow. For inversion layers with an electron density around 10^{18} cm^{-3} the typical well width is of the order of 10 to 100 Å for a given material. A typical value for the de Broglie wavelength of conduction electrons is 50 Å. Therefore, quantum effects can be very important, and we have to solve Schrödinger's equation self-consistently with Poisson's equation, much in the way as was done in Chapter 6. However, the well cannot be treated as a small perturbation, and therefore a numerical procedure has to be used to obtain a solution.

As shown in Eq. (1.32), the electrons in a quantum well populate only discrete energy levels owing to size quantization. However, our well is not purely one dimensional. Parallel to the interface the electrons can move uninhibited, and in the effective mass approximation we have for the parallel kinetic energy of the electrons

$$E_{\parallel} = \hbar^2 k_{\parallel}^2 / 2m^* \quad (10.32)$$

Thus, the electrons are still in a “band.” However, because their perpendicular energy is quantized into levels with energy $E_n (n = 0, 1, 2, \dots)$, one speaks about subbands. The energies E_n are the solutions of the one-dimensional Schrödinger equation

$$\left(-\frac{\hbar^2}{2m^*} \frac{d^2}{dz^2} + \phi_{\text{tot}}(z) \right) \zeta_n(z) = E_n \zeta_n(z) \quad (10.33)$$

The potential $\phi(z)$ that confines the electrons consists of several contributions:

$$\phi_{\text{tot}} = \phi(z) + \Delta E_c H(z) + \phi_{\text{im}}(z) + \phi_{\text{ex}}(z) \quad (10.34)$$

$\phi_{\text{ex}}(z)$ is a “many-body” contribution and comes from the fact that electrons are Fermions and are correlated owing to their spin and the Pauli principle. The functional form of $\phi_{\text{ex}}(z)$ is known (see the review of Stern and Das Sarma [3]), and it can easily be included in the calculation; for our purpose here we disregard it. $\phi_{\text{im}}(z)$ represents the contribution of the image force (i.e., the image potential energy), which can be significant; we are discussing the boundary of two dielectrics. For the AlGaAs-GaAs system the image force is negligible because the dielectric constants of both materials are very similar. Values for $\phi_{\text{im}}(z)$ are also given by Stern and Das Sarma [3]. $H(z)$ is the step function with $H(z) = 1$ for $z \geq 0$ and $H(z) = 0$ otherwise, and $\phi(z)$ is the electrostatic potential, which in turn is a function of charge density and therefore of E_n and $\zeta_n(z)$. Analogous to Eq. (6.16) we can write the charge density ρ owing to the inverted (accumulated) free carriers as

$$\rho_{\text{inv}} = e \sum_{k_{\parallel}} \sum_n \frac{|\zeta_n(z)|^2 f(k_{\parallel})}{A} \quad (10.35)$$

The sum over k_{\parallel} can be replaced by a twofold integral analogous to Eq. (5.9) as

$$\sum_{k_{\parallel}} \rightarrow \frac{2A}{(2\pi)^2} \int dk_x dk_y \quad (10.36)$$

Here A is the interference area. If $f(k_{\parallel})$ is simply the Fermi distribution, which depends only on energy, the integration in polar coordinates is over the polar angle and gives 2π . The \mathbf{k} integration is easily transformed to an energy integration, which gives for a spherical parabolical band

$$\sum_{k_{\parallel}} \rightarrow \frac{Am^*}{\pi\hbar^2} \int_{E_n}^{\infty} dE \quad (10.37)$$

The integration of Eq. (10.35) gives then

$$\rho_{\text{inv}} = e \sum_n N_n |\zeta_n(z)|^2 \quad (10.38)$$

with

$$N_n = \frac{m^* k T_c}{\pi \hbar^2} \ln \left[1 + \exp \left(\frac{E_F - E_n}{k T_c} \right) \right] \quad (10.39)$$

Therefore, Poisson's equation reads

$$\frac{\partial^2 \phi}{\partial z^2} = +\frac{e}{\epsilon \epsilon_0} \left(N_A^- + e \sum_n N_n |\zeta_n(z)|^2 \right) \quad (10.40)$$

Here we have assumed a background of acceptors instead of donors, that is, we have treated the inversion case. Holes have again been neglected, although they can easily be included in Eq. (10.40).

This equation shows that the carrier concentration is not a simple function of the potential but rather a functional; that is, the charge density does not depend only on the potential $\phi(z)$ at a certain point z but depends on the potential as a whole. Each different function $\phi(z)$ will result in different $\zeta_n(z)$ and therefore in different ρ_{inv} . Thus, we cannot find an equation corresponding to Eq. (10.29), which completed the classical solution. What we have to do instead is solve Eqs. (10.33) and (10.40) iteratively. This proceeds as follows.

We start with a guessed potential ϕ_{tot} in Eq. (10.33) and solve it to obtain E_n and $\zeta_n(z)$ by using, for example, the Ritzit method. Then, using these results, we integrate Eq. (10.40) twice to obtain $\phi(z)$. A numerical procedure that accomplishes just that is listed in Appendix F.

Although it is relatively easy to develop a numerical code, a few analytical considerations are in order. A useful analytical approximation to $\zeta_n(z)^2$ has been given by Stern and Howard [4]:

$$\zeta_n(z)^2 = \frac{1}{2} b^3 z^2 \exp(-bz) \quad (10.41)$$

with $3/b$ being the average extension of the inversion layer in the z -direction. The expression for b will not be given here; it can be found in the work of Stern and Howard [4]. A typical value for $3/b$ is 5 Å. Equation (10.41) is only for the lowest subband, and we have omitted the subscript n .

Equation (10.41) permits analytical solution of the Poisson equation and also an approximate solution for E_n . There is, however, only limited use for these solutions from our point of view; for many device applications the classical solution is sufficient, and for calculations of the electron mobility, which will be discussed in the next section, one prefers either a simple $\delta(z)$ function approximation to $\zeta_n(z)^2$ or the complete numerical solution as described in Appendix F. This appendix also deals with the complications that arise from the many-valley character of the silicon conduction band.

The classical solution is sufficient for most device applications connected to the calculation of Q_{tot} . The reason is the following. Equation (10.30) is also valid from a quantum point of view because it merely represents one formal integration of Poisson's equation (Gauss's law). Assume now for simplicity that the AlGaAs is free of charge and that the potential at the surface of the AlGaAs is fixed. This corresponds to the situation in a metal-insulator semiconductor device where on top of the insulator (AlGaAs, SiO₂) a metal is placed as a "gate" and the gate voltage V_G is fixed. A twofold integration of Poisson's equation in the charge-free AlGaAs gives then

$$\phi_R = C_1 z + C_2 \quad (10.42)$$

with C_1 and C_2 being integration constants. With the interface at $z = 0$, the boundary condition at the gate gives

$$V_G = C_1 d + C_2 \quad (10.43)$$

where d is the AlGaAs thickness. The connection rules Eq. (10.21) and (10.22) give

$$-\epsilon_L F_i^L = \epsilon_R C_1 \quad (10.44)$$

and, because our solutions for ϕ have been obtained from doping and free carrier charge, only Eq. (10.23) applies (we can disregard Δ_c), which gives

$$C_2 = \phi_i^R \quad (10.45)$$

Equations (10.44) and (10.45) together with Eqs. (10.43) and (10.30) yield

$$V_G = \frac{Q_{\text{tot}} d}{\epsilon_0 \epsilon_R} + \phi_i^R \quad (10.46)$$

where $\epsilon_0 \epsilon_R / d \equiv C_{\text{ins}}$ is the (insulator) capacitance of the AlGaAs layer and ϕ_i^R the interface potential. We remember that for the classical case we have derived an equation for $\phi_i^L = \phi_i^R$ as a function of $F_i^L = -Q_{\text{tot}} / \epsilon_0 \epsilon_R$, and therefore we can calculate Q_{tot} as a function of V_G . A complete quantum treatment is more involved. However, the classical treatment is approximately valid for the following reason.

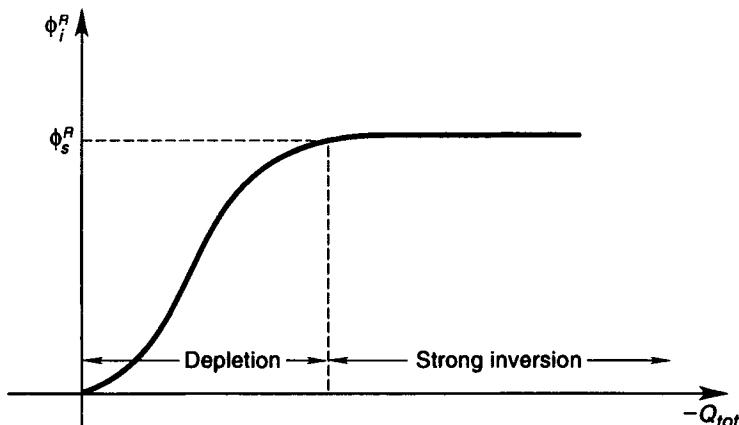


Figure 10.7 Interface potential as a function of total charge Q_{tot} . This figure corresponds precisely to Figure 10.6.

The interface potential rises first linearly with Q_{tot} and then saturates at the onset of strong inversion (accumulation) of free charge at the interface as shown in Figure 10.7. The reason is that the free charge screens the potential at the interface and prevents any further rise. The potential at the onset of strong inversion, ϕ_s^R , has been given in Eq. (10.31). This equation is also valid from a quantum point of view because before strong inversion occurs, the potential well is not very deep and the potential varies slowly enough to make size quantization unimportant. Beyond the onset of strong inversion quantization occurs. However, ϕ_s^R stays approximately constant and one therefore obtains from Eq. (10.46)

$$Q_{tot} = C_{ins}(V_G - \phi_s^R) \quad (10.47)$$

which is approximately valid from both a classical and a quantum point of view. This equation is basic to the operation of the important metal–oxide–silicon semiconductor (MOS) field effect transistor, discussed in Chapter 15.

The general validity of Eq. (10.47) is the precise reason that the inclusion of quantum effects is mostly unnecessary in order to obtain a semiquantitative picture of MOS transistor operation.

Quantitatively, quantum effects are important for two reasons. First, it is not the total charge that one needs to know to characterize a transistor completely but rather the spatial distribution of free and fixed charge. This distribution differs somewhat classically and quantum mechanically. From Eq. (10.41) it can be seen that the wave function vanishes at the interface. This approximation is only good for the MOS system and not quite as good for AlGaAs-GaAs. The reason is that in the latter system the band edge discontinuity is rather small and the wave function penetrates the AlGaAs somewhat. Therefore, the probability to find mobile interface charge Q_i directly at the interface is quantum mechanically much smaller than one would expect from the classical theory [Eqs. (10.29) and (10.30)]. Figure 10.8 shows this effect.

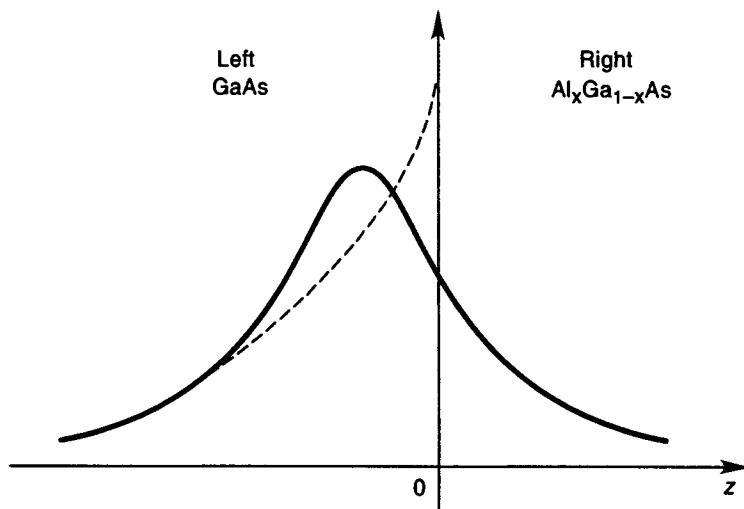


Figure 10.8 Classical (---) and quantum mechanical (—) distribution of free charges close to the heterointerface of AlGaAs-GaAs.

The quantum distribution of charge leads to slightly different device characteristics than expected from classical reasoning. We will see later that the insulator can be regarded as capacitance and so can the depletion region at the left side. This is schematically shown in Figure 10.9.

As is evident from Figure 10.9, the quantum mechanical calculation gives an additional almost insulating range of length $d_1 - d$. The dielectric constant in this range is approximately the average between the left and right side dielectric constants; that is,

$$\bar{\epsilon} \approx (\epsilon_L + \epsilon_R)/2$$

The length $d_1 - d$ is roughly one-third of the extension of the inversion layer, which is $3/b$. This gives the quantum capacitance

$$C_{QM} \approx \epsilon_0 \bar{\epsilon} b \quad (10.48)$$

Because $3/b$ is typically 5 \AA , the series capacitance C_{QM} is only significant for very thin insulator (AlGaAs or SiO_2) thicknesses.

Thus, concerning charge distribution, quantum effects will be typically small. The major effects of size quantization are related to the mobility of electrons parallel to the layer, which will be treated in the next section. Some of these ideas are due to H. Shichijo (unpublished).

Before discussing mobilities in size-quantized systems, a few remarks on the insulating qualities of AlGaAs should be made. We have referred to the AlGaAs as an insulator; however, the energy gap of the AlGaAs is only slightly larger than the gap of the GaAs and the intrinsic concentration of AlGaAs is usually not negligible. Furthermore, if we apply a voltage, a thermionic emission current

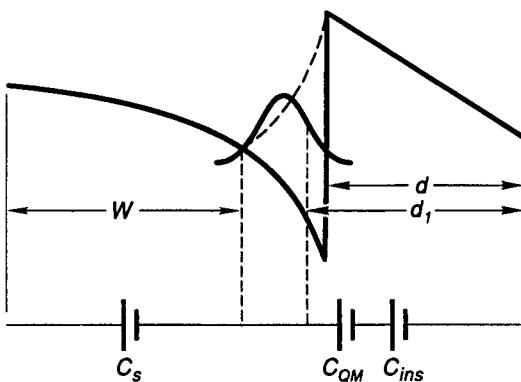


Figure 10.9 Insulating regions and corresponding capacitance for the classical and quantum mechanical calculations at a heterojunction.

will flow according to Eqs. (10.6) and (10.7) (with the appropriate quasi-Fermi levels). We will see from Eqs. (11.36) and (11.38) that undoped semiconductors can carry currents that are much larger than expected from their intrinsic concentration. This is even true for different neighboring semiconductors, depending on doping of the neighbor and the band edge discontinuity. Notice also that electrons penetrate the barrier because of the extension of the wave function beyond the interface, which represents still another “influence” of a neighboring material that can change the insulating qualities (tunneling).

10.5 PRONOUNCED EFFECTS OF SIZE QUANTIZATION AND HETEROLAYER BOUNDARIES

We have seen from the derivation of Eq. (10.47) that the quantum corrections to the total charge density at interfaces will usually be small. However, the effects of size quantization on the electronic properties at interfaces can be significant. As we have seen in Chapter 8, the density of states enters into the scattering rates and collision operators. In fact the two-dimensional constant density of states has helped us solve some problems that are not tractable in three dimensions. Therefore, mobility, energy loss, diffusion constant, and other transport properties will be different at interfaces as size quantization leads to a quasi-two-dimensional behavior of the electrons. These differences are more pronounced than the contributions of the capacitance of Eq. (10.48). They are not, however, order of magnitude changes. The mobility may change by a factor of 2 or so, but will not change drastically except under unusual conditions.

These unusual conditions are mostly connected to a separation of scattering centers and electrons, as has been illustrated in Figure 10.3. In this figure dopants are only on the AlGaAs (right) side, while all the electrons have left their parent donors and reside in the GaAs, where they are confined by the heterointerface

and are remote from the scattering centers.

Remote and confined are the key words that indicate unusual transport effects (such as the enhanced mobility) at interfaces. It is not only the impurities that can be remote, but also phonon modes characteristic of only one medium can be separated from the electrons. We will discuss some questions related to phonons toward the end of the section. Now we will focus our attention on the detailed treatment of impurity scattering under conditions as shown in Figure 10.3, which is the circumstance known as modulation doping (Dingle, et al. [1]). Below we will derive the scattering rate for electrons interacting with remote impurities. This example gives us an excellent opportunity to repeat and illustrate the principles discussed in Chapters 6 and 7.

Because the motion of electrons is confined in the z -direction, the wave function is more complicated than the free electron wave function and we have to use the solution $\zeta_n(z)$ of Eq. (10.33) or approximations to this solution. The perturbed wave function (by the impurity charge) is still given by Eq. (6.14) but the coefficient $b_{\mathbf{k}+\mathbf{q}}$ now takes the form

$$b_{\mathbf{k}+\mathbf{q}} = \frac{\int e^{-i(\mathbf{k}_{||} + \mathbf{q}_{||}) \cdot \mathbf{r}_{||}} \zeta_l^*(z) eV_{\text{tr}} \zeta_m(z) e^{-i\mathbf{k} \cdot \mathbf{r}_{||}} d\mathbf{r}_{||} dz}{E(\mathbf{k})_l - E(\mathbf{k} + \mathbf{q})_m} \quad (10.49)$$

Performing the $r_{||}$ integration and denoting the two-dimensional Fourier transform of V_{tr} by $V_{\text{tr}}^{q_{||}}(z)$, we have

$$b_{\mathbf{k}+\mathbf{q}} = \frac{e \int_{-\infty}^{\infty} \zeta_l^*(z) \zeta_m(z) V_{\text{tr}}^{q_{||}}(z) dz}{E(\mathbf{k})_l - E(\mathbf{k} + \mathbf{q})_m} \quad (10.50)$$

Here we have implicitly assumed that the indices l and m run over all subbands in the conduction band. If the valence band also contributes (e.g., for the calculation of the dielectric constant of the semiconductor, and not only the contribution of the free electrons), then the sum runs also over all valence band subbands.

If $V_{\text{tr}}^{q_{||}}(z)$ varies only slowly with z , we can approximate Eq. (10.50) by

$$b_{\mathbf{k}+\mathbf{q}} = \frac{eV_{\text{tr}}^{q_{||}}(0) \int_{-\infty}^{\infty} \zeta_l^*(z) \zeta_l(z) dz}{E(\mathbf{k})_l - e(\mathbf{k} + \mathbf{q})_l} \quad (10.51)$$

because the integral vanishes for $l \neq m$ as the $\zeta_l(z)$ are orthogonal functions. Analogous to the derivation of Eq. (6.19) (with l being now the subband index), we obtain

$$\delta\rho = e^2 V_{\text{tr}}^{q_{||}}(0) e^{-i\mathbf{q}_{||} \cdot \mathbf{r}_{||}} \int_{-\infty}^{\infty} \zeta_l^*(z) \zeta_l(z) dz \sum_{k_{||}, l} \frac{f^l(\mathbf{k}) - f^l(\mathbf{k} + \mathbf{q})}{E(\mathbf{k})_l - E(\mathbf{k} + \mathbf{q})_l} + \text{cc} \quad (10.52)$$

where cc stands for complex conjugate and the additional ζ s are introduced as in Eq. (6.17). The sum can be converted into an integral and expanding $f^l(\mathbf{k} + \mathbf{q})$

and $E(\mathbf{k} + \mathbf{q})$ for small q , one obtains

$$\delta\rho = e^2 V_{\text{tr}}^{q\parallel}(0) e^{-i\mathbf{q}\parallel \cdot \mathbf{r}\parallel} 2 \left[\sum_l \int \left(-\frac{\partial f'}{\partial E} \right) g(E) dE \right] \times \int_{-\infty}^{\infty} \zeta_l^*(z) \zeta_l(z) dz \cdot \zeta_l^*(z) \zeta_l(z) \quad (10.53)$$

To simplify further, we consider now the case when only one subband is populated and assume that the electron gas is essentially two dimensional. In this spirit we approximate

$$\zeta_l^*(z) \zeta_l(z) \sim \delta(z) \quad (10.54)$$

which will give us good results only if the electrons are confined in a very narrow well.

The potential can be written in its original form (before Fourier transformation)—that is, $V_{\text{tr}}^{q\parallel}(0) e^{-i\mathbf{q}\parallel \cdot \mathbf{r}\parallel} = V_{\text{tr}}(z=0)$ —and we have

$$\delta\rho = e^2 V_{\text{tr}}(z=0) 2 \int_0^{\infty} \left(-\frac{\partial f}{\partial E} \right) g(E) \delta(z) dE \quad (10.55)$$

where $g(E)$ is the two-dimensional density of states.

We proceed now to calculate the scattering rate of two-dimensional electrons for scattering by screened impurities at an arbitrary distance z_0 from the sheet of electronic charge. From perturbation theory, Eq. (7.13), we know that all we need is the (two-dimensional) Fourier transform of the scattering potential, that is, $V_{\text{tr}}^{q\parallel}(0)$. To obtain $V_{\text{tr}}^{q\parallel}(0)$, we need to Fourier-transform Eq. (6.20) in two dimensions. Multiplying Eq. (6.20) by $e^{i\mathbf{q}\parallel \cdot \mathbf{r}\parallel}$ and integrating over $r\parallel$ gives

$$\left(\frac{\partial^2}{\partial z^2} - q_{\parallel}^2 \right) V_{\text{ad}}(\mathbf{q}_{\parallel}, z) = -\frac{1}{\epsilon_0} \delta\rho \quad (10.56)$$

The differential operator $\frac{\partial^2}{\partial z^2} - q_{\parallel}^2$ has the Green's function

$$-G(z, z') = \frac{\exp(-q_{\parallel}|z - z'|)}{2q_{\parallel}} \quad (10.57)$$

which means

$$V_{\text{ad}}(\mathbf{q}_{\parallel}, z) = -\frac{1}{\epsilon_0} \int_{-\infty}^{\infty} dz' G(z, z') \delta\rho \quad (10.58)$$

as can be proven by inspection [insertion of Eq. (10.58) into (10.56)]. Because, according to Eq. (6.10), $V_{\text{ad}} = V_{\text{tr}} - V_{\text{ap}}$ and because

$$\nabla^2 V_{\text{ap}} = -\frac{e}{\epsilon_0} \delta(r - r_0) \quad (10.59)$$

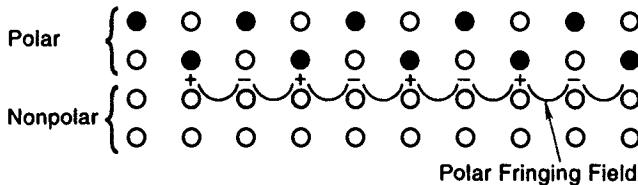


Figure 10.10 Polar-nonpolar material on top of each other. Notice that the phonons can create a long-range fringing field in the nonpolar material.

for a point charge located at $r_0 = (0, 0, z_0)$, one obtains

$$V_{\text{tr}}(\mathbf{q}_{\parallel}, 0) = -\frac{1}{\epsilon_0} \int_{-\infty}^{\infty} dz' G(0, z') \left(\delta\rho + \frac{e\delta(z' - z_0)}{A} \right) \quad (10.60)$$

This equation is a linear algebraic equation for $V_{\text{tr}}(\mathbf{q}_{\parallel}, 0)$ (which appears also in $\delta\rho$) and can easily be solved.

Defining the two-dimensional screening constant by

$$q_s^{\parallel} = \frac{e^2}{\epsilon_0} \int_0^{\infty} \left(-\frac{\partial f}{\partial E} \right) g(E) dE \quad (10.61)$$

one arrives at the matrix element for the scattering probability as outlined in Chapter 7, Eqs. (7.13)–(7.17):

$$|M_{\mathbf{k}\mathbf{k}'}| = \frac{e^4}{4(q_{\parallel} + q_s^{\parallel}/\epsilon)^2 (\epsilon\epsilon_0)^2 A^2} e^{-2q_{\parallel}|z_0|} \quad (10.62)$$

where A is the area of the two-dimensional electron gas, z_0 is the distance of the impurity from the electrons in z -direction, and $\mathbf{q}_{\parallel} = \mathbf{k}_{\parallel} - \mathbf{k}'_{\parallel}$. A corresponding relaxation time can again, as in Chapter 7, be derived explicitly only for limiting cases such as strong or weak screening (large or small q_s^{\parallel}).

The most important feature of the result is the exponent in Eq. (10.62), which tells us that the scattering matrix element decreases exponentially with the distance from the impurity. The concept of modulation doping makes use of this decrease: One dopes the higher energy gap layers such as AlGaAs neighboring to the GaAs. The electrons then leave their parent donors and accumulate at a distant place of lowest potential energy (GaAs). There the scattering is much reduced and, if phonon scattering is unimportant (low electron temperatures), enormous mobilities can be reached. In the GaAs-AlGaAs system mobilities well above $10^6 \text{ cm}^2/\text{Vs}$ have been achieved. Phonon scattering in quasi-two-dimensional systems has also been treated in detail and shows interesting features because of size quantization.

A key to understanding the existing literature is that, although the dimensionality of electrons is reduced by the presence of interfaces, the phonons may

(or may not) still propagate in three dimensions. One can therefore have scattering of two-dimensional electrons by three-dimensional phonons but also scattering of the electrons by the surface (fringing) fields of remote phonons, which can be understood from Figure 10.10. For more details the reader is referred to the review by Hess [2].

PROBLEMS

10.1 Derive the Green's function of Eq. (10.57) from its definition

$$\left(\frac{\partial^2}{\partial z^2} - q_{||}^2 \right) G(z, z') = \delta(z, z')$$

10.2 Derive Eq. (10.62) from Eq. (10.60) using the methods of Chapter 7.

10.3 Show the validity of Eq. (10.24) by balancing field and diffusion current.

10.4 Solve the expression for the built-in voltage (Eq. 10.15) in the presence of exponential doping distributions, that is, $N_D^+ = N_D^0 \exp(z/L)$ with L being a fixed length and the highest dopant occurring at $z = 0$.

REFERENCES

- [1] Dingle, R., Störmer, H. L., Gossard, H. C., and Wiegmann, W. "Electron mobilities in modulation-doped semiconductor superlattices," *Applied Physics Letters*, vol. 33, 1978, p. 665.
- [2] Hess, K. "High Field Transport in Semiconductors," in *Advances in Electronics and Electron Physics*, ed. P. W. Hawkes. New York: Academic, 1982.
- [3] Stern, F., and Das Sarma, S. "Electron energy levels in GaAs-GaAlAs heterojunctions," *Physical Review B*, vol. 30, 1984, p. 840.
- [4] Stern, F., and Howard, W. E. "Properties of semiconductor surface inversion layers in the electric quantum limit," *Physical Review*, vol. 163, 1967, p. 816.

CHAPTER 11

THE DEVICE EQUATIONS OF SHOCKLEY AND STRATTON

To calculate the electronic current in a semiconductor device, we need to solve the Boltzmann equation subject to (usually very complicated) boundary conditions. From the Boltzmann equation we can obtain n , Eq. (5.7), and the current density \mathbf{j} , Eq. (8.41), and all other quantities we may need in order to characterize a device. There is, however, a complication, which we have discussed already in Chapter 6. The potentials and therefore the electric fields depend on the charge density via the Poisson equation and have to be determined self-consistently. That is, we have to solve the Boltzmann equation coupled with the Poisson equation, Eq. (6.20) or even with Maxwell's equations.

This is, of course, very complicated; therefore simpler ways have been sought. It turns out that with very few approximations simpler formulas can be found for simple band structures. If these approximations are not good enough (generally they are not when electrons reach energies far above the band edges), the previously discussed Monte Carlo approach to solve the Boltzmann equation is more appropriate.

In this chapter we use in many instances the effective mass approximation $E = \hbar^2 k^2 / 2m^*$, and derive simpler equations that follow from the Boltzmann equation and can be solved easier and in combination with the equation of Poisson or even Maxwell's equations, at least by use of large computational resources.

11.1 THE METHOD OF MOMENTS

One way simpler equations than the Boltzmann equation can be found is called the *method of moments*. To illustrate this method (which is described in more detail in Appendix E), we multiply the Boltzmann equation by $1/(4\pi^3)$ and integrate it over all \mathbf{k} -space. This is done below for each term of the BTE:

1.

$$\frac{1}{4\pi^3} \int_{-\infty}^{\infty} \frac{\partial f}{\partial t} d\mathbf{k} = \frac{\partial}{\partial t} \frac{1}{4\pi^3} \int_{-\infty}^{\infty} f d\mathbf{k} = \frac{\partial n}{\partial t} \quad (11.1)$$

2.

$$\frac{1}{4\pi^3} \int_{-\infty}^{\infty} \mathbf{v} \cdot \nabla f d\mathbf{k} = -\frac{1}{e} \nabla \mathbf{j} \quad (11.2)$$

which follows from the definition of \mathbf{j} , Eq. (8.41).

3.

$$\frac{1}{4\pi^3} \int_{-\infty}^{\infty} e \mathbf{F} \cdot \nabla_{\mathbf{k}} f d\mathbf{k} = 0 \quad (11.3)$$

as can be found by using the fact that f_0 is an even and $\nabla_{\mathbf{k}} f$ is an odd function of \mathbf{k} . Inserting Eq. (8.16) into Eq. (11.3), one can also show that the integral over f_1 vanishes. (One also has to use partial integration to get rid of $\nabla_{\mathbf{k}}$, and the fact that f_0 vanishes exponentially as $|\mathbf{k}| \rightarrow \infty$.)

4. We have calculated the integral over the collision operator in Chapters 8 and 9. The integral over the f_1 terms vanishes as f_1 is odd. The rest of the scattering operator can be written as

$$\left. \frac{\partial f_0}{\partial t} \right|_{\text{coll}}^{\text{sc}} + \left. \frac{\partial f_0}{\partial t} \right|_{\text{coll}}^{\text{GR}}$$

where the subscript "sc" denotes scattering by imperfections (impurities, phonons, etc.) and GR denotes scattering owing to generation-recombination. For special cases, one can show [using, e.g., Eq. (8.21)] that the integral over the scattering term vanishes and only the generation-recombination terms, which we have calculated in Chapter 9, are left.

For electrons recombining via a trap (Shockley-Read-Hall) we have

$$\nabla \mathbf{j}_n - e \frac{\partial n}{\partial t} = e(R(n) - G(n)) \quad (11.4)$$

This is the well-known *equation of continuity* for electrons. Similarly one obtains for holes:

$$\nabla \mathbf{j}_p + e \frac{\partial p}{\partial t} = -e(R(p) - G(p)) \quad (11.5)$$

In steady state and for low frequencies one often uses

$$G(n) - R(n) = G(p) - R(p) = -U_s \quad (11.6)$$

The equations of continuity guarantee the correct counting of particles during their supply and removal. They are therefore of very general validity and follow also from quantum mechanical considerations. In fact, we have calculated $G(n)$ and $R(n)$ from the Golden Rule anyway. If in addition an appropriate approximation is found for the current density, then the equations of continuity will

also be valid from a quantum point of view. This approximation for the current density does not necessarily have to be a physical one. A valid polynomial expansion may do. This fact is used extensively in conventional device simulations as we will describe them in Chapter 12. These simulations solve the equations of continuity together with the equation of Poisson. We will refer to this system of equations as the Shockley equations because Shockley used them extensively in his celebrated treatment of *p-n* junctions and transistors. The Shockley equations form a complete and exact system for the description of semiconductor devices if the current densities are known as a function (or functional) of the electric field. In what follows, we will derive the current densities from the Boltzmann equation. To do this we also will need the charge carrier temperatures T_c . It is then necessary to derive the next higher moments of the Boltzmann equation and also to make an assumption of closure (the neglection of still higher moments by some ad hoc rule or assumption). We use here the system developed by R. Stratton. This can be obtained by multiplying the Boltzmann equation by \mathbf{k} and by the energy $E(\mathbf{k})$ and again integrating over all \mathbf{k} -space. We show this here first in a more general form for any generic function $Q(\mathbf{k})$. Using the definition for the average of $Q(\mathbf{k})$

$$\langle Q(\mathbf{k}) \rangle = \frac{\int_{-\infty}^{\infty} Q(\mathbf{k}) f(\mathbf{k}) d\mathbf{k}}{\int_{-\infty}^{\infty} f(\mathbf{k}) d\mathbf{k}} \quad (11.7)$$

one obtains a general equation that is valid for any scalar “moment” $Q(\mathbf{k})$:

$$\begin{aligned} \frac{\partial}{\partial t} (n \langle Q(\mathbf{k}) \rangle) + \frac{\hbar}{m^*} \nabla (n \langle Q(\mathbf{k}) \mathbf{k} \rangle) - \frac{e}{\hbar} F n \langle Q(\mathbf{k}) \frac{\nabla_{\mathbf{k}} f}{f} \rangle \\ = \frac{1}{4\pi^3} \int_{-\infty}^{\infty} d\mathbf{k} Q(\mathbf{k}) \left. \frac{\partial f_0}{\partial t} \right|_{\text{coll}} - \frac{1}{4\pi^3} \int_{-\infty}^{\infty} d\mathbf{k} \frac{f_1}{\tau_{\text{tot}}(\mathbf{k})} Q(\mathbf{k}) \end{aligned} \quad (11.8)$$

It can be shown by partial integration that the term $\langle Q(\mathbf{k}) \nabla_{\mathbf{k}} f / f \rangle$ is equal to $-\langle \nabla_{\mathbf{k}} Q(\mathbf{k}) \rangle$. On occasion, this second form is more convenient to use. For some purposes it is also more appropriate to rewrite Eqs. (11.7) and (11.8) with respect to the appearance of $\tau_{\text{tot}}(\mathbf{k})$. Because Eq. (11.8) is derived from the Boltzmann equation and subsequent integration, it is possible first to multiply the equation by $\tau_{\text{tot}}(\mathbf{k})$ and to integrate subsequently. Then $\tau_{\text{tot}}(\mathbf{k})$ will appear as a multiplicative factor for each $Q(\mathbf{k})$ except for the last term on the right-hand side, where it cancels out.

To illustrate the intricacies that are inevitable if one wants to derive device equations from Eq. (11.8), we discuss the case of $Q(\mathbf{k}) = k_z$. Consider, for the moment, a distribution function that does not explicitly depend on time. We can then rewrite Eq. (11.8) to read

$$\frac{\hbar}{m^*} \nabla (\langle \mathbf{k} k_z \tau_{\text{tot}}(\mathbf{k}) \rangle n) - \frac{e}{\hbar} F n \langle k_z \tau_{\text{tot}}(\mathbf{k}) \frac{\nabla_{\mathbf{k}} f}{f} \rangle = -\frac{1}{4\pi^3} \int_{-\infty}^{\infty} d\mathbf{k} k_z f_1 \quad (11.9)$$

The term containing $\partial f_0 / \partial t|_{\text{coll}}$ vanishes because f_0 is even in \mathbf{k} . To simplify Eq. (11.9) further it is convenient to neglect f_1 compared to f_0 in the second term

and in all the averages. Precise solutions of the Boltzmann equation show that this approximation is good for the scattering mechanisms, which are important in silicon (except in extremely high electric fields) but is basically invalid in GaAs because of the peculiarity of polar optical phonon scattering to prefer small angles and thus enhance the streaming terms (terms that are odd in \mathbf{k}).

The integrations (averages) can then be performed as in Eqs. (8.44) through (8.46), and the right-hand side of Eq. (11.9) is equal to the z component of the current density multiplied by a factor $m^*/(\hbar e)$.

One then obtains

$$\mathbf{j} = en\mu\mathbf{F} + e\nabla Dn \quad (11.10)$$

with

$$D = \mu k T_c / e \quad (11.11)$$

Here we have assumed that f_0 is a Maxwell-Boltzmann distribution at electron temperature T_c and f_1 is given by Eq. (8.18). We also have assumed that $\tau_{\text{tot}}(\mathbf{k})$ does not explicitly depend on the space coordinate. (It may depend on T_c , which may in turn depend on the space coordinate.)

Equation (11.11) is the Einstein relation. Notice, however, that the diffusion constant D is a function of T_c and therefore belongs after the gradient in Eq. (11.10), as T_c may be a function of the space coordinate. The second term in Eq. (11.10) is the well-known diffusion current. If the time derivative is not neglected in Eq. (11.8), a term $\bar{\tau} \partial \mathbf{j} / \partial t$ arises in Eq. (11.10) in addition to \mathbf{j} . $\bar{\tau}$ is a certain average of τ_{tot} which cannot be determined within our formalism because we do not know the precise form of f . Only if τ_{tot} is independent of \mathbf{k} do we have $\bar{\tau} = \tau_{\text{tot}}$.

Equation (11.10) can be used in the Shockley equations and would complete the system if we knew T_c . However, T_c can only be obtained from the next higher moment using the definition $3k_B T_c = 2\langle E \rangle$. This next higher moment is fairly complicated to calculate particularly when the electric field varies rapidly in space as it usually does in devices. The results for this case are derived in Appendix E. Here, however, we proceed with a simple physical description, first for constant electric fields, and subsequently giving a glimpse of what happens when the electric field varies in space (velocity overshoot). Any more precise treatment can only be achieved by detailed computer simulations as they are described in Chapter 12.

11.2 MOMENT FOR THE AVERAGE ENERGY AND HOT ELECTRONS

Here we discuss an equation for T_c for electric fields that vary very slowly with the space coordinate. We derive the current density as a function of T_c first for the steady state and in the following section including time dependence.

11.2.1 Steady-State Considerations

Neglecting the spatial gradients, we obtain from Appendix E

$$\frac{\partial \langle E \rangle}{\partial t} = \frac{1}{n} \mathbf{F} \cdot \mathbf{j} - \left. \frac{\partial \langle E \rangle}{\partial t} \right|_{\text{coll}} \quad (11.12)$$

where

$$\left. \frac{\partial \langle E \rangle}{\partial t} \right|_{\text{coll}} \equiv \frac{1}{8\pi^3} \int d\mathbf{k} E \left. \frac{\partial f_0}{\partial t} \right|_{\text{coll}} \quad (11.13)$$

The term $\partial f_0 / \partial t|_{\text{coll}}$ gives an important contribution. Without this term the average energy diverges with time, because

$$\langle E \rangle = \frac{t}{n} \mathbf{F} \cdot \mathbf{j} + \text{const} \quad (11.14)$$

It is thus the term $\partial \langle E \rangle / \partial t|_{\text{coll}}$ that keeps the energy low, and it therefore represents the heat (Joule's heat) given to the crystal lattice.

One can prove in general (for Boltzmann statistics) that for small $|\mathbf{F}|$

$$\left. \frac{\partial \langle E \rangle}{\partial t} \right|_{\text{coll}} = \frac{\langle E \rangle - \frac{3}{2} k T_L}{\tau_E} \quad (11.15)$$

where τ_E is a time constant called energy relaxation time and T_L is the temperature of the crystal lattice. Often one defines an electron temperature independent of the actual form of the nonequilibrium distribution function by

$$\langle E \rangle = \frac{3}{2} k T_c \quad (11.16)$$

Then

$$\left. \frac{\partial \frac{3}{2} k T_c}{\partial t} \right|_{\text{coll}} = \frac{3}{2} k (T_c - T_L) / \tau_E \quad (11.17)$$

This gives

$$\frac{\partial T_c}{\partial t} = \frac{2}{3k} \frac{1}{n} \mathbf{F} \cdot \mathbf{j} - (T_c - T_L) / \tau_E \quad (11.18)$$

from which one can see that the temperature of the electron gas is always raised when a finite electric field is applied. If $T_c \gg T_L$, one calls the electrons "hot electrons."

For silicon at room temperature ($T_L = 300\text{K}$), the energy relaxation time is almost entirely due to the interaction with optical phonons and can roughly be approximated by

$$\tau_E \approx 4 \cdot 10^{-12} \sqrt{\frac{T_c}{T_L}} \text{ sec} \quad (11.19)$$

The general calculation of $\partial\langle E \rangle / \partial t|_{\text{coll}}$ including all inelastic scattering mechanisms is usually connected with tedious algebra and has been explained in detail by Conwell [1].

For the special case of optical deformation-potential scattering, the calculation is straightforward, at least in principle: For any absorption event, the energy $\hbar\omega_0$ is gained; for emission the same energy is lost. The net energy gain or loss per unit time $\partial E / \partial t|_{\text{coll}}$ is therefore given by the difference of absorption and emission rates. Using Eq. (7.35) one obtains

$$\frac{\partial E}{\partial t} \Big|_{\text{coll}} = \frac{Z_0^2 m^{*3/2} \omega_0^2}{\sqrt{2\pi\hbar^2\rho v_s^2}} \left[N_q \sqrt{E + \hbar\omega_0} - (N_q + 1) \sqrt{E - \hbar\omega_0} \right] \quad (11.20)$$

To obtain $\partial\langle E \rangle / \partial t|_{\text{coll}}$ we have to average Eq. (11.20) using Eq. (11.7) and recognize that the integral over f vanishes. The integration involves modified Bessel functions and the result will not be given here (see Conwell [1]). For large energies, however, we can neglect $\hbar\omega_0$ in the square roots of Eq. (11.20) and obtain

$$\frac{\partial E}{\partial t} \Big|_{\text{coll}} = \frac{Z_0^2 m^{*3/2} \omega_0^2}{\sqrt{2\pi\hbar^2\rho v_s^2}} \sqrt{E} \quad (11.21)$$

which gives

$$\frac{\partial\langle E \rangle}{\partial t} \Big|_{\text{coll}} = \frac{Z_0^2 m^{*3/2} \omega_0^2 \sqrt{2}}{\pi \sqrt{\pi\hbar^2\rho v_s^2}} \sqrt{kT_c} \quad (11.22)$$

Comparing this equation with Eq. (11.17) for large T_c (and neglecting T_L versus T_c), we can obtain the energy relaxation time τ_E . It is seen that τ_E still depends (although weakly) on the average energy and is therefore not a relaxation time in the strict sense. For T_c approaching T_L , however, τ_E approaches a constant, as can be seen from the exact integration of Eq. (11.20) (Conwell [1]).

This derivation of $\partial\langle E \rangle / \partial t|_{\text{coll}}$ is a more intuitive one. Actually, the definition Eq. (11.13) involves the complicated collision operator $\partial f_0 / \partial t|_{\text{coll}}$. One can derive the energy loss also from this definition and get the same result, Eq. (11.22), if the same approximations ($E \gg \hbar\omega$) are made.

To illustrate some implications of these equations, we focus our attention on the current density in a homogeneous semiconductor. From Eqs. (11.9), (11.10), (11.18), and (11.19), it is clear that the current density is a nonlinear function of the applied voltage (electric field), because the mobility μ is a function of T_c (we neglect diffusion in this discussion). Using Eqs. (11.18) and (11.19) and putting $\partial T_c / \partial t = 0$ (steady state), we have

$$e\mu F^2 \approx \frac{C}{(300\text{K})} (T_c - T_L) \sqrt{\frac{T_L}{T_c}} \quad (11.23)$$

where $C \approx 10^{-9}$ W (the approximation holds for silicon at room temperature).

The mobility μ can often be approximated by

$$\mu = \mu_0 \sqrt{\frac{T_L}{T_c}}$$

as has been derived for scattering by acoustical phonons (also by optical phonons at high energy) in Eqs. (8.52) and (8.53). Therefore we obtain

$$T_c = T_L + e\mu_0 F^2 \cdot \frac{(300\text{K})}{C} \quad (11.24)$$

Inserting this value of T_c into the equation for the mobility and using $\mathbf{j} = en\mu_0\mathbf{F}$, we have

$$\mathbf{j} = en\mu_0\mathbf{F} \sqrt{\frac{1}{1 + e\mu_0 F^2 \left(\frac{300\text{K}}{T_L C} \right)}} \quad (11.25)$$

As can be seen, \mathbf{j} follows Ohm's law for small \mathbf{F} because the term proportional to F^2 can be neglected. For large \mathbf{F} , however, \mathbf{j} is constant (it saturates), because then the term proportional to F^2 dominates and F cancels, giving

$$j_{\text{sat}} = \frac{en\mu_0}{\sqrt{e\mu_0 \left(\frac{300\text{K}}{T_L C} \right)}} \quad (11.26)$$

Dividing j_{sat} by en , we obtain the saturation velocity v_{sat} of the electrons.

The physical explanation of this phenomenon is, of course, that electrons are scattered more frequently as they are accelerated to higher velocities and therefore the mobility decreases. The fact that the mobility decreases proportional to just $1/F$ at high fields F is more general than suggested by the above derivation. The term $\sqrt{T_c}$ in the above expressions arises from the density of states $g(E)$, which is proportional to \sqrt{E} . In general, $g(E)$ increases much more steeply than \sqrt{E} at higher energies in the bands (see Figure 5.2). Even for steep increases in the density of states, the proportionality of the mobility to $1/F$ holds as is shown below.

Assuming that $g(E) \propto E^P$, and assuming $E \gg \hbar\omega_0$ and $T_c \gg T_L$, one obtains [see Eqs. (11.21) and (11.22)]

$$e\mu F^2 \propto \left(\frac{T_c}{T_L} \right) \quad (11.27)$$

This is because the scattering rates are roughly proportional to the final density of states, and \sqrt{E} in Eq. (11.21) becomes E^P . Note, however, that this is only exact for the deformation potential interaction with optical phonons, which are the most important scattering agents for the energy loss in silicon. For the same reasons one obtains approximately

$$\mu \propto \left(\frac{T_c}{T_L} \right)^{-P} \quad (11.28)$$

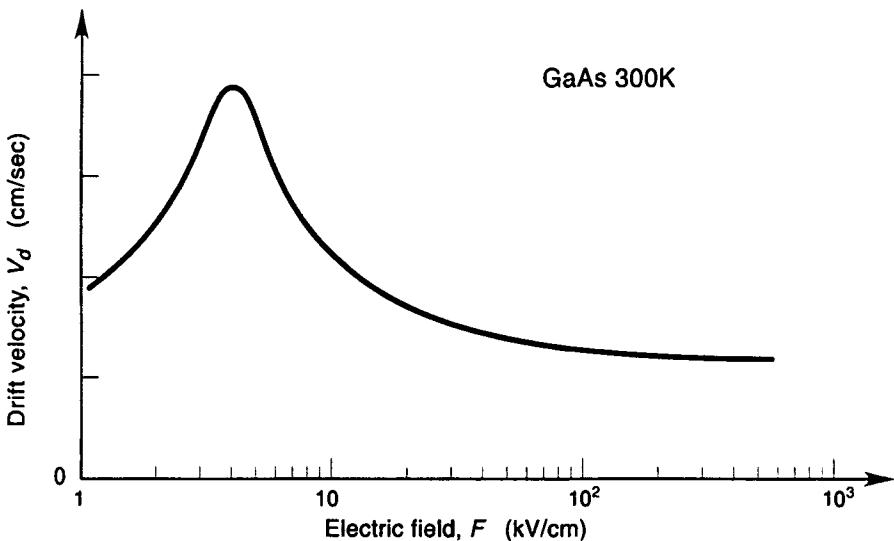


Figure 11.1 Calculated electron drift velocity in GaAs at room temperature. The calculations have been performed by Shichijo and Hess [4].

and thus using Eq. (11.27)

$$\frac{T_c}{T_L} \propto F^{1/p} \quad (11.29)$$

Inserting Eq. (11.29) into Eq. (11.28), we have

$$\mu \propto 1/F \quad (11.30)$$

It has been tacitly assumed in this derivation that $E \gg \hbar\omega_0$. This is only the case if the electrons are very hot (i.e., T_c is very high), which implies high electric fields. We therefore have shown that $\mu \propto 1/F$ (which is equivalent to velocity saturation) under rather general conditions.

The electron velocity saturates at high fields in silicon, germanium, and many other semiconductors. However, some III-V compounds, having the conduction band minimum at Γ , are special in certain respects. At Γ , polar optical scattering is predominant and the effective mass is small. This means that the mobility for small electron energies and low electric fields will be high. As the electron energy increases with increasing field, the L minima (Germanium-type) and the X minima (Silicon-type) are also populated by intervalley scattering. Electrons in L and X valleys experience deformation-potential scattering and have a large mass. Thus, their mobility is much smaller compared to the Γ valley. This leads to a drop in the average velocity of the carriers above a certain field, as shown in Figure 11.1. At very high fields the velocity saturates approximately. This effect is the basis for the Gunn effect, which is discussed in Chapter 13.

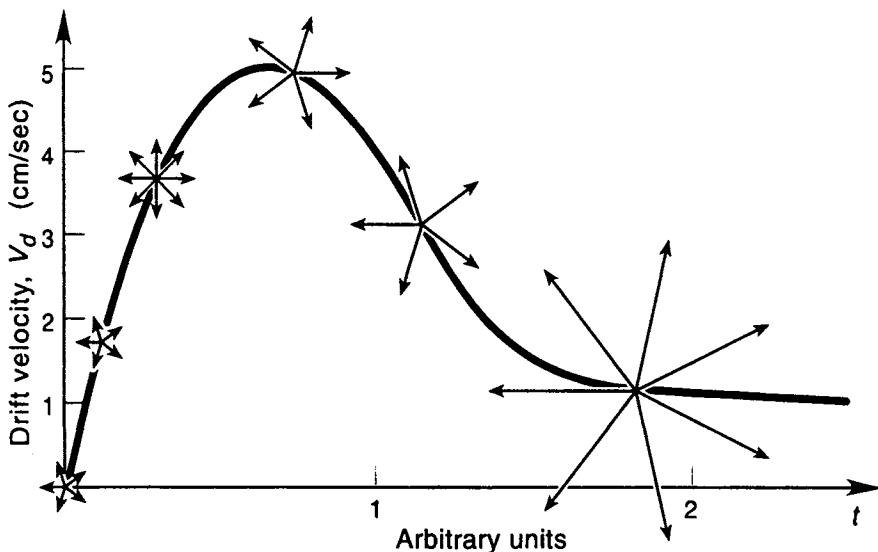


Figure 11.2 Velocity overshoot in semiconductors. (The time scale is typically in picoseconds, the velocity scale typically in 10^7 cm/s.)

11.2.2 Velocity Transients and Overshoot

In the above derivations it has been assumed that $\partial T_c / \partial t = 0$. If this is not the case, the situation becomes more complicated, as another differential equation needs to be solved, Eq. (11.18), to obtain the electric current. A very important effect connected to the time development of the electron temperature is *velocity overshoot*. To understand this effect, assume that all electrons start at $\mathbf{k} = 0$ at time $t = 0$ (at which we apply a high electric field). The electrons are then accelerated and for a short time period do not scatter ($T_c = 0$). Because all electrons go in the forward direction, very high average velocities can be reached. As time goes on scattering events occur, and the electron energy is partially randomized, which also means that the average velocity decreases.

As shown by Ruch [3], this leads to the velocity-versus-time curve shown in Figure 11.2. In some instances (e.g., electrons in GaAs) this effect is very much enhanced because electrons start in the Γ valley and are accelerated beyond the L and X valleys. Then (after, typically, 10^{-13} s or less) they scatter into the L and X valleys, where they have a large mass. This effect decreases the velocity further (as it does for the steady state, Figure 11.1) and therefore leads to a significant overshoot effect, as shown in Figure 11.2.

The physical mechanism leading to the overshoot is also illustrated in Figure 11.2. In this figure a distinction is made between the drift velocity and the instantaneous random velocity of the electron gas. The drift velocity is the average over all random velocities of the electrons that point in forward and backward direction. The random velocity is indicated by the arrows plotted at a number of

points. At $t = 0$, the electrons are in equilibrium and the drift velocity is zero. After a short time electrons are accelerated in the forward direction and very few scattering events occur. The drift velocity therefore increases while the random velocity still stays small. As time goes on the electrons are scattered and their forward velocity is randomized, leading to a large random velocity and to a smaller saturated drift velocity. Typical values in this range are 10^8 cm/s for the random velocity and 10^7 cm/s for the drift velocity. Velocities much higher than 10^8 cm/s do not usually occur in semiconductors because of Bragg refraction (the mass becomes negative).

Note that Eq. (11.18) is valid only if ∇f (or ∇n) can be neglected. In the presence of spatial gradients additional terms need to be added (see Appendix E). These terms need to be included when overshoot phenomena are possible, because overshoot implies automatically space-dependent velocities. Electrons start with small drift velocities in the contact areas and once accelerated overshoot their steady-state velocity and finally approach steady state typically after a distance of the order of 500 to 5000 Å.

It is also important to realize that overshoot phenomena can and do happen under steady-state conditions if the electric field varies rapidly. Consider, for example, a step in the electric field. As the electrons (holes) encounter the step, the situation is very similar to the case when the electric field is just switched on. The step thus corresponds to $t = 0$ in Figure 11.2. The electrons now faced with a higher electric field will overshoot before returning to the steady-state velocity that corresponds to the new higher electric field. (Similarly, if the step is toward a lower electric field an undershoot will occur). This effect can increase the average velocity in a transistor even for dc electric fields. It was suggested that this effect may have beneficial effects for transistor performance. Indeed there is evidence for such effects but the benefits to transistor speed are often rather small. For more details, the reader is referred to the review by Hess and Iafrate [2].

11.2.3 Equation of Poisson and Carrier Velocity

The patterns of space-dependent carrier distribution that can be connected to the dependence of the drift velocity on the space coordinate can be illustrated by the example of the simplest possible semiconductor device, an n^+nn^+ junction (semiconductor sandwiched between semiconductor contacts, the heavily doped n^+ , or p^+ for holes, regions).

If the n region of the semiconductor is very long, then the equations for the current that have been developed above for a constant carrier concentration apply. For very short lengths of the n region, a new effect becomes important. Electrons can be pulled out of the contacts, and the electron concentration in the semiconductor increases independent of the doping of the semiconductor. The current associated with this effect is called the *space-charge limited current*, because the electron concentration is in excess of the concentration of positively

charged donors. The electric field is then given by the Poisson equation [see Eq. (6.20)]:

$$\frac{\partial F}{\partial x} = \frac{\delta p}{\epsilon \epsilon_0} \quad (11.31)$$

Here δp is the excess charge that can be approximated by the total free carrier concentration n if the donor (acceptor) charge is small. If the electric field is high and diffusion currents are also neglected, the current density is $j = en\mu F$ and n becomes

$$n = \frac{j}{e\mu F} \quad (11.32)$$

Below we discuss three separate cases for the solution of Eqs. (11.31) and (11.32).

First we assume that the mobility is constant $\mu = \mu_0$. Then

$$\frac{\partial F}{\partial x} = \frac{j}{\epsilon \epsilon_0 \mu_0 F} \quad (11.33)$$

and

$$F^2 = \frac{2jx}{\epsilon \epsilon_0 \mu_0} \quad (11.34)$$

Here we have also assumed that $j = \text{constant}$ (steady state) and that $F = 0$ at $x = 0$. This boundary condition is appropriate if we place the end of the n^+ contact at $x = 0$. The field is necessarily small in the contact region because the resistance of this region and, therefore, the external voltage drop over this region are small.

Assuming a length L for the n region and using

$$-\int_0^L F dx = V_{\text{ext}} \quad (11.35)$$

where V_{ext} is the external voltage, and using Eq. (11.34), one obtains the Mott-Gurney law:

$$j = \frac{9 \epsilon \epsilon_0 \mu_0}{8 L^3} V_{\text{ext}}^2 \quad (11.36)$$

As discussed at the start of this calculation, the carrier concentration is space dependent. This can be seen from Eqs. (11.36), (11.34), and (11.32). Furthermore, a current will flow independently of the doping of the semiconductor. Even if there are no carriers at all in the semiconductor, for $V_{\text{ext}} = 0$ there will be a current for $V_{\text{ext}} \neq 0$, the reason, of course, being that electrons are supplied by the contacts. This is a strange result because it seems to contradict the existence of thin insulating films in nature, at least if they are sandwiched between contacts. However, close to a conducting contact made out of the same material, and differing only in doping, an insulator does not exist. Even if we do not apply an external voltage, electrons will spill over to the lesser doped semiconductor

from the highly doped contacts. The spilling length is equal to the Debye length, which was derived in Chapter 6. The Debye length decreases with increasing carrier concentration. This does not mean, however, that a contact with infinite carrier concentration will not spill electrons into the neighboring material. On the contrary, it will spill the most. The Debye length represents in this case the penetration length for any given carrier concentration.

Second, we assume that the velocity is saturated, that is, $\mu F = \mu_{\text{sat}}$. Equation (11.33) then becomes

$$\frac{\partial F}{\partial x} = \frac{j}{\epsilon \epsilon_0 v_{\text{sat}}} \quad (11.37)$$

Integrating this equation (the right-hand side is constant) and using Eq. (11.36), one obtains for the current density j

$$j = -2\epsilon \epsilon_0 v_{\text{sat}} V_{\text{ext}} / L^2 \quad (11.38)$$

It is important to note that in this case the current obeys Ohm's law. However, the velocity of the electrons is saturated and the increase of j with V_{ext} arises from an increase in charge density.

If μF does not follow a simple law as assumed for the derivations above, a numerical integration is necessary to obtain the current density j . Especially when velocity overshoot is involved, the problem can be quite complicated and a precise solution can only be achieved by a many-particle Monte Carlo calculation.

There is still one more important case that can be solved explicitly. In the case of negligible energy loss of the electrons, the velocity can be calculated from the kinetic energy and the potential energy $eV(x)$

$$m^* v^2 / 2 = eV(x) \quad (11.39)$$

and we have

$$\mu F = v = \sqrt{2eV(x)/m^*} \quad (11.40)$$

In this case we obtain for the current density

$$j = \frac{4}{9} \left(\frac{2e}{m^*} \right)^{1/2} \frac{\epsilon_0 \epsilon}{L^2} V_{\text{ext}}^{3/2} \quad (11.41)$$

which is known as Child's law and was known for electrons emitted into a vacuum as early as about 1900.

Equations (11.31) through (11.41), along with Eqs. (11.23) through (11.30), demonstrate how complicated the simplest semiconductor "device" (a piece of semiconductor between two ohmic contacts) is and prepares us for further complications in the next chapters on $p-n$ junction diodes and transistors. Before dealing with these, however, we give a description of general ways to solve the device equations numerically.

PROBLEMS

- 11.1 Prove Eq. (11.3).
- 11.2 Use Eq. (8.21) to show that $\int d^3k \frac{\partial f_0}{\partial t} \Big|_{\text{coll}}^{\text{ph}}$ vanishes. Be sure to state explicitly any assumptions you make.
- 11.3 Derive the term corresponding to $\nabla(n(Q(\mathbf{k})\mathbf{k}))$ in Eq. (11.8) for $Q(\mathbf{k}) = E$.
- 11.4 Assume that the n -type region of an n^+nn^+ diode is doped with N_D donors. Calculate the range of V_{ext} for which the current is space charge limited [i.e., follow Eqs. (11.36), (11.38), and (11.41)] and determine for which range the current is given by $j = \mu V_{\text{ext}}/L$. Distinguish carefully between the different laws for μF .

REFERENCES

- [1] Conwell, E. M. "High Field Transport in Semiconductors," in *Solid State Physics*, ed. F. Seitz, D. Turnbull, and H. Ehrenreich. New York: Academic, 1967 pp. 155–160.
- [2] Hess, K., and Iafrate, G. J. "Theory of applications of near ballistic transport in semiconductors," *Proceedings of the IEEE*, vol. , 1988, pp. 519–532.
- [3] Ruch, J. G. "Electron dynamics in short channel field-effect transistors," *IEEE Transactions on Electron Devices*, vol. ED-19, 1972, pp. 652–659.
- [4] Shichijo, H., and Hess, K. "Band structure dependant transport and impact ionization in GaAs," *Physical Review*, vol. B 23, 1981, pp. 4197–4207.

CHAPTER 12

NUMERICAL DEVICE SIMULATIONS

12.1 GENERAL CONSIDERATIONS

It is clear from the previous chapters, that the simultaneous solution of the equations of Poisson and continuity for the electron and hole current densities requires elaborate numerical methods, particularly for realistic structures. We have also seen that, although essential parts of the theory of semiconductor devices can be understood analytically, the multitude of nonlinear transport effects make special numerical calculations (the solution of the Boltzmann equation including the band structure) necessary if quantitative understanding is to be achieved. In addition there are important devices such as laser diodes and high-speed transistors that operate under circumstances requiring the simultaneous solution of Maxwell's equations in addition to the Boltzmann equation (or the equations of continuity); in other words, the equation of Poisson is not sufficient. Whenever such a combination of powerful but complicated equations is necessary, numerical simulation is the only path to quantitative answers; this is particularly true for the usually complicated geometry of semiconductor devices.

Large-scale computer simulations have become a cost-effective and predictive tool of device design and related computer-aided engineering questions (Technology Computer Aided Design or TCAD). Our treatment of principles was aimed toward highlighting the most precise treatments and the various conditions for excellent approximations, and we will give a short summary in this section. Table 12.1 shows the "ultimate" system of equations, which of course, already contains numerous approximations. For example, the Boltzmann equation is semiclassical and does not account for the uncertainty principle or tunneling. Note, however, that tunneling can be included as a scattering mechanism by using the Bardeen Transfer Hamiltonian formalism from Appendix A. The Uncertainty Principle can also be included in an approximate fashion by considering collision broadening (as shown in Chapter 8). Thus this set of equations is extremely complete and it is difficult to think of a situation in which it is completely invalidated. For example, even if all dimensions of the device are

Table 12.1 Optimum Set of Device Equations

Maxwell equations	Including space-dependent complex dielectric constant
Boltzmann equation	Including band structure, scattering mechanisms and Pauli principle; can be replaced by Monte Carlo simulations (which are totally equivalent)
Schrödinger equation	Used indirectly to determine properties of electronic states, band structure, effects of size quantization (quantum wells) and of wave functions for scattering matrix elements

very small, as they will be in the metal–oxide–silicon (MOS) transistors of the future, the Boltzmann equation may still be applicable at least for reasonably high supply voltages. Remember that the scattering rate goes to 10^{14} s^{-1} for energies higher than 1 eV above the silicon conduction band edge. The mean free path drops then to about 10^{-6} cm , which is below the size of MOS transistors that can currently be envisioned. Transport for large transistors (larger than 10^{-6} cm channel length) would then still be semiclassical. Even for nanostructures, which show quantized resistance, the Boltzmann equation can still be applicable as shown in Chapter 8. Similarly, the Maxwell equations are fully valid, as long as one does not consider single photon events or the like.

Unfortunately this system is beyond what can be solved by current computers. Laser diodes and high-speed transistors do require such a system, and certain approximations of it have indeed been shown to be soluble [3] by use of large-scale computation, at least in two dimensions of real space. In general, it is necessary to solve greatly simplified systems.

Such a simplified set of device equations is shown in Table 12.2. This set represents intermediate applicability because the Maxwell equations are replaced by Poisson's equation. This still automatically includes the displacement current $\partial \epsilon \epsilon_0 F / \partial t$ in all space covered by the solution of the set of differential equations. The Boltzmann equation (Monte Carlo) is replaced by a set of moment equations, including energy balance (Appendix E), and the Schrödinger equation is used only conceptually. For example, one could model the electrons at the interface or in a quantum well as a two-dimensional electron gas. This set of equations is not only of intermediate flexibility but also of intermediate status with respect to computer time consumption.

It should be noted that the numerical methods that can be applied to this set of equations converge only if the electron energy increases owing to the electric field are having an upper bound. It is “physical” to set as the upper bound of this energy the energy of steep increase in the density of states (steep increase in scattering). For example, in GaAs, this could be (for practical transistor applications) the energy of the X valleys. The intermediate set can include velocity overshoot and real space transfer, but is often seen as numerically too laborious, particularly when compared with the next set.

Because of this intermediate status, the description of this approach has been

Table 12.2 Intermediate Set of Device Equations

Poisson's equation	(Displacement current included automatically)
Equations of continuity	Eq. (11.4)
Equation for current	Eqs. (11.9) and (11.10), plus term $\tau \partial j / \partial t$, as described in context with Eq. (11.10)
Energy balance equation	Eqs. (11.12) and (11.18) and Appendix E

Choice of constants (μ , D , etc.) in accordance with scattering theory (perturbational solution of the Schrödinger equation). One band approximation [Eqs. (3.25) and (3.26)]. Size quantization effects are included phenomenologically, for example, by declaring the electrons in a quantum well two dimensional.

Table 12.3 Shockley Set of Device Equations

Poisson's equation	(Displacement current included automatically)
Equation of continuity	Eq. (11.4)
Equation for current and μ and D chosen as functions of the electric field, as in Eq. (11.25)	Eq. (11.9), (11.10)
Effective mass approximation	Eq. (3.31)

cut short here and is mostly relegated to Appendix E. The interested reader is referred to the literature on energy transport and hydrodynamic equations [7]. The energy transport approach is in essence identical to the approach of Stratton given in Appendix E.

Table 12.3 shows the simplest device equations that still contain a maximum of physics. It is these equations that are most frequently used and whose numerical solutions are most developed. These equations have been solved using finite difference methods [9], which will be described below. The corresponding numerical programs are used in various computer-aided engineering tools for device design. Examples are the MINIMOS, DESSIS, and the PISCES codes (search the Internet to find details). These codes are very effective in modeling silicon devices with significant accuracy. They can deal with almost arbitrary geometries (two-dimensional, three-dimensional) and are reasonably accurate, with some corrections even much below micrometer feature sizes. The reason for the great effectiveness is threefold: For one, the equations of continuity (zero moment equations) are extremely general. The current density equations are just approximated by phenomenological expressions. As long as these expressions are reasonable, the results are essentially exact and no higher moments are necessary. The only problems that arise are connected with the effective mass approximation. Effects that occur high in the bands such as impact ionization and hot electron degradation need to be treated separately, for example, by full-band Monte Carlo [10] approaches. Numerical approaches that accomplish this in devices exist (e.g., IBM's DAMOCLES). A second problem is connected with velocity overshoot and real space transfer. The current density then becomes a

functional (not merely a function) of the electric field, and approximations of the current density then become challenging. As a minimum, one would need to approximate the current density as a function of electric field plus its gradient, as often done under such circumstance. Third, the equation of Poisson becomes insufficient at very high speeds of device operation. However, this is currently only of concern for high-speed transistors such as the MODFET's or resonant tunneling diodes in high-speed circuits.

Because of the great importance (and still great generality) of this system, we describe below the numerical solution of the Shockley set in detail. All considerations are presented in one dimension for simplicity, but using a formalism that can be generalized to two and three dimensions with ease and indeed has been, as described by Selberherr [9].

12.2 NUMERICAL SOLUTION OF THE SHOCKLEY EQUATIONS

We wish to solve the equations of continuity and Poisson numerically, assuming a certain relation between current, electric field, and carrier concentrations. This assumption closes the infinite system of moment equations as described in Chapter 11. The results of our calculation are then as valid as this assumption is. Remember the equation of continuity is quite exact, even from a quantum point of view, if only the current densities are expressed correctly. Researchers in device simulation have experimented quite a bit in deriving and using expressions for the current density. In fact these phenomenological derivations are often a considerable part of device engineering and one could write a whole chapter on this topic alone. Commercial device simulators contain descriptions of these approximations. The expressions are often only valid in limited parameter ranges and need to be changed when the device structure changes greatly. We assume here that

$$j_n = en\mu_n(\mathbf{F})\mathbf{F} + e \frac{\partial D_n(\mathbf{F})n}{\partial x} \quad (12.1)$$

with $\mu_n(\mathbf{F})$ and $D_n(\mathbf{F})$ being arbitrary functions of the electric field \mathbf{F} that make Eq. (12.1) correct in the device parameter range of interest.

As we know from the previous chapters, the mobility and diffusion constant indeed are functions of the field, if \mathbf{F} does not vary too rapidly which causes, for example, overshoot phenomena. In this latter case, μ_n and D_n become functionals of the electric field; that is, they depend in a nonlocal way on the electric field. One then can approximate μ_n and D_n as functions of the electric field \mathbf{F} and its derivative $\nabla\mathbf{F}$ as mentioned before. However, in more than one dimension there are complications because the delayed heating of the electrons that causes the overshoot is due to the field in the direction of the current density and the derivative therefore needs to be taken in this direction. In addition, the electron heating leads to real space transfer effects (the electrons spill out of potential wells), which are difficult to describe by simple approximations.

Here we assume, however, that Eq. (12.1) is correct and we have a corresponding equation for the holes:

$$j_p = ep\mu_p(\mathbf{F})\mathbf{F} - e \frac{\partial D_p(\mathbf{F})p}{\partial x} \quad (12.2)$$

Then the Shockley equations form a closed system that describes the device. We repeat them here in one dimension again using quasi-steady-state assumptions; we neglect as already done before the term $\tau \partial j / \partial t$ and put $R(n) - G(n) = U_s$. Therefore:

$$\frac{\partial j_n}{\partial x} - e \frac{\partial n}{\partial t} = eU_s \quad (12.3)$$

and

$$\frac{\partial j_p}{\partial x} + e \frac{\partial p}{\partial t} = -eU_s \quad (12.4)$$

and the Poisson equation reads:

$$\frac{\partial F}{\partial x} = e \frac{p - n + N_D^+ - N_A^-}{\epsilon \epsilon_0} \quad (12.5)$$

These five equations now need to be solved subject to the actual boundary conditions. Naturally, for any realistic conditions and functions this can be done only numerically. To accomplish this, one discretizes space (i.e., one chooses a set of points on the x -coordinate and rewrites the differential equations in finite difference form). If one wants to accomplish solutions in more than one dimension one needs to generate a grid of points over the device, and these mesh points typically have irregular spacing. They must be chosen more densely where the unknowns vary rapidly and less densely in regions where nothing much happens. As we will see, the number of finite difference equations to solve is proportional to the number of mesh points and therefore, one wants to avoid unnecessary close spacing. The finite difference expressions for the differential equations on such a nonuniform mesh needs to be derived with great care and consistent error truncation. This is then a much more complicated procedure (particularly in more than one dimension) than to use an equal mesh spacing, but is necessary from an engineering point of view to produce efficient simulation tools. A detailed treatment of this has been given by Selberherr [9]. Here we proceed in one dimension and with an equal mesh point spacing and follow the pioneering work of Scharfetter and Gummel [8].

Assume then a mesh of equally spaced points numbered by integers m such that every point on the x -axis is given by $m \Delta x$. We ignore for simplicity the fact that there are boundary conditions for certain values of m . These can be easily taken care of and we derive here only the bulk of the finite difference equations. We thus obtain the finite difference equations of continuity:

$$\frac{j_n(m+1) - j_n(m-1)}{2 \Delta x} - e \frac{\partial n(m)}{\partial t} = eU_s(m) \quad (12.6)$$

where we have taken the derivative involving three neighboring meshpoints. This little trick gives a better approximation to the space derivatives. For the holes we have

$$\frac{j_p(m+1) - j_p(m-1)}{2\Delta x} + e \frac{\partial p(m)}{\partial t} = -eU_s(m) \quad (12.7)$$

and for the electric field

$$\frac{F(m+1) - F(m-1)}{2\Delta x} = e \frac{p(m) - n(m) + N_D^+(m) - N_A^-(m)}{\epsilon\epsilon_0} \quad (12.8)$$

Disregarding for the moment the complications of time dependencies, one is faced with the following facts. We obtain a system of equations for the unknowns $n(m)$, $p(m)$ and $F(m)$. Notice that the total number of unknowns is therefore three times the number of mesh points. To proceed with the solution we still need to express the current densities in term of the unknowns. This is a nontrivial task because the carrier concentrations depend (as we will see) essentially exponentially on the electric field. It was therefore proposed by Scharfetter and Gummel [8] to regard Eqs. (12.1) and (12.2) as differential equations and solve them over a mesh section by assuming that μ_n , μ_p , F , j_n , and j_p are all constant just there. This turns out to be an excellent procedure that guarantees numerical stability and, as shown in [9], also precision. In fact, no reasonable and numerically stable scheme can be developed without this procedure; it has therefore been named *Scharfetter-Gummel discretization*.

The integration between the mesh point m and $m+1$ is a bit laborious, although one solves only a differential equation of first order with constant (between the mesh points) coefficients (see problems). One obtains

$$j_n(m) = eF(m) \left[\frac{n(m+1)\mu_n(m)}{1 + \exp(-A)} + \frac{n(m)\mu_n(m)}{1 + \exp(A)} \right] \quad (12.9)$$

with

$$A = \frac{\mu_n(m)F(m)\Delta x}{D_n(m)} \quad (12.10)$$

and a similar equation for the holes with $p(m)$ instead of $n(m+1)$ and $p(m+1)$ instead of $n(m)$, respectively. This completes our system of equations which is, however, nonlinear. Thus our progress so far is to have transformed the differential equations in a large number of nonlinear algebraic equations. The next step is to solve these equations. To do this we first deal with the time dependence and rewrite the equations in a shorthand type of way. Thus Eq. (12.8) reads

$$0 = \mathbf{F}_1(F(m), n(m), p(m)) \quad (12.11)$$

Of course, also indices $m+1, m+2$ appear in the argument of the unknowns. Similarly we write for the equations of continuity

$$e \frac{\partial n(m)}{\partial t} = \mathbf{F}_2(F(m), n(m), p(m)) \quad (12.12)$$

and

$$e \frac{\partial p(m)}{\partial t} = \mathbf{F}_3(F(m), n(m), p(m)) \quad (12.13)$$

where $\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3$ are functions that can be easily derived from the original discretized equations. The so-called direct forward solution of the time dependence can now be obtained by discretization in time using $e \partial n / \partial t = (n(l+1, m) - n(l-1, m)) / (2 \Delta t)$. Notice that we have now added the index l (which is also to be added on the right-hand side of the above equations) because we have the additional dimension of time. The resulting equations, however, do not lead to a numerically stable scheme. One needs instead to go backward in time and rewrite Eqs. (12.12) and (12.13) as

$$\frac{n(l+1, m) - n(l, m)}{\Delta t} = \mathbf{F}_2(F(l+1, m), n(l+1, m), p(l+1, m)) \quad (12.14)$$

and

$$\frac{p(l+1, m) - p(l, m)}{\Delta t} = \mathbf{F}_3(F(l+1, m), n(l+1, m), p(l+1, m)) \quad (12.15)$$

to obtain always stable solutions. The proof of this has to be left to the numerical mathematicians but can be made plausible by simple examples (try to numerically integrate $\tau \partial Y / \partial t = y + \sin(\omega t)$).

The above system of equations needs to be solved and contains a large number of unknowns; all variables are unknown at every point of the mesh except at the boundaries. To complicate things further, the system of algebraic equations for the unknowns is nonlinear. We therefore need to employ methods to solve such a large system of nonlinear equations. A well-established and well-working method is Newton's method. We refer the interested reader to Selberherr [9] for the details and mention here only the flow of the process. The nonlinear system of equations is transformed into a linear one by calculating derivatives of the functions $\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3$ with respect to the unknown electric field F , electron density n , and hole density p . These derivatives are then assembled into the so-called Jacobian matrix, which now gives a linear system for the unknowns. The matrix is still very large but has only finite entries along certain diagonal chains. All the other matrix elements are zero (because we had only integer indices $m, m+1, m-1, m+2, m-2$ in the equations). These large but sparse matrices are easily solved by standard routines (e.g., the Yale Sparse Matrix solver, which is described on the Internet).

We have discussed here the Newton method only very briefly because it is a standard tool to solve nonlinear algebraic equations. In one dimension it simply amounts to picking a point at a nonlinear curve, plotting the tangent and determining where the tangent crosses the x -axis. This is then the solution in first approximation, and so on. Next, one uses this new x -value and plots now at this point the tangent of the curve which crosses the x -axis at the solution in the second approximation and so on. The advantage of this procedure is that it

converges quadratically. It also is easily generalized to more than one unknown. One then forms and solves the Jacobian matrix as mentioned above and as described in detail by Selberherr [9]. The assembly and solution of the Jacobian matrix in more than one dimension on an irregular grid and with more than one unknown is a laborious task and is the core of many commercial software packages that deal with the device equations (e.g., MINIMOS, PISCES, DESSIS). The development of these and equivalent packages was a major achievement and they are routinely used in computer-aided design. It is highly recommended to use these instead of trying to produce homemade versions. It takes enormous numbers of work hours to construct such a code. There are, of course, still some shortcomings of the available software packages. For example, the formation of impurity bands with higher doping densities, and the band-tailing that we described in Chapter 5, are not contained in most simulators. This certainly will be rectified in the future, because devices have higher doping densities as their dimension shrinks. We will give results of numerical solutions for diodes and transistors in the next chapters.

12.2.1 Numerical Simulation Beyond the Shockley Equations

The procedure described above stands and falls with the validity of Eq. (12.1), the expression for the current as a function of the applied field. There are cases when such an expression is difficult to find, all related to velocity overshoot and real space transfer. In addition, under certain circumstances, one is not interested in the average response of the ensemble of electrons or holes such as the average drift velocity, current density, and so on. Reliability considerations, for example, deal mostly with electrons in the high energy tail of the distribution function. It is known that transistors age because hydrogen bonded to silicon at the silicon-silicon-dioxide interface is released owing to collisions with very hot electrons having an energy between 1.5 eV and 4 eV. This mechanism, which will be described in Chapter 15, requires then the knowledge of the energy distribution at extremely high energies, which can only be obtained reliably by solving the Boltzmann equation including a full-band structure. Such a solution can be and has been achieved by the so-called full-band Monte Carlo method [4, 10] that has already briefly been described in Chapter 8. The method essentially consists in following numerically a large ensemble of electrons or holes (10,000 or more, as necessary to obtain convergent results) and monitor their free flights and their scatterings by simulation exactly as would happen in nature. The charge carriers are propagated according to the one-band approximation by using Eqs. (3.35) and (3.36); that is,

$$\mathbf{v} = \frac{1}{\hbar} \nabla_{\mathbf{k}} E(\mathbf{k}) \quad (12.16)$$

and

$$\hbar \frac{d\mathbf{k}}{dt} = -e\mathbf{F} \quad (12.17)$$

At appropriately chosen times (see Chapter 8) electrons and holes are scattered according to the Golden Rule result for $S(\mathbf{k}, \mathbf{k}')$. Thus one follows the motion of the charge carriers exactly as an observer would if located in a real crystal, taking notes of (storing the values of) their energy, momentum, and so on. By taking suitable averages in narrow energy bins, one can thus obtain the desired energy distribution of the charge carriers. One can also easily include effects into this type of simulation that only occur at high energies. Examples are impact ionization and hydrogen desorption related transistor aging.

This procedure is physically very precise, and one can expect excellent results. However, the required computational resources are considerable. There are other methods to solve the Boltzmann equation and variations of the Monte Carlo Method [1, 4]. For a precise evaluation of the high energy tail of the distribution function, however, all of them require large numerical efforts. To do justice to devices, one needs in addition to seek the solutions consistent with the equation of Poisson. This can be done to various degrees of accuracy. A relatively simple procedure is the following. One can calculate the solution of the Shockley equations and argue that this solution is good for the bulk of the charge carriers. The failures of the Shockley equation system are mostly related to the few electrons in the high energy tail. Because these few will not carry much charge, one may be able to assume that Poisson's equation will not be changed significantly by them. Therefore, the electric field as obtained from the Shockley system will be mostly correct. Then one can take this solution for the electric field and use it as a basis for the accelerations in the Monte Carlo method and obtain from this method the high energy tail energy distribution separately. This, in turn, can be used to compute details of impact ionization, hot electron degradation, and the like, and represents an excellent procedure, except in the presence of massive velocity overshoot and real space transfer. In the presence of these two effects and whenever the high energy tail of the energy distribution is essential at the same time, a full band Boltzmann-Poisson solution is necessary. This is accomplished, for example, by the IBM simulator DAMOCLES [4]. Figure (12.1) shows an example of the high energy tail of the electron distribution for a constant electric field in bulk silicon as calculated by full-band Monte Carlo methods [4]. Note that for a constant field very high energies can be reached by the electrons, although with decreasing probability. In a device, however, the field is encountered only over a finite distance and the electrons cannot easily acquire energies larger than the applied voltage multiplied by the elementary charge. Then the high energy tail will be decreasing as $\exp(-E/kT_L)$ above this critical energy of the applied voltage. This is shown in Figure (12.2). These high energy electrons are mainly there because of the possibility of phonon absorption; therefore, the decrease corresponds to the phonon temperature T_L . Note,

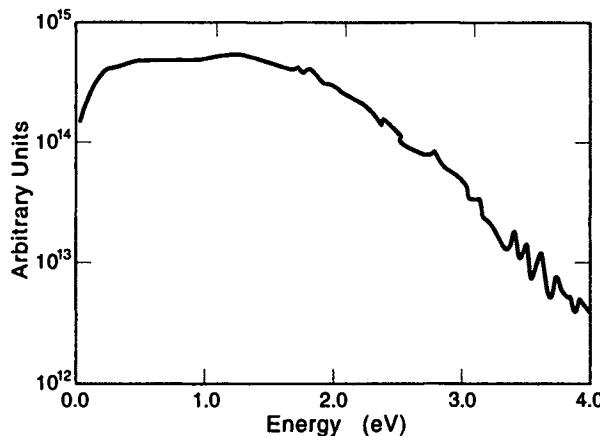


Figure 12.1 Energy distribution in bulk silicon for a constant electric field of 300 kV/cm as obtained from typical full-band Monte Carlo.

however, that electron-electron interactions and hot phonons [6] can contribute to a high energy tail even above the energy corresponding to the applied voltage. This is still an area of much research.

If one is not interested in the high energy distribution function but wants to account for velocity overshoot and real space transfer (the redistribution of hot electrons in potential wells and emission out of the wells) only, then one also can add to the Shockley equations the next higher moment; that is, one regards mo-

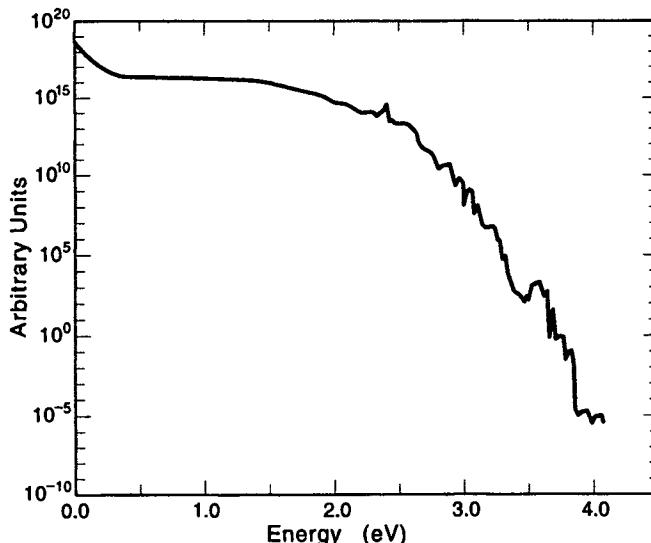


Figure 12.2 Typical energy distribution in a short channel MOSFET close to the drain for a drain voltage of 3 V. (Courtesy of A. Duncan, U. Ravaioli, and J. Jakumeit [2].)

bility and diffusion content a function of the charge carrier temperature, which is calculated from the next higher moment equation as, for example, shown in Appendix E. The particular approach of Appendix E is due to Stratton. There are other ways of closure for this next higher moment (the way that higher moments are neglected), and the approaches are then known under various names (e.g., hydrodynamic simulations). The method of Stratton and variations of it are now known as energy transport [5]. Overshoot and real space transfer can, in this way, be accounted for by an intermediate numerical effort.

PROBLEMS

- [12.1] Derive Eq. (12.9) by integrating the equation for the current Eq. (12.1) between the meshpoints $m, m + 1$ assuming constant coefficients and a given value $n(m)$ of the electron concentration.
- [12.2] Derive the function F_2 of Eq. (12.13) by using Eqs. (12.6) and (12.9).
- [12.3] Write a numerical program to solve by forward and backward integration the equation $\tau \partial Y / \partial t = Y + \sin(\omega t)$ with $Y = 0$ for $t = 0$. Run the programs and check numerical stability.

REFERENCES

- [1] Alarm, M. A., and Lundstrom, M. S. "Table-based Monte Carlo simulation of electron, phonon, and photon dynamics in quantum well laser," in *Compound Semiconductors*, ed. H. Goronkin, and U. Mishra, Institute of Physics Publishing, Bristol and Philadelphia, pp. 765–768, 1994.
- [2] Duncan, A., Ravaioli, U., and Jakumeit, J. "Full-band Monte Carlo investigation of hot carrier trends in the scaling of metal–oxide–semiconductor field-effect transistors," *IEEE Transactions on Electron Devices*, vol. 45, 1998, p. 45.
- [3] Grupen, M., and Hess, K. "Simulation of carrier transport and nonlinearities in quantum well laser diodes," *IEEE Journal of Quantum Electronics*, vol. 34, 1998, pp. 120–140, (see errata p. 384).
- [4] Hess, K., ed. *Monte Carlo Devices Simulation: Full Band and Beyond*, Boston: Kluwer, 1991.
- [5] Kan, E. C., et al. "Formulation of macroscopic transport models for numerical simulation of semiconductor devices," *VLSI Design*, vol. 3, 1995, pp. 211–224.
- [6] Poetz, W., and Kocevar, P. "Cooling of highly photoexcited electron–hole plasma in polar semiconductors and semiconductor quantum wells, a balance equation approach," in *Hot Carriers in Semiconductor Nanostructures*, ed. J. Shah, New York: Academic, 1992, pp. 87–120.
- [7] Ravaioli, U. "Hierarchy of simulation approaches for hot carrier transport in deep sub-micron devices," *Semiconductor Science and Technology*, vol. 13, 1998, p. 1.
- [8] Scharfetter, D. L., and Gummel, H. K. "Large signal analysis of silicon Read diode oscillator," *IEEE Transactions on Electron Devices*, vol. ED-16, 1969, pp. 64–77.
- [9] Selberherr, S. *Analysis and Simulation of Semiconductor Devices*. New York: Springer-Verlag, 1984.

- [10] Shichijo, H., and Hess, K. “Band structure dependant transport and impact ionization in GaAs,” *Physical Review B*, vol. 23, 1981, pp. 4197–4207.

CHAPTER 13

DIODES

We have already discussed simple diodes (two-terminal devices) in Chapters 11 and 12, and have developed some of the theoretical concepts that are important for their understanding. In this chapter we discuss several types of diodes, which represent useful semiconductor devices. Their use in electronics will not be specifically emphasized here because it is outlined in many other texts (Streetman [20]) and is also, at times, obvious (rectification, light generation, etc.). The basic physical concepts necessary for the understanding of those diodes that have not been discussed in the previous two chapters are treated here in connection with the description of diode operation.

We know from Chapter 9 that the generation and recombination terms can be derived by integrating the scattering integral of the Boltzmann equation over all \mathbf{k} assuming that f_0 is of Fermi- or Boltzmann-type. The integration of the other terms in the Boltzmann equation gives a term $\partial n/\partial t$ (from $\partial f/\partial t$) and $\nabla \mathbf{j}_n/e$ (from $-v\nabla f$). We also have obtained in the previous chapters the equation of continuity for the conduction band

$$\nabla \mathbf{j}_n - e \frac{\partial n}{\partial t} = e(R(n) - G(n)) \quad (13.1)$$

and for the valence band

$$\nabla \mathbf{j}_p + e \frac{\partial p}{\partial t} = e(G(p) - R(p)) \quad (13.2)$$

Although we derived these equations from the Boltzmann equation, their basic validity goes beyond classical derivations. As a consequence, these equations assume a central importance in the physics of semiconductor devices. This is emphasized by our particular treatment of the $p-n$ junction diode and has been discussed in Chapter 12. An approximation that is often made and that we also make in this chapter is $G(n) - R(n) = G(p) - R(p) = -U_s$. This is only valid close to steady-state conditions, as described in Chapter 9.

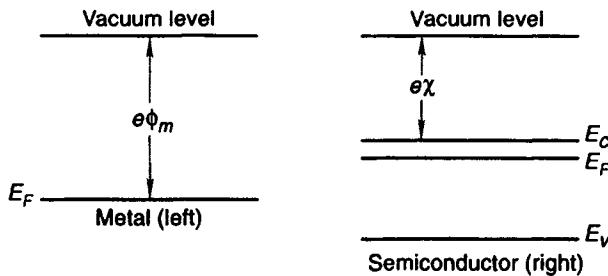


Figure 13.1 Band structure of semiconductor and metal when they are still separated, $e\phi_m$ is the metal work function and $e\chi$ the semiconductor electron affinity.

13.1 SCHOTTKY BARRIERS—OHMIC CONTACTS

Section 10.1 dealt with the ideal lattice-matched semiconductor heterojunction. Similar techniques apply, of course, for other heterojunctions, such as metal-semiconductor contacts. In fact, historically these have been investigated first and a detailed understanding has been provided by Schottky.

The band structure of a separated semiconductor and metal is shown in Figure 13.1. Imagine that we bring the semiconductor and metal closer together so that charge can flow from one to the other. The Fermi levels will then line up, and if we start with a Fermi level that is higher in the semiconductor, electrons will flow from the semiconductor to the metal. This will cause depletion of the semiconductor and a potential change between the semiconductor (at the end of the depletion width) and the metal. This is shown in Figure 13.2. Finally when metal and semiconductor are in contact, the situation of Figure 13.3 emerges.

The quantity $e(\phi_m - \chi)$ is usually called the Schottky barrier height, denoted by $e\phi_{Bn}$. The subscript n is for electronic semiconductors. If we had considered a junction with a *p*-type semiconductor (holes), then we would have found a barrier height

$$e\phi_{Bp} = E_G - e(\phi_m - \chi) \quad (13.3)$$

which gives us

$$e(\phi_{Bp} + \phi_{Bn}) = E_G \quad (13.4)$$

This result applies only to ideal junctions.

In a real semiconductor-metal junction, a new phenomenon occurs: The Fermi level can be fixed around a certain energy, which is typically close to the middle of the gap of the semiconductor material. The reason for this “pinning” is that high densities of electrons can be trapped close to the interface. These electrons form an interface charge that determines the barrier height, which corresponds then to the “pinning” energy, as shown in Figure 13.4. J. Bardeen suggested that surface states of the kind shown in Figure 4.4 are responsible for the pinning. However, numerous recent investigations have shown that the problem is more

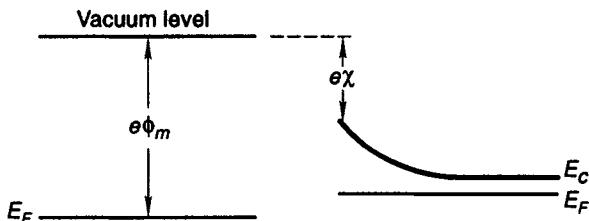


Figure 13.2 Same as Figure 13.1, but for closer distance of metal and semiconductor.

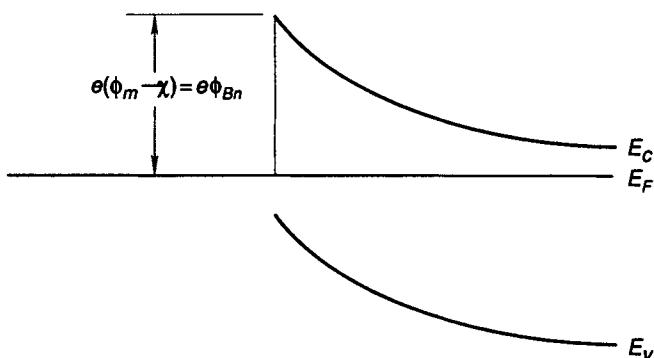


Figure 13.3 Same as Figure 13.2, but for contact between metal and semiconductor.

involved, and the explanation of the pinning depends on the actual preparation of the Schottky barrier contact.

An ideal semiconductor surface that has reconstructed (Figure 4.5) does not necessarily pin the Fermi level at all because the bonds are saturated. Such an ideal surface can only be generated by cutting (cleaving) the semiconductor in ultrahigh vacuum. As soon as the semiconductor comes in contact with air, oxides form at the surface and oxygen remnants remain no matter how carefully the semiconductor is treated afterward. The surface contamination and subsequent metal deposition will “undo” the reconstruction, and interface (not surface) traps of various kinds will form and have to form because the metals are not lattice matched to the semiconductors as demonstrated by Woodall. This is schematically shown in Figure 13.5.

It is not clear at present what precisely pins the Fermi level, which is understandable from the complicated variety of possibilities inherent in a picture such as Figure 13.5. Most probably there exists no universal explanation of pinning energies. The experimental findings indicate that the explanation of pinning will depend on the particular technology of barrier formation. For our purpose, we can accept these trapping or pinning energies as given facts and insert into our transport theory experimental values such as those shown in Table 13.1. The barrier heights in the table are approximate and typical for “clean” interfaces. Contamination of the interface may lead to widely varying results. The barrier



Figure 13.4 Pinning of the Fermi level by localization of electrons at the interface.

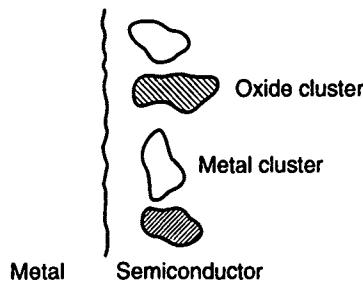


Figure 13.5 Nonideal metal–semiconductor interface with metal and oxide clusters present.
[After Woodall.]

heights also apply only to equilibrium (no applied electric field).

The addition of an external field leads to a barrier lowering (the Schottky effect) caused by the image force, which is equal to

$$\frac{e^2}{16\pi\epsilon_0\epsilon z} \quad (13.5)$$

as is well known from electrostatics. The total potential energy $PE(z)$, including the applied field, is then

$$PE(z) = \frac{e^2}{16\pi\epsilon_0\epsilon z} + eFz \quad (13.6)$$

This function has a maximum that is lowered by

$$\Delta\phi_B = \sqrt{\frac{eF}{4\pi\epsilon_0\epsilon}} \quad (13.7)$$

from the original barrier height. The maximum is also some distance from the interface, which can be easily found by equating $dPE(z)/dz = 0$. Deducting Eq. (13.7) from the given equilibrium barrier height gives the barrier height ϕ'_B , which is appropriate for transport calculations. The barrier lowering is illustrated in Figure 13.6.

There is an additional term that contributes to the Schottky barrier height. This term arises from the penetration of the metal wave functions into the semiconductor and the corresponding penetration of metal charge. This term is easy to add to Eq. (13.6), and analytical expressions have been given by Dutton and co-workers. The interested reader is referred to Shenai et al. [17].

Table 13.1 Typical Schottky Barrier Heights (volts at 300 K)

Semiconductor	Ag	Au
Si <i>n</i>	0.78	0.80
Si <i>p</i>	0.54	0.34
Ge <i>n</i>	0.54	0.59
Ge <i>p</i>	0.50	0.30
GaAs <i>n</i>	0.97	1.02
GaAs <i>p</i>	0.63	0.42

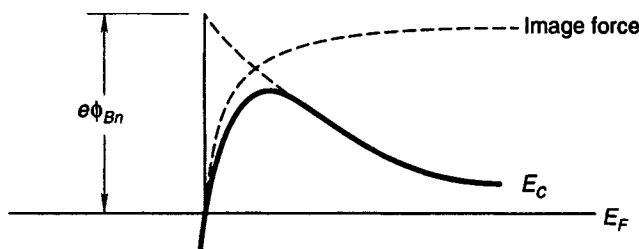


Figure 13.6 Schottky barrier lowering by applied (built-in) electric field. The dashed line is the semiconductor conduction band edge without the image force contribution, and the solid line is with the image force contribution. The image potential alone is also indicated.

We can now proceed to calculate the current over a Schottky barrier. The Schottky barrier from the metal to the semiconductor can be treated in the same way as we have treated the semiconductor-semiconductor junction of Figure 10.1. If we measure the barrier height from the Fermi energy of the metal, putting $E_F^L - E_c^L$ in Eq. (10.6) equal to 0, and replace $|\Delta E_c|$ by $e\Phi'_B$, we obtain

$$J_{LR} = A^* T^2 e^{-e\Phi'_B/kT} \quad (13.8)$$

for the equilibrium current j_{RL} from the semiconductor to the metal. The total current is the difference between j_{RL} and j_{LR} . To calculate this difference, we need to digress briefly and discuss the quasi-Fermi levels of a barrier structure when a voltage is applied.

Assume that we have a very long semiconductor that is, far to the right, essentially unperturbed by the presence of the junction with the metal. If we bring an electron from the quasi-Fermi level of the metal to the quasi-Fermi level (which is constant away from the junction) of the semiconductor, we need to perform a certain amount of work that needs to be done by external voltages. We therefore conclude with the important notion that the difference in quasi-Fermi levels is equal to eV_{ext} :

$$|E_F^L - E_F^R| = |eV_{ext}| \quad (13.9)$$

The applied voltage will mainly drop in the depletion region (another important

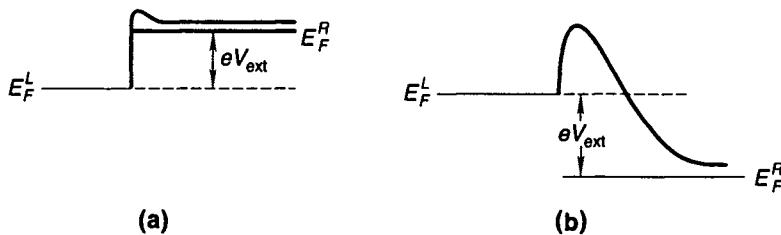


Figure 13.7 Schottky barrier under forward (a) and reverse(b) bias.

notion), and therefore we will change the built-in potential V_{bi} of Eq. (10.18) just by eV_{ext} .

Figure 13.7 shows the typical form of the quasi-Fermi levels of a Schottky barrier under bias. Figure 13.7a is plotted for forward bias, the direction of voltage that causes a large current to flow [positive \bar{V}_{ext} in Eq. (13.12)], whereas Figure 13.7b shows reverse bias [negative \bar{V}_{ext} in Eq. (13.12)].

A simple analogy of the current flow can be obtained by comparing the semiconductor side with a mountain that can be varied from flat plateau form (Figure 13.7a) to peaked form (Figure 13.7b). The electron flow corresponds, then, to the way the atmosphere (air) would flow if the mountain were moved. In Figure 13.7a, we have moved the air up, and high-density air at the right will flow to the low-density side (left), and vice versa in Figure 13.7b. It is important to note that this Schottky barrier current is totally different (in its origin) from the current in Eq. (8.42); this current was a field current owing to the drift (odd) part f_1 , of distribution function f . Here we can have a current entirely owing to the even part f_0 of the distribution function . The current arises from the difference of the quasi-Fermi levels relative to the band edges of the different materials, whereas for a current as in Eq. (8.42), the distance of the quasi-Fermi level and band edge change by eV_{ext} in the same way). In the case of a Schottky barrier, all the external voltage drops in the depletion region of the semiconductor and modulates the built-in voltage V_{bi} to $V_{bi} + V_{ext}$ while the quasi-Fermi level stays approximately constant on either side.

The current density from the right to the left of our junction is therefore equal to

$$j_{RL} = A^* T^2 e^{-e\phi'_B/kT} e^{eV_{ext}/kT} \quad (13.10)$$

with an external voltage applied [see Eq. (10.19)]. This makes the total current density $j = (j_{RL} - j_{LR})$ equal to

$$j = A^* T^2 e^{-e\phi'_B/kT} (e^{eV_{ext}/kT} - 1) \quad (13.11)$$

In the following it is convenient to denote eV_{ext}/kT by \bar{V}_{ext} and analogous $\bar{\phi}$ so that we obtain

$$j_{RL} = A^* T^2 e^{-\bar{\phi}} (e^{\bar{V}_{ext}} - 1) \quad (13.12)$$

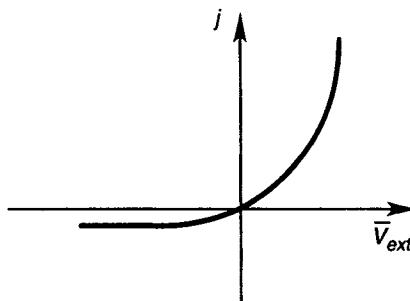


Figure 13.8 Schematic plot of the current in a Schottky barrier demonstrating the rectifying property.

Figure 13.8 shows schematically this current density and demonstrates the use of Schottky barriers as rectifiers. Notice that the application of an external voltage to a Schottky barrier diode also changes the depletion width. From Eq. (10.17), the depletion width, including external potentials (see Eq. (10.15) and its deviation), is

$$W = \left(\frac{2\epsilon\epsilon_0(V_{bi} \pm V_{ext})}{eN_D^+} \right)^{1/2} \quad (13.13)$$

The minus sign applies for barrier lowering; that is, for negative polarity of V_{ext} at the *n*-type semiconductor side.

We turn now to the speed limitations of Schottky barrier diodes, which will also give us a better understanding of the assumptions involved in our discussion. The ultimate speed for switching of a Schottky barrier diode is given by the time t_{ext} , which is the time the bulk of the carriers need to transfer from the semiconductor over the semiconductor depletion width W to the metal. Within the framework of our derivation the speed of the carriers can roughly be put equal to v_{z0} as given in Eq. (10.2) (a more precise evaluation would involve appropriate averaging over the velocities). Using Eq. (10.17), one obtains

$$t_{ext} = \frac{W}{v_{z0}} \approx \left(\frac{\epsilon\epsilon_0 m^*}{e^2 N_D} \right)^{1/2} \quad (13.14)$$

which is of the order of 10^{-13} s for typical values of parameters (such as $N_D = 10^{16} \text{ cm}^{-3}$). It is this extraordinary high speed that makes Schottky barriers interesting.

Of course, other factors are involved in the actual speed of Schottky barrier diodes. A trivial limitation is the product of resistance and capacitance, which includes the total semiconductor resistance and the depletion width capacitance (which is given by $\epsilon\epsilon_0/W$ for current purposes). Another speed limitation would be the time the electrons need to lose energy after going over the barriers (so they cannot return), which is typically equal to the inverse scattering rate by phonons (10^{-13} s). However, even more important (for speed limitations) is the fact that

the transport of electrons has a diffusive nature. This is because of the scattering and because Eq. (10.1) holds only if the electrons that transfer to the metal are replenished from lower energies. For this replacement of electrons it is necessary that electron-electron interactions and phonon absorption events are frequent enough to maintain a Maxwellian distribution, which we have assumed in the deviation. The electrons also must be replenished from the right side, which can represent a bottleneck if the distances are longer than the scattering length, because the transport then becomes diffusive. Indeed, if $W > 1000\text{\AA}$, transport by diffusion dominates and the thermionic emission process that takes place close to the interface represents just a boundary condition for the diffusion process (the bottleneck for electron supply). We give a treatment of this diffusion process in Section 13.2 in the context of the p - n junction, and also in Appendix G. This appendix gives also an explanation of the constancy of the quasi-Fermi levels on each side, as shown in Figure 13.7.

Before leaving this section, we need to discuss another mode of transport through Schottky barriers, which is tunneling. Electrons can cross barriers quantum mechanically by tunneling through them. Tunneling is, as we will see, especially important for thin barriers. For thick barriers the thermionic current is larger, except if the temperature becomes very small, as can be seen from Eq. (10.6). The tunneling current can be derived from Eqs. (1.50) and (A.16) as well as (A.19).

To calculate the current from the left ($z < z_a$) to the right ($z > z_a$), we proceed as follows: We multiply the probability per unit time of a tunneling transition by the probability to find an electron at the left (which is equal to the energy distribution f_L on the left side) and by the probability that the state is not occupied at the right ($1 - f_R$) and then sum over all states on the right. This gives the current from the left to the right owing to one state. The total current is the sum over all left states multiplied by the elementary charge. Denoting \mathbf{k} by \mathbf{k}_L at the left and by \mathbf{k}_R at the right and including a factor of 2 for the spin, we get

$$j_{LR} = \frac{4\pi e}{\hbar} \sum_{\mathbf{k}_L} \sum_{\mathbf{k}_R} \left[\frac{|\bar{k}_{z_a}|}{L_a} \frac{|\bar{k}_{z_b}|}{L_b} \exp \left(-2 \int_{z_a}^{z_b} |\bar{k}_z| dz \right) \delta(E_R - E_L) f_L (1 - f_R) \right] \quad (13.15)$$

Now transforming the sums into energy integrals, one obtains

$$j_{LR} = \frac{\pi e}{\hbar} \int \int \left[\frac{|\bar{k}_{z_a}|}{L_a} \frac{|\bar{k}_{z_b}|}{L_b} \exp \left(-2 \int_{z_a}^{z_b} |\bar{k}_z| dz \right) \delta(E_R - E_L) f_L (1 - f_R) g_L(E_L) g_R(E_R) dE_L dE_R \right] \quad (13.16)$$

where $g_L(E_L)$ and $g_R(E_R)$ represent the density of states on the left and right sides, respectively. The notation is otherwise the same as in Appendix A. Because these density-of-states functions each include a factor of 2 for the spin,

which has already been accounted for above, we have divided Eq. (13.15) by a factor of 4.

The current from the right to the left can be calculated in analogous fashion. The integral is easily evaluated. One integration involving the function simply replaces all energies by E_L or E_R . The other integration can be performed for special cases (triangular barrier, etc.) or can be done with a pocket calculator for general potential forms.

After the discussion of the metal-semiconductor contact and its rectifying properties, it seems difficult to understand how an ohmic contact is formed. Of course, there is the possibility that the Schottky barrier height is zero or such that electrons are transferred to the *n*-type semiconductor and no depletion region forms (see also Streetman [20]). There is, however, another possibility that is of greater importance. The metal can be enriched by donor material (or acceptors) and, by alloying, a thin layer of highly doped semiconductor will form in the proximity of the metal. If this is achieved (by whatever technology), the depletion layer width W will be very thin. Using $z_b - z_a = W$ in the tunneling formula, we can see that the tunneling current will be large. In fact, it can be large enough to satisfy the most stringent requirements of low-contact resistance. Note that doping fluctuations will enhance the current at certain places and reduce it at others, the effect being an overall enhancement by a factor of 2 to 10 over the computation for homogeneous material, as has been shown by McGill and co-workers.

13.2 THE *p-n* JUNCTION

13.2.1 Introduction and Basic Physics

We saw in Chapter 10 the importance and significance of heterojunctions, that is, junctions in which the semiconductor and its band structure vary. However, as we know from Chapter 4, minute changes in the chemical composition can drastically alter the electronic properties of semiconductors. For this reason, the homojunction—a junction of identical semiconductor but different doping (typically, donors on one side, acceptors on the other)—has its own special importance. The doping agents can be introduced by diffusion at high temperatures or by ion implantation. The simplistic approach of a theorist “to glue” two semiconductors of different types together does not work because of the interface states between glue and semiconductors. We do not attempt to discuss the experimental details and assume an ideal abrupt junction (Figure 13.9) where the doping changes suddenly at $z = 0$ from *p*-type at the left and *n*-type at the right. Gradual junctions can be treated similarly (Streetman [20]). We first discuss the band diagram of such a junction in equilibrium (no external forces applied) and list the rules to graph such diagrams.

The voltages and band edge energies are measured relative to a fixed point; for example, the point c , where c is far away from the junction (far enough not

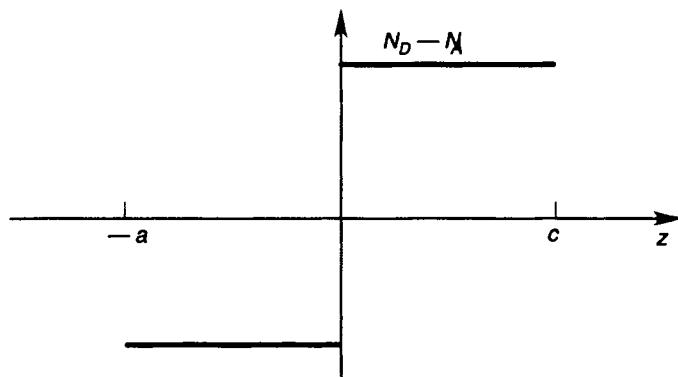


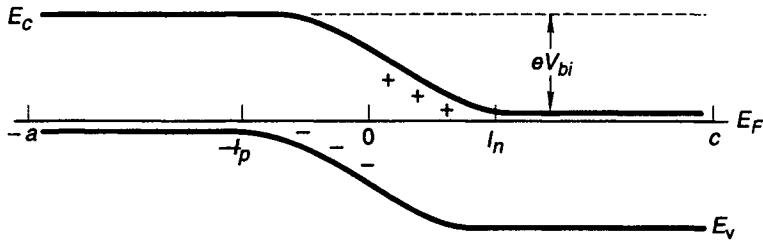
Figure 13.9 Doping concentration in an ideal abrupt homojunction. Notice that $N_D \neq N_A$ at the different sides.

to be disturbed by the presence of the junction). The rules for plotting band diagrams are as follows.

1. The band edges follow the additional potential V_{tr} (Chapter 6) caused by impurity charges and external voltages. This rule is a consequence of the effective mass theorem [Eqs. (3.32) and (3.34)]. Notice that the potential energy varies with the opposite sign of the electrical potential.
2. The Fermi energy is constant and far away from the junction, located at the bulk equilibrium distance from conduction and valence band edges [see Figure 13.10 and Eq. (5.23)].
3. The energy gap E_G is the same throughout except in regions of very heavy doping or large electron (hole) density [see Figure 5.4 and Eq. (5.16)].
4. At the junction, space-charge regions form (see Eq. (10.17) and Figure 10.2), and both conduction and valence band edges vary considerably.

Consequently, the diagrams are plotted best by first drawing a straight horizontal line representing the Fermi level. Next the junctions are marked and the band edges are drawn away from the junctions. On the sides with donor doping (n sides), the conduction band edge is marked as a line close and parallel to the Fermi level. Similarly, on the sides with acceptor doping (p sides), the valence band is marked close to the Fermi energy and (far away from the junction) parallel to it. Knowing E_c , we can then plot parts of the conduction band on the p sides and of the valence band on the n sides. Smooth (except for band edge discontinuities in heterojunctions) connections finish the diagram. The diagram for our abrupt $p-n$ junction is shown in Figure 13.10.

Figure 13.10 shows clearly that the Fermi level is distant from both the conduction and valence band edges for the range $l_p \leq z \leq l_n$. According to Eqs. (5.23) and (5.30), this range is therefore depleted of electrons and holes, and for some purposes we can approximate $n = p = 0$ in this range. In the same

Figure 13.10 Band diagram for an abrupt p - n junction.

spirit the electric field is put at zero for $z < l_p$ or $z > l_n$. In the depleted range there is fixed space charge, the positively charged donors and negatively charged acceptors as shown in Figure 13.12. The voltage difference from the n to the p side is denoted by V_{bi} , which is the built-in voltage totally analogous to the built-in voltage in heterojunction barriers [Eqs. (10.15) through (10.17)]. To obtain $-l_p$ and l_n as a function of the built-in voltage, we proceed as in Eqs. (10.15) and (10.17) and integrate Eq. (10.15) from $-l_p$ to l_n .

In addition we use the fact that the electric field at l_n , $F(l_n)$, or beyond it to the right, is equal to zero. From Eq. (10.12) we have

$$F(z) = -\frac{e}{\epsilon \epsilon_0} \int_{-l_p}^z (p - n + N_D^+ - N_A^-) dz \quad (13.17)$$

Then we use the depletion approximation $p = n \approx 0$ to obtain

$$0 = F(l_n) = \frac{e}{\epsilon \epsilon_0} (N_A^- l_p - N_D^+ l_n) \quad (13.18)$$

Using $N_A^- \approx N_A$ and $N_D^+ \approx N_D$ in the depletion region, we further have

$$l_p N_A = l_n N_D \quad (13.19)$$

which we also could have guessed from charge neutrality considerations. This, together with the equation for the built-in voltage [Eq. (10.15)] and $p = n \approx 0$, gives

$$l_p = \left(V_{bi} \frac{2\epsilon \epsilon_0}{e} \frac{N_D}{N_A(N_A + N_D)} \right)^{1/2} \quad (13.20)$$

and

$$l_n = \left(V_{bi} \frac{2\epsilon \epsilon_0}{e} \frac{N_A}{N_A(N_A + N_D)} \right)^{1/2} \quad (13.21)$$

It is easy to verify that for $N_A \gg N_D$ we are led back to Eq. (10.17) by using $W = l_p + l_n$. Our results are summarized in Figure 13.11. The free carrier concentration is approximately constant outside the depletion region and denoted by p_p , p_n (holes on the p and n sides, respectively), as well as by n_n , n_p (electrons on the n and p sides, respectively). If the doping is homogeneous, the fixed

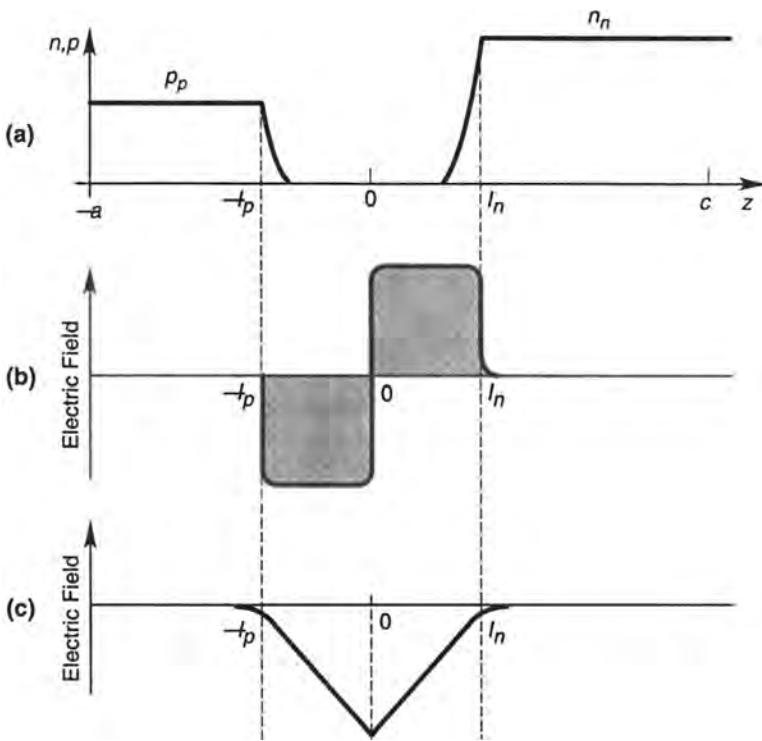


Figure 13.11 (a) Concentration of free carriers (electrons and holes) in a p - n junction at equilibrium; (b) distribution of fixed charge in the depletion layer (notice that charge neutrality requires equal areas); (c) electric field in the depletion region.

charge is approximately constant between $-l_p$ and 0 , as well as between 0 and l_n . The electric field has a maximum, F_{\max} at $z = 0$, and is triangular within the depletion approximation.

From Eq. (13.17), F_{\max} is found to be

$$F_{\max} = -\frac{e}{\epsilon \epsilon_0} N_D l_n = -\frac{e}{\epsilon \epsilon_0} N_A l_p \quad (13.22)$$

These equations and figures form the basis for an ideal model of the p - n junction away from equilibrium, that is, with external voltage applied. We will see that the current in a p - n junction can best be understood from the following analogy of high-energy physics.

Consider the box shown in Figure 13.12 and a beam of electrons incident from the right side and a beam of positrons (antiparticles) from the left. The electrons and positrons meet in the box, annihilate each other, and emit radiation. Therefore a current is flowing continuously from place a to c , but neither electrons nor positrons make it all the way through the box. It is precisely the same mechanism that explains the current in a p - n junction. The fact that the

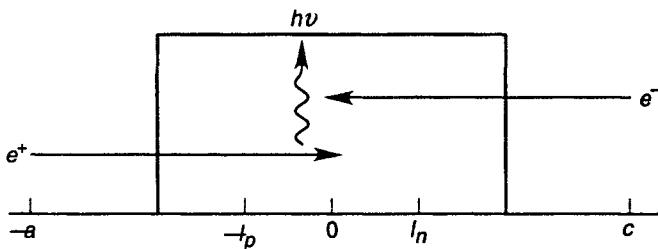


Figure 13.12 High-energy physics analogy for p - n junction.

electrons and holes in a p - n junction need to overcome a potential barrier before recombining is also important and leads to exponential dependencies in the current. The depletion region is, of course, formed by the recombination of electrons and holes. If external voltages are applied, the equilibrium is perturbed and the current is maintained by continuous additional recombination (or generation).

Although the details require more discussion, we can already plot the non-equilibrium band diagram by applying two principles that we developed previously. The first principle we use is that the difference in the quasi-Fermi levels for electrons and holes is equal to eV_{ext} , far away from the junction. Second, we have discussed in Appendix G that the quasi-Fermi levels are also constant in the depletion region provided that certain criteria (for the replenishing of charge carriers) are fulfilled. These criteria (Appendix G) are somewhat different for p - n junctions than for the case of Schottky barriers, because the carriers are not “lost” by emission but by recombination. However, by reasoning as in Appendix G, one easily sees that as long as the recombination rate (time) is weaker (longer) than the rate (time) that is necessary to replenish the carriers, the quasi-Fermi levels will be constant in the depletion region. We will derive these rates below. A detailed examination (which is left to the reader) shows that for small V_{ext} the assumption of constant quasi-Fermi levels is justified. We denote the quasi-Fermi levels of electrons by E_F^n and of holes by E_F^p (the notation involving L far left and R far right is not sufficient now because we will need the extension of electron quasi-Fermi levels for the hole side, and vice versa).

The above discussion leads to the band diagram shown in Figure 13.13. Beyond the depletion region the quasi-Fermi levels have to merge because far away from the junction electrons *and* holes approach their equilibrium density.

The external voltage appears across the depletion region and, depending on its polarity, adds or subtracts from the built-in voltage. It reduces the barrier if negative voltage is applied to the n side and positive to the p side. This is therefore called forward bias (the opposite is reverse). Note that the depletion width $W = l_n + l_p$ changes with external bias. From the derivation [Eq. (10.13)] it can be seen that the external voltage adds (subtracts) to the built-in voltage. In the formulas for l_n and l_p , Eqs. (13.20) and (13.21), we need to replace V_{bi} by

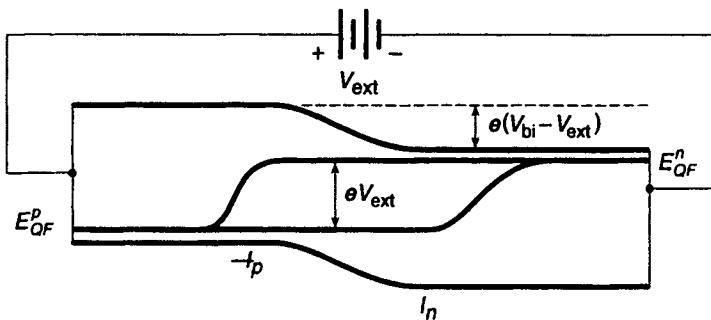


Figure 13.13 Band diagram of p - n junction with external voltage applied.

$V_{bi} + V_{ext}$. Therefore

$$l_n = \left((V_{bi} + V_{ext}) \frac{2\epsilon\epsilon_0}{e} \frac{N_A}{N_A(N_A + N_D)} \right)^{1/2} \quad (13.23)$$

Equation (13.23) is of importance for many applications.

Of equal importance is the product of electron and hole concentrations when a voltage is applied. We have learned from Eq. (5.31) that this product is proportional to $\exp(-E_G/kT_L)$ (in this equation the lattice temperature must be used because the electrons are excited by lattice vibrations). Following the derivation of Eq. (5.31) with quasi-Fermi levels, we see that

$$n \cdot p = n_i^2 \exp(E_F^n - E_F^p)/kT_L \quad (13.24)$$

which gives us (by use of Figure 13.13)

$$n \cdot p = n_i^2 \exp(eV_{ext}/kT_L) \quad (13.25)$$

If the bias is in forward direction, eV_{ext} is positive and increases the product.

Equations (13.24) and (13.25) can also immediately be deduced from Figure 13.13, which shows that the energy gap is effectively reduced (forward bias) by eV_{ext} or increased by this amount in reverse bias. From Eqs. (9.36) and (13.25) we obtain the steady-state recombination rate

$$U_s = n_i^2 \bar{c} (\exp(eV_{ext}/kT_L) - 1) \quad (13.26)$$

where the quantity \bar{c} depends on the carrier concentrations but for the time being is assumed to be constant; the consequences of the dependence on n and p are discussed later. According to our analogy from high-energy physics, the current density is then

$$j = \int_{-a}^c eU_s dz \quad (13.27)$$

assuming that all electrons or holes recombine in the device. If they do not, they usually recombine at the contacts and all one has to do to make Eq. (13.27) more generally valid is to insert a contact recombination term U_c . For a long device

as we are considering it throughout, contact recombination can be disregarded because all nonequilibrium charge recombines a short distance beyond l_n (or $-l_p$). Then U_s can be expressed in a range of width $W \approx (l_n + l_p)$ by Eq. (13.26) and zero otherwise to give

$$j = W e n_i^2 \bar{c} (\exp(eV_{\text{ext}}/kT_L) - 1) \quad (13.28)$$

This is very close to the actual current density in a typical *p-n* junction. In the following we will derive more exact expressions for the dc and also ac current densities for long and fully symmetric diodes ($N_A = N_D$, equal electron and hole mobilities etc.). This symmetry assumption is important for the development of a consistent theory. It will be dropped at the end of the section, which gives a short description of the asymmetric case. The complications of this case give a good motivation for the use of complete numerical treatments as they are now commercially available. The explicit treatment below is necessary for the understanding of the device physics.

13.2.2 Basic Equations for the Diode Current

For long and symmetric diodes there is a particularly illuminating way to integrate the equations of continuity Eqs. (13.1) and (13.2) in one dimension [13]:

$$j_n(z) = \int_0^z (eU_s + e\frac{\partial n}{\partial t}) dz + j_n(0) \quad (13.29)$$

and

$$j_p(z) = \int_0^z (-eU_s - e\frac{\partial p}{\partial t}) dz + j_p(0) \quad (13.30)$$

Here we have put $G(n) - R(n) = G(h) - R(h) = U_s$. This is commonly assumed to be true also for finite frequency conditions and is included in all commercial simulators. However, in general (and particularly for high frequencies) one needs to solve Eq. (9.30) in addition to the equations for continuity of electrons and holes. Strictly speaking, we are limited therefore to low frequencies ω as compared to the generation and recombination rates R and G .

Because, for symmetric diodes $j_n(0) = j_p(0)$, and because, for long diodes, no hole current flows at c (and no electron current at $-a$), we obtain for the terminal current $j(c)$:

$$j(c) = j_n(c) = \int_0^c (eU_s + e\frac{\partial n}{\partial t}) dz + j_n(0) \quad (13.31)$$

which gives after using $j_p(c) = 0$

$$j(c) = \int_0^c (2eU_s + e\frac{\partial n}{\partial t} + e\frac{\partial p}{\partial t}) dz \quad (13.32)$$

or rewritten to separate the displacement current

$$j(c) = \int_0^c (2eU_s + 2e\frac{\partial p}{\partial t} + e\frac{\partial n - p}{\partial t}) dz \quad (13.33)$$

The last term is equal to the displacement current $\epsilon\epsilon_0\frac{\partial F}{\partial t}$ at $z = 0$ because from Eq. (6.2) we have

$$\frac{\partial}{\partial t} \int_0^c e(n - p + N_A^- - N_D^+) dz = \frac{\partial}{\partial t} \int_0^c e(n - p) dz \quad (13.34)$$

Thus the total diode current of Eq. (13.33) appears as the sum of particle and displacement currents at the junction $z = 0$. Equation (13.33) therefore also reflects implications of Gauss's law, although it does not solve it; up to now only the two continuity equations are integrated. To actually calculate $j(c)$ one needs to obtain n and p which necessitates the integration of Gauss's law. Whenever this is necessary we recommend to the reader to seek a numerical solution using commercial software. For our discussion we solve Gauss's law only in the depletion approximation and assume quasi-neutral conditions outside the depletion range. This leads sometimes to logical contradictions, which we will discuss toward the end of the section. Next we discuss only the physical meaning of the various terms in Eq. (13.33).

The first term (which we name the " U_s -term") describes the complete dc current density as already discussed in Eqs. (13.27) and (13.28). The U_s -term also depends on time via $n(t)$ and $p(t)$ and therefore contributes to the ac current density; this is important, but has often been neglected in the past. It contributes, as we will see, a negative capacitance that is caused by the unique physics of the p - n junction. For an ordinary metal-insulator-metal capacitor, the charge is supplied and removed through the contact wires. In a p - n junction, minority charge traverses the depletion region by diffusion and can be stored by a mechanism corresponding to a "diffusion capacitance," which is described mathematically by the second term in Eq. (13.33). However, this charge now can be removed in two ways to the contacts at $-a$ or c , by direct transport within the band (as usual), and also by recombination and subsequent transport in the other band. The charge so removed cannot be counted as stored (or reclaimable, as usually termed in connection with a capacitance) and needs to be deducted, which is mathematically taken care of by the first term, the U_s -term that thus contains a negative capacitance. This fact has been overlooked in some texts and has only recently been fully acknowledged (see, e.g., Tyagi [22]).

The second term represents the diffusion capacitance C_{diff} as seen from

$$\frac{\partial p}{\partial t} = \frac{\partial p}{\partial V} \frac{\partial V}{\partial t} \quad (13.35)$$

where V is the appropriate voltage drop. The concept of diffusion capacitance is a complex and relatively new one in electronics and deserves a few comments.

The origin of this term is the minority carrier density diffusing against the built-in potential reduced (in forward bias) by a voltage close to the external applied voltage V_{ext} . It should be noted that the net charge storage on one side of the diode is described by the third term, the displacement current term that usually is also described by a capacitance, the depletion capacitance C_{dep} . The diffusion capacitance is not related to any net charge but to the total minority carrier density. (Of course, the equation of Poisson still needs to be fulfilled. However, this requires only global neutrality of all charges in the device including N_D^+ and N_A^- and no local charge neutrality.) We will derive the injected minority carrier density below in some detail. Assuming constant quasi-Fermi levels throughout the depletion region and no voltage drops for $l_n < z < c$ (and none on the left side of the junction as shown in Figure 13.13), we get for the hole density at l_n

$$p(l_n) \approx p_{no} \exp(eV_{\text{ext}}/kT_L) \quad (13.36)$$

where p_{no} is the equilibrium hole density on the n -side (right side in Figure 13.13). This follows naturally from Figure 13.13 and Eq. (13.24). This concentration decays within the well-known diffusion length (also derived below). Again, the true value of $p(l_n)$ can only be determined by also solving the equation of Poisson in addition to the continuity equations, which will then give the exact value of the voltage drop in the depletion region. This indeed is accomplished exactly by commercial software, and we also will discuss below a more careful analytical approach. For now we disregard these complications and use the following approximation. We assume the region $l_n < z < c$ to be well conducting (like a metal); any positive injected charge δp will reorder the majority of electrons to completely compensate for it and achieve charge neutrality (i.e., $\delta p = \delta n$). This, of course becomes increasingly incorrect as we approach the depletion region, which starts nominally at l_n . As δp increases exponentially, so does δn , compensating δp exactly. Therefore δp can be stored as in an ordinary capacitor but the equation of Poisson can be disregarded for all purposes other than increasing δn . This type of capacitance represents a new concept that does not exist for metals. Consider an intrinsic semiconductor; if you illuminate it, electron hole pairs are created increasing both n and p and the conductance. Charge carriers are “stored” but because $n = p$, one does not need to consider the equation of Poisson. The same situation underlies the concept of diffusion capacitance, which therefore might be more justly termed a plasma storage capacitance. This is a truly beautiful concept and very important for semiconductor devices and their ac response. This concept, however, has not the general validity that textbooks usually ascribe to it; it is only valid for short junctions. All the stored minority carriers recombine in long junctions, by definition, before reaching the contacts. The negative capacitance of the U_s -term therefore cancels part (one-half to a good approximation) of the diffusion capacitance. Another portion (one-half for fully symmetric diodes) is cancelled by the minority carrier density in the third term. In addition, voltage drops, for $l_n < z < c$ and $-a < z < l_p$, reduce the injected minority densities and reduce the diffusion capacitance further.

This results in an always decreasing capacitance in forward bias, which means that a diffusion capacitance does not exist. The situation is completely different in short diodes, too short for recombination to occur to a significant extent and for a voltage drop to be significant either. Then the concept of plasma storage is completely operational and the diffusion capacitance becomes prominent.

The third term represents the displacement current which is usually described by a depletion capacitance that is often approximated by

$$C_{\text{dep}} = \frac{\epsilon \epsilon_0}{l_n + l_p} \quad (13.37)$$

This approximation is only valid for negligible minority charge (i.e., low forward bias). It is based on the picture of a $p-n$ junction that is an insulator for $-l_p < z < l_n$ and a metal otherwise. Then capacitance charging occurs just by positional shifts of l_n and l_p with the external voltage and the depletion capacitance can be derived in the following way

$$C_{\text{dep}} \equiv \frac{\partial Q}{\partial V_{\text{ext}}} \quad (13.38)$$

where Q is the charge per unit area on one side of the junction

$$Q \equiv - \int_{-l_p}^0 \rho dz = + \int_0^{l_n} \rho dz \quad (13.39)$$

From Poisson's equation [Eqs. (10.12) and (10.15)], we have

$$V_{\text{bi}} + V_{\text{ext}} = \frac{1}{\epsilon \epsilon_0} \int_{-l_p}^{l_n} z \rho dz \quad (13.40)$$

Differentiating V_{ext} , by using the chain rule and assuming that only l_n and l_p depend on V_{ext} , on the right-hand side of Eq. (13.40) (this assumption breaks down in forward bias as will be pointed out later), we obtain

$$\epsilon \epsilon_0 = \frac{\partial}{\partial l_n} \left[\int_{-l_p}^{l_n} z \rho dz \right] \frac{\partial l_n}{\partial V_{\text{ext}}} + \frac{\partial}{\partial l_p} \left[\int_{-l_p}^{l_n} z \rho dz \right] \frac{\partial l_p}{\partial V_{\text{ext}}} \quad (13.41)$$

which is equivalent to

$$\epsilon \epsilon_0 = l_n \rho(l_n) \frac{\partial l_n}{\partial V_{\text{ext}}} + l_p \rho(-l_p) \frac{\partial l_p}{\partial V_{\text{ext}}} \quad (13.42)$$

From the definition of the capacitance, we have

$$\begin{aligned} C_{\text{dep}} &= \frac{\partial Q}{\partial V_{\text{ext}}} = \frac{\partial}{\partial V_{\text{ext}}} \left[- \int_{-l_p}^0 \rho dz \right] \\ &= -\rho(-l_p) \frac{\partial l_p}{\partial V_{\text{ext}}} = \rho(l_n) \frac{\partial l_n}{\partial V_{\text{ext}}} \end{aligned} \quad (13.43)$$

Use of Eqs. (13.42) and (13.43) results in

$$\epsilon\epsilon_0 = \left[l_p \frac{\partial Q}{\partial V_{\text{ext}}} + l_n \frac{\partial Q}{\partial V_{\text{ext}}} \right] \quad (13.44)$$

and

$$C_{\text{dep}} = \frac{\epsilon\epsilon_0}{W} \quad (13.45)$$

where W is a function of V_{ext} as shown in Eq. (13.23). This result is, of course, expected and the derivation has only been given to stress its generality, as we have made only the assumption that the charge density $\rho(z)$ does not depend on V_{ext} . This means that the result applies in reverse bias for junctions with arbitrary transitions from *n* to *p* doping. However, the general frequency response of a *p*-*n* junction needs more detailed treatment because of the existence and dynamics of generation-recombination and because of the dependence of $\rho(z)$ on V_{ext} in forward bias, which have not been included into the above derivation.

We will now derive a more detailed description of the dc and ac responses of long *p*-*n* junctions and first treat only the region of $l_n < z < c$ which gives the most important contributions for long diodes because $0 < z < l_n$ covers only a short region. There are, however, exceptions for semiconductors with larger energy gap, in which the region $0 < z < l_n$ (and also $-l_p < z < 0$) can be important as the minority carrier densities and their product may be exceeding the corresponding quantities in the metallic regions. We ignore this for the moment and also ignore the transition region between the insulating and metallic regions. Both approximations will be discussed later.

13.2.3 Steady-State Current in Forward Bias

The following treatment follows that given in Landsberg [12]. In the range $-a \leq z \leq -l_p$, the electric field is negligible. The electronic current in the conduction band is therefore a diffusion current that has its origin in the concentration gradient caused by the recombination (see also Appendix G). Therefore,

$$j_n \approx eD_n \frac{\partial n}{\partial z} \quad (13.46)$$

We now have to use the subscript *n* for the diffusion constant because the constant for holes is different, and we have both types of carriers present. Using Eq. (13.1), we obtain

$$eD_n \frac{\partial^2 n}{\partial z^2} = eU_s \quad (13.47)$$

This second-order differential equation has no explicit solution for the general expression of U_s . We therefore use the approximation of Eq. (9.37), that is,

$$U_s \approx \frac{n - n(-a)}{\tau_n} \quad (13.48)$$

where $n(-a)$ is the electron concentration on the p side. This gives

$$\frac{\partial^2 n}{\partial z^2} = \frac{n - n(-a)}{\tau_n D_n} \equiv \frac{n - n(-a)}{L_n^2} \quad (13.49)$$

Because of the constant coefficients in this equation,

$$n(z) = C_1 e^{z/L_n} + C_2 e^{-z/L_n} + C_3 \quad (13.50)$$

C_1 , C_2 , and C_3 are constants that can be determined by the following boundary conditions: The electron concentration approaches small values at $z = -a$; because a/L_n is large and the exponent becomes very large, C_2 must be 0. Furthermore, because e^{z/L_n} becomes extremely small at $z = -a$, we have $C_3 \approx n(-a)$. Therefore, we simply have to determine C_1 . This can be obtained from Eq. (13.25), which reads (at $z = l_p$)

$$n(-l_p)p(-l_p) = n_i^2 e^{\bar{V}_{\text{ext}}} \quad (13.51)$$

Here $\bar{V}_{\text{ext}} = eV_{\text{ext}}/kT$ and because $p(-l_p) = p(a)$, one obtains

$$n(-l_p) = \frac{n_i^2}{p(-a)} e^{\bar{V}_{\text{ext}}} \approx n(-a) e^{\bar{V}_{\text{ext}}} \quad (13.52)$$

which completes the boundary conditions and results in

$$C_1 = n(-a)(e^{\bar{V}_{\text{ext}}} - 1)e^{l_p/L_n} \quad (13.53)$$

This ends the calculation of the carrier concentration and along with Eq. (13.48) gives

$$U_s = \frac{n(-a)}{\tau_n} (e^{\bar{V}_{\text{ext}}} - 1) e^{(z+l_p)/L_n} \quad (13.54)$$

which permits us to calculate the dc current through a $p-n$ junction.

The first contribution to the current is then

$$\int_{-a}^{-l_p} eU_s dz = \frac{eL_n n(-a)}{\tau_n} (e^{\bar{V}_{\text{ext}}} - 1) \left(1 - e^{(-a+l_p)/L_n} \right) \quad (13.55)$$

Here the last factor is approximately equal to one for long diodes. For short diodes, the boundary conditions need to be changed as discussed below.

The contribution of the depletion region to the electron current is easier to calculate because in the space-charge region the product $n \cdot p$ is about constant and is given by Eq. (13.25). Therefore we can derive this portion of the current using the general expression of Eq. (9.34). The concentrations n and p in the denominator of Eq. (9.34) can be neglected under weak forward or reverse bias conditions. Then U_s is constant over the depletion range, and as indicated in Eq. (13.26)

$$\int_{-l_p}^{l_n} eU_s dz = \frac{ec_n c_p n_i^2 N_{\text{TT}} W}{e_n + e_p} (e^{\bar{V}_{\text{ext}}} - 1) \quad (13.56)$$

Notice that $W = l_n + l_p$ depends on the external voltage according to Eq. (13.23). The contribution of the minority holes for $l_n < z < c$ is obtained in the same way as Eq. (13.55) to give

$$\int_{l_n}^c eU_s dz = \frac{eL_p p(c)}{\tau_p} (e^{\bar{V}_{\text{ext}}} - 1) \left(1 - e^{(-c+l_n)/L_p} \right) \quad (13.57)$$

Note again that the last factor can be approximated by the one for long diodes. The total current is then

$$j = j_{rs} (e^{\bar{V}_{\text{ext}}} - 1) \quad (13.58)$$

where j_{rs} is the sum of the constant factors of Eqs. (13.55)–(13.57) and is called the reverse saturation current (because $j = j_{rs}$ for reverse bias). This is the famous Shockley equation.

13.2.4 AC Carrier Concentrations and Current in Forward Bias

The derivation of the electron concentration under ac conditions proceeds along the lines shown in the previous section. We consider only low frequencies ω and write

$$n(z, t) = n_{dc}(z) + n_1(z)e^{i\omega t} \quad (13.59)$$

Inserting this concentration into the equation of continuity (now including the term $\partial n / \partial t$) and writing n_{dc} for the dc concentration that has been derived before and was denoted by $n(z)$ one gets

$$\frac{\partial^2 n_1}{\partial z^2} = n_1 \left(\frac{1 + i\omega\tau_n}{L_n^2} \right) \quad (13.60)$$

Actually, Eq. (13.60) is also almost identical to this equation and therefore has the solution

$$n_1(z) = \bar{C}_1 e^{z/\bar{L}_n} + \bar{C}_2 e^{-z/\bar{L}_n} + \bar{C}_3 \quad (13.61)$$

where \bar{C}_1 , \bar{C}_2 , and \bar{C}_3 are constants of integration and $\bar{L}_n = D_n \tau_n / (1 + i\omega\tau_n)$. If we now determine the constants of integration as in Eqs. (13.51) and (13.52), then we could arrive at an equation for the ac carrier concentration in precisely the same way as in Eq. (13.55) and can write

$$[(n_{dc}(-l_p) + n_1(-l_p)e^{i\omega t})[p_{dc}(-l_p) + p_1(-l_p)e^{i\omega t}] = n_i^2 e^{\bar{V}_{\text{ext}}} \quad (13.62)$$

as in Eq. (13.51) under the condition that the quasi-Fermi levels in the depletion region follow instantaneously the ac voltage. \bar{V}_{ext} represents now the sum of dc and ac voltage

$$\bar{V}_{\text{ext}} = \bar{V}_{\text{ext}}^{\text{dc}} + \bar{V}_{\text{ext}}^{\text{ac}} \exp(i\omega t) \quad (13.63)$$

The supply of charge carriers in this region is mainly by diffusion through the region of width W . Diffusion over such a distance will typically take the time t_{diff}

$$t_{\text{diff}} \approx W^2/D_n \quad (13.64)$$

This can be seen from solving the equations of continuity and Eq. (13.46) (neglecting the generation-recombination term) and is well known from problems of heat conduction. If the inverse frequency of the ac voltage is longer than t_{diff} , electrons (holes) follow the ac voltage in the depletion region, and we can indeed use boundary conditions analogous to those used for Eq. (13.50). The coefficient \bar{C}_1 can be obtained from Eq. (13.53). Thus we have for

$$n_1(z, \omega) = n_1(-l_p) \exp((l_p + z)/\bar{L}_n) \quad (13.65)$$

and

$$p_1(z, \omega) = p_1(l_n) \exp((l_n - z)/\bar{L}_p) \quad (13.66)$$

with

$$n_1(-l_p)p_{\text{dc}}(-l_p) = n_i^2 e^{\bar{V}_{\text{ext}}^{\text{dc}}} \bar{V}_{\text{ext}}^{\text{ac}} e^{i\omega t} \quad (13.67)$$

where we have assumed $\bar{V}_{\text{ext}}^{\text{ac}} \ll 1$ to obtain linear response. Notice also that \bar{L}_n and \bar{L}_p are complex quantities as opposed to the real diffusion length (denoted without the bar).

It is now advisable to return to the time domain, because we like to insert the results for $n_1(z, t)$ and $p_1(z, t)$ into the various terms of Eq. (13.33), which is also in the time domain. Restoring the time dependence in Eq.(13.65) and evaluating the real part for long diodes and low ω one obtains (in terms of the now real diffusion length L_n and L_p):

$$n_1(z, t) = n_1(-l_p) \exp\left(\frac{(l_p + z)}{L_n}\right) \left[\cos\left(\frac{(l_p + z)\omega\tau_n}{2L_n}\right) \cos(\omega t) + \sin\left(\frac{(l_p + z)\omega\tau_n}{2L_n}\right) \sin(\omega t) \right] \quad (13.68)$$

and proceeding the same way for minority hole on the other side of the junction

$$p_1(z, t) = p_1(l_n) \exp\left(\frac{(l_n - z)}{L_p}\right) \left[\cos\left(\frac{(l_n - z)\omega\tau_p}{2L_p}\right) \cos(\omega t) + \sin\left(\frac{(l_n - z)\omega\tau_p}{2L_p}\right) \sin(\omega t) \right] \quad (13.69)$$

Of course, for the symmetric case $l_n = l_p$, $L_n = L_p$, and so on.

We are now in a position to evaluate the total ac current density at c by integrating Eq. (13.33) over the range $l_n < z < c$ (neglecting for the moment $0 < z < l_n$). The U_s -term can be evaluated using the approximate expression

$$U_s = (p(z, t) - p(c))/\tau_p \quad (13.70)$$

For low frequencies ω the integral then gives

$$2 \int_{l_n}^c e U_s dz \approx - \int_{l_n}^c \frac{\partial p}{\partial t} dz \quad (13.71)$$

with

$$2e \int_{l_n}^c \frac{\partial p}{\partial t} dz \approx -ep_{dc}(l_n) e^{\bar{V}_{ext}} L_p i \omega \bar{V}_{ext}^{ac} e^{i\omega t} \quad (13.72)$$

We see that the first term in Eq. (13.33) is exactly 1/2 and the negative of the second. This means that 1/2 of the diffusion capacitance is cancelled by the recombination that empties the capacitor and is mathematically expressed by the U_s -term. The evaluation of the third term is accomplished in the following way. The minority charge cancels the second remaining 1/2 of the diffusion capacitance, and the majority carrier contribution has been derived approximately in Eq. (13.45) which means

$$2e \int_{l_n}^c \frac{\partial n}{\partial t} dz \approx \epsilon \epsilon_o / l_n \left(\frac{\partial V}{\partial t} \right) \quad (13.73)$$

This equation with l_n from Eq. (13.21) is only valid for $V_{ext} < V_{bi}$ because l_n otherwise vanishes or becomes imaginary. An exact numerical evaluation of this term gives about the same value as obtained for small V_{ext} with a slight increase close to $V_{ext} \approx V_{bi}$ and a decrease for still larger forward bias.

Note that in long diodes the diffusion capacitance is practically cancelled by the U_s -terms and the total capacitance per unit square is given approximately by Eq. (13.73). Long diodes do exist in nature. They require a strong recombination mechanism as it occurs by photon emission in III-V semiconductors or when a high density of Shockley-Read-Hall centers is present; otherwise minority carriers will always reach the contacts and our derivations are invalid, particularly for high forward bias. In most practical cases, there will be a transition from long to short diode behavior as the forward bias is increased. This has masked the obvious contradiction in many previous texts that the existence of a diffusion capacitance was postulated for long diodes. What was actually measured if a diffusion capacitance was seen was short diode behavior. Because silicon is an indirect semiconductor with no first-order radiative (light) emission, most silicon diodes will show short diode behavior as the forward bias increases. We therefore add here a section for short diodes as a tribute to this most important semiconductor material.

13.2.5 Short Diodes

Most silicon diodes produced these days are short. This is therefore a very important special case. Short diodes are by definition short enough to permit injected minority carriers to reach the contacts (holes at c and electrons at $-a$), meaning

that under the given forward bias the minority density close to the contacts exceeds significantly the equilibrium density. The boundary condition that leads to a vanishing constant C_2 in Eq. (13.50) must therefore be changed.

Ohmic contacts can often be idealized by attributing to them infinite recombination speed, which can also be expressed by a vanishing lifetime τ_n or τ_p that leads to a vanishing minority concentration at the contact (but still finite concentration close to the contact). That is,

$$n_1(-a) = p_1(c) = 0 \quad (13.74)$$

which gives then in the frequency domain (complex diffusion length!)

$$n_1(z, \omega) = n_1(-l_p) \frac{\sinh(-a-z)/\bar{L}_n}{\sinh(-a-l_p)/\bar{L}_n} \quad (13.75)$$

which is well known and found in many texts.

The calculation of the total current must now contain an additional term, the minority carrier current at the contact, which can be usually (not too high forward bias) approximated as a diffusion current only

$$j_n(c) = eD_n \frac{\partial n_1(z, \omega)}{\partial z} \quad (13.76)$$

and

$$j_p(-a) = -eD_p \frac{\partial p_1(z, \omega)}{\partial z} \quad (13.77)$$

with n_1 from Eq. (13.75) and a similar equation for p_1 .

Using now Eq. (13.33) and the additional terms of Eqs. (13.76) and (13.77), one can see that a finite and exponentially increasing diffusion capacitance does now exist because the U_s -terms are small owing to the shortness of the diode. The minority carrier concentration can now rise significantly, even close to the contact, because the exponential decay of the long diodes is replaced by the hyperbolic functions. The minority charge is neutralized in the metallic regions by the majority carriers; the equation of Poisson does therefore not set any limitations, and the junction capacitance rises exponentially with forward bias V_{ext} . Exact numerical results for a typical (but still symmetric) $p-n$ junction are shown in Figure 13.14. The transition between short and long diode is simulated by varying the minority lifetime $\tau_n = \tau_p$. For infinite lifetime, the diode is, of course, short, no matter what its physical length really is and for the smallest values of lifetimes the diode is long even if its physical length is only one micrometer.

13.2.6 Recombination in Depletion Region

We have up to now considered only so-called ideal diodes that are defined by the neglect of the integration over the depletion region in Eq. (13.33). The complication in this range is that U_s is not given by Eq. (13.54). Nor can it be approximated by Eq. (9.35) if the forward bias becomes large. However, for a symmetric

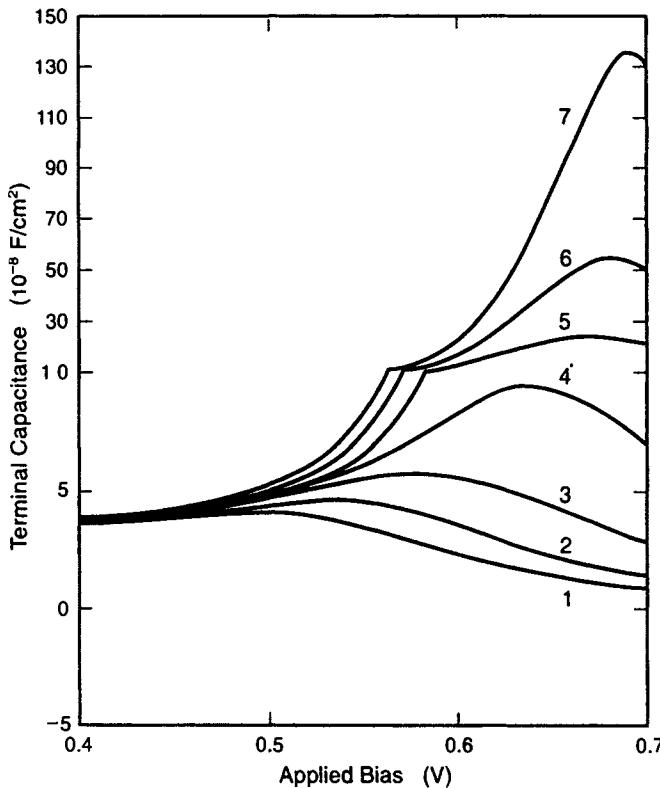


Figure 13.14 Terminal capacitance of a $p-n$ junction versus applied bias for various minority lifetimes $\tau_n = \tau_p = \tau = 0.01, 0.03, 0.1, 0.3, 1, 3$ and ∞ nsec (corresponding to curve 1-7, respectively).

diode, we can rewrite Eq. (9.35) under these conditions as

$$U_s = \frac{c_n N_{TT} np}{n + p} \quad (13.78)$$

with $c_n = c_p$ because of the symmetry. It is then a simple exercise to find

$$\int_0^{l_n} U_s dz \propto p(0) \quad (13.79)$$

and because

$$p(0) \propto \exp(eV_{ext}/2kT_L) = \exp(\bar{V}_{ext}/2) \quad (13.80)$$

we obtain a current contribution proportional to $\exp(\bar{V}_{ext}/2)$.

The factor $1/2$ that now appears in the exponent for the terminal current has been called non-ideality factor. This contribution of the recombination in the depletion region can be dominant and, in close approximation, equal to the total

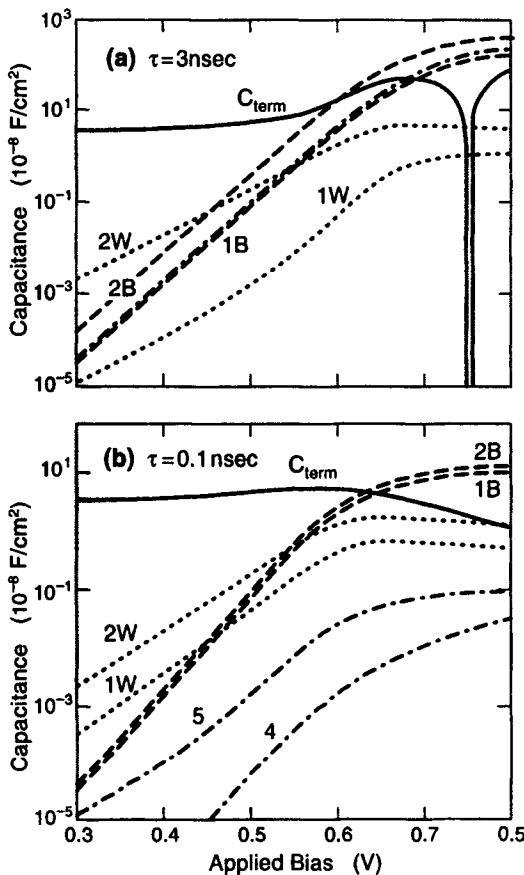


Figure 13.15 Contributions of various terms (defined in the text) to the terminal capacitance of a $p-n$ junction.

diode current for semiconductors with relatively large band gap E_G (e.g., Silicon, GaAs). For large E_G we have over a wide range of applied bias

$$p(0) \gg p(l_n) \quad (13.81)$$

as is obvious from Figure 13.13. Therefore Silicon $p-n$ junctions can show a pronounced non-ideality in their forward current-voltage characteristic if recombination in the depletion region is important, while Germanium diodes do not. Note, however, that recombination is often not important in very short diodes.

Naturally, the non-ideality (the contribution of the range $-l_p < z < l_n$) becomes important for the diode capacitance under exactly the same conditions. Figure 13.15 shows contributions of the range $0 < z < l_n$ to the diffusion capacitance (labeled 2W) and the U_s -term (labeled 1W), which represents a negative capacitance. The contributions of the region $l_n < z < c$ (labeled 2B for the diffusion capacitance and 1B for the negative U_s -term) are shown for comparison,

all again for the parameters of our typical diode. One can see that, depending on bias, either the B- or the W-terms dominate, with the B-terms winning as the forward bias becomes very large. However, for low bias the W-terms are the most significant.

13.2.7 Extreme Forward Bias

In extreme forward bias, when the external voltage exceeds the built-in voltage, an effect becomes dominant that we have not yet discussed: it can no longer be assumed that all the external voltage drops in the depletion region. Of course, for extreme forward bias, no really depleted region exists; and the region at the junction that might still show some depletion becomes very thin. The question becomes then, how does one calculate the important quantity $p(l_n)$ that has entered Eqs. (13.30), (13.31), and (13.35) as well as Eq. (13.51)? The relationship for $p(l_n)$ (and also $n(-l_p)$) is so important that it is sometimes called the “law of the junction.” In Eq. (13.35), this law stated that the minority carrier charge at l_n was equal to the equilibrium value multiplied by a Boltzmann-type factor $\exp(eV_{\text{ext}}/kT_L)$. This was based on the assumption that the external voltage was reducing the built-in voltage by V_{ext} thus leading to the exponential increase. What law of the junction do we have to use if the external voltage does not drop in its entirety in the range $-l_p < z < l_n$? The answer to this is more complicated, as one would expect. One could easily argue that one just has to deduct the voltage drop in the regions $-a < z < -l_p$ and $l_n < z < c$. One could also argue that one deducts the quasi-Fermi level difference of the same regions because these play a major role in calculating current and carrier density. Trying both by exact numerical solution, one can see that only for extremely long diodes does one get the same result; however, in general and, particularly for short diodes, the result is different. Assuming $T_c = T_L$ and using the Einstein relation, we can write

$$j_n = en\mu F + \mu kT_L \frac{\partial n}{\partial z} \quad (13.82)$$

Because in forward bias the fields are usually small, the electron temperature will indeed be close to the lattice temperature as assumed. (As a side note, if $T_c > T_L$ or $T_c < T_L$ —and both are possible—the electrons can give energy to the phonon system or take it from it, respectively. One then needs to include also the phonon fluxes, and then the so completed theory of *p*-*n* junctions becomes a PhD thesis problem.)

Using Eq. (5.26), but for a general conduction band energy E_c [assumed to be 0 in Eq. (5.26)]

$$n = N_c \exp((E_{\text{QF}}^n - E_c)/kT_L) \quad (13.83)$$

we obtain by using Eq. (13.82)

$$j_n = -e\mu n \frac{\partial E_{\text{QF}}^n}{\partial z} \quad (13.84)$$

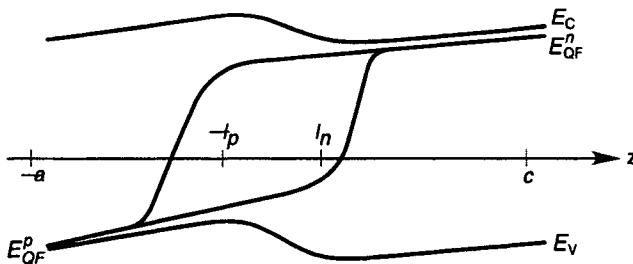


Figure 13.16 Band edges and quasi-Fermi levels of a p - n junction in extreme forward bias; no symmetry assumed here.

and a similar expression for holes. From this we can see that the change of the quasi-Fermi levels and of the electrostatic potential are, in general, not equal

$$E_{\text{QF}}^n(z) - E_{\text{QF}}^n(0) = - \int_0^z j_n(z') / (e\mu n(z')) dz' \quad (13.85)$$

whereas the difference in electrostatic potential is given by

$$V(z) - V(0) = - \int_0^z F(z') dz' \quad (13.86)$$

which is only for very long diodes equal to the difference of the quasi-Fermi levels of Eq. (13.85) because the diffusion current density can then be neglected over most of the range $l_n < z < c$.

Figure 13.16 shows the band edges and quasi-Fermi level differences in an exaggerated way. Clearly, the band edges of conduction and valence bands, that are proportional the electrostatic potential, have two extrema (close to $-l_p$ and l_n), whereas the quasi-Fermi levels increase (decrease) monotonically. Figure 13.16 also holds the answer to our question of how to obtain the voltage V_n that determines the density of injected minority electrons (V_p for holes) if we still want to use the law of the junction

$$p(l_n) \approx p_{no} \exp(eV_p/kT_L) \quad (13.87)$$

as before. Figure 13.16 makes it plausible that at the point l_n , we have to add to the energy eV_{ext} the difference of quasi-Fermi levels $E_{\text{QF}}^p(-a) - E_{\text{QF}}^p(l_n)$. We also have to deduct the electrostatic potential energy difference $e(V(l_n) - V(c))$ from eV_{ext} , both to obtain the now relevant energy difference (relative to equilibrium) that determines the minority carrier concentration at l_n

$$eV_p = eV_{\text{ext}} + E_{\text{QF}}^p(-a) - E_{\text{QF}}^p(l_n) - e(V(l_n) - V(c)) \quad (13.88)$$

which can also be derived algebraically as shown by Laux and Hess [13]. eV_n , which determines the electron minority injection, can be derived similarly and is for asymmetric diodes different from eV_p . This brings home the profound difference between quasi-Fermi levels and the electrostatic potential energy, which becomes very important for large forward bias and not-too-long diodes.

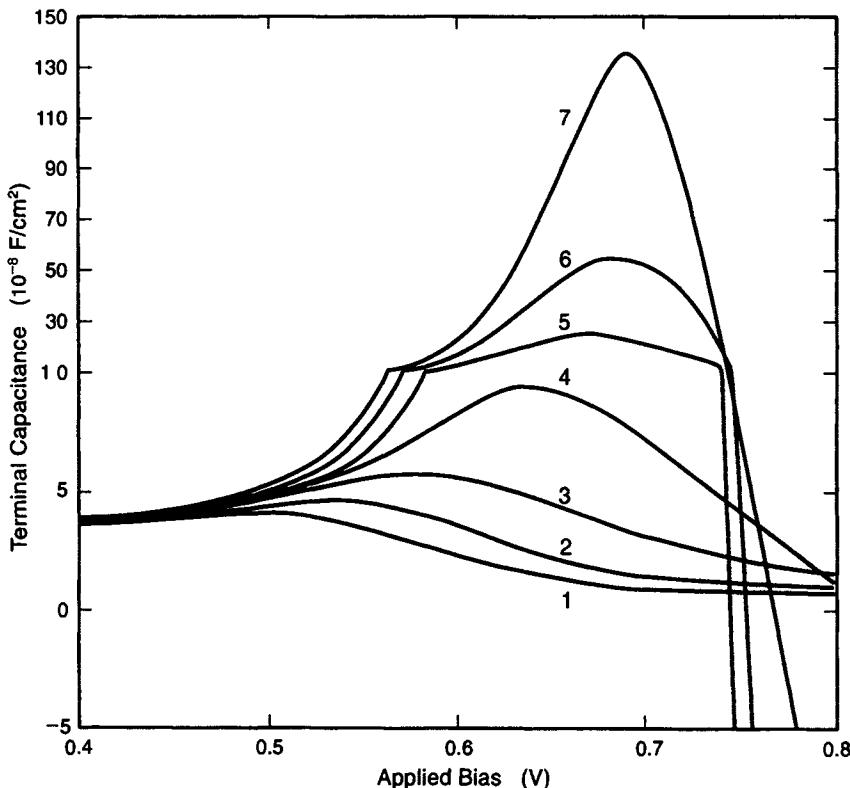


Figure 13.17 Same as Figure 13.14, but for a larger range of applied forward bias.

These voltage and quasi-Fermi level drops across $-a < z < l_p$ and $l_n < z < c$ are also modulated by alternating currents. For example, when increasing the forward bias, these drops act against minority carrier injection and according to Eq. (13.88) significantly reduce injection. This opposition to current flow represents an inductance that indeed is always observed in strongly forward biased $p-n$ junctions. We call it here the sfb-inductance. The complete impedance at all biases for our typical $p-n$ junction is shown in Figure 13.17.

13.2.8 Asymmetric Junctions

The case of asymmetric $p-n$ junctions (i.e., $N_A \neq N_D$ and $\mu_n \neq \mu_p$) can be treated similarly as we have done above with one major difference: The symmetric treatment had the nice advantage that the displacement current and the diffusion capacitance could be clearly separated by the exact expression of Eq. (13.33), which has these two contributions separated in the second and third terms. For asymmetric junctions, an exact expression like this that separates the displacement current cannot be found; the important equality $j_n(0) = j_p(0)$ that was used

to derive Eq. (13.33) is not valid. Therefore, the integrations in Eq. (13.33) cannot be performed from the junction to the point c at the contact. Such an integration performed for the asymmetric diode (see, e.g., Sah [16]) is not only invalid, but also mixes hopelessly the contributions of diffusion capacitance and depletion capacitance (corresponding to the displacement current). One can, however, obtain an equation exactly as Eq. (13.33), if one performs the integrations not from the junction at $z = 0$ but from a point $z = b$ that is defined by

$$j_n(b) = j_p(b) \quad (13.89)$$

which unfortunately depends on the current densities that just have to be determined. It is also possible that no such point exists. Laux and Hess [13], however, found through numerical experimentation that an approximately valid result is obtained if the integrations are all performed from the point b at which we have $n = p$ in equilibrium. The reason for this can be found in Eq. (13.79), which suggests that, at least for large band gaps, the current densities and carrier densities are closely related. The point $z = b$ is not at the junction and, in fact, can be considerably far away from the junction (i.e., $b \neq 0$). Using this new integration limit, all the qualitative features we have described stay then the same, but the treatment loses some of its mathematical rigor and transparency.

In passing, we discuss briefly a problem of rigorous $p-n$ junction theory that is often disregarded because it calls for a numerical solution and is difficult to explain explicitly. The problem is posed by the fact that the $p-n$ junction cannot be strictly divided into a depleted and a metallic part. There needs to be a transition region, and this region can actually be rather wide. This transition region may be defined by the distance between l_n as defined above and the actual position of zero electric field l_n' . We have shown in Figure 13.16 that l_n is not the place where $F = 0$. There exists this “murky” semiconducting range $l_n < z < l_n'$ (and a similar range on the p side) over which local charge neutrality is significantly violated. This range is much wider than the Debye length, L_D , often more than an order of magnitude larger as shown by numerical simulations. The reason is that the Debye length is only derived for one type of charge carrier (majority carriers) and does not deal with the subtleties around the extrema of the potential energy—the place where the field vanishes (see Figure 13.16). It is important to note that the changes in mobile carrier densities δn and δp owing to external bias are not equal as often assumed. In fact

$$\delta n \neq \delta p \quad \text{for} \quad -l_p' < z < l_n' \quad (13.90)$$

which is important to know when calculating differences between field and diffusion currents. If such a need exists, the reader is advised to use numerical simulators.

13.2.9 Effects in Reverse Bias

Reverse bias is easier to understand than forward bias and has been treated extensively in elementary texts. The diffusion capacitance in this case is negligible and the displacement current is well described by the depletion approximation. We therefore deal here only with special effects as they arise if one switches from forward to reverse and observes the transients. These current transients and capacitance transients have been pioneered by Tasch and Sah, as described in [16] and have been developed by Lang to the so-called *deep level transient capacitance spectroscopy* that is described below.

If we assume that all the traps are filled by electrons at $t = 0$ (because of the forward bias), then the solution of Eq. (9.29) is straightforward, and

$$n_T = N_{TT} \left[\frac{1}{1 + e_n/e_p} + \left(1 - \frac{1}{1 + e_n/e_p} \right) e^{-(e_n + e_p)t} \right] \quad (13.91)$$

Should the traps be filled to any other degree at $t = 0$, similar equations can be obtained with ease. For $e_p \approx 0$ and $n_T = N_{TT}$ at $t = 0$ we have the simple result

$$n_T = N_{TT} e^{-e_n t} \quad (13.92)$$

which will be used in the following for demonstration.

Consider now the switching cycle shown in Figure 13.18. At $t < 0$ the junction is forward biased, as shown in Figure 13.18a. The donors (acceptors) are partially filled with electrons in accord with the position of the quasi-Fermi levels. The trap, which is shown in the middle of the gap, is almost totally filled with electrons because the conduction band is flooded with electrons that can be captured. At $t = 0$ the junction is switched to reverse bias. The electrons are then removed from the conduction band within the dielectric relaxation time, which for higher electron densities is of the order of 10^{-12} s. Electrons in shallow donors are also emitted with about the same time constant and leave the junction region. To simplify the calculation we assume, as stated above, that $e_p = 0$ (electron trap only) and furthermore $N_A \gg N_D$, which means we consider an abrupt p^+ -*n* junction (or Schottky barrier) whose depletion width can be approximated by $W \approx l_n$. We also assume that the trap density is much smaller than the density of shallow donors that form the junction. At $t = 0$, all traps are filled. In the following we will calculate the capacitance change due to the dynamics of electron release.

The depletion width in the absence of these traps is denoted by W_0 . Then we have

$$V_{bi} + V_{ext} = -\frac{e}{\epsilon \epsilon_0} \int_0^{W_0} z N_D^+ dz \quad (13.93)$$

where N_D^+ denotes the density of positively charged donors (the conduction band and donors are emptied immediately after $t = 0$). Our goal is to calculate the

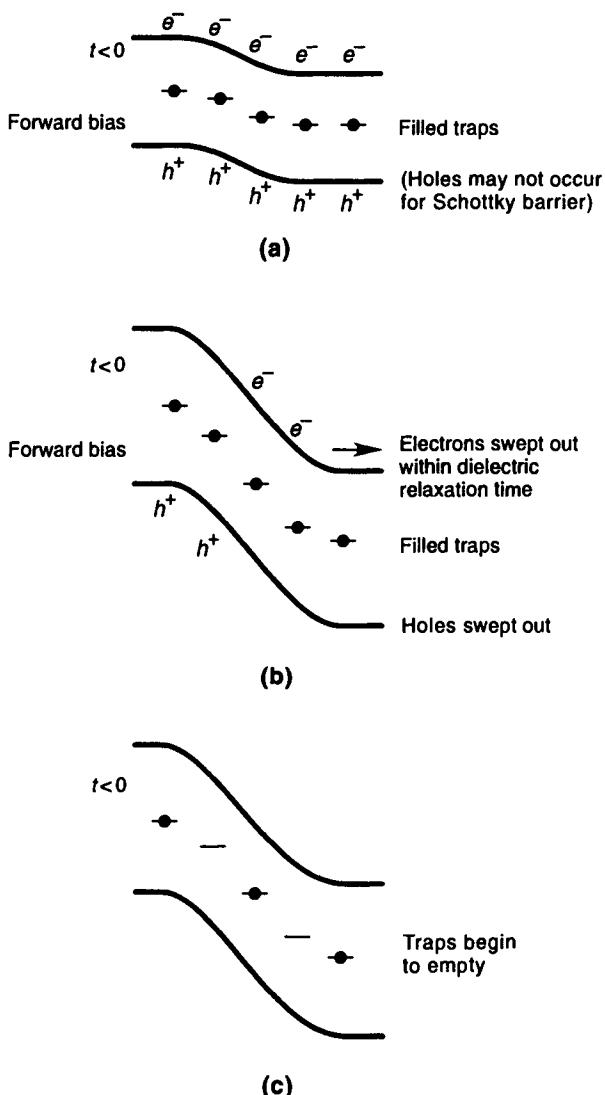


Figure 13.18 Switching cycle of a $p-n$ junction (or Schottky barrier) and the corresponding filling and emptying of electron traps: (a) forward bias (filling); (b) reverse bias (electrons are swept out of the conduction band); and (c) release of electrons from traps. (Typical time constant for a trap at midgap in silicon at $T = 300$ K is ≈ 1 ms.)

depletion width W , including the time-dependent filling of traps, which is given by the equation

$$V_{\text{bi}} + V_{\text{ext}} = -\frac{e}{\epsilon \epsilon_0} \int_0^{W_0} z(N_D^+ + N_{\text{TT}}(1 - e^{-e_n t})) dz \quad (13.94)$$

We have tacitly assumed that the traps are also donorlike; that is, the positively charged traps are the empty traps whose density is $N_{\text{TT}}(1 - e^{-e_n t})$.

Because N_{TT} is small compared to N_D , we have $W = W_0 + \Delta W$ with $\Delta W \ll W_0$ (this can also be seen from our result), and therefore we can use perturbation theory [as for Eq. (1.34)]:

$$\begin{aligned} -\frac{\epsilon \epsilon_0}{e} V_{\text{bi}} + V_{\text{ext}} &= -\int_0^{W_0} zN_D^+ dz + \int_{W_0}^{W_0 + \Delta W} zN_D^+ dz \\ &\quad + \int_0^{W_0} zN_{\text{TT}}(1 - e^{-e_n t}) dz \end{aligned} \quad (13.95)$$

where we have neglected the second-order term:

$$\int_{W_0}^{W_0 + \Delta W} zN_{\text{TT}}(1 - e^{-e_n t}) dz \quad (13.96)$$

Using Eq. (13.95), we obtain

$$\int_{W_0}^{W_0 + \Delta W} zN_D^+ dz = -\int_0^{W_0} zN_{\text{TT}}(1 - e^{-e_n t}) dz \quad (13.97)$$

which gives for small ΔW

$$W_0 \Delta W N_D^+(W_0) = -\int_0^{W_0} zN_{\text{TT}}(1 - e^{-e_n t}) dz \quad (13.98)$$

We can now use the mean value theorem of integral calculus to evaluate the right-hand side of Eq. (13.98), we can numerically integrate it, or, in the case of constant N_{TT} , explicitly integrate. If W_1 is a suitable mean value of z , the result is

$$W_0 \Delta W N_D^+(W_0) = -W_0 N_{\text{TT}}(W_1)(1 - e^{-e_n t}) W_1 \quad (13.99)$$

Using the capacitance formula, Eq. (13.45), one gets for the relative capacitance change

$$\frac{\Delta C}{C_0} = \frac{N_{\text{TT}}(W_1)}{N_D^+(W_0)} (1 - e^{-e_n t}) \frac{W_1}{W_0} \quad (13.100)$$

In other words, the time-dependent capacitance gives us information on the density of deep traps as well as the emission rates e_n . This fact is used to determine e_n and N_{TT} by so-called transient capacitance methods that measure capacitance versus time and temperature. The temperature is varied to put e_n into the appropriate range accessible to measurement. Remember that e_n depends exponentially on temperature.

13.3 HIGH-FIELD EFFECTS IN SEMICONDUCTOR JUNCTIONS

In the previous sections, we have largely ignored the problem of electron heating, and we have just pointed out occasions where temperature—the temperature of the charge carriers or the temperature of the crystal lattice—must be used. In this section we treat the high-field effects in more detail. On the whole there are four major consequences of electron heating that are of importance in semiconductor devices:

1. The carrier mobility and diffusion constants change because of the higher scattering rates at higher electron temperatures.
2. The electrons (holes) are redistributed in \mathbf{k} space among the various valleys of the band structure, which leads to changes in the effective mass and therefore to changes in all the transport coefficients.
3. Hot electrons can escape over barriers by thermionic emission easier than cold electrons and transfer to neighboring layers. We call this effect real space transfer.
4. Finally, electrons (holes) can, if their energy is very high, impact ionize and even create crystal defects. The most important effect is impact ionization. In very high fields ionization can also occur by band-to-band tunneling (the Zener effect).

All of these effects will be treated in this section by discussing certain examples. We start with the increase of the phonon scattering rate in high fields as a consequence of the higher carrier temperature T_c . This case has already been extensively treated via Eqs. (11.23) through (11.30). However, at junctions there is the additional problem of having a built-in field as well as an external field, and the question arises as to whether the built-in field can contribute to electron heating. The answer to this question is not straightforward and will be discussed below.

13.3.1 Role of Built-In Fields in Electron Heating and $p-n$ Junction Currents

There has been some controversy in the literature regarding the effects of built-in fields arising from (mobile or fixed) space charge. The question is: Can a built-in field heat up the energy distribution of the electrons? Additional questions not unrelated to the above have arisen concerning the diffusion current and the modification of the Einstein relation in the presence of high electric fields. Although a precise answer to this question will in most cases require Monte Carlo simulations, one can formulate the balance equations in an appropriate way and derive approximate answers on a physical basis. It is the purpose of this section to clarify how these equations should be used in the study of hot electrons in the space-charge region of Schottky barriers or $p-n$ junctions. The general princi-

ples are inherent in Eq. (11.18). From this equation one can see that it is the term $\mathbf{F} \cdot \mathbf{j}$ that is responsible for the heating of the electrons and not the electric field alone. If $\mathbf{F} \cdot \mathbf{j}$ is zero, then $T = T_L$, as can be seen from Eq. (11.18). The presence of the built-in field alone is therefore not sufficient to heat electrons. As long as the field and diffusion current balance each other and the total current (field plus diffusion) is zero, no average heating effect will be achieved. True, some electrons will slip down the energy band to lower potential energy and get hot in the process. Others, however, will diffuse against the barrier and lose kinetic energy. If the system is not perturbed from the outside, phonons will be emitted and absorbed at the same rate, and a dynamic equilibrium will be achieved. However, as soon as this equilibrium is perturbed by an external voltage (or other factors such as illumination) the situation changes, a current flows, and the electrons can be heated. However, Eq. (11.23) cannot be used now, and we have to return to Eq. (11.18) to calculate the electron temperature. It is important, then, to include the diffusion term in the current. For steady state ($\partial T_c / \partial t = 0$), Eq. (11.18) reads (in one dimension)

$$\left(en\mu F + e \frac{\partial nD}{\partial z} \right) F = \frac{3}{2} \frac{k}{\tau_E} n(T_c - T_L) \quad (13.101)$$

where τ_E is the energy relaxation time, given by Eq. (11.19), and Eq. (11.10) has been used for the current density j . Equation (13.101) does not in general have an explicit solution for T_c because the carrier concentration can be a complicated function of z , and the mobility and diffusion constant are also functions of the electron temperature. The point we want to make here is very clear, however: If the diffusion current balances the field current, the left-hand side of Eq. (13.101) vanishes and $T_c = T_L$, as is the case for the unbiased $p-n$ junction.

We have ignored here yet another problem. Because, in $p-n$ junctions, the electric field is also dependent on the space coordinate (see Figure 13.11), the whole derivation of Eq. (11.18) does not hold. The drifting carriers transport energy from places of high field to places of low field, and electronic heat conduction plays a role if T_c becomes a function of z . This gives rise to additional terms in the energy balance equation, which are derived in Appendix E. These terms complicate the calculation of T_c greatly, and the numerical solution of the differential equation together with the other device equations is time consuming (in some instances, it is almost as long as the more precise Monte Carlo method). Approximations are possible and have been discussed in the literature (see Higman and Hess [10] and references therein).

For the calculation of the current in a $p-n$ junction, including hot electron effects, there exists the additional problem that according to Eq. (8.39) the quasi-Fermi level becomes a function of z as the electron temperature does. All of these taken together require a numerical treatment if a precise solution for j in a $p-n$ junction is desired. However, because j depends weakly on T_c not much has been done with this problem and little appears in the literature.

The reason that j does not depend strongly on T_c is the following. For for-

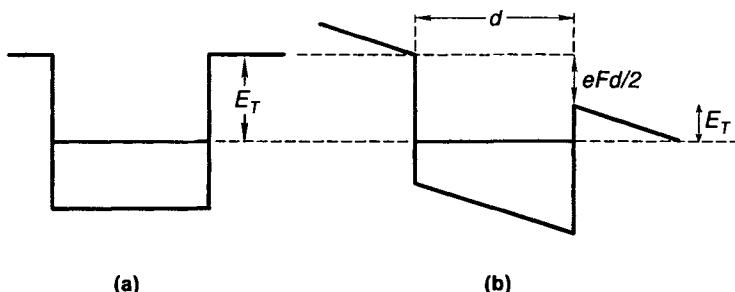


Figure 13.19 Square well trap energy E , without (a) and with (b) an external electric field.

ward bias the electric fields are small and $T_c \approx T_L$. For reverse bias the fields are quite large and the electrons are heated greatly. However, in this case the speed of the electrons through the depletion layer (which depends on T_c) is not the limiting factor for the current. As can be seen from Eq. (13.56) it only matters how fast carriers are generated. The speed of the carriers does not even enter Eq. (13.56) because the field current is not involved in its derivation within the approximations that lead to Eq. (13.56). As we will see in Chapter 15 the situation is very different for transistors where the increase in T_L influences the current significantly.

For $p-n$ junctions there is, however, a direct influence of high fields on the current via the generation (emission) rates. As can be seen from Eq. (9.27), the emission rate e_n (and the same is true for holes) depends exponentially on the energy difference E_T between the trap level and the conduction band edge. This difference depends on external fields, as can be seen from Figure 13.19 for the case of a trap with a square well-like potential. The energy level does not change with external fields in first-order perturbation theory. (M_{mm} in Eq. (1.38) does not change with external fields, as one can see easily by putting the zero of the energy at a reference point in the middle of the well.) However, the distance of the level to the conduction band edge is reduced by the amount $eFd/2$ for wells of width d . This can be seen from Figure 13.19b. The amount of reduction is different for different forms of well potentials, but can always be calculated by searching for the maximum of the potential at the right-hand side of the well. In three dimensions the problem is a little more involved. This effect of emission over lowered barriers is sometimes called the Poole-Frenkel effect. The electron can, of course, also tunnel directly out of the well, and this adds to the rate e_n if the tunneling rate becomes high (in high fields). The total rate is then the sum of the Poole-Frenkel emission and tunneling, which can be calculated as shown in Appendix A.

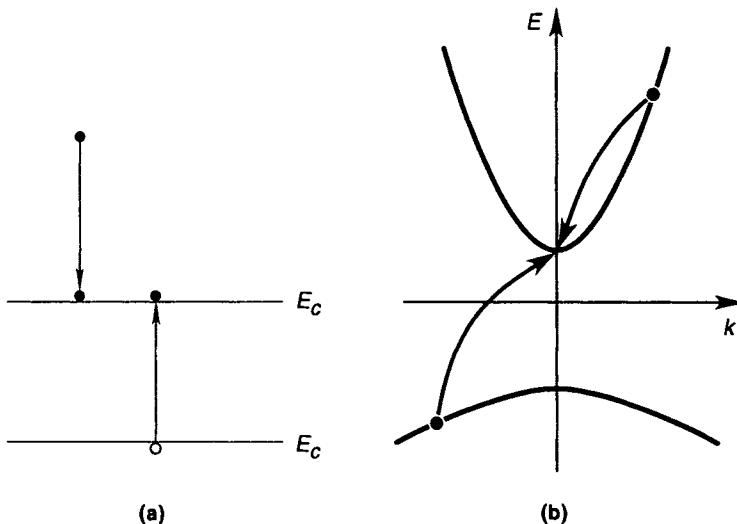


Figure 13.20 The impact ionization process in (a) real space and in (b) \mathbf{k} space.

13.3.2 Impact Ionization in $p-n$ Junctions

Impact ionization is not an effect specific to $p-n$ junctions. However, because, in Ge, Si, and GaAs, it occurs only at very high electric fields ($> 10^5$ V/cm), it is mostly observed in reverse-biased $p-n$ junctions, and we therefore treat it in this chapter.

The impact ionization process is the inverse of the Auger process; that is, an electron (hole) in the conduction (valence) band gains energy by some means and becomes so highly energetic that it can create an electron-hole pair by colliding with an electron in the valence band and exciting it to the conduction band. The process is schematically shown in Figure 13.20. The process can be viewed as a scattering event; and as long as the Golden Rule can be applied, we expect momentum and energy to be conserved. Energy conservation demands that the energy of the ionizing particle is at least as large as the energy gap E_G . From Figure 13.20b, however, we can immediately see that if momentum is to be conserved too, the energy of the ionizing particle must be substantially larger than E_G .

For the simplest case of parabolic isotropic bands with equal mass, one would obtain a threshold energy $E_{th} = 3/2E_G$ for ionization to be possible. For single-valued (but otherwise arbitrary) band structures (conduction and valence bands) it has been shown by Anderson and Crowell [2] that momentum and energy conservation are equivalent to the requirement that the sum of group velocities of the final particles is zero. Using this criterion, threshold energies have been calculated for various materials by Bude [4]. These energies become a function of crystallographic direction and assume very complicated shapes if plotted in \mathbf{k} space in accord with the complicated $E(\mathbf{k})$ relations (see Chapter 3). Notice,

however, that the $E(k)$ relation is not single valued for most semiconductors in the interesting energy range. Furthermore, the impact ionization rates may be high at high energies, making collision broadening important. We know from the uncertainty principle that as soon as \hbar/τ_{tot} reaches values comparable to the energy scale that is interesting in the given problem, the δ function approximation in the Golden Rule becomes questionable [see Eq. (1.41)], and the δ function has to be replaced by a broadened Lorentzian as is well known from optics (see Bethe and Jackiw [3]). The total scattering rate (owing mostly to phonons and impact ionization) can be enormous at the energies typical for the ionization process ($\approx 2\text{--}3$ eV in silicon), as is seen from Figure 7.8. The broadening effect will then smear out the threshold for impact ionization. One can therefore relax the momentum conservation requirement. Actual calculation, however, requires the treatment we have discussed in connection with Monte Carlo simulations. Kane [11] has numerically estimated that the ionization rate $1/\tau_I$ for electrons in the conduction band of silicon has a rather soft threshold above the energy $E_{\text{th}} = E_G$, as is shown in Figure 13.21. His results are in reasonable agreement with experiments.

Even more complicated than the calculation of the threshold is the calculation of the ionization rate per unit time $1/\tau_I$ above threshold. For not too large rates, $1/\tau_I$ can be calculated using the Golden Rule (as we did for impurity scattering) but now including transitions in between different bands.

A transparent treatment has been given by Ridley [15], who also discusses a formula previously derived by Keldysh. For real band structures these formulas have only a limited validity, and a numerical calculation (as performed by Kane) is necessary to obtain $1/\tau_I$. We use in the following the Keldysh formula for $1/\tau_I$. However, we regard the constants in the formula as freely adjustable parameters, not given by first principle theory.

The Keldysh formula suggests

$$\frac{1}{\tau_I} = \frac{B}{\tau_I(E_{\text{th}})} \left(\frac{E - E_{\text{th}}}{E_{\text{th}}} \right)^p \quad \text{for } E > E_{\text{th}}, \text{ and zero otherwise} \quad (13.102)$$

Here B is a dimensionless parameter and $1/\tau_I(E_{\text{th}})$ is the scattering rate at threshold. It may be lumped together with B to form an adjustable parameter. The second parameter, the exponent p , is typically between 1 and 2. For more precise purposes this parameterization is insufficiently accurate and the full treatment (as given by Kane [11]) is necessary. A recent description of these rates and thresholds has been given by Bude [4].

The energy dependence of the ionization rate is not the only quantity that is needed in detail to understand devices involving impact ionization. What one needs to understand devices is the increase in current I with time (or over distance) owing to ionization. This increase is defined by coefficients α_t and α_z as

$$\frac{\partial I}{\partial t} = \alpha_t I \quad (13.103)$$

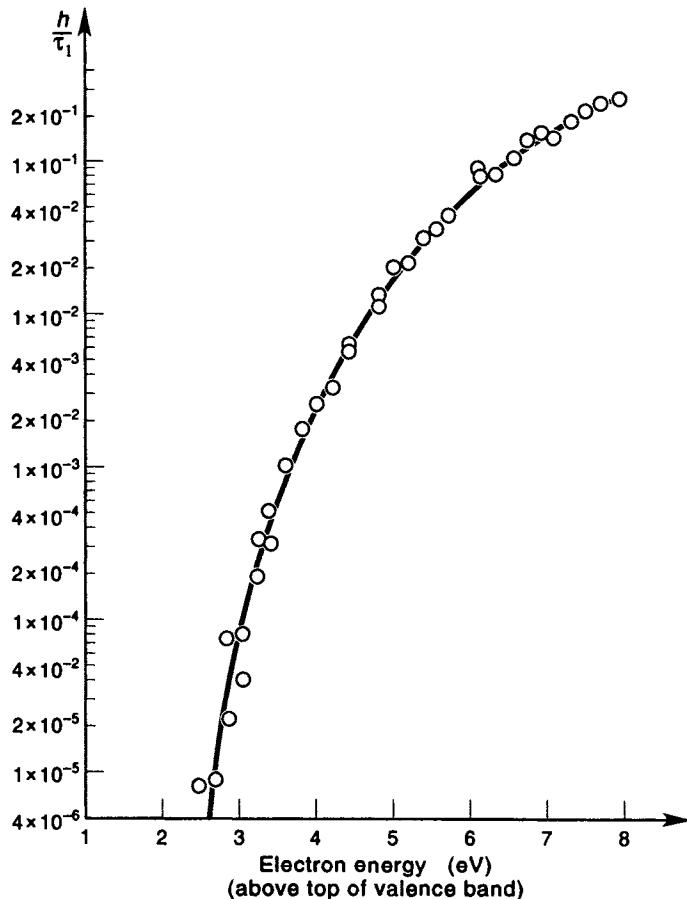


Figure 13.21 Impact ionization rate for electrons in silicon as a function of energy. [Source: After Kane [11].]

and

$$\frac{\partial I}{\partial z} = \alpha_z I \quad (13.104)$$

From these definitions it follows that α_t represents an average rate of ionization per unit time or per unit distance (α_z). The average is to be taken over the energy distribution and we therefore have

$$\alpha_t = \int_{-\infty}^{\infty} d\mathbf{k} \frac{1}{\tau_i} f \Big/ \int_{-\infty}^{\infty} d\mathbf{k} f \quad (13.105)$$

with $1/\tau_i$ being given by Eq. (13.102) or from Figure 13.21 and f being the electron (hole) energy distribution function. For the special case of steady state and time and space independence, α_z can be calculated from α_t by dividing α_t

by the average drift velocity. This follows from the chain rule of differentiation and the definitions. Therefore, $\alpha_z \propto \alpha_t$.

An inspection of Eq. (13.105) shows that the functional dependence of $1/\tau_I$ on the energy influences α_t only rather weakly as long as $1/\tau_I$ is given by Eq. (13.102), whereas the energy dependence of the distribution function above E_{th} is exponential and influences α_t greatly. One can put this fact into other words: An electron will go through the history of scattering events and accelerations before the threshold energy is overcome. It is this history that determines the probability of finding an electron above threshold. Then after the electron reaches some energy above threshold, it impact ionizes. It is therefore the history of the electron at the energies below threshold that is of utmost importance for α_t , because it determines the high-energy tail of the distribution function. This high-energy tail will be discussed now from various points of view.

Wolff assumed that the distribution function in Eq. (13.105) can be approximated by a spherical symmetric distribution (i.e., the drift term does not matter). We derived f_0 in Chapter 11 by using the electron temperature approximation. For simple parabolic bands and phonon scattering, we can therefore use the carrier temperature as given by Eq. (11.24). However, this approach is not self-consistent; we need to include impact ionization itself as a scattering mechanism. In other words, one needs to include an average energy loss owing to impact ionization into the energy balance equation. If the electrons lose all their energy owing to the ionization (which is only approximately true) the loss is given by

$$\int_{-\infty}^{\infty} dk \frac{E}{\tau_I} f_0 \quad (13.106)$$

This term is easy to evaluate, at least for steplike ionization rates $1/\tau_I$, and can then be added to the energy balance equation (see Problems section). Following this approach, one finds

$$\alpha_t \propto \exp\left(-\frac{\text{const}}{F^2}\right) \quad (13.107)$$

where the square of the electric field enters as ordinarily does the carrier temperature. This result compares reasonably with experiments only for very large F .

For small F a large number of experimental results exhibit a dependence

$$\alpha_t = \exp\left(-\frac{\text{const}}{F}\right) \quad (13.108)$$

This dependence was explained by Shockley. Shockley's idea was that only those electrons that do not interact with phonons contribute to impact ionization. This means that impact ionization is not caused by the spherical part of the distribution function, or by high electron temperature, but by those electrons that

stream only in the field direction and do not scatter until they are at the ionization threshold where they ionize instantaneously (in Shockley's opinion). The probability P that an electron is not scattered is given by Eq. (8.70). We rewrite Eq. (8.70) by coordinate transformation, and use also (in one dimension)

$$\frac{dE}{dt} = \frac{dE}{dk} \frac{dk}{dt} = \frac{dE}{dk} \cdot \frac{eF}{\hbar} \quad (13.109)$$

This gives, with the definition of the total scattering rate by phonons $1/\tau_{\text{tot}}^{\text{ph}}$ and Eq. (7.38), the probability

$$P = \exp \left\{ -\frac{\hbar}{eF} \int_0^{E_{\text{th}}} \left(\frac{dE}{dk} \right)^{-1} \frac{1}{\tau_{\text{tot}}^{\text{ph}}} dE \right\} \quad (13.110)$$

Taking the lower limit of zero energy was sufficiently accurate for Shockley's purpose. This assumes that the electrons start at $k = 0$ at $t = 0$. However, the solution of the equation of motion Eq. (3.26) is

$$k = eFt/\hbar + k_0 \quad (13.111)$$

Trivial as the addition of k_0 is, the consequences are very important. The electron does not start from $k = 0$ but rather from k_0 , corresponding to some average energy E of the electrons in the electric field. An energy histogram is shown in Figure 13.22 to illustrate this point. (Note that the curve lies high above $E = 0$.) The figure shows that the electron energy first increases rapidly within a very short time. At higher energies the electron is scattered frequently.

Whenever it is scattered against the field, it loses much energy and restarts the process. Overall a heated electron gas is established, and at instances electrons escape in field direction toward high energies (Shockley's so-called lucky electrons). However, the electrons do not start at $E = 0$ but rather at some average energy that in GaAs and Si is typically 1 eV at $F = 500$ kV/cm. One therefore should replace the lower limit of integration in Eq. (13.110) by an average energy equal to $\frac{3}{2}kT_c$, where kT_c can be calculated from the formalism proposed by Wolff.

There exists a theory of impact ionization (due to Baraff) that accomplishes this inclusion of the starting energy in a more precise way by solving the Boltzmann equation. Baraff's theory improves Shockley's and Wolff's and is very useful; the results are available in terms of polynomial expansions [21]. It does not do, however, complete justice to all impact ionization effects because the consequences of band structure, which enter into the electron equation of motion and total scattering rate, are difficult to include in Baraff's formalism. The greatest influence of band structure comes from the steep rise in the density of states at higher energies, which is shown in Figure 5.2. This increase leads in turn to a steep increase of the phonon scattering rate as given in Figures 7.5 and 7.6. As can be seen from both the Wolff and the Shockley formalisms, the phonon scattering rate influences the electron dynamics and ionization probability very sensitively. It is this fact (together with the material dependence of

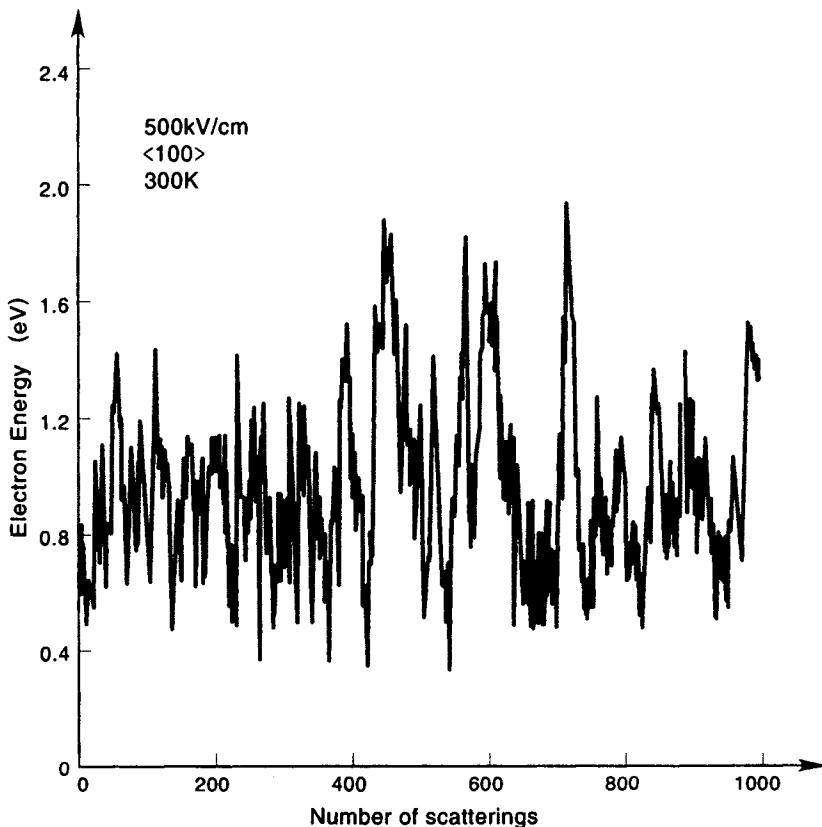


Figure 13.22 Variation of electron energy in GaAs after scattering events at 300 K for $F = 500 \text{ kV/cm}$. [Source: After Shichijo and Hess [18].]

E_{th}) that leads to the different ionization rates in different materials. Only Monte Carlo simulations (Shichijo and Hess [18]) can include all these factors. However, a reasonable idea of the influence of band structure can also be obtained from Eq. (13.110). We therefore continue to calculate the ionization coefficient according to Shockley and calculate α_z . The coefficient α_z is, according to Eq. (13.104), the inverse of the mean free distance L_I^* that an electron travels without ionizing. The distance L_I that is necessary to reach threshold without scattering is obtained from

$$E_{\text{th}} = e \int_0^{L_I} F(z) dz \quad (13.112)$$

To proceed we need to specify the z dependence of the electric field F , which in $p-n$ junctions is given by Eq. (10.12). To avoid the complications of z -dependent F , which also enters Eq. (13.110), we neglect the z dependence here ($F = \text{const.}$

stant) and discuss its consequences briefly below. Then we have

$$L_I = E_{\text{th}}/eF \quad (13.113)$$

If we divide L_I by the probability that an electron travels without collision, we obtain L_I^* and α_z :

$$\alpha_z = 1/L_I^* = \frac{eF}{E_{\text{th}}} \exp \left\{ \frac{\hbar}{eF} \int_{\frac{3}{2}kT_c}^{E_{\text{th}}} \left(\frac{dE}{dk} \right)^{-1} \frac{1}{\tau_{\text{tot}}^{\text{ph}}} dE \right\} \quad (13.114)$$

The z dependence of F is difficult to include in explicit treatments. Here we discuss two limiting cases.

First, if F increases steeply on a length scale much smaller than the mean free path for ionization, α_z will not immediately follow the field because it takes time to accelerate more electrons to higher energies, and there will be a range of no change in the ionization that can be termed dead space.

Second, if F changes slowly on the scale of L_I^* , then Eq. (13.114) is essentially valid. To illustrate the consequences of the z dependence in this case assume that F is given by

$$F = F_0 + F_1 \cos(2\pi z/L_p) \quad (13.115)$$

that is, a constant with a periodic component (period L_p), which can be small ($F_1 < F_0$). The average field is F_0 (averaged over distance). However, the average α_z is not equal to $\alpha_z(F_0)$ because F enters into the exponent and

$$\int_0^{L_p} dz \exp \left(-\frac{C}{F_0} \right) < \int_0^{L_p} dz \exp \left(-\frac{C}{F_0 + F_1 \cos(2\pi z/L_p)} \right) \quad (13.116)$$

The proof of this inequality is easily given for $F_1 \ll F_0$ because then

$$\frac{C}{F_0 \left(1 + \frac{F_1}{F_0} (\cos 2\pi z/L_p) \right)} \approx \frac{C}{F_0} \left(1 - \frac{F_1}{F_0} \cos(2\pi z/L_p) \right) \quad (13.117)$$

and

$$\frac{1}{L_0} \int_0^{L_p} dz \exp \frac{CF_1}{F_0^2} \cos(2\pi z/L_p) = I_0 \left(\frac{CF_1}{F_0^2} \right) \quad (13.118)$$

Here I_0 is a modified Bessel function that is always larger than one (Abramowitz and Stegun [1]).

This result is at first sight strange because it is counterintuitive that the average field is the same but the average ionization rate is increased. This is, however, a very general feature of nonlinear transport and applies also for the electron temperature, which in our approximations varied with F^2 . It is therefore possible to enhance the ionization by superposed fields that actually average to zero. This can be achieved by varying, for example, the alloy composition of a material and

changing the band edge, and this can lead to a band-structure engineering of the ionization coefficient (see review by Capasso [6]).

Let us consider once again the influence of band structure on the ionization coefficient. For this purpose we consider the band structure of GaAs and the influence that a single scattering event can have on the ionization probability. Our sample electron starts at the Γ point and moves along the [111] direction. At point A $\{k = (\pi/\alpha)(0.3, 0.3, 0.3)\}$, the energy is at the conduction band maximum in this direction, but it is still much less than the threshold energy, and therefore impact ionization is impossible. Now assume that the electron is scattered (by a phonon or impurity) to some other point in the Brillouin zone, point B , for example. Following this scattering event, the [111] component of the electron wave vector continues to increase. However, the wave vector points in a direction different from [111], so that the electron can now reach a higher energy because the band is wider in this direction. As shown in Figure 13.23, the electron can actually exceed the threshold energy for impact ionization (≈ 2 eV) and subsequently can impact ionize. This example shows that, in some crystallographic directions, electrons can impact ionize only if they are scattered, which contradicts Shockley's assumption. It also illustrates the importance of the inclusion of k_0 in the starting conditions of the electrons. This, of course, does not mean that Shockley's ideas are wrong. It only means that they cannot be stretched too far.

Ionization coefficients as obtained experimentally for electrons and holes in GaAs are shown in Figure 13.24. The measurements have been taken in the [100] crystallographic direction and are virtually independent of the orientation. Monte Carlo simulations fit the experimental data almost perfectly and show little if any dependence on the crystal direction in which the electric field points. Finally, it should be noted that the two-carrier transport $p-n$ junctions adds a new feature to the ionization process. If an electron excites another electron to the conduction band a hole is created, which now moves in the opposite direction, is then again accelerated, and itself can ionize, thus creating a second hole and an electron in the conduction band. In this way, if the device is long enough, infinite multiplication of one primary carrier can occur, which is termed avalanche breakdown (Streetman [20]).

13.3.3 Zener Tunneling

Very heavily doped reverse biased $p-n$ junctions also show significant current increase, which, however, is not caused by impact ionization but rather by the tunneling of electrons from the valence band to the conduction band because of high electric fields (the Zener effect). The principle of the Zener effect is illustrated in Figure 13.25. The electron tunnels from point z_a to point z_b , and changes bands during this process. The effect is, of course, only an extension of the field emission, discussed in Chapter 1 (see Figure 1.8). The calculation of the tunneling current has been outlined in Appendix A and in Eqs. (13.15) and

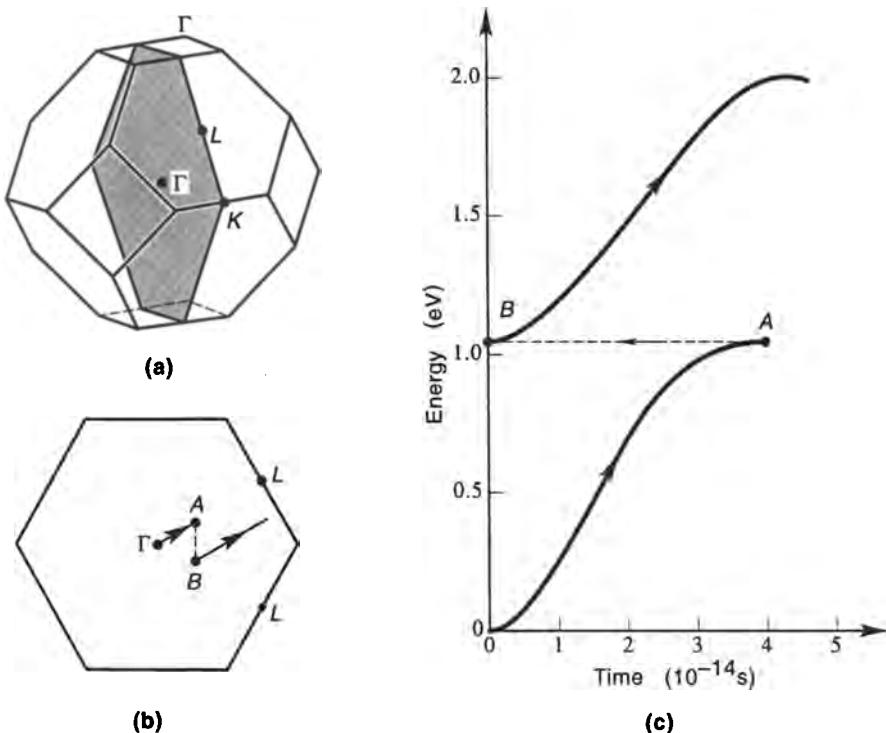


Figure 13.23 (a) The [110] section through the first Brillouin zone. (b) Wavevector trajectory of electron in this plane with F in the [111] direction. The electron is scattered from point A to point B . The energy change in the scattering process has been neglected. (c) Variation of electron energy with time for scattering process illustrated in part b. [Source: After Shichijo and Hess [18].]

(13.16). Here we evaluate only the exponent (the dominant term) of these equations, which also represents the tunneling probability of a single electron, and put $z_a = 0$ and $z_b = E_G/eF$ (Figure 13.25). This exponent becomes (measuring the energy from the top of the valence band at z_a)

$$\exp \left\{ -2 \int_0^{E_G/eF} \frac{\sqrt{2m}}{\hbar} \sqrt{E_G - eFz} dz \right\} \quad (13.119)$$

The integration gives

$$\exp \left\{ -\frac{4}{3} \frac{\sqrt{2m}}{eF\hbar} E_G^{3/2} \right\} \quad (13.120)$$

This formula describes the Zener tunneling probability. Its limitations can be seen from the appearance of the mass of the electron, which is generally not the same in the conduction and valence bands. The simple appearance of the

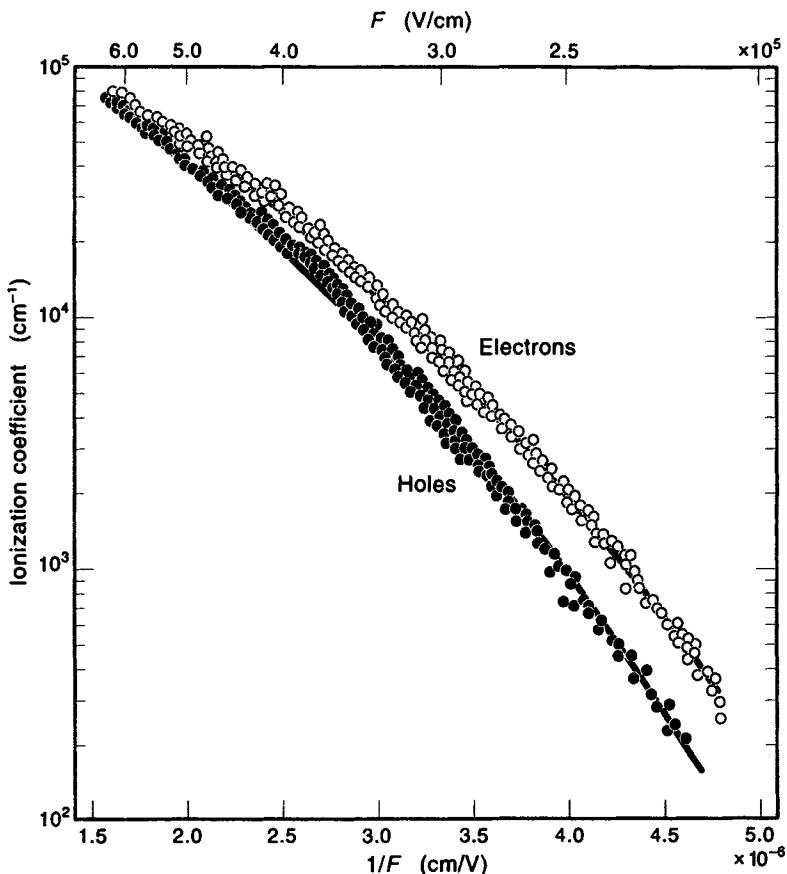


Figure 13.24 Electron and hole ionization coefficient α_z at $T_L = 300\text{K}$ versus electric field. [Source: After Stillma et al. [19].]

free electron mass in Eq. (13.120) arises from our idealization of the potential and the neglect of the atomic potentials and Bragg reflection. Kane has developed an elegant method of including these effects by viewing the $E(\mathbf{k})$ relation as an analytical function of complex variable \mathbf{k} (for details, see Burstein and Lundquist [5], and Duke [7]).

The tunneling process can be assisted by photons (or phonons), as shown schematically in Figure 13.26, and then gives rise to optical absorption below the band edge ($\hbar\omega < E_G$), as seen in the figure. This effect is called the Franz-Keldysh effect. Note, however, that the electron-hole interaction (there is also a hole created in the valence band in addition to the electron in the conduction band) introduces additional subtleties (excitonic effects), which go beyond the scope of this book.

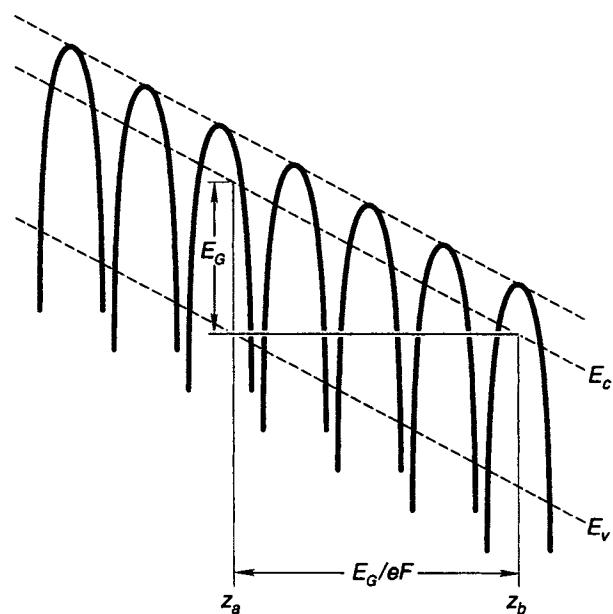


Figure 13.25 Conduction and valence band edges as well as atomic potentials in an electric field. The triangle with height E_G and base E_G/eF marks an idealized potential for the tunneling of an electron from point z_a to point z_b .

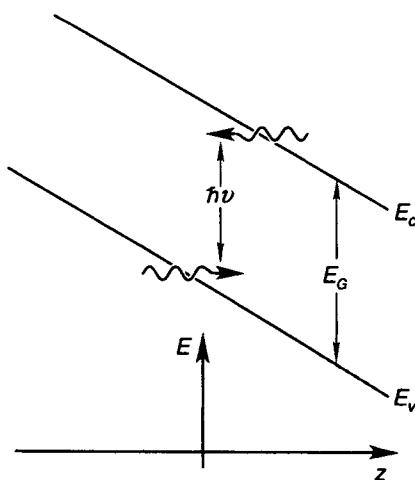


Figure 13.26 Illustration of the possibility of an optical transition for energies $\hbar\omega < E_G$ owing to the effect of the electric field and tunneling.

13.3.4 Real Space Transfer

The transfer of electrons between two different solids was discussed in Section 13.1 for the case of electric fields perpendicular to the solid–solid interfaces. This section deals with the effects again, but from a more general viewpoint. It will be shown that transfer of electrons can also be accomplished if the field is parallel to the interface. The electric field heats the electrons, and the increased electron temperature enables them to overcome barriers. This effect is called real space transfer (RST) [9] and is shown schematically in Figure 13.27.

The emission of hot electrons over barriers (or tunneling through them) is more complicated and more difficult to understand than other basic effects in semiconductor device operation. The reason is that RST can be visualized only by the combination of two concepts concerning the energy distribution of electrons. The first concept is the concept of quasi-Fermi levels, and the second is the concept of a charge carrier temperature T_c (different from the equilibrium lattice temperature T_L). We have outlined these concepts and their interconnections in Chapter 8 [see Eq. (8.39)]. For RST problems, both concepts matter, and both the carrier temperature and the quasi-Fermi levels are a function of space coordinate and time.

Imagine, for example, electrons residing in a layer of high-mobility GaAs neighboring on either side two layers of low mobility AlGaAs. The GaAs equilibrium distribution function f_0 is

$$f_0 = \exp(-E/kT_L) \quad (13.121)$$

whereas in the AlGaAs we have

$$f_0 \propto \exp[-(\Delta E_c + E)/kT_L] \quad (13.122)$$

Here the energy is measured from the GaAs conduction band edge and ΔE_c is the band edge discontinuity between AlGaAs and GaAs. If now the electrons are heated by an external field parallel to the layers, we have to replace T_L in Eqs. (13.121) and (13.122) by a space-dependent carrier temperature T_c . It is clear that for $T_c \rightarrow \infty$ the difference between the AlGaAs and the GaAs popula-

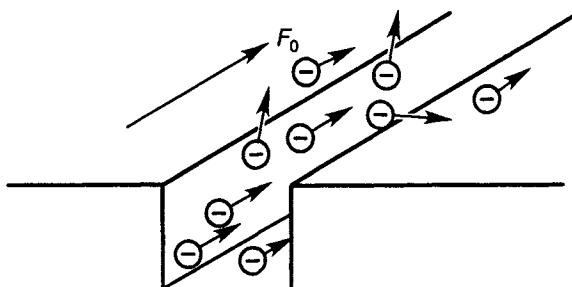


Figure 13.27 Emission of electrons heated by an electric field parallel to a heterolayer interface.

tion density vanishes. In other words, the electrons will spread out into the Al-GaAs layers. This also means that even perpendicular to the layers (z -direction) a constant Fermi level cannot exist, and E_F has to be replaced by the quasi-Fermi level $E_{QF}(z)$ as the density of electrons becomes a function of $T_c(z)$. This is unusual because the quasi-Fermi levels commonly differ only in the direction of the applied external voltage V_{ext} (by the amount eV_{ext}). In the present case, a voltage is applied parallel to the layers, the electrons redistribute themselves perpendicular to the layers, and a field (and voltage perpendicular to the layers) develops caused by the carrier redistribution. Basic to the calculation of this process are the thermionic emission currents of hot electrons to the other j_{RL} and j_{LR} , which have been derived in Eqs. (10.6) and (10.7). Because the external voltage is applied parallel to the layers, we have in steady state

$$j_{RL} = j_{LR} \quad (13.123)$$

from which we can determine the quasi-Fermi levels if the carrier temperatures are calculated from the power balance equation. If precision is needed, then one needs to use the space-dependent power balance equation derived in Appendix E. For rough estimates, T_c can be obtained from Eq. (11.24) for each layer separately. A further complication is presented by the necessity (in most cases) of having to also solve Poisson's equation self-consistently as charge is transferred.

The time constants for the real space transfer process can be calculated as in Eqs. (10.8) through (10.11). For typical parameters of the GaAs-AlGaAs material system and electric fields of the order of 10^3 – 10^4 V/cm parallel to the layers, one obtains time constants of the order of picoseconds, which makes the RST effect attractive for device applications (switching and storage in between the layers). The real space transfer effect is also of general importance in all situations when electrons are confined in potential wells and parallel fields are applied (accelerate the carriers), even if the electrons do not get out of the wells but merely redistribute themselves within the well. This is important for the understanding of the influence of transverse fields (such as the “gate field”) in a transistor (see Chapter 15) on nonlinear transport parallel to interfaces. The RST effect and the spreading of the electrons are then determined by the transverse field. The quantum analog of this classical picture is the redistribution of hot electrons in the different size quantized subbands. Remember that the carrier concentration in the various subbands depends in principle on the electron temperature [Eq. (10.39)] and therefore on the parallel electric field, as well as the perpendicular electric field, which determines the energy of the subbands [E_n in Eq. (10.33)].

13.4 NEGATIVE DIFFERENTIAL RESISTANCE AND SEMICONDUCTOR DIODES

In Section 13.3.3 we discussed Zener tunneling, which becomes important in very high fields and especially for cases of high built-in fields as they occur in

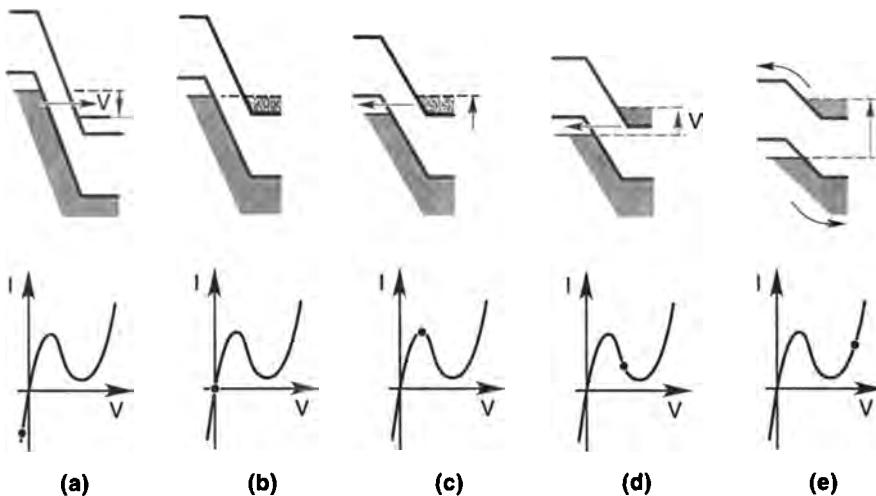


Figure 13.28 Simplified energy band diagrams of tunnel diode at (a) reverse bias; (b) thermal equilibrium, zero bias; (c) forward bias such that peak current is obtained; (d) forward bias such that valley current is approached; and (e) forward bias with thermionic current flowing. [Source: After Hall, ©1960 IRE [8] (now IEEE).]

heavily doped $p-n$ junctions. This process gives rise to an entirely new feature of the $p-n$ junction current: a negative differential resistance, shown in Figure 13.28. This negative differential resistance has been explained by Esaki. The Nobel Prize-winning explanation is illustrated in Figure 13.28.

In reverse bias, electrons can tunnel from the valence band to the conduction band and a relatively large current is flowing. In forward bias, electrons tunnel from conduction states into the hole states of the valence band and the current is large. Then tunneling is prohibited because the electron and hole states are not energetically aligned (to first order only, energy-conserving transitions are possible). Therefore the current decreases and a negative differential resistance (dI/dV is negative) occurs. Finally in extreme forward bias the current increases again.

As is well known, negative differential resistance can be used to convert dc power into ac power and negative differential resistance devices can therefore (and indeed are, if fast enough) be used as microwave generators. The Esaki diode is used for low power applications only; the voltage range in which the negative differential resistance occurs is small (and fixed by the value of E_G/e of the semiconductor). If higher power output is required, two other kinds of negative resistance effects are used, which are discussed below.

We have treated the negative resistance of GaAs already in Chapter 11 (Figure 11.1). GaAs exhibits this negative differential resistance because of its inherent band-structure properties owing to the transfer of electrons in k space. No special doping profile or $p-n$ junction is necessary in this case. The effect was

predicted by Ridley and Watkins and by Hilsum. Devices based on this effect are called transfer devices. When Gunn measured (without knowing the previous theoretical work) the current-voltage characteristic of *n*-GaAs, he found microwave oscillations related to the transit time of electrons through his GaAs samples. This transit time effect is based on the negative differential resistance and can be explained as follows.

Consider a section of semiconductor in the range of negative differential resistance and assume a small increase in resistance in this section, let's say because of fluctuations in the doping. Then, because of its higher resistance, a still higher voltage drops in this section and the electric field is increased. This in turn leads to a lower current, a still higher resistance, and a still larger voltage drop. The process continues until a "domain" of very high field is formed. This domain can "hang" at the positive contact or, if it forms at the negative contact, it will move through the semiconductor with the high field drift velocity ($\approx 10^7$ cm/s). After the domain vanishes at the positive contact, the process starts again and can lead, if the semiconductor is short enough, to microwave oscillations, as observed by Gunn.

Because the effect is dynamic and involves carrier density changes, we add here a few comments on high-field domain formation. The fundamentals proceed very much along the line of our deviations of the space-charge limited current, which starts from Eq. (11.31). Because now we are also interested in time development, we write the equation of continuity (Eq. (11.4), with generation-recombination neglected) in one dimension

$$\frac{\partial n}{\partial t} = \frac{1}{e} \frac{\partial j}{\partial z} \quad (13.124)$$

and use Eqs. (11.31) and (11.32), which give

$$\frac{\partial n}{\partial t} = F \frac{\partial}{\partial z} (n\mu) + \frac{en\mu}{\epsilon\epsilon_0} \delta n \quad (13.125)$$

Here $\delta n = n - n_0$ and n_0 is the equilibrium carrier density. The factor $(en\mu/\epsilon\epsilon_0)$ has the dimension of time, and

$$\tau_R = \frac{en\mu}{\epsilon\epsilon_0} \quad (13.126)$$

is called the dielectric relaxation time. This time constant appears in all problems of similar type. As can be seen from the form of Eq. (13.125), the carrier concentration n will not change significantly for times much shorter than the dielectric relaxation time. High-field domains will therefore only build up on time scales shorter or comparable to τ_R . The transit time τ_{tr} of a domain is approximately given by

$$\tau_{tr} \approx L/v_d \quad (13.127)$$

where v_d is the drift velocity of the domain. If this transit time is shorter than τ_R , obviously domains will not form. Therefore there is a condition for domain

formation

$$\tau_R \ll \tau_{tr} \quad (13.128)$$

which is equivalent to

$$L \cdot n \gg \frac{v_d \epsilon \epsilon_d}{e\mu} \quad (13.129)$$

All of these processes are very complicated because of effects occurring simultaneously in k space and real space, and a proper description needs involved computation. My preference is again the Monte Carlo method if a quantitative understanding is needed. Note that although many semiconductors have satellite minima, only those with relatively large band gaps are eligible for Gunn devices, since in other materials impact ionization interferes with the transfer effect.

The real space transfer effect mentioned in the previous section can also give rise to negative differential resistance, and real space transfer diodes have also shown current oscillations.

For high-power applications, important device types are impact ionization avalanche transit time (IMPATT) devices. These devices are based on negative resistance (not merely negative differential resistance), which is achieved by the generation of a carrier avalanche (by impact ionization) and its subsequent transit through a drift region.

The detailed theory of this type of devices is based on the following idea. Consider a structure consisting of an “electron injector” (a range in which electrons are supplied depending on the external voltages, e.g., a $p-n$ junction with high built-in fields in which electrons are generated by impact ionization once an external voltage is applied) and an electron drift region. This latter region is biased toward saturation of the current, and a smaller additional external voltage will not change the drift velocity. Assume, then, that an ac voltage is applied to the structure. This ac voltage will modulate the injection (increase it at a certain phase). Subsequently the injected electron will go through the drift region and increase the current there even if by now the ac voltage lowers the injection and acts against the drift (but does not change the saturated drift velocity). Therefore the device will exhibit truly negative resistance during certain periods of the ac cycle. A comprehensive theory of this process has been developed by Read [14].

PROBLEMS

- 13.1** Derive Eq. (13.7) from Eq. (13.6).
- 13.2** Calculate the current of Eq. (13.16) for a triangular potential.
- 13.3** Derive j_{rs} in Eq. (13.58) from Eqs. (13.55) through (13.57). Discuss the limiting cases of short devices ($a/l_n, c/L_p \leq 1$) and long devices ($a/l_n, c/L_p \geq 1$).
- 13.4** Derive Eq. (13.100) for a graded profile of N_{TT} . Assume linearly decreasing and exponentially decreasing deep trap concentrations N_{TT} .

- 13.5 Evaluate Eq. (13.106) using Eq. (13.102) for $p = 1$. Discuss the influence of impact ionization on the electron temperature.
- 13.6 Calculate the depletion width in a $p-n$ junction that contains an additional “doping spike” $N_D^{\text{ad}}\delta(\epsilon)$, where ϵ is a short distance from the junction at the n side.

REFERENCES

- [1] Abramowitz, M., and Stegun, I. A. *Handbook of Mathematical Functions*. New York: Dover, 1965.
- [2] Anderson, C. L., and C. R. Crowell “Threshold energies for electron–hole pair production by impact ionization in semiconductors,” *Physical Review B*, vol. 5, 1972, p. 2267.
- [3] Bethe, H. A., and Jackiw, R. *Intermediate Quantum Mechanics*. New York: Benjamin, 1973.
- [4] Bude, J., and Hess, K. “Thresholds of impact ionization in semiconductors,” *Journal of Applied Physics*, vol. 72, 1992, p. 3554.
- [5] Burstein, E., and Lundquist, F. *Tunneling Phenomena in Solids*. New York: Plenum, 1969.
- [6] Capasso, F. *Physics of Avalanche Photodiodes in Semiconductors and Semimetals*, vol. 22, ed. R. K. Willardson and A. C. Beer. New York: Academic, 1985.
- [7] Duke, C. B. *Tunneling in Solids*. New York: Academic, 1969.
- [8] Hall, R. N. “Tunnel diodes,” *I.R.E. Transactions on Electron Devices*, vol. ED-7, 1960.
- [9] Hess, K. “Lateral transport in superlattices,” *Journal de Physique* vol. C7, Supplement 10, 1981, pp. C7-3–C7-17.
- [10] Higman, J. M., and K. Hess “Use of the electron temperature concept for nonlinear transport problems in semiconductor $p-n$ junctions,” *Solid-State Electronics* vol. 29, 1986, p. 915.
- [11] Kane, E. O. “Electron scattering by pair production in silicon,” *Physical Review*, vol. 159, 1967, p. 624.
- [12] Landsberg, P. T. *Solid State Theory*, New York: Wiley/Interscience, 1969, pp. 305–306.
- [13] Laux, S. E., and Hess, K. “New quantitative theory of $p-n$ junction impedance: analytical resolution of past misconceptions,” *IEEE Transactions on Electron Devices*, vol. ED-46, 1999, pp. 396–412 (1999).
- [14] Read, W. T., “A proposed high frequency, negative-resistance diode,” *Bell System Technology Journal*, vol. 37, 1958, p. 401.
- [15] Ridley, B. K. *Quantum Processes in Semiconductors*. Oxford: Clarendon, 1982, pp. 251–263.
- [16] Sah, C. T. *Fundamentals of Solid State Electronics* New Jersey: World Scientific, 1991.
- [17] Shenai, K., et al. “Modeling and Characterization of dopant redistributions in metal and silicide contacts,” *IEEE Transactions on Electron Devices*, vol. ED-32, 1985, p. 793.
- [18] Shichijo, H., and Hess, K. “Band structure dependant transport and impact ionization in GaAs,” *Physical Review B*, vol. 23, 1981, pp. 4197–4207.
- [19] Stillman, G. E., Robbins, V. M., and Hess, K. “Impact ionization in InP and GaAs,” *Physica*, vol. 134 B–C 1985, p. 241.

- [20] Streetman, B. G. *Solid State Electronic Devices*, 2nd ed. Englewood Cliffs, N.J.: Prentice-Hall, 1980.
- [21] Sze, S. M. *Physics of Semiconductor Devices*, New York: Wiley, 1981.
- [22] Tyagi, M. S. *Introduction to Semiconductor Materials and Devices*, New York: Wiley, 1991.

CHAPTER 14

LASER DIODES

Semiconductor laser diodes, and particularly those containing a quantum well as the active region for light generation, form a prime example of the combination of classical and quantum transport physics as well as electromagnetic theory. Quantum well laser diodes (QWLDs) are, in our opinion, among the most complex semiconductor devices. This complexity arises from the fact that in addition to the current continuity equations one needs now to solve Maxwell equations (and not just the equation of Poisson). Electronic transport is also complicated by the existence of heterojunctions, quantum wells, and transport over as well as capture into the wells (heterojunctions in transistors form mainly barriers of confinement). The overriding principle of designing QWLDs is the localization of electrons, holes, and the electromagnetic field (of the light generated by stimulated emission) within the very same region of space: the quantum well and its immediate surroundings.

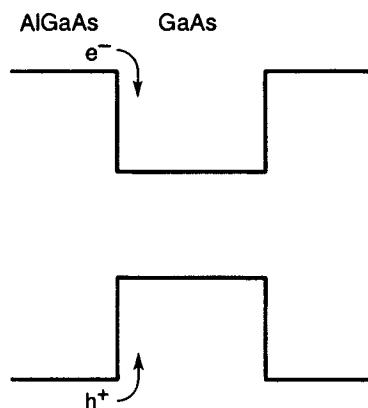


Figure 14.1 AlGaAs/GaAs quantum well collecting both electrons and holes in the same region of space.

It is intrinsic to the properties of heterojunctions of III-V (and other) compound semiconductors that this localization is indeed possible. Figure 14.1 shows a heterojunction of GaAs and AlGaAs with two different aluminum mole fractions. From the laws of acceleration for electrons and holes, we know that both types of carriers will mostly accumulate in the quantum well (center region) with some still populating the adjacent so-called separate confinement region. In current implementations, the quantum well is typically 100Å or less (and several such wells can be used) and the separate confinement region is of the order of 500Å or more. The refractive index is larger in the regions of the smaller gap (see Chapter 6) and therefore, according to standard electromagnetic theory, the light is also concentrated in these regions. However, because the wavelength of the light is typically much larger than the width of the quantum well, this well cannot confine the light and for efficient confinement the separate confinement region is necessary. This is all mentioned up front to show that heterostructures are key for prime laser performance. The inexperienced reader is referred to Chuang [2] for more detail.

These complications have made QWLDs inaccessible to complete analytical theory as still was (mostly) possible for ordinary $p-n$ junction diodes. Limited theories for certain effects and certain spacial regions have been possible and can be found in the literature. The combination of these facts and regions, however, has been phenomenological (i.e., with so-called rate equations that were coupled with phenomenological constants instead of physical equations). It is only recently that computer simulation has permitted a more complete theoretical approach with a physical description of the diode as a whole and with predictive equations and numerical procedures that lend themselves to engineering optimizations. The key for these optimizations is of course, the mentioned confinement of electrons, holes, and the light in the same region of space. This leads as we will see, to low threshold currents (for lasing) and high reliability as well as high-speed modulation response; all very desirable properties.

In the following we describe the basic equations, their coupling and many aspects of their numerical solution. We also present results that go beyond some of the elementary descriptions (e.g., we show deviations from the usually assumed charge neutrality).

14.1 BASIC GEOMETRY AND EQUATIONS FOR QUANTUM WELL LASER DIODES

Figures 14.2 and 14.3 show typical structures of edge- and surface-emitting laser diodes. In edge emitters, light is propagating parallel to the interface layers. To confine the light to the region of diode current flow, insulating layers with lower dielectric constant are often used as indicated in Figures 14.2 and 14.3. That is, confinement is not only desired in z -direction (perpendicular to the quantum well and separate confinement region) but also to the section of the x -direction

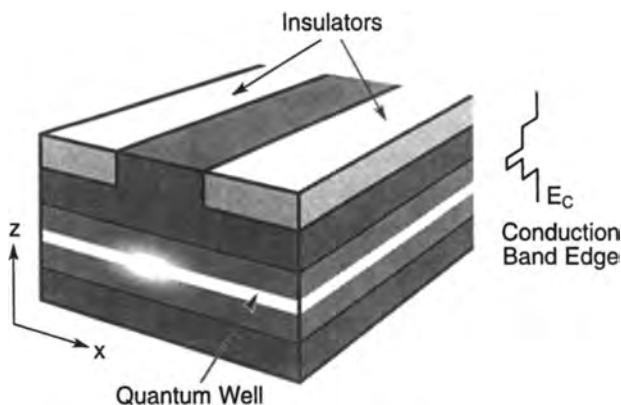


Figure 14.2 Edge-emitting laser diode. The quantum well is sandwiched between the separate confinement regions; these are in turn sandwiched by even higher gap materials as indicated by the energy diagram (right upper side). The light propagates parallel to the layers and exits the front and back surfaces as indicated. The $p-n$ junction is in the z -direction

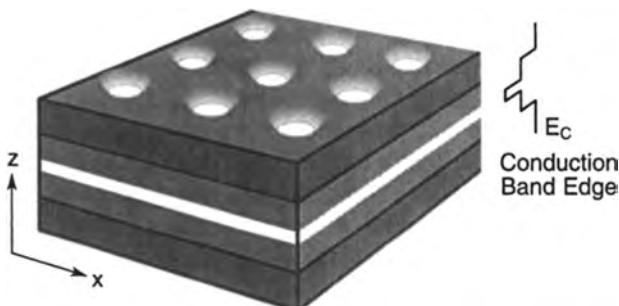


Figure 14.3 Surface-emitting laser diode. The light exits the top surface. Several light patterns are indicated to illustrate the integrability of this device. Otherwise, same as Figure 14.2.

over which the electrons and holes flow. The light must therefore be guided by spacial variations of the refractive index in more than one direction. The fact itself that the diode current leads, as we will see, to stimulated light emission would also confine the light to the region where current flows; this confinement is called gain guiding. Most edge-emitting QWLDs are mainly index guided.

Figure 14.3 shows a surface emitting laser structure (vertical cavity surface emitting laser or VCSEL). Here the light propagates perpendicular to the hetero-layers. High reflectivity Bragg reflectors are necessary in this case to limit light transition out of the structure and thus increase stimulated emission. Notice that surface emitters have only a very small active (light emitting) range, the quantum well width, over all the lengths of light propagation whereas in edge emitters the

light propagates along the QW and the propagation length within the semiconductor can be made in principle arbitrarily long. Therefore, for edge emitters to work, the reflectivity at the end of the semiconductor can be rather low and high losses by the out-propagation of light can be tolerated.

Again one wants to confine the electromagnetic field of the light to the region of current flow. This can be achieved by oxide rings confining the emitted light as indicated in Figure 14.3. The use of alumina oxide that can be grown as a “noose” around the lasing region has been a major development for VCSEL’s (See Huffaker et al. [4]).

These structures of edge and surface emitters determine the essential equations that need to be solved and coupled. The electrical and quantum mechanical equations and solutions are very similar for both laser types and will be treated first. The necessary electromagnetic solutions are distinctly different and are discussed subsequently.

14.2 EQUATIONS FOR ELECTRONIC TRANSPORT

The standard current equations for the $p-n$ junction diode have been treated in the previous chapter [see Eqs. (13.29) and (13.30)]. These equations also apply for the laser diode; special care needs to be taken at the heterojunctions. The heterojunctions of the separate confinement region can be treated separately from each other and separately from the quantum well because the distances of interface separation are much longer than the quantum mechanical dephasing length. Therefore these heterojunctions can be treated with the standard thermionic emission equations that we have described in Chapter 10, particularly Section 10.1. Numerical implementations of these equations into simulators are not entirely trivial but have been described in detail by Grupen and Hess [3]. This reference will also be useful to answer other questions concerning numerical simulation of laser diodes and also concerning some of the relevant equations. Transport over the quantum well and capture into the well needs special considerations that go beyond standard thermionic emission theory. The coupling of classical and quantum transport is a difficult problem, as we know from Chapter 8. If the transition region from coherent to dephased (classical) transport can be neglected, the Bethe approach is justified and one can couple reservoirs and quantum region as we have done in Chapter 8. The reservoir is now the whole laser diode including the separate confinement region but excluding the quantum well, and we need to carefully describe the communication of the separate confinement region with the quantum well.

A complication arises from the fact that we are not only interested in transport over or through the quantum region as in Chapter 8 but also, and very much so, in capture into the well. For only electrons and holes captured into the well will finally contribute to stimulated emission and lasing. A complication for the description of this arises from the fact that around the well there are “well

states" and also three-dimensional propagating states. Capture from the three-dimensional states into the well and subsequent cascading down in energy in the well need to be considered. The matrix elements for the scattering can easily be calculated once the wave functions of initial and final states are known, and have been calculated by various authors [5]. A distinctive feature of capturing by emission of polar optical phonons are resonances (i.e., the matrix element shows strong peaks at certain energies). These resonances are somewhat artificial because the actual calculation of the wave functions is assumed to be unchanged by the capture process itself. However, the capture is inelastic and dephases the wave function. There are also other scattering events such as electron-electron interactions. As a consequence of all of this, the resonances are damped out and the capture is often well described by a process that does not strongly depend on energy. The phonons are, for the reasons described in Chapter 10, assumed to be three-dimensional, an approximation that is often excellent.

The use of the transition from three-dimensional to two-dimensional electron states by use of the three-dimensional wave function

$$\frac{1}{\sqrt{V_{\text{ol}}}} \exp(i\mathbf{k} \cdot \mathbf{r}) \quad (14.1)$$

and the two-dimensional wave function (see Chapter 10)

$$\frac{1}{\sqrt{A}} \exp(i\mathbf{k}_{||} \cdot \mathbf{r}_{||}) \zeta_n(z) \quad (14.2)$$

gives then transition rates similar to those we know from the three-dimensional treatment in Chapter 7 with pre-factors not too different from one. If we absorb these and possible energy (\mathbf{k}) dependences, then we arrive at the simplified equation for electron (hole is similar) capture into the well

$$U_{\text{cap},i}^{\text{el-LO}} = \int_{E_c^0}^{\infty} \left(dE_{3D} g_{3D}(E_{3D}) \Delta E_i \frac{g_i}{L_{\text{norm}}} s_{3D,i}^{\text{el-LO}} \delta(E_{3D} - E_i - \hbar\omega_{\text{LO}}) \right. \\ \left. \times [f_{3D}(1-f_i)(N_q + 1) - (1-f_{3D})f_i N_q] \right) \quad (14.3)$$

Here $U_{\text{cap},i}^{\text{el-LO}}$ denotes the net capture into a small energy range ΔE_i of the confined two-dimensional states in the well. The subscript LO denotes longitudinal polar optical phonons, and the integration goes over the three-dimensional states above the well. L_{norm} is the distance over which the confined wavefunctions are normalized (essentially the width of the well). The other symbols have the usual meaning. The derivation is entirely analogous to the derivations in Chapter 9 and is left to the reader. For emitting or absorbing phonons within the well one gets a similar expression, now only including the two-dimensional well states.

To actually implement such a calculation into a device simulator, one needs to draw a line between three-dimensional and two-dimensional states. The most

plausible division is to count all states energetically above the well as three-dimensional and all below the energy of the confining band edge as two dimensional.

A further complication in the understanding of QWLD transport is the necessity to include electron-electron and electron-hole interactions. When we calculated the mobility, it was clear that these interactions can influence the energy distribution. However, because energy and momentum of the ensemble of electrons are conserved by that particular scattering mechanism, it acts only indirectly (via the distribution function) on mobility and conductivity and we have neglected its influence.

For the QWLD, carrier-carrier interaction is vital. Electrons and holes recombine optically from any given energy in the well. However, polar optical phonons have a distinct given energy and can only supply electrons (holes) to certain given well energies that are eventually a multiple n of the phonon energy $\hbar\omega$ measured from the top of the well. Phonons alone can therefore have a problem to resupply the charge carriers that recombine and emit light. Carrier-carrier interaction takes care of this. Unfortunately carrier-carrier interaction is too complicated a scattering mechanism to be included in most engineering simulations. An artificial mechanism that emulates the true carrier-carrier scattering very well has been implemented and extensively used by Grupen and Hess [3].

The injection of energetic carriers into the well leads to heating of the electron and hole gas through the carrier-carrier interactions. Phonon scattering at the higher energies is mostly by phonon emission. Polar optical phonons, however, cannot propagate out of the well and decay only after relatively long times by scattering and conversion into acoustic phonons; at room temperature this takes about 5 ps in n -type and closer to 1 ps if both electrons and holes are present in high concentrations. The accumulated phonons can be reabsorbed by electrons and holes, and the complete process leads to a heating of the electron hole gas that follows any modulation of the carrier input and output within a picosecond time scale. Therefore, even for modulation response in the gigahertz range, this heating can be regarded as instantaneous.

The details of these processes and a complete solution of a time-dependent Boltzmann-type equation have been given in [3]. For further understanding of the consequences of this heating we need only a relationship between temperature increase of the electron (hole) gas and the number of photons S_v emitted by stimulated emission into a laser mode labeled by v . Because, for every emitted photon, we need to resupply an electron-hole pair into the well, which in turn will heat the "cold" well carriers by carrier-carrier interactions and the accumulation of phonons (hot phonons) in the well, we can expect that the heating is monotonically increasing with the number of emitted photons S_v . If we assume that the laser is operated at a given emission of S_v photons per unit time and we modulate the laser by some means slightly around this number (change S_v by

δS_v), then the temperature of the electron (or hole) gas will follow the relation

$$\delta T_c = D_c(S_v) \delta n \quad (14.4)$$

where D_c is a monotonically increasing function that is different for electrons and holes. The proof of this relation goes beyond the scope of this text. As a zero-order approximation, one could use a function linear in S_v , which should be valid over some range of photon number S_v emitted per unit time. One therefore gets

$$\delta T_c = D_{co} S_v \delta n = D_{c1} S_v \delta E_{QF} \quad (14.5)$$

with D_{co} being a constant.

The lattice temperature T_L is, of course, increased also. This increase can be calculated in the classical transport regions by the increase owing to Joule's heat and use of a heat conduction equation. In the quantum well the increase of energy corresponds exactly to the number of net emitted phonons that can be calculated by use of Eq. (14.3) in a Boltzmann-type of equation [3]. A simple formula for T_L as a function of S_v (or Joule's heat in the classical region) cannot be given. It should be noted, however, that T_L is not modulated on the picosecond time scale but follows the light output very slowly, on a nano- or microsecond timescale, and can therefore for fast modulation be taken as constant in time.

As can be seen from the above discussion, the electrical conduction, carrier capture, and energy exchange dynamics of QWLSs are complicated and the interested reader is encouraged to study more detailed accounts such as given in [3].

14.3 COUPLING OF CARRIERS AND PHOTONS

We have already in described Chapter 9 some of the elements of the theory of optical transitions and pointed out the importance of the optical matrix element as defined in Eq. (9.1). This assumes that the electron–photon interaction is described by the Golden Rule. One might doubt that the Golden Rule is valid at high light intensities, when stimulated emission becomes very strong. However, even at the highest light intensities of QWLDs, no evidence has been found that the Golden Rule breaks down for this reason by itself. The only problem with the Golden Rule is the δ shape for energy conservation when the scattering is completed. There exists considerable broadening of the energy levels; that is, one has to include a shape with finite width that replaces the δ -function. This effect is caused by the scattering by photons, phonons, and other carriers. This broadening would be of no importance in ordinary transport problems. However, in QWLDs there is a special effect that makes it important. If carriers are annihilated by the electron–hole recombination and emit light, then they need to be resupplied to maintain lasing. However, the annihilation becomes extremely effective if the electron (hole) is just taken out of a single state with given

discrete energy; it is then difficult to resupply the electron (hole) in time and one says a “spectral hole” is burned (i.e., no electron–hole pair is available for recombination and stimulated emission is diminished or stops). One therefore has to replace the δ -function in the Golden Rule by a better approximation, for example, by a Lorentzian function

$$\delta(E_c - E_v - \hbar\omega) = \frac{\gamma}{(E_c - E_v - \hbar\omega)^2 + \gamma^2} \quad (14.6)$$

Here $\hbar\omega$ is the photon energy and E_c and E_v denote as usual the conduction and valence band energies. The quantity γ is representing half the complete scattering rate that contributes to the broadening (e.g., phonon scattering plus electron–electron interactions). The theory of broadening, particularly when electronic transport is involved, is not trivial and we refer the interested reader to Chuang [2]. For all our purposes here the broadening just serves to avoid the burning of spectral holes and is otherwise of no significance; for example, the broadening of the laser line itself has different reasons and is given by the theory of Henry as discussed in Chuang [2]. We thus have settled on using the Golden rule to essentially couple the light to the carriers. The perturbing Hamiltonian H (owing to the interaction with light) is given in all electromagnetic texts to be:

$$H' \approx -\frac{e\mathbf{A} \cdot \mathbf{p}}{m_o} \quad (14.7)$$

where \mathbf{A} is the vector potential in the Coulomb gauge [2]. If the photon wavelength can be approximated as infinite, then the vector potential can be approximated by a constant vector \mathbf{A} pointing in the direction of the optical electric field and thus indicating a polarization dependence (see Chuang [2]). The matrix element appearing in the Golden rule is then the momentum matrix element as described in Chapter 9 with a prefactor of $(Ae/(2m_o))$. The probabilities per unit time for optical absorption as well as emission (stimulated and spontaneous) follow immediately from this by similar algebra as used in Chapter 9.

However, unlike the treatment of the generation (G) and recombination (R) terms in Chapter 9, we include here the conservation of the wave vector in the optical transition as stated in Eq. (9.8)—assuming negligible wavevector of the light and taking the vector potential as spacially constant. One then gets for the net stimulated emission (stimulated emission minus absorption)

$$R_v = C_o \int_{\mathbf{k}_c \mathbf{k}_v} \delta_{\mathbf{k}_c, \mathbf{k}_v} M_{c,v}^2 \delta(E_c(\mathbf{k}_c) - E_v(\mathbf{k}_v) - \hbar\omega_{\text{phot}})(f_c - f_v) d\mathbf{k}_v d\mathbf{k}_c \quad (14.8)$$

with

$$C_o = S_v \left(\frac{eAV_{\text{ol}}}{m_o \hbar} \right) \left(\frac{1}{128\pi^5} \right) \quad (14.9)$$

Notice the quantity S_v in the prefactor C_o . This is the photon number. The index is given because we will have the possibility of different optical modes v in a

laser structure and we refer to the photon number of a given mode. For more than one mode a sum over all modes has to be performed. $M_{c,v}$ is the matrix element from Eq. (9.1). R_v corresponds to the difference $G(n) - R(n)$ as derived in Chapter 9.

Because the integrand contains the Kronecker symbol δ_{k_c, k_v} , the double integration can be performed explicitly if the matrix element is otherwise independent of the wave vector. The result is

$$R_v = B_v S_v g_v (E_c^v - E_v^v) (f(E_c^v) - f(E_v^v)) \quad (14.10)$$

Here B_v is the so-called Einstein coefficient (from the celebrated Einstein 1917 paper), which is

$$B_v = \frac{(M_{c,v})^2 e^2 \pi}{\epsilon \epsilon_0 \omega_v (m_e m_h)^2} \quad (14.11)$$

and g_v is the so-called reduced density of states. The mass is now a composite of electron m_e and hole m_h masses

$$m_r = \left(\frac{1}{m_e} + \frac{1}{m_h} \right)^{-1} \quad (14.12)$$

A note must be added here on the band structure. If we have more than one band for holes, we must sum, of course, over all the bands. Also, if there were more than one conduction band within the optical range, it would have to be added into our considerations. We have throughout the book discussed the band structure only within the empirical pseudopotential model. This is a great model, if one wants to know about electronic conduction in the various minima at Γ , X , and L . However, it does not do entire justice to the valence band because it considers only two bands; one for heavy and one for light holes. There is, however, often a third band very important for the optical transition which arises from the effect of spin orbit splitting that had not been included in our empirical pseudopotential approach. To include this band another method of band-structure calculation is often used, which is called the $\mathbf{k} \cdot \mathbf{p}$ method. This is the method of choice for the simulation of optical transitions and has been described in detail by Chuang [2]. This method also can be used to calculate band-structure and optical matrix elements in quantum wells by assuming a periodic structure (super cell) in which the quantum well and neighboring layer are periodically repeated as a giant unit cell. The method is also suited to include the often important effects of strain (arising from the use of layers of different material) and many body effects as we have discussed them in Eq. (10.34). All of this is an area of research on its way to some maturity of understanding.

Equation (14.10) was for the net stimulated emission (stimulated emission minus absorption). For spontaneous emission (and emission alone) a slightly different equation is obtained in a completely analogous way. The difference of distribution functions, however, needs to be replaced by $f_c(E_c(v))(1 - f_v(E_v(v)))$

and the photon number S_v needs to be replaced by the average number of photons now in a broad spectrum. This gives for the total spontaneous emission R_v^{sp}

$$R_v^{sp} = B_v g_v (E_c^v - E_v^v) (f_c(E_c^v) - f_v(E_v^v)) D(\hbar\omega_{\text{phot}}) \quad (14.13)$$

where

$$D(\hbar\omega_{\text{phot}}) = N_{\text{phot}} 8\pi \hbar \omega_{\text{phot}} \left(\frac{m_r}{hc} \right)^3 \quad (14.14)$$

is the density of states of photons and N_{phot} is the photon occupation number defined exactly as in Eq. (7.24) for the phonons (lattice vibrations).

To complete all necessary portions of a laser diode theory, we still need to know how S_v is to be calculated. To achieve this, we first deal with the mode structure. Because only optical modes with certain geometrical confines can be emitted in a QWLD, we need to calculate the mode as a function of space and time coordinates. This means we actually have to solve Maxwell's equation to obtain the optical field ϕ_v of the given mode. This field requires in general a vector solution of Maxwell's equations with a complex dielectric constant in the region of net stimulated emission (the gain region). Such a complete approach is a complicated numerical project for practical device geometries, and requires very large computational resources. In addition, the solution depends on R_v , which in turn depends on the energy distribution function of the electrons and holes. This means the whole problem—Maxwell equations and electronic transport—needs to be calculated self-consistently, which means numerical iteration. At this time, no existing computer system is capable of solving this problem in all generality. There are therefore usually a few simplifications made. First, it is assumed that the optical and electrical problem are linked by a polarization and dielectric constant that depends only on the charge carrier density at any given instant. This is an excellent approximation even if the carrier density is modulated on time scales as small as 10ps. A breakdown of this assumption would require a still more elaborate framework including the so-called Boltzmann-Bloch equations as described by Chow, et al. [1]. Second, we assume that the electromagnetic field distribution, i.e., its spacial form (not its intensity), does not depend much on the gain (i.e., on R_v). This means mathematically that the average field square

$$(\phi_v^{\text{av}})^2 = \frac{\phi_v^2}{\int \phi_v^2 dr} \quad (14.15)$$

does not depend on the gain or stimulated emission during lasing. One calls such a laser diode strongly index guided and many of the existing QWLDs are indeed strongly index guided, which requires a certain geometrical arrangement of materials. Under these circumstances then, S_v can be calculated from a rate equation, the form of which is intuitively obvious

$$\frac{dS_v}{dt} = (R_v + R_v^{sp}) (\phi_v^{\text{av}})^2 - \frac{S_v}{\tau_v} \quad (14.16)$$

where τ_v is the photon lifetime. It depends on photon losses out of the reflecting surfaces of the laser diode and on absorption and scattering losses that we will not describe here in any more detail. As obvious as this equation appears to be, its precise derivation takes detail beyond our scope and we refer the interested reader to Chuang [2]. The term $R_v(\phi_v^{av})^2$ is usually called the gain and denoted by G_v (not to be mixed up with the generation rate). Note in this context, that the gain necessitates electron hole recombination. Therefore a corresponding net recombination term due to spontaneous and stimulated photon emission needs to be included into the equations of continuity. While these terms cause a gain of photons, they cause a loss (recombination) of charge carriers. We have thus described all the basic equations that describe QWLDs that are essentially the Boltzmann equation for the electronic transport (or simplified equations as obtained by the method of moments and thermionic emission theory), the Schrödinger equation for the quantum well and the bandstructure, Maxwell's equations for the electromagnetic field and the rate equation of Eq. (14.16) which couples all of these equations through the Golden Rule (or its extension with broadening of the energy levels). The validity of this system is proven in wide ranges and agrees very well with experiments within the assumptions stated above.

14.4 NUMERICAL SOLUTIONS OF THE EQUATIONS FOR LASER DIODES

The numerical procedure to solve the preceding equations follows the general principles that have been described in Chapter 12. Currently available advanced laser simulators [3] solve the coupled electrical and optical problems in the following way. The electrical problem is solved in the drift-diffusion approximation using the equations of continuity given in Chapter 13, typically in two dimensions, everywhere outside the quantum well. Inside the well, transport occurs only parallel to the interface and is treated separately, also within the drift-diffusion approximation. The coupling of the classical portion of the diode to the quantum well is a complicated problem. A reasonable way to solve it is to adopt the Landauer-Büttiker picture described in Chapter 8. In its simplest form with a quantum transmission coefficient equal to unity, this corresponds to Bethe's thermionic emission theory. Separating the quantum well states in propagating three-dimensional states and two-dimensional confined states, one connects the three-dimensional states by thermionic emission formulae to the classical bulk region. Then the electrons (holes) of the three-dimensional states above the well are permitted to scatter into the two-dimensional confined states. The quantum well is thus treated like a (big) Shockley-Read-Hall trapping center with a continuum of states [3]. The equations for the electron (hole) capture into the well can be obtained by integrating Eq. (14.3). Of course a realistic approach also needs to include electron-electron interactions for the capture into the well. The opti-

cal recombination in the well is then described by using Eq. (14.8). As a result of all of these considerations one obtains the quasi-Fermi levels. These, in turn, are needed in Eq. (14.8), which thus couples electrical and optical problems. Note that the quantum well has its own quasi-Fermi level, which in general differs from the bulk quasi-Fermi levels even at the points where classical region and well touch. This discontinuity arises from the abruptness of the heterojunction.

The optical recombination depends, of course, on S_V , which is obtained from the rate equation given in Eq. (14.16). This, in turn, necessitates a solution of Maxwell's equations. Note, however, that in the approximation of strong index guiding and neglecting the mode changes owing to the contribution of the free charge carriers to the index of refraction (as treated in Chapter 6), this solution is reasonably simple and needs to be obtained only once. Nevertheless, the use of a commercial code is recommended for this; doing everything from scratch will turn fast into a Master's or even PhD thesis.

As an aside, it should be pointed out that Eq. (14.16) is highly singular. In steady state it is equivalent to

$$S_V = \frac{R_V^{\text{sp}}(\phi_V^{\text{av}})^2}{G_V - 1/\tau_V} \quad (14.17)$$

Because S_V is very large under lasing conditions, one operates close to the pole of Eq. (14.17) where the gain G_V equals the losses (i.e., $R_V = 1/\tau_V$). This requires numerical care and indicates a strong coupling of all the involved equations. These are therefore solved simultaneously by using Newton's method to transform the nonlinear algebraic equations into linear ones, as described in Chapter 12.

The differential equations are turned into nonlinear algebraic equations for each mesh point by use of the finite box or other finite difference methods. Then a choice of independent variable is made. This choice is of numerical importance and is described in Chapter 12 for laser diodes. A typical choice would include: the electron concentration n , the hole concentration p , the carrier (electron, hole) temperature T_c , the lattice temperature T_L , the number of photons S_V , and, there may be other variables. With those five unknowns, the Newton method is then employed (a so-called 5 times 5 Newton) to transform the nonlinear algebraic equations into linear ones, which then are assembled into one big (Jacobian) matrix equation for all the unknowns ($5 \times$ number of mesh points in our case).

This matrix is ideally sparse and can be solved using well-known methods, as described in Chapter 12. However, the introduction of thermionic emission and complicated geometries with thermionic coupling in more than one direction can introduce many elements into the Jacobian matrix and fill many places that would be otherwise zero. Grupen and Hess [3] give a detailed account of this complication (which is of no concern when thermionic emission can be neglected as is often the case in the silicon–silicon dioxide system).

Numerical laser simulators are therefore more complex than their electronic counterparts; this is also the case because of the necessity to use more indepen-

dent variables than in electronic calculations (e.g., the additional S_V).

In the following we describe a few important numerical results that distinctly differ from standard analytical theory. The standard analytical theory solves the continuity equations for electrons and holes together with the rate equation for the photons and is described in every text on laser diodes. The unfamiliar reader is referred to Chuang [2]. This approach assumes strict local charge neutrality and ignores solutions of the equation of Poisson. It also assumes $T_c = T_L$ and both equal to the temperature of the surroundings. Internal heating of the laser diode crystal lattice or the charge carriers are neglected. It also neglects all the detailed dynamics of the charge carrier energy distribution, which are assumed to be of Fermi-type. Questions of carrier supply and demand are not addressed. For example, if the light emission becomes very strong, is the electron-electron interaction sufficient to supply charge carriers or will the charge carriers be depleted and thus a "spectral hole" develop that impedes further stimulated emission? Nor does the standard approach include diode properties such as a diffusion capacitance. The most significant approximation of the standard model is the assumption of a gain that is linear in the carrier concentration, which is assumed to consist of an equal number of electrons and holes. This basically neglects all the nonlinearities in the relations of spontaneous and stimulated emission to the energy distribution and quasi-Fermi levels of electrons and holes. We show below the standard (idealized) result for the modulation response for a QWLD and then discuss the severity of deviations from the standard assumptions and the true modulation response.

Figure 14.4 shows the modulation response of the light intensity modulation by a small (linear response) ac modulation of the diode current plotted as a function of modulation frequency ω for various values of the dc diode current, all above lasing threshold. For small frequencies, the standard analytical model predicts a monotonous increase in the light response with frequency to a maximum value and from this a decrease which is 40dB per decade of frequency in the limit of high frequencies. The maximum occurs because of a "resonance" in the supply (by electronic transport) and demand (by stimulated emission and carrier recombination) of the charge carriers. Below we will describe deviations from this idealized behavior as obtained by numerical simulation.

We begin by discussing how standard $p-n$ junction concepts need to be included and influence the modulation response. We have discussed the ac response of $p-n$ junctions in Chapter 13 and have seen that in forward bias one needs to include a depletion and diffusion capacitance. The diffusion capacitance of short diodes can become very large and therefore would degrade the modulation response [i.e., would lead to a decrease of response with increasing frequencies particularly for small frequency (for higher frequencies the diffusion capacitance decreases)]. The question is then, is the laser diode short or long? We know from Chapter 13 that "long" also means that the lifetime of charge carriers is extremely short (the recombination is very efficient). In laser diodes this lifetime depends on the magnitude of the stimulated emission and therefore on

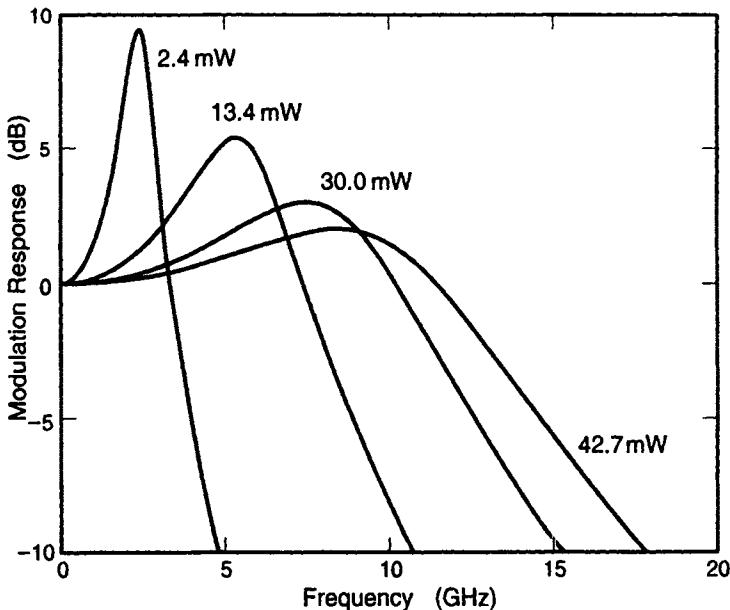


Figure 14.4 Typical modulation response of a laser diode, neglecting nonlinearities due to carrier transport and heating.

S_V and the dc current density that determines the magnitude of S_V . Depending on the actual diode length (and particularly the length of the separate confinement region), a laser diode will therefore be “short” for low current densities (just above threshold) and then become “long” at high current densities. Correspondingly it may have a large diffusion capacitance at low current densities and therefore show a decrease of response with frequency at low frequencies (a so-called “low frequency roll off”). At higher frequencies (not considered in Chapter 13) the diffusion capacitance will be diminished and the modulation response will increase to a maximum that will be reduced from its ideal value, depending how much diffusion capacitance remains. At high current densities the diode is “long” because of the short recombination times and no low frequency roll off will be observed.

Next we discuss the necessity of including the equation of Poisson. The assumption used in analytical theories is $n = p$ and absolute local charge neutrality. This, however, is not justifiable for three reasons. First, even in extreme forward bias there are remnants of charged donors and acceptors particularly close to the quantum well that can capture, for example, electrons from neighboring donors as we know from the concept of modulation doping. Second, there is a huge capacitance between neighboring quantum wells and large changes in electron and hole densities can occur with relatively small voltage drops. Third, the transport of charge carriers between the quantum wells occurs mostly by thermionic emission because the barriers between the wells are usually too thick to permit much

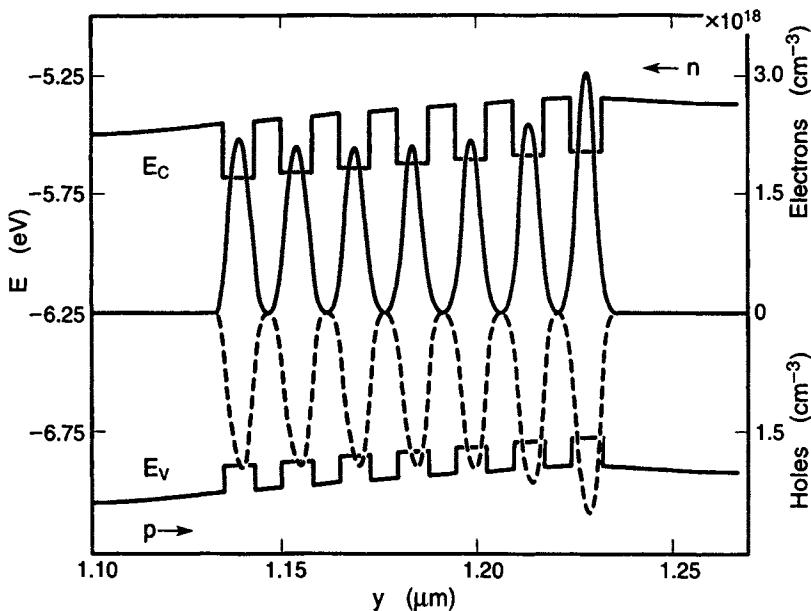


Figure 14.5 The conduction and valence band edges and electron (solid) and hole (dashed) carrier distributions for a QWLD at a dc bias of ~ 24 mW single facet output power. Electrons are injected from the right and holes from the left.

tunneling. Therefore the current between the well depends exponentially on the barrier height to get out of the well and over the band edge discontinuity. This, however, is known to be in general different for electrons and holes, which will give rise to tendencies of distributing charge carriers more easily if they have a smaller barrier to overcome. In other words, if the band edge discontinuity for electrons is much larger than that for holes we would expect that electrons are less likely to propagate from quantum well to quantum well, and are less likely to be resupplied in the wells further away from the n -region and therefore decrease in density there.

All these features are demonstrated in Figure 14.5, which shows electron and hole densities in several quantum wells during laser operation. A careful look at the figure also shows that the electron and hole densities in a given quantum well are not equal, which causes the slight distortions of the band edges from horizontal. Note that to obtain the results of Figure 14.5 all equations including current continuity Poisson and the Schrödinger equation (the charge distribution in the quantum wells) have been solved self-consistently. The most important fact of this result is that the gain (stimulated emission) that depends sensitively on the distribution function (and therefore carrier concentration) is different in different wells. Actually, some wells may only exhibit a net loss of photons and thus mostly undo the gain of others. Multiple wells are used to increase efficiency. However, one can see that increase above a certain number (typically

five) is counterproductive. This is also well documented by experiments. Remember that the quasi-Fermi levels change discontinuously between the wells and between well and separate confinement region because the coupling is by thermionic emission over abrupt barriers (see, e.g., Grupen and Hess [3]).

Next we discuss the consequences of the rise in the temperatures T_c and T_L . Joule's heat, or more generally speaking, the net emitted phonons lead to a rise in the lattice temperature T_L . This rise is relatively slow (nano- to microseconds), usually much slower than the dynamic response of laser diodes. T_L therefore can be used as a constant in simulations of the diode dynamics. This constant is determined from solving the parabolic differential equation of heat conduction. Such a solution is available in several commercial software packages.

The carrier temperature T_c , on the other hand, rises and falls much faster when the diode bias is changed; typically in a few picoseconds. The cause for the heating of the electrons is not the electric field, which is typically very small when the diode is in strong forward bias. The electron (hole) gas is heated because electrons enter the quantum well at relatively high energies and are taken out of the well by optical recombination from states close to the lowest confined well states. The energy difference raises the average energy above equilibrium until this added energy is offset by the loss to the phonons, almost exactly as described before when dealing with hot electrons. There is, however, the additional complication that energy can also be dissipated from the electrons to the holes (and vice versa). Also, the phonon distribution can now be considerably disturbed and may need to be calculated by a separate rate equation [3].

In spite of all these complications, one typically arrives at the simple result of Eq. (14.5), which describes the small increment of T_c when the photon number is given by S_v and changes by the small amount δS_v . How does this influence the gain G_v and therefore lasing? From Eq. (14.8) we can see that the gain depends directly on the energy distribution of the charge carriers. Let's assume for simplicity that this energy distribution is even in the wavevector k (because the electric field is negligible), and can be approximated by a Fermi distribution with quasi-Fermi level E_{QF} and carrier temperature T_c . E_{QF} and T_c enter the equation of the gain therefore in the form of the exponent:

$$\exp\left(\frac{E_{QF}}{kT_c}\right) \quad (14.18)$$

If we modulate the laser diode by varying the applied voltage, then we also modulate E_{QF} because the supply of electrons and holes to the quantum well is varied. This modulation expressed in terms of the photon density gives the additional term $\delta S_v C_{QF}$ where C_{QF} is a constant (see the problems section). In other words, the modulation of the quasi-Fermi level is independent of the actual power at which the laser operates in contrast to the modulation of T_c which follows Eq. (14.4). Therefore Eq. (14.18) is replaced by

$$\exp\left(\frac{E_{QF} + C_{QF} \delta S_v}{k(T_c + C_{c1} C_{QF} S_v \delta S_v)}\right) \quad (14.19)$$

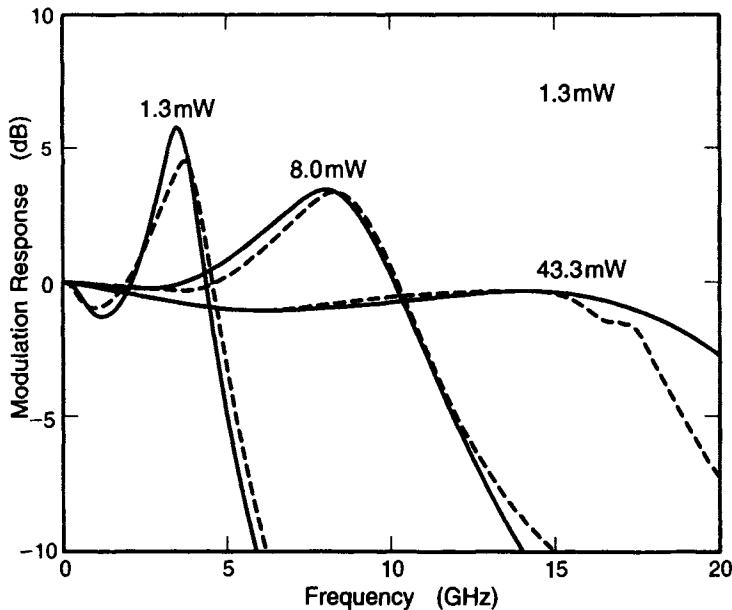


Figure 14.6 Modulation response including nonlinear gain effects and comparison with experiments. (For a detailed description see Grupen and Hess [3].)

We can see from this equation that the change in carrier temperature becomes large as S_V increases and can then counteract the modulation of the quasi-Fermi level. In other words, there exists a point, at high photon densities, beyond which the modulation of temperature and carrier concentration cancel each other and the gain becomes highly nonlinear; modulation of the diode is no longer possible. This point also sets the high-frequency limit of laser modulation because high-frequency response is only possible for high power densities as is known from standard analytical theory (see Figure 14.4). A modulation response curve as obtained when including this gain nonlinearity (and other effects discussed above such as the diffusion capacitance) agrees well with experimental results and is shown in Figure 14.6. Notice the strong damping of the maximum of the response as the lasepower increases and as predicted by the above arguments.

To summarize, we have presented in this chapter some of the basic equations for QWLDs and effects of laser diode response that can only be understood by extensive numerical simulation solving several systems of differential equations self-consistently. Laser diode theory and simulation is currently an active field of research particularly in the area of vertical cavity surface emitting lasers or more generally microcavity lasers. These types are three-dimensional in nature and make extensive numerical treatments a necessity.

PROBLEMS

- 14.1 Derive Eq. (14.3) by the methods developed in Chapter 9.
- 14.2 Assume a gain G_V proportional to the electron density and derive the infinitesimal increase of the quasi-Fermi level that follows an infinitesimal increase of S_V . Use Eq. (14.16) for steady-state conditions.
- 14.3 Design a flowchart that describes the self-consistent solution of the equations of continuity with Eq. (14.16) to determine S_V to determine S_V and the quasi-Fermi levels.

REFERENCES

- [1] Chow, W. W., Koch, S. W., and Sargent, M. III *Semiconductor-Laser Physics*, Berlin: Springer-Verlag, 1994.
- [2] Chuang, S. L. "Physics of Optoelectronic Devices," New York: Wiley, 1995.
- [3] Grupen, M., and Hess, K. "Simulation of Carrier Transport and Nonlinearities in Quantum-Well Laser Diodes," *IEEE Journal of Quantum Electronics*, vol. 34, 1998, pp. 120–140, (see errata p. 384).
- [4] Huffaker, D. L., et al. "Native-oxide defined ring contact for low-threshold vertical-cavity lasers," *Applied Physics Letters*, vol. 65, 1994, pp. 97–99.
- [5] Register, L. F., and Hess, K. "Simulation of Carrier Capture in Quantum Well Lasers due to Strong Inelastic Scattering," *Superlattices and Microstructures*, vol. 18, 1995, pp. 223–228.

CHAPTER 15

TRANSISTORS

Transistors (transfer resistors) are the most important of all solid-state devices and distinguish themselves from the diodes by having a third terminal. In 1947, John Bardeen and Walter Brattain identified minority carrier injection and invented the point contact transistor. Their invention had an unprecedented impact on the electronic industry and on solid-state research.

Diodes and their use rely generally on a material having a nonlinearity with a fairly distinct threshold, which will not respond to an input below a certain strength and which turns on above it. Both signal and power supply come into the same port to achieve the desired function. Transistors, on the other hand, have a third terminal and, in their digital application, can be compared to a mechanical switch. Transistors have, of course, enormous advantages over mechanical switches. They can be very fast (with a switching time constant of the order of picoseconds), they can be produced in large numbers (many millions on a chip the size of 1 cm^2), and they can be made very small (0.1 micrometer feature size or smaller). Transistors also are cheap because of the elaborate photolithographic techniques that are available to produce them. As a consequence of all of these advantages, transistors are widely used in analog and digital circuits.

The transistors of today are typically too small or too fast to be described by simple analytical theories and require elaborate models. Nevertheless, one needs to obtain some analytical understanding to appreciate the numerical results. Therefore, simplified transistor models are presented in the next section. The remainder of this chapter deals with aspects of transistor theory that cannot be analyzed without the help of computational resources. The basis of the theory presented next is the theory of $p-n$ junction diodes which has been described in great detail.

15.1 SIMPLE MODELS

Although the models of transistors presented here are highly simplified, they work astonishingly well as long as they are not pushed too far. The reason is the following. The equations that are actually used are greatly simplified (assuming, for example, constant field-independent mobility). However, if the various physical constants are used as parameters, the equations often are sufficiently accurate because of the mean value theorem of integral calculus. This theorem is the savior of many device models and sometimes works so well that the scientists who develop the models tend to forget the simplifications. Then, only when some features (size, temperature) are changed drastically does the model fail, and one needs to go back to the drawing board and investigate the neglected physics.

The use of the mean value theorem is so powerful because in the method of moments, Eqs. (11.7) and (11.8), mean values are taken. If these mean values are used as parameters, the Boltzmann equation can be highly simplified for narrow (more or less) ranges of variables. An additional help for simple device models is given by “sum rules,” such as overall charge neutrality, Eq. (13.19), and the theorem of Gauss, Eq. (6.1). It is no accident that the best models have been developed by noted scientists who knew the multitude of aspects that are important for semiconductor devices.

15.1.1 Bipolar Transistors

The idea of a transistor composed of $p-n$ junctions can be understood from Figure 15.1, which shows a reverse-biased $p-n$ junction into which carriers are injected optically. Obviously, the light source can cause a large generation rate and corresponding current, and therefore can switch on the reverse current to high levels. The light source can be represented by another $p-n$ junction (light-emitting diode) in forward bias, as shown in Figure 15.2. (In fact, a device like this has been proposed, called a beam-of-light transistor.)

The charge flow in Figure 15.2 is as follows. Electrons are pushed from the left side of the left-hand $p-n$ junction to its p side, where they recombine and generate light. Then the light creates electrons at the p side of the reverse-biased $p-n$ junction that contribute to the current. The question arises of why we need the light generation and absorption process. Can't we bring the $p-n$ junctions close enough together so that the electrons from the forward biased junction are directly injected and switch on the current in the reverse-biased junction? Indeed, one can, and the injection of electrons (minority carriers) into a p -type base is the fundamental transistor principle found by Bardeen and Brattain. The resulting band structure is shown in Figure 15.3, which represents the band structure of a bipolar transistor with its three parts emitter (of electrons), base, and collector (of electrons). The voltages $V_{\text{ext}}^{\text{e}}$, and $V_{\text{ext}}^{\text{c}}$ are called the emitter base and collector base voltage and are sometimes also denoted as V_{EB} and V_{CB} , respectively.

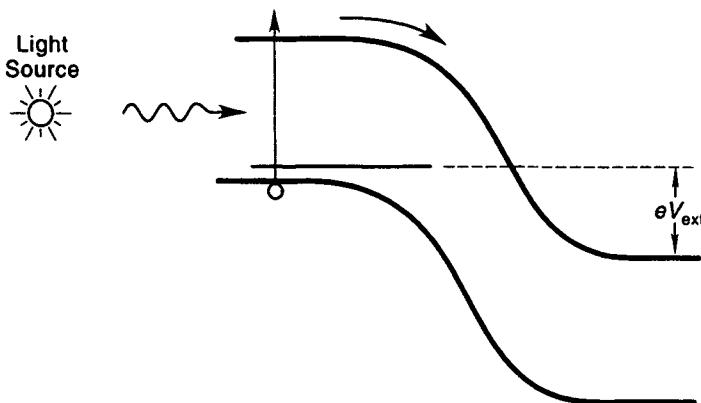


Figure 15.1 $p-n$ junction in extreme reverse bias. The reverse current is caused by thermally generated carriers plus by electrons that are generated optically by a light source and the absorption of light.

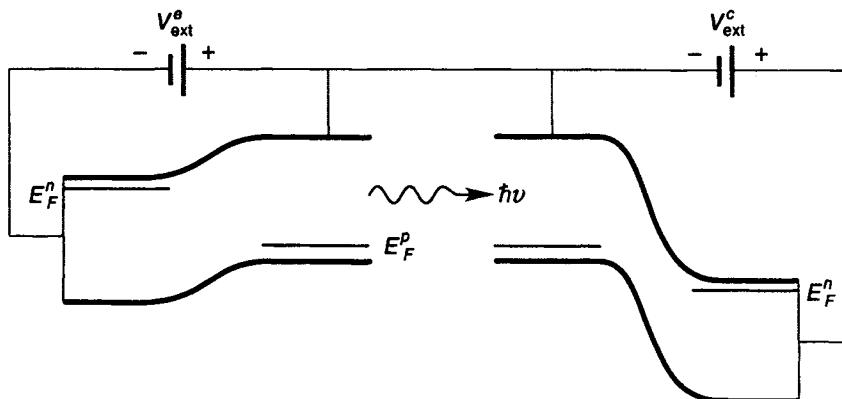


Figure 15.2 Forward-biased (light-emitting) $p-n$ junction triggering current in a reverse-biased $p-n$ junction through light emission.

The theory of this device develops essentially in the same way as the $p-n$ junction theory with the significant difference that in the base the boundary conditions change for the injected electrons. Obviously, for transistor action to occur, the base must be short enough to let electrons (minority carriers) be injected into the collector junction without recombination with the holes (majority carriers) in the base. Therefore, the integration constants need to be recalculated for Eq. (13.50), as a/L_n cannot be regarded as large (infinity) any longer. Also C_2 is no longer equal to zero. Our new boundary conditions are now determined by the presence of the collector junction. To facilitate the discussion, we repeat Eq. (13.50)

$$n(z) = C_1 e^{z/L_n} + C_2 e^{-z/L_n} + C_3 \quad (15.1)$$

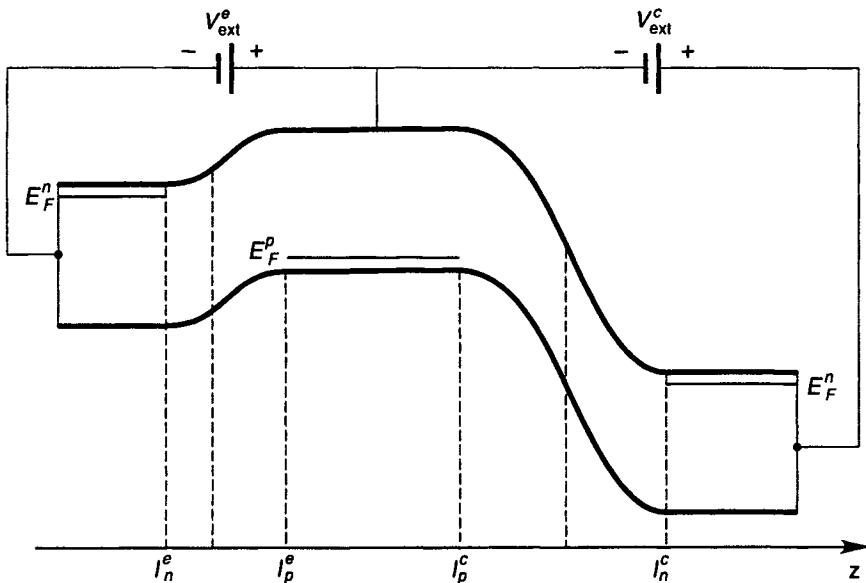


Figure 15.3 Energy band diagram of bipolar transistor under “normal” bias.

At the emitter base borderline l_p^e , we still have [see Eq. (13.52)]

$$n(l_p^e) = n_{\text{base}} e^{\bar{V}_{\text{ext}}^e} \quad (15.2)$$

where we have denoted the equilibrium concentration of electrons in the base by n_{base} and the external voltage and depletion distance carry now a label e for emitter. The notation is also explained in Figure 15.3.

Another boundary condition is obtained from the requirement that C_3 be equal to the equilibrium electron concentration at the p side [$C_3 = n_{\text{base}}$], as can be seen from Eq. (13.49) and replacing $n(-a)$ by n_{base} . The third boundary condition comes from the determination of the electron concentration at the collector base junction l_p^c . This latter concentration is in general not easy to find with precision; its determination requires the detailed knowledge of electronic transport in the reverse-biased collector with the simultaneous injection of electrons. The standard transistor theory inserts for the change Δn from the equilibrium $\bar{V}_{\text{ext}}^c = 0$ concentration

$$\Delta n(l_p^c) = n_{\text{base}}(e^{\bar{V}_{\text{ext}}^c} - 1) \quad (15.3)$$

without much justification. This value represents the true value only if no carriers are injected from the emitter and the collector diode is essentially independent [see Eq. (13.53)]. To highlight the essence of transistor theory, we put $n(l_p^c) = 0$ for the moment, chose l_p^e (i.e., $l_p^e = 0$) as the zero of the coordinate system and denote l_p^c as the (voltage-dependent) base width W_b . Then

$$\Delta n(0) = n_{\text{base}}(e^{\bar{V}_{\text{ext}}^c} - 1) = C_1 + C_2 \quad (15.4)$$

and

$$C_1 e^{W_b/L_n} = -C_2 e^{-W_b/L_n} \quad (15.5)$$

which gives

$$C_1 = \frac{n_{\text{base}}(e^{\bar{V}_{\text{ext}}^e} - 1)}{e^{2W_b/L_n} - 1} \quad (15.6)$$

and

$$C_2 = \frac{n_{\text{base}}(e^{\bar{V}_{\text{ext}}^e} - 1)}{1 - e^{2W_b/L_n}} \quad (15.7)$$

Insertion of Eqs. (15.6) and (15.7) into Eq. (15.1) or Eq. (13.50) and integrating over the generation-recombination rate U_s as in Eq. (13.55) leads to an emitter current contribution that is quite similar to Eq. (13.55) and for $W_b \ll L_n$ practically identical. Notice that the condition $W_b \ll L_n$ is typical for transistors because the electrons need to be injected into the collector without much recombination in the base. This means that the emitter diode functions basically independent of the collector (it does not care much where its electrons are dumped). The collector current density, on the other hand, can be significantly increased. The increase owing to emitter injection can be calculated by assuming that the additional collector current density j_c is a diffusion current at $z = W_b$ and, therefore,

$$j_c = -eD_n \frac{dn(z)}{dz} \Big|_{W_b} \quad (15.8)$$

where the subscript W_b means that the derivative is taken at W_b . $n(z)$ can be obtained from Eqs. (15.1), (15.6), and (15.7). For Eqs. (15.6) and (15.7), it is immediately evident that

$$j_c \propto (e^{\bar{V}_{\text{ext}}^e} - 1) \quad (15.9)$$

which shows that the collector current density is controlled by the emitter voltage (or emitter current).

One can, therefore, write for the collector current I_C (I denotes current in contrast to j , which denotes current density) of an arbitrary bipolar transistor

$$I_C = -I_{CO}(e^{\bar{V}_{\text{ext}}^e} - 1) + \alpha_N I_E \quad (15.10)$$

where α_N is a transfer coefficient, which in principle can be obtained from the theory outlined above and I_{CO} is the reverse saturation current of the collector (with zero emitter base).

In this discussion one can interchange emitter and collector and then obtain

$$I_E = I_{EO}(e^{\bar{V}_{\text{ext}}^e} - 1) + \alpha_I I_C \quad (15.11)$$

where we have included an influence term of the collector on the emitter. This time a coefficient α_I is used where the I stands for inverted (because in the normal operation electrons are not transferred from collector to emitter). Equations (15.10) and (15.11) together with Kirchhoff's law

$$I_B = I_E - I_C \quad (15.12)$$

(where I_B is the base current) are the so-called *Ebers-Moll equations*. The four constants α_N , α_I , I_{EO} , and I_{CO} are usually regarded as parameters and determined by experiments. The Ebers-Moll equations represent an example for an excellent device model that is greatly simplified but retains much of the device physics. It can be used also for realistic device geometries because the geometrical factors are automatically included in the parameters of the theory.

The ac characteristic of the bipolar transistor is more difficult to understand and involves the (not entirely independent) emitter and collector capacitance (see Section 13.2). The interested reader is referred to the charge control model of Gummel and Poon as described by Streetman [8].

We finish this section by discussing some principles that are important for the operation of bipolar transistors. Device performance necessitates high doping in the base because the resistance-capacitance products need to be kept small if high speed is to be achieved. High doping leads to a narrowing of the band gap for several reasons. In Chapter 5 we discussed the effect of band tailing, Eq. (5.16), and its consequences for the effective narrowing of the band gap. The high density of holes or electrons in the base leads to an additional band gap shrinkage owing to many body interactions that have briefly been described in Chapter 10. This many-body shrinkage must be added to the tailing of Chapter 5. We have not placed much emphasis on many body effects in this book. One exception has been the treatment of screening effects. These many-electron (hole) effects are always important for scattering problems, as we saw from the treatment of impurity scattering. The screening of impurities gives us a natural idea of the many body effects that have bearing on the total energy of an electron gas and therefore also to the value of the band gap. Charged impurities are not any different from the electrons except that they are localized. The coulombic repulsion will therefore also lead to screening of the electron potential itself. In other words, each electron will attempt to correlate the other electrons (pushing them away and thus create a "correlation hole" around itself). This of course will lead to differences in the total energy of the system. The Pauli principle adds to this classical coulombic effect as an electron repulses another electron with equal spin. The total effect, the "exchange correlation effect," has been subject to many studies and is by now reasonably well understood [4].

Exchange correlation effects can be approximated by adding an exchange correlation potential as in Eq. (10.34). The simultaneous presence of crystal imperfections, however, gives additional complications to the theory and consequently the experimental results (shown in Figure 15.4) are not so well understood. The puzzling fact of the experiments is that the change in band gap ΔE_G

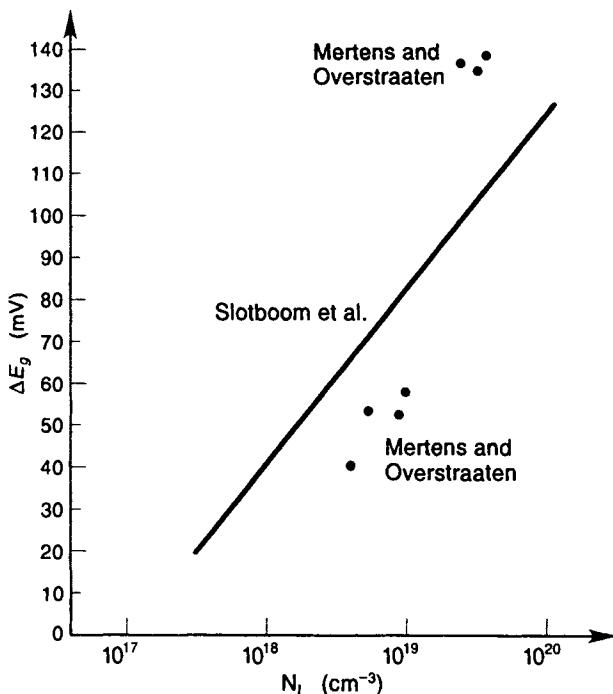


Figure 15.4 Measured band gap narrowing as a function of impurity concentration. [Source: After Mertens and van Overstraaten, ©1978 IEEE.]

increases much more rapidly than with the third root of the doping density N_D , whereas all “simple” theories give an approximate $N_D^{1/3}$ dependence because the average distance between dopants is $\approx N_D^{1/3}$.

Band gap differences in the bipolar transistor, for example, the special case of a smaller band gap in the base relative to the emitter, have numerous consequences on transistor operation. The effects are especially pronounced in heterolayer transistors, which feature large band gap differences. If, for example, the emitter is $n\text{-Al}_x\text{Ga}_{1-x}\text{As}$ and the base $p\text{-GaAs}$ with a smaller band gap, then holes are blocked to flow from the base to the emitter (holes are minority carriers for the emitter), which gives rise to high “emitter efficiencies.” By the same token the base can be doped heavier (lower base resistance) without sacrificing emitter efficiency. As a consequence, large current gain factors

$$\beta = \frac{I_C}{I_B} \quad (15.13)$$

can be obtained.

Hot electron effects can also be of some importance to bipolar transistor operation as the velocity at the base boundaries (immediately to the right of I_p^c in Figure 15.2) can exhibit saturation and even overshoot effects owing to

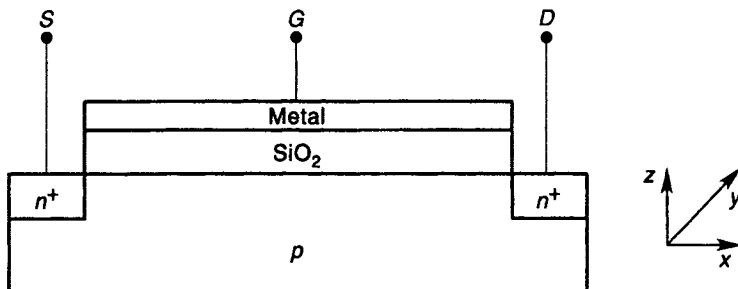


Figure 15.5 Schematic plot of a metal–insulator semiconductor (MIS) field effect transistor.

the high electric fields. The effects, however, are not as large and significant as hot electron effects are in the field effect transistors. Of course, in extreme reverse bias impact ionization can play some role and hot electrons can even be emitted into neighboring oxides. If such effects must be included in the transistor theory, elaborate numerical calculations are necessary. Let us note finally that for high doping levels, Zener tunneling can also interfere with normal transistor operation.

For more advanced treatments we refer the reader to Sze [9] and to commercial simulation tools that can deal with two- and three-dimensional aspects of transistors. Such aspects will be discussed in a little more detail below, in connection with field effect transistors.

15.1.2 Field Effect Transistors

We have treated the field effect in Sections 10.3 and 10.4.1 and have shown in Eq. (10.47) that the application of a voltage to a metal gate will lead to charge accumulation in a semiconductor that is separated from the metal by another semiconductor or (preferably) insulator with larger energy gap. This effect can be used as second transistor principle and is illustrated in Figure 15.5. Ideas of this type were enunciated by Lilienfeld and Heil and later by Shockley in his search for a solid-state device that would replace the vacuum tube.

Two heavily doped n^+ regions (contacts) are diffused (or implanted) into a p -type semiconductor. These contacts are usually denoted by S (for source) and D (for drain). Whatever bias is applied to these contacts, one $n^+ - p$ junction will always be in reverse bias and only a small reverse saturation current will flow. However, as soon as the gate electrode G is given a positive bias, a sheet of inversion electrons (see Figure 10.7) will form at the interface and source and drain will be connected by this sheet, which now forms an $n^+ - nn^+$ diode (see Chapter 11), which for long and homogeneous n regions represents a conducting semiconductor element. In other words, we can switch from a reverse-bias operation to a conducting channel. This is the field effect transistor principle added by the important concept of the inversion layer that was suggested by Bardeen.

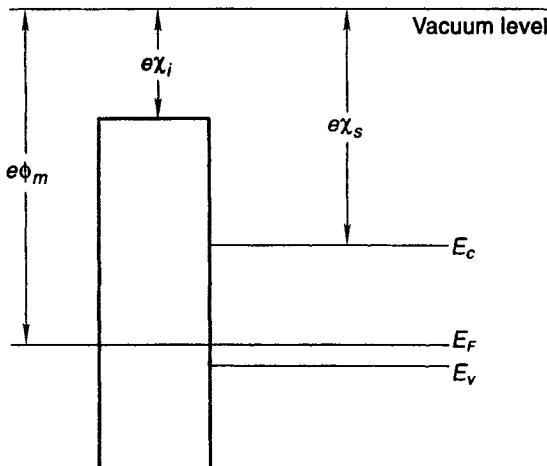


Figure 15.6 Band structure of an ideal metal–insulator–semiconductor structure. Ideal means that the Fermi energy is lined up (straight line) automatically without charge transfer and band distortions (such as in Figure 13.3).

This combination of inversion layer and field effect is an absolute necessity in today's integrated electronics because one desires high on/off ratios of transistor currents ($\approx 10^5$) and low current densities in the off state for reasons of power consumption. All of today's metal–oxide–silicon (MOS) transistors involve an inversion layer.

The transistor type shown in Figure 15.5 is called an *n* metal–insulator–semiconductor (NMIS), or if the insulator is silicon dioxide, an *n*-MOS (NMOS) transistor. It is the insulating quality of SiO_2 (quartz) that makes the MOS system so valuable. Also, the system can be fabricated with few interface states (see Figure 4.6). No semiconductor other than silicon has a comparable oxide. With *n* and *p* interchanged, it would be a *p*-MOS (PMOS) transistor. Of special importance for logic applications are circuits that contain both *n*-MOS and *p*-MOS devices. One then talks about complementary MOS devices or *c*-MOS (CMOS).

Before we proceed with the theory of *n*-MOS transistors, we ask ourselves the question: Where do the inversion electrons (the electrons that switch the transistor on) come from? In the good old days, people talked about “induced” electrons at the surface. To understand what “induced” means, we discuss again the band structure (see Chapters 10 and 13) for the various regions separately (Figures 10.1, 10.4, and 13.1). An ideal combination for all regions (MIS) is shown in Figure 15.6.

Application of a positive voltage eV_{ext} to the metal gate will separate the Fermi energies (now quasi-Fermi energies) by eV_{ext} and will also lead to a redistribution of charge in the semiconductor as shown in Figure 15.7 and discussed in Chapter 10.

The quasi-Fermi level in the semiconductor is a straight line because in the

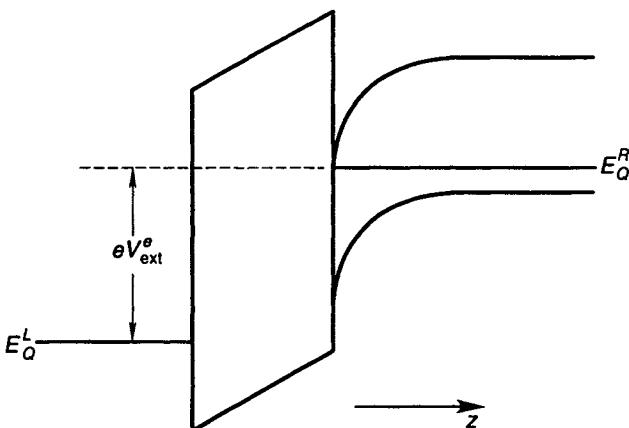


Figure 15.7 Metal-insulator semiconductor (MIS) structure with applied voltage.

ideal case no current flows through the insulator which gives a constant quasi-Fermi level according to Eq. (13.84). As already known from Chapter 10, electrons accumulate at the interface and an electron channel forms at the interface. If the structure consists only of MIS, the electrons at the interface have to be generated thermally (or optically), as discussed in Chapter 9 on generation-recombination. This process is, of course, slow. Especially at low temperatures it would take years to establish equilibrium. Such a switch would therefore be useless. Fortunately, there are electrons available in a transistor structure in the n^+ contact regions (Figure 15.5) and these can be pulled into the channel so that the switch on time in the ideal case is equal to channel length divided by an average electron velocity. This time is for channel length of 10^{-4} cm of the order of picoseconds.

The total charge “induced” by the gate voltage has been derived previously in Eq. (10.47). We subdivide this charge now into two contributions. Q_i is the charge of the mobile inversion layer electrons. It is this charge that contributes to the current. The second part is the fixed charge of acceptors Q_A in the $p-n$ junction depletion layer, which is formed in the region where E_{QF}^R is close to the middle of the band gap in Figure 15.7. Because we need to apply a voltage between source and drain in addition to the gate voltage, the surface potential becomes a function of x (the direction of current parallel to the interface). Equation (10.47) reads then

$$-Q_i - Q_A = C_{\text{ins}}(V_G - \phi_s^R(x)) \quad (15.14)$$

It is important that relatively high voltages V_G can be applied to the SiO_2 without breakdown and, therefore, high densities of inversion charge can be achieved. This is unique for the MOS system.

We now assume that the source electrode is grounded and the gate voltage is such that we have achieved inversion. The interface potential is then equal to the saturated value ϕ_s^R as given by Eq. (10.31) plus an external x -dependent potential

equal to 0 at the source, equal to the drain voltage at the drain, and x -dependent $V(x)$ in between such that

$$\phi_s^R(x) = 2kT_c \frac{\ln(N_A/n_i)}{e} + V(x) \quad (15.15)$$

and the acceptor charge is

$$Q_A = -eN_A W = -\sqrt{2\epsilon\epsilon_0 e N_A \left(V(x) + 2kT_c \frac{\ln(N_A/n_i)}{e} \right)} \quad (15.16)$$

where use has been made of Eq. (10.17) for the depletion width W with the built in potential replaced by $\phi_s^R(x)$. The channel (inversion layer) conductance Δg_c of a small section of length x then becomes

$$\Delta g_c = \mu |Q_i| \frac{W_y}{\Delta x} \quad (15.17)$$

where W_y is the width of the transistor in y -direction. We now assume that the mobility μ is constant (independent of the electric field) and neglect all hot electron effects and the diffusion current. Then we have from Ohm's law

$$\Delta g_c \Delta V(x) = I_D \quad (15.18)$$

where I_D is the drain current, which is in steady state constant and the same everywhere. Collecting all terms in differential form ($\Delta \rightarrow d$), we get

$$W_y \mu [C_{ins}(V_G - \phi_s^R - V(x)) + Q_A] dV(x) = I_D dx \quad (15.19)$$

Because I_D is constant the integration of this equation is straightforward (see Problems section) and gives for small drain voltages V_D

$$I_D \approx \frac{W_y}{L} \mu C_{ins} (V_G - V_{th}) V_D \quad (15.20)$$

where L is the channel length in x -direction and

$$V_{th} = 2kT_c \frac{\ln(N_A/n_i)}{e} + \sqrt{2\epsilon\epsilon_0 e N_A \left(2kT_c \frac{\ln(N_A/n_i)}{e} \right)} \quad (15.21)$$

V_T is called the threshold voltage, which represents the voltage below which the channel is essentially nonconducting.

There is a subthreshold current, however, which is not included in Eq. (15.20) because it is essentially a diffusion current and diffusion has been neglected so far. For large V_D values, I_D approaches a saturated value as soon as the inversion charge goes to zero at the drain. It is clear that as V_D increases to more positive values, the potential difference between channel and gate approaches zero and the carrier concentration at the drain side decreases. Therefore the resistance increases steeply, giving rise to I_D values below the Ohmic line. It is easy to show

from the integration of Eq. (15.19) that the saturation current $I_{D_{\text{sat}}}$ is proportional to

$$I_{D_{\text{sat}}} \propto (V_G - V_{\text{th}})^2 \quad (15.22)$$

For small devices the saturation is strongly influenced by hot electron phenomena, which are described in Section 15.2, and Eq. (15.22) has little validity. However, the reason for the saturation of Eq. (15.22) without hot electron effect deserves discussion. We have mentioned already that the density of inverted electrons decreases as the drain voltage approaches the gate voltage. At $V_D = V_G - V_{\text{th}}$, the concentration of mobile carriers at the drain side approaches zero. One calls this phenomenon the “pinch-off” effect. Beyond this pinch-off point, the current I_D saturates. In this regime, the channel can be essentially subdivided into two parts: the part that still contains inversion electrons and is described by the above equations and the pinched-off region, which, according to the above calculation, does not contain much mobile charge. If we view the existing electron channel (inversion layer) as an extension of the source n^+ contact, then the channel represents essentially a n^+nn^+ diode with the n region being eventually depleted. We have treated the n^+nn^+ diode in Chapter 11 and found the space charge limited current density with [Eq. (11.36)] and without [Eq. (11.38)] hot electron effects included.

The voltage V_{sp} , which controls the space-charge limited current, is given by the drain voltage minus the voltage $V(x)$ at the pinch-off point, which is denoted by V_{pi}

$$V_{\text{sp}} = V_D - V_{\text{pi}} \quad (15.23)$$

We now perform a “gedanken” experiment. If the space-charge limited current were smaller than the saturation drain current, we could have the case of negative differential resistance, which means that as we increase the drain voltage, the drain current decreases. In such a situation, however, the voltage drop would continue to increase over the region of negative resistance and decrease in the rest of the channel region, leading to an increase in space-charge limited current up to the value of the saturation current. An increase beyond $I_{D_{\text{sat}}}$ is limited by the inversion channel resistance. In other words, the pinched-off range will adjust itself such that the current through the device stays saturated and we arrive at a transistor current-voltage characteristic as shown in Figures 15.8 and 15.9. Without the possibility of a space-charge limited current in the pinch-off region, the current would vanish instead of saturate which is, of course, never observed. Any negative differential resistance in transistors that is (and indeed can be) observed is related to spatial transfer of electrons out of the channel, for example, to real space transfer caused by hot electrons. In MOSFETs, this would indicate excessive interface traps.

Finally we briefly discuss the current below the onset of inversion. In the subthreshold region $V_G < V_{\text{th}}$, the channel is essentially a n^+pn^+ structure with p being close to the intrinsic concentration. Therefore, the conducting part of

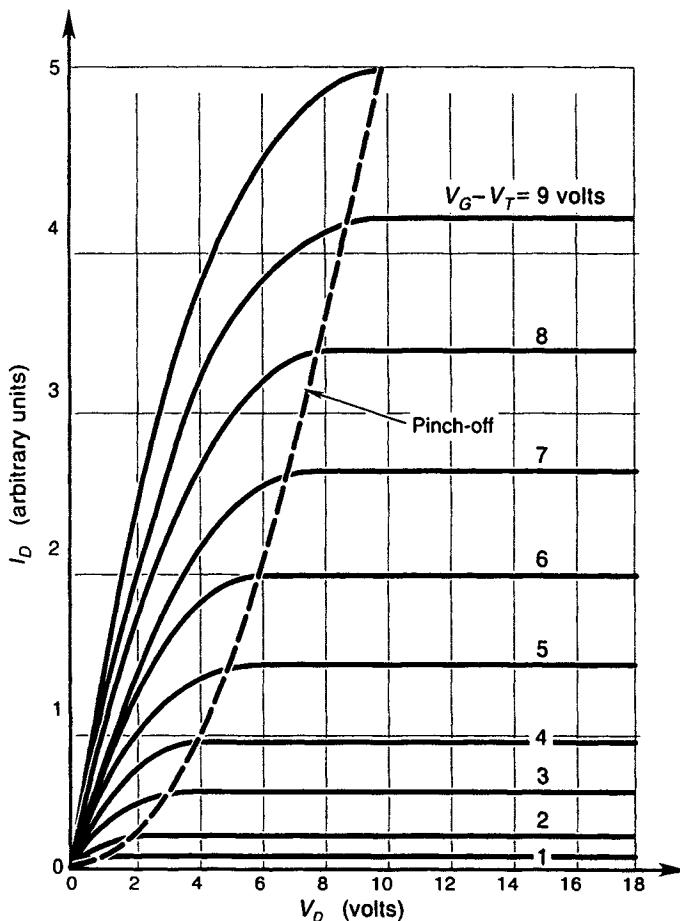


Figure 15.8 Idealized drain current versus drain voltage characteristics of a MOS transistor for various values of the gate voltage V_G . These are classical curves corresponding to large transistors.

the MOS transistor is actually a bipolar transistor with the bulk of the channel forming the base. The current in the bipolar structure is a diffusion current and can be calculated from Eq. (15.8). Notice, however, that the carrier concentration depends on the surface potential.

It is clear that for a small device the calculation has to be performed numerically because the results will depend sensitively on the charge distribution around the source and drain contact and not only at the interface. The appropriate device model has to be, at least, two-dimensional.

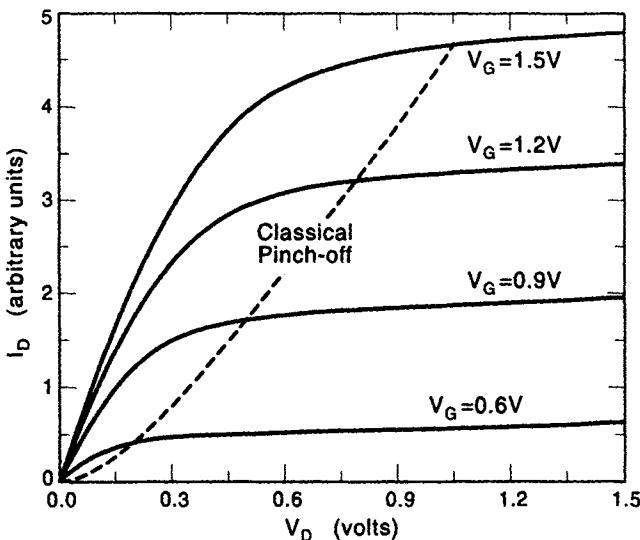


Figure 15.9 Drain current versus drain voltage for a submicrometer channel length transistor.

15.2 EFFECTS OF REDUCTION IN SIZE, SHORT CHANNELS

The reduction in device size as required by attempts to integrate large circuits and systems on a chip, and the desire of high performance, lead to complications on the device level. For example, if the device should have a size of $0.1 \mu\text{m}$ and the various applied voltages are of the order of 1 to 2 V, electric fields of up to $2 \times 10^5 \text{ V/cm}$ and more will be present in the device. The idealized theory of the previous sections will therefore be invalidated because the mobility is not constant and other high field effects become important. In this section we discuss some of the effects associated with short channel length and high electric fields. We limit ourselves mostly to field effect transistors because this is a vast field, still subject to much research.

15.2.1 Scaling Down Devices

How many devices fit on the tip of a pin? We can give a rough estimate as follows. Most semiconductor devices contain at least one $p-n$ junction. Any $p-n$ junction has at least a length of its depletion layer width W , which is for a p^+-n junction given by Eq. (10.17). The integration of devices uses mostly planar technology; that is, the devices are essentially in a plane and the third dimension is not as critical and can only help (by stacking up devices) for size reductions. If we assume, then, a minimum device area of W^2 , we can fit on a square centimeter $N_{\max} = 1\text{cm}^2/W^2$ devices, and N_{\max} is given by

$$N_{\max} = \frac{e^2 N_D}{2\epsilon\epsilon_0 E_G} \quad (15.24)$$

Table 15.1 Scaling Schemes

Variable	Constant Field	Constant Voltage	Mixed Scheme
L	l/K	l/K	l/K
d	l/K	l/C	l/K
N	K	K	K
V	l/K	l	l/C

We have replaced here the built-in voltage by its approximate value of E_G/e . Equation (15.24) demonstrates the dependence of “integrability” on material constants (E_G , ϵ , and the donor concentration). In particular, the maximum number of devices on a chip is directly proportional to the doping concentration. Therefore, if we know an ideal chip that makes optimal use of the area and if we want to increase the number of devices by a factor of K , then we have to scale (increase) the doping by the same scaling factor as long as Eq. (15.24) determines the maximum number of devices.

Scaling requirements may be derived from other considerations as well. If, for example, our devices were all designed to follow the ideal device equations with constant mobility and if we have a working chip, then we might postulate that it is scaled by keeping the electric field constant because the mobility will then stay invariant. This means that if we scale the device area down by a factor of K^2 , the voltages, oxide thickness d , and channel length L have to be reduced by a factor of K . The doping needs to increase by K [Eq. (15.24)] and the transit time through the channel will decrease by K automatically (higher speed). Also, the power dissipation per device will be reduced by K^2 . These latter facts are, of course, very desirable.

Although scaling rules as discussed have been useful guides, they never have been strictly adhered to in practice because scaling has been influenced by limitations in fabrication technology. Table 15.1 lists three scaling schemes that summarize the rules introducing a constant C with $1 < C < K$ [5].

These rules have no strict theoretical truth content. For example, the electric field in the channel is a highly nonlinear function of x and the maximum field in the channel therefore does not scale linearly with the gate length as implicitly assumed. Thus a general theory of scaling does not exist. However, numerical simulations with the elaborate commercially available tools do give significant help and can be used to confirm or discard scaling strategies. Simple phenomenological models have also been developed [5].

15.2.2 Short Gates and Threshold Voltage

Real MOS devices look considerably different than our sketch in Figure 15.5. The gate metal is often replaced by highly conducting polycrystalline silicon (POLY), and there are typically many layers of metalization. These are a necessary part of the metal interconnections of devices that are arranged in the third

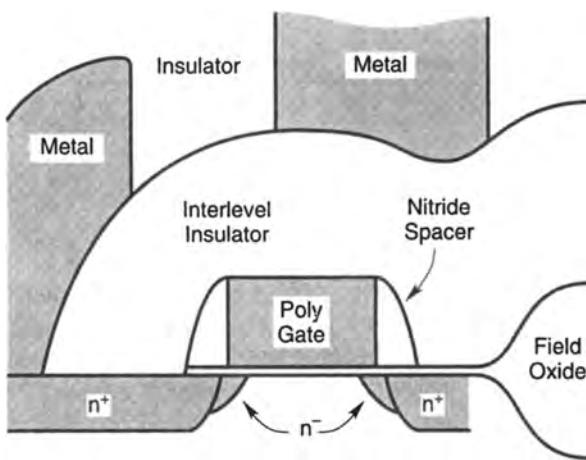


Figure 15.10 MOSFET device structure illustrating the complicated geometry and embedding in various layers.

dimension (z -direction). The device is also embedded in various insulators and can have spacers (nitride) on the side of the gate. The actual gate oxide between gate and channel is extremely thin (50 Å or less) and needs to be grown with great care, as we will discuss in Section 15.3. The channel is typically engineered to reduce the maximum electric field which can be accomplished by the introduction of a separate lightly diffused drain (LLD) region. All of this is shown in Figure 15.10, the lightly diffused region is indicated by n^- . As a consequence of these complexities in device geometry, there are several additional effects that have not been considered in our simplified treatment. All of these effects are easily calculated numerically and, in fact, it is advisable to use commercial simulators (MINIMOS, PISCES, DESSIS, etc.) whenever such effects are of interest.

The short channel effect arises from the fact that the charge under the gate is to some extent controlled by the source and drain junctions. In the simplified treatment we had assumed that the gate controls the entire bulk charge Q_A from Eq. (15.18). This indicates the importance of the two-dimensional nature of the charge and field distributions, which is best dealt with device simulators that solve the Shockley equations. Note, however, that simplified computer aided design models are based on charge sharing assumptions and do not exactly solve the equation of Poisson.

The narrow-width effect (of lesser importance) is a consequence of the gradual transition of the gate oxide to the thicker field oxide as indicated in Figure 15.10. This effect of oxide encroachment gives rise to an extra depletion charge caused by the fringing fields around the gate edges. Again a charge additional to Q_A must be accounted for. Both effects influence the value of the threshold voltage V_{th} .

Another well-understood effect, observed in very small devices, is the drain-induced barrier lowering (DIBL). For higher drain voltages, the high electric field at the drain end starts to penetrate toward the source. This lowers the electrostatic barrier at the source. The existence of such a barrier is general and is illustrated in Figure 8.5 at the left contact (reservoir). This again leads to a threshold voltage shift and can only be quantitatively understood by numerical simulation.

15.3 HOT ELECTRON EFFECTS

Hot electron effects can be divided into mobility related phenomena (effects that depend mainly on how the bulk of the charge carriers behave) and effects of the high-energy tail of the distribution function (effects that are mostly related to reliability questions). We treat both areas separately, although self-consistency often will demand that they are dealt with together.

15.3.1 Mobility in Small MOSFETs

The mobility in MOSFETs depends, of course, on all the scattering mechanisms that we have discussed in Chapter 7 and becomes a function of the heating electric field as discussed in Chapters 13 and 8 as well as Chapter 11. Scattering by ionized impurities can be by channel impurities and follows for reasonably high inversion densities the Brooks Herring formula. It can also be by the remote ionized impurities residing in the oxide, which then requires the treatment described in Section 10.5. In addition to all these scattering mechanisms, one must include surface roughness scattering. As described in Chapter 10, the wave function is confined to a narrow well at the interface, which becomes narrower as the gate voltage causes higher inversion layer densities. Imperfections of the interface become then very strong scattering agents. Interface scattering (often also called surface scattering) is different when a half-space bulk semiconductor is considered or, as another extreme, the electrons are narrowly confined as a quasi-two-dimensional gas at the interface. The effects of the random interface potential fluctuations on a size quantized electron gas have been described by Ferry, et al. [1]. The classical treatment of scattering at interfaces for charge carriers in a half space has a long history. Depending on interface defects and the \mathbf{k} -vector of the charge carriers the scattering after the interface is hit can be diffuse (i.e., randomizing) or specular. If the de Broglie wavelength is smaller than the scattering object it will be diffuse otherwise more specular.

From all of this, it becomes clear that the mobility depends in a very complex way on the electric field. The field component perpendicular to the interface tends to confine the electrons and compress the wave function toward the interface and thus increase the interface (surface) roughness scattering. The field parallel to the interface, on the other hand, tends to heat the electrons. Actu-

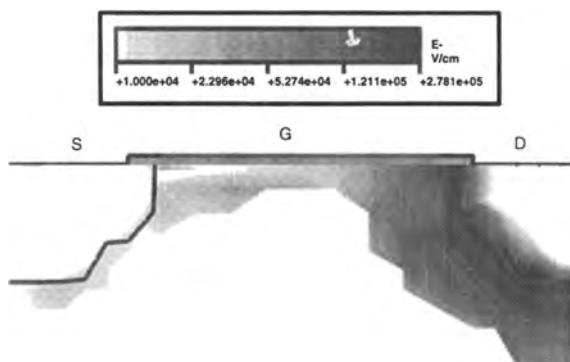


Figure 15.11 F_x as function of space coordinates x, z in the silicon with a maximum at the silicon dioxide boundary.

ally, as we know from Chapter 13, it is the electric field parallel to the current that heats the electrons and this is, in general, not just F_x . The heating of the electrons leads to a redistribution in the potential well that confines the charge carriers owing to real space transfer effects. Higher drain voltage means more heating, which ultimately leads to a transition from size quantized state to a three-dimensional deconfined electron gas on the drain side. Correspondingly, the size quantized interface scattering also encounters a transformation to classical half-space scattering. This in turn is specular for the lower electron energies but becomes diffuse at higher electron energies.

It is clear that such a complicated scattering problem calls for Monte Carlo simulations. However, the transitions from quantized to classical and from specular to diffuse are not easily included even into such an advanced approach. Therefore phenomenological fits to experimental data are used in device simulations. Examples are given in Ko [5]. Figure 15.11 shows a relief of F_x as a function of x, y in a short channel MOSFET under typical operating conditions as calculated from a Shockley set by a commercial simulator.

It can be seen that the field increases in a nonlinear fashion toward the drain reaching a peak of $\approx 10^5$ V/cm. This means that the electrons are heated far above room temperature to thousands of degrees Kelvin (or Celsius, or Fahrenheit). Electron temperature effects are therefore very important. The distribution of the perpendicular field F_z is shown in Figure 15.12. This shows that there exists strong size quantization at the source, but at the drain side, the electrons are essentially three-dimensional, also because of heating and real space transfer. These effects cause, of course, significant deviations from the formula for the drain current as given in Eq. (15.23). Only numerical simulation can give a just account of all of this and the solution of the Shockley equations with a mobility μ depending on both F_x, F_z , as obtained from fits to experimental data, does give excellent results even for extremely small (submicrometer) feature sizes of the MOSFETs.

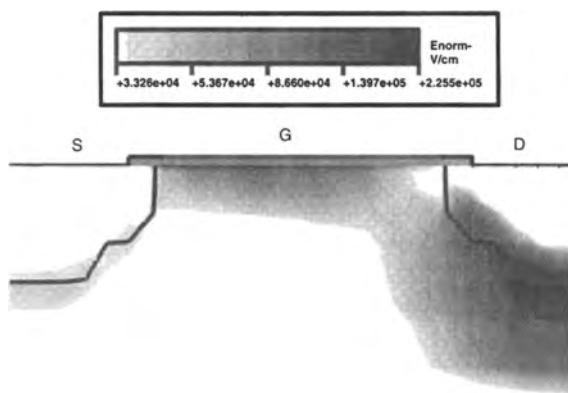


Figure 15.12 Same as Figure 15.11, but for F_z .

Why is the numerical solution of the Shockley equations that good? We have already discussed that the Shockley set would be exact if we knew the current as function of field and carrier density. This is accomplished to a fair extent by using a mobility and diffusion constant obtained from fits to the experiments. Luckily, it turns out that the diffusion constant does not vary much with the electric field in silicon; according to the Einstein relation, it is proportional to $T_c\mu$ and T_c increases with the electric field F_x while μ decreases. There is still the transition from size quantized carrier gas at the source to the heated classical three-dimensional gas at the drain. However, this again can be included by using appropriate fitting formulae for the mobilities and, as pointed out in Chapter 10, by including a quantum capacitance. I would like to emphasize here again that, owing to the heating, the mean free path of the charge carriers decreases significantly. Close to the drain the scattering rate will typically approach 10^{14} s^{-1} corresponding to a mean free distance between the scatterings of 10^{-6} cm . Therefore, the wave function will be dephased within this distance and the transport will be quasiclassical down to device sizes much larger than this distance, (i.e., down to typical channel length of 0.1 micrometer). For overall extremely high fields, the drain current is then limited by the saturation velocity v_s in all the channel and one obtains as a limiting equation

$$I_D \approx eC_{\text{ins}}[V_G - \phi_s^R(O)]v_s \quad (15.25)$$

This limiting case is included in the Shockley equations through the field dependent mobility. Figure 15.9 does indeed show an early saturation of the current as predicted by Eq. (15.25).

Velocity overshoot and real space transfer are serious causes for a breakdown of the validity of the Shockley transport equations because the mobility and diffusion constant become then functionals of the electric field instead of functions (i.e., depend on the field in a nonlocal fashion). Then the mobility depends on the whole field profile in the device and not only on the local fields. It

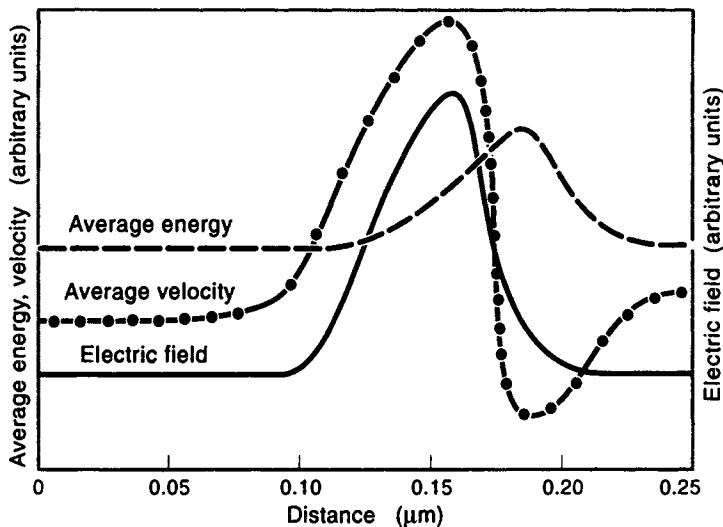


Figure 15.13 Typical electron temperature and drift velocity for rapidly varying electric fields.

is, of course, thinkable to parametrize the mobility for all possible field distributions. However, this certainly is complicated and the Stratton approach or a Monte Carlo simulation is probably more satisfactory. A typical case of velocity overshoot is shown in Figure 15.13. At the region of rapidly rising electric field, the average energy of the electron gas (electron temperature) rises with a slight delay because the evolution of a carrier temperature takes time. Because the mobility decreases with the electron temperature (for predominant phonon scattering), the drift velocity $v = \mu F$ overshoots the value that it would have if the electric field were constant. However, the effect is small; at most a factor around two and occurs only over the short range of rapidly varying field. Well-designed MOSFETs also have lower field gradients which further reduces the effect.

Real space transfer out of the confining well toward the bulk is also an effect of smaller significance but is of interest under some circumstances. For very strong heating, the electrons can even be emitted thermionically into the silicon dioxide. This effect was known long before the term real space transfer was used [7]. It can be understood, however, on this basis.

15.3.2 Impact Ionization, Hot Electron Degradation

The high-energy tail of the energy distribution of electrons and holes has two major consequences relevant for MOSFETs: impact ionization and hot electron structural damage. Impact ionization has been described in some detail in Chapter 13. We add here only few specifics and then turn to structural damage.

The main differences of impact ionization in MOSFETs compared to our previous treatment arise from the rapidly varying electric field and the finite volt-

age. If one assumes a constant field, the heating of the electrons is according to this field and has no principal upper limit. If, however, the field varies, then the heating is delayed as seen above. For a given drain voltage V_D the electrons cannot be energized by the electric field above eV_D . Phonon absorption and electron-electron interactions can create a high-energy tail above eV_D . However, any strong scattering rate, such as impact ionization, can readily empty the high-energy tail. This has already been illustrated in Figures 12.1 and 12.2. Therefore, a dead space for impact ionization needs to be added to our previous treatment to account for the delayed heating of electrons in rapidly varying fields and an upper limit of the electron energy somewhere around eV_D needs to be included. Both effects call for Monte Carlo solutions as obtained, for example, by the commercial simulator DAMOCLES.

Impact ionization in MOSFETs leads to a substrate current I_{sub} that is caused by the generated minority carriers. Majority electrons in n -channel MOSFETs create minority holes and vice versa. This substrate current is easy to measure and can be measured with great sensitivity (e.g., between the source and a bulk silicon contact). I_{sub} therefore represents a favorite indicator for hot electron effects. Whenever they become significant, hot electron effects lead to a dangerous change in device characteristics. In turn, they can give clues about more complicated other hot electron effects as long as these have a similar soft threshold behavior. Certain types of hot electron structural degradation are of this type and will be described next.

Several important structural degradation effects that lead to transistor aging are linked to the gate oxide and the interface between the silicon channel and the gate oxide. The gate oxide and its reliability are of great importance for the operation of MOS transistors. One would desire a perfectly lattice matched insulator to avoid strains and defects. Unfortunately, silicon dioxide is far from lattice-matched and the fit of it on top of silicon necessarily involves bond angle distortions as shown in Figure 4.6. Thermodynamic considerations show then, that there must be defects of various forms present at such an interface. One of the most investigated defects is the dangling bond shown in Figure 4.4. Indeed it is known that a relatively high density of the silicon interface atoms cannot find oxygen or another silicon as bonding partner because at some point the bond angle distortions would be too large. Naturally, the gate oxides are grown in such a way that interface imperfections are minimized. Nevertheless, no matter how carefully the oxide is grown, a density of approximately 10^{11} cm^{-2} of dangling bonds will remain. These then behave like Shockley-Read-Hall centers and can capture electrons and holes depending on the location of the quasi-Fermi levels. If an electron is captured from the channel, this electron is counted as interface charge but does not contribute to the current. In other words, this will cause a threshold voltage shift. If N_{it} electrons per square centimeter are captured by these interface traps, then the threshold voltage shifts by $\Delta V_{\text{th}} = N_{\text{it}}/C_{\text{ins}}$. Because the distribution of the interface traps is statistical, this can lead to unacceptable threshold voltage fluctuations on a chip. Therefore an “annealing” pro-

cedure involving hydrogen is generally performed as one of the final steps in the transistor and chip fabrication processes. The chips are baked in an atmosphere containing hydrogen at a temperature of around 450°C . The hydrogen diffuses easily through the silicon dioxide within a short time (minutes to hours depending on the number of layers it has to go through). It finally leads to a saturation of the dangling bonds by forming Si-H at the interface as shown schematically in Figure 15.14. This stabilizes the interface electronically and prevents capture of channel electrons. In other words the interface trap level disappears. There are also defects within the oxide that communicate with the channel by tunneling or thermionic emission. Naturally the trapping times involved are much longer than for the traps directly at the interface. The defects in the oxide can be of various forms and probably are often related to oxygen vacancies. We will not deal here with defect chemistry, which would deserve a separate book. We are here only interested in hot electron aging related to the defects directly at the interface because this process has reached a degree of understanding that will permit the development of first principle theories.

The bonding energy between silicon and hydrogen is rather small. If one wants to remove the hydrogen from the crystal, an energy around 3.6 eV is necessary. To remove the hydrogen from its bond and put it somewhere within the silicon or the silicon dioxide takes still less energy. Structural calculations show that this energy can be 2 eV and less. The channel electrons do have such an energy available as can be seen from the distribution functions in the previous chapters. Even if the supply voltages drop below 2 eV, electron-electron interactions will create a high-energy tail above 2 eV. As a consequence, the energy necessary for desorption of the hydrogen is available during operation and there-

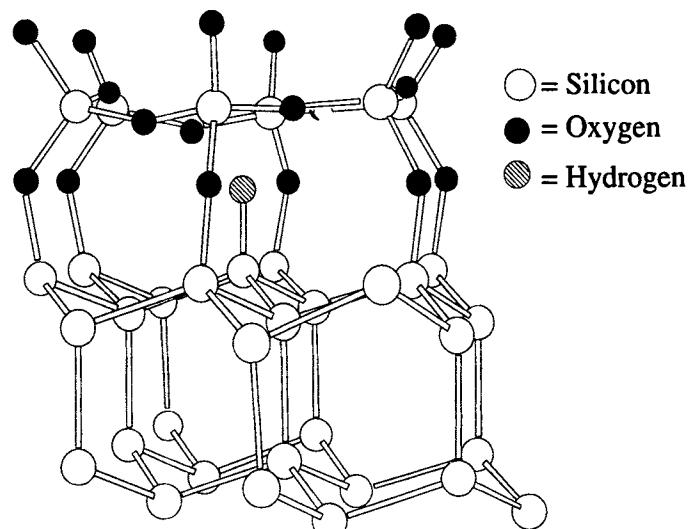


Figure 15.14 Hydrogen passivated interface between silicon and silicon dioxide.

fore transistors can age. Chemical reactions of this type at interfaces are well known, although the energy is usually supplied by photons and the reactions are described by photochemistry. Electrons do usually not have such high energies in materials, and semiconductor devices present here something new: hot electron desorption chemistry.

How does the hot electron desorption work? An important clue to this was found recently when aging tests of chips were performed with hydrogen anneals and with deuterium anneals. Deuterium is a stable isotope of hydrogen found in great abundance in the water supplies of the world. Its mass is twice that of hydrogen, but otherwise it is identical. It has the same electronic energy levels and chemical binding energies. Nevertheless it was found that chips treated with deuterium did not age as much during operation involving hot electrons as the hydrogen annealed chips did [2]. This isotope effect hints toward the involvement of lattice vibrations and distortions in the degradation mechanism. The hot electron cannot kick the much heavier hydrogen directly off the interface as in a game of billiards; energy and momentum conservation would not allow this. However, we do know that large amounts of energy can be transferred from the electrons to the crystal atoms, for example, by the electron phonon interaction. This, in turn, depends on the atomic mass and will therefore show an isotope effect. A detailed theory has not been developed at this time; however, two possibilities are described in Hess et al. [2]. One has been advanced by C. G. van De Walle and is based on the fact that Si-H vibrations cannot dissipate their energy to the bulk silicon phonons because of a large mismatch in the vibrational energies. The Si-D vibrational energies are lower owing to the larger mass of the deuterium (D) and are better matched to the bulk and dissipate the energy easier. This makes it plausible that collisions with hot electrons can lead to very high vibrational excitation of Si-H and to desorption, with a much smaller effect for Si-D. Indeed orders of magnitude larger lifetimes have been found for deuterium treated transistors.

The determination of transistor and chip lifetime is a difficult problem. We know already how difficult it is to obtain a precise distribution of hot electrons at very high energies. This involves the band structure and Monte Carlo solutions of the Boltzmann equation. Then one needs, obviously, the vibrational modes of Si-H (Si-D) and their coupling to the bulk materials as well as the desorption energies necessary for removing hydrogen (these energies are the same for H and D). This latter information can only be obtained from structural calculations. Simulations involving the atomistic structure have made large progress. One typically uses local density functional theory as described in Inkson [4] and pioneered by Kohn and Sham. Commercial codes for this are available but require large computational resources. Altogether, one can see that transistor aging of this type presents a challenge for device theory. A precise lifetime prediction is beyond current theoretical means. Experimentally, one also cannot determine transistor lifetime. For example, if one requires a lifetime of 10 years, one would have to run at least a number of transistors for 10 years. The cycles of new prod-

ucts are, as is well known from Moore's law, around 18 months. Even if one requires only an 18-month lifetime, one would have to measure for 18 months before marketing a technology. This is, of course, out of the question.

What one wants therefore is an accelerated stress procedure that ages the transistor very fast and from which one can then interpolate the actual lifetime. Such a procedure is in place and works as follows. The transistor degradation during operation can be characterized by the threshold voltage shifts that go along with hydrogen desorption and the creation of dangling bonds owing to hot electrons. One can then determine from circuit design considerations an unacceptable threshold voltage shift. If this is around 10 mV, the transistor is pronounced dead when this shift is reached and its lifetime is determined accordingly. Therefore one can stress the transistor at higher voltages, plot the lifetime and then extrapolate to lower voltages. But how does one extrapolate? It was noted [3] that the substrate current owing to impact ionization I_{sub} is not only correlated clearly to hot electrons but also to the degradation by hot electrons. As we know from the previous treatment, impact ionization has a soft thresh-

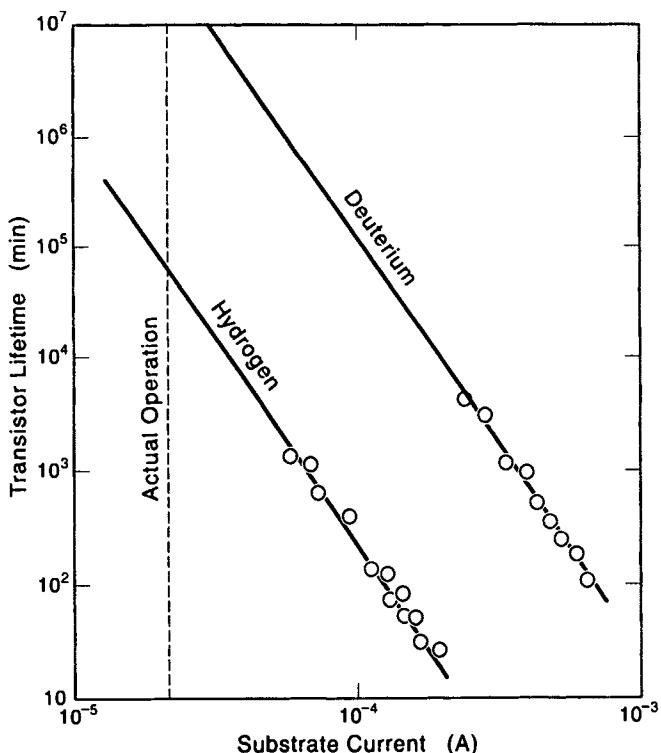


Figure 15.15 Transistor lifetime versus substrate current under accelerated stress conditions and extrapolated to actual operating conditions. The lifetime under actual operation is desired to be around 10 years and determined from this extrapolation.

old behavior involving threshold energies above the value of the band gap. It becomes effective 2 eV above the band gap. This is an energy comparable to the desorption energies for hydrogen removal. It is therefore customary to plot the transistor lifetime τ_{tra} versus the substrate current I_{sub} , all measured at high substrate current corresponding to short lifetimes. Then an extrapolation is made to the substrate current under actual operating conditions. If this is around ten years, the chips are declared reliable. A typical result is shown in Figure 15.15 for both deuterium and hydrogen treatment and shows the advantages of using deuterium. The precision of the extrapolation is not entirely clear, and a complete theory of desorption as outlined above could significantly improve the lifetime predictions. Finally, we note that there also exist other mechanisms of hot electron degradation, such as hole trapping in silicon dioxide.

REFERENCES

- [1] Ferry, D. K., Hess, K., and Vogl, P. "Physics and modeling of submicron insulated gate field effect transistors," in *VLSI Electronics*, vol. 2, pp. 68–104, eds. N. G. Einspruch and G. Sh. Gildenblatt. New York: Academic, 1981.
- [2] Hess, K., Kizilayalli, I. C., and Lyding, J. W. "Giant isotope effect in hot electron degradation of metal–oxide–silicon devices," *IEEE Transactions on Electron Devices*, vol. 45, 1998, pp. 406–416.
- [3] Hu, C., et al. "Hot-electron-induced MOFSET degradation: model, monitor, and improvement," *IEEE Transactions on Electron Devices*, vol. ED-32, 1985, pp. 375–385.
- [4] Inkson, J. C. *Many-Body Theory of Solids*. New York: Plenum, 1984.
- [5] Ko, P. K. "Approaches to scaling in advanced MOS device physics," *VLSI Electronics*, vol. 8, pp. 1–37, eds. N. G. Einspruch and G. Sh. Gildenblatt. New York: Academic, 1989.
- [6] Mertens, R., and Van Overstraaten, R. "Measurement of the minority carrier transport parameters in heavily doped silicon," *IEEE Technical Digest, International Electron Devices Meeting*, 1978, p. 322.
- [7] Ning, T. H. "Hot-electron emission from silicon into silicon dioxide," *Solid State Electronics*, vol. 21, pp. 273–282, 1978.
- [8] Streetman, B. G. *Solid State Electronic Devices*, 2nd ed. Englewood Cliffs, N.J.: Prentice-Hall, 1980.
- [9] Sze, S. M. *Physics of Semiconductor Devices*, New York: Wiley, 1981.

CHAPTER 16

FUTURE SEMICONDUCTOR DEVICES AND THEIR SIMULATION

16.1 NEW TYPES OF DEVICES

Semiconductor devices have been developed and proposed in the first half of the twentieth century with the declared goal to replace the vacuum tubes and devices of gaseous electronics by counterparts in semiconductors or, generally, solids. It was hoped from the very beginning that the higher density of the solids would permit the construction of much smaller devices. This hope is generally seen as fulfilled, and the ever shrinking size of devices and increasing number of devices in integrated circuits appears to provide ample of confirmation. A careful examination, however, reveals that the dream was not fulfilled and the high density of atoms in solids is only partially used to miniaturize devices. The semiconductor provides the possibility of electron and hole conduction and the existence of an energy gap, all consequences of the principles of quantum mechanics. The semiconductor also is used as a matrix onto which devices are printed. Heterojunctions are used and are the basis for the MOSFET as well as for the quantum well laser diode (QWLD). However, all semiconductor devices use dopants and the vast majority of them is based on $p-n$ junctions; every MOSFET contains at least two $p-n$ junctions. The dopants have typically a density of 10^{18} cm^{-3} and may be even an order of magnitude higher. They do not approach the density of atoms in the semiconductor, which is of the order of 10^{23} cm^{-3} . In other words, only every one hundred thousandth atom is a doping atom and the dopants thus constitute a dilute gas within the solid. Semiconductor electronics is based on this dilute gas; this fact presents a challenge for the future.

16.1.1 Extensions of Conventional Devices

The novel device types that have successfully made the step to commercial use show, in some respect, the trends toward higher densities and a modified use (or lack of) of dopants. An example is the resonant tunneling diode. The basic principle of this device is illustrated in Figure 16.1. It consists of two paral-

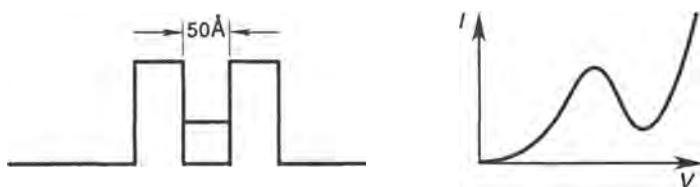


Figure 16.1 Double barrier structure, the electronic analogy to a Fabry-Perot interferometer, and corresponding I-V characteristic of a diode combining such barriers.

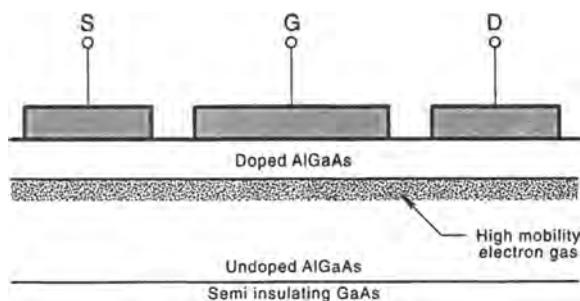


Figure 16.2 The high electron mobility transistor. The AlGaAs is heavily doped, whereas the GaAs is not intentionally doped.

lel heterojunction barriers separate by a distance of 50 Å or even smaller. The current voltage characteristic of such a structure exhibits pronounced negative differential resistance because there exists a tunneling resonance when the energy of the incident electron coincides with the size quantized state between the two barriers. The theory of this device is by now well understood [1] and simulation tools (e.g., the quantum simulator NEMO [4]) exist that permit quantitative simulation of realistic device structures. In principle, dopants are not needed to construct such a device and it therefore could be a candidate for a future device of ultrasmall dimension. However, it is a two terminal device only and the third terminal, which controls the current flow in a transistor, is missing. There also are problems with achieving a very high ratio of on and off currents. In addition, the layered structure can only be made thin enough by crystal-growth techniques such as molecular beam epitaxy. Whether or not such a device function can form the basis of future integrated circuits with ultrahigh density is therefore in question. The device has shown already superior high frequency performance in the terahertz range [6].

Another novel device that extends conventional schemes is the high electron mobility transistor (HEMT) also called modulation doped field effect transistor (MODFET), which uses the principle of modulation doping. Its schematics are shown in Figure 16.2.

This device is a composite of MOSFET and metal semiconductor field effect transistor (MESFET). The MESFET uses Schottky metal barriers as gate elec-

trode and the gate electrode of the MODFET is also a Schottky barrier metal on top of the doped AlGaAs layer. This layer is ideally free of conduction electrons because these transfer to the lower gap material next to the AlGaAs. The material with lower gap can be GaAs but is now preferentially an InGaAs ternary compound that has an even smaller gap and also smaller effective mass. We have already described the lower impurity scattering rate and higher mobility in such structures. The heating of the electron gas, however, degrades the mobility considerably and k -space and real space transfer are at work simultaneously in this device type. Simulation is therefore best done by powerful simulation packages using Monte Carlo (e.g., DAMOCLES) or a Stratton-type of approach (e.g., MINIMOS NT). The doping in the AlGaAs layer can be extremely high, without a lowering in the mobility, and the use of heterolayers permits the fabrication of very small devices. Again, however, there are serious disadvantages for use in very large scale integrated circuits mostly connected to the use of Schottky barriers, which are difficult to control and often show barrier height fluctuations. These barriers are also typically lower than the barriers created by built-in voltages, and the on-off current ratios are therefore not as large as desired depending, of course, on the material system that is used. Again this transistor type has found great use in high frequency application, and has graduated to a valuable commercial device for these special applications [6]; it does not provide, however, a basis for ultrahigh integration.

16.1.2 Future Devices for Ultrahigh Integration

One could finish this section easily by just stating that no such device currently exists. However, ideas for such devices do exist, although they are vague and not well developed. Devices of very small feature size in all dimensions, nano-structure devices (NDs) and nanodevice integrated circuits (NICs) are subject of much research, and at least two possible ways to achieve ultrahigh integration densities have been delineated and investigated on the device and integrated circuit level. The idea of transistors based on the transfer of single electrons [2] is based on the Coulomb blockade effect that is illustrated in Figure 16.3.

The figure shows a capacitor of very small size and with extremely thin insulator so that electrons can tunnel through the insulator. One might ask, what is new with such a structure? Indeed, it is only the small size. This size prevents a current flow below a certain critical voltage V_c . The reason is simply the following. The voltage on the metal contacts of the capacitor is generated by the displacement of electrons and positive nuclei in the metallic conductor and can have any value. However, according to the laws of electrostatics, it takes a finite energy to bring an electron from one plate of the capacitor to another. Given a capacitance C_{tu} for the tunneling capacitor this energy eV_c is given by

$$eV_c = \frac{e^2}{2C_{tu}} \quad (16.1)$$

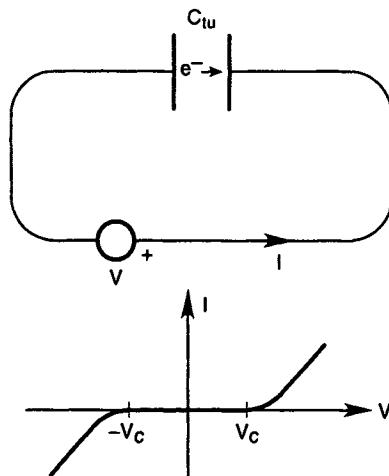


Figure 16.3 A finite energy is needed to complete the tunneling process which results in the existence of a minimum voltage V_c necessary for current flow.

which explains that no tunneling current can flow below this critical voltage in either direction. This is a manifestation of the quantization of charge. Equation (16.1) is valid only under certain circumstances. For example, if

$$eV_c \ll kT_L \quad (16.2)$$

the effect is then smeared out and unobservable. This means that only extremely small capacitors will show Coulomb blockade at room temperature; the capacitance must be of the order of Atto Farad. If semiconductor materials are used as capacitance plates, size quantization within the semiconductor will play a role in determining the capacitance. The capacitance then is determined by quantum contributions as already described in Chapter 10. If the quantization is in all three dimensions, the quantum contributions to the capacitance begin to exhibit effects known from atoms. It was shown by Macucci et al. [5] that shell-filling effects may become important. As in the case of atomic physics, a certain number of electrons can be placed on the capacitor (which now can be seen as a large atom) with relatively little energy needed. This number depends on the symmetry of the capacitor. It is higher for symmetric structures, and equal to two (two electrons with opposite spin) for completely asymmetric structures. If more electrons are added a significant energy is needed for the next electron; the addition of further electrons again requires little energy until the next shell is filled. This effect of atomistic capacitance features was experimentally found by Tarucha [7]. The Coulomb blockade effect, modified by the shell effects, permits the controlled transfer of single electrons onto small capacitors. Ultimately the plate of a capacitor can be so small that it becomes a quantum dot (QD).

Quantum dots are investigated in current research with many purposes in mind. They can replace the quantum well in laser diodes or they can serve as an

entity that stores charge and can be thought of as a memory element with single electrons transferred on and off the QD. Thus the QD becomes in a way the prototypical device of the future, although no function has yet been found that could replace the MOSFET in logic circuits. One idea to create logic with QDs is based on the charge distribution in the dots that can be influenced by neighboring dots. For example, in a dot of square shape with two electrons, these two electrons will be found with high probability in opposite corners. A neighboring dot with fixed charges can influence then in which two opposite corners the electron will be. It has been shown that such a system can in principle create logic functions [3]. The mechanism of operation is so different from MOSFET operation that the potential for large-scale integration is currently difficult to assess. One fact, however, is certain. We need great advances in the material science of QD creation and great advances in the engineering of dot patterns and the corresponding science of lithography on ultrasmall scales. It appears that the engineering of small structures and the chemical science of large molecules will need to merge to create a NIC of quantum dots with logic or at least memory function. Very interesting research, involving atomic probe microscopy, is currently being pursued in several laboratories.

16.2 CHALLENGES IN NANOSTRUCTURE SIMULATION

As has been described in several previous chapters, the ever-increasing power of computational resources and speed has brought forth the ability to find numerical solutions of important device equations including the equations of Shockley, Maxwell's equations, Boltzmann's equation, and the Schrödinger equation for given potentials, all with realistic boundary conditions and even in three dimensions. One can now treat the device as a whole and the interactions between the various device functions, parts, and sections can be optimized numerically resulting in important improvements that are really needed from an engineering point of view. Some problems, however, can still be solved only with aid of certain approximations or in reduced dimensions. For example, a laser diode demands the solution of the Shockley equations, the Maxwell equations, and the Schröedinger equation for the quantum well; all fully and self-consistently coupled. Even if such a solution were possible with current computers (in three dimensions it is not), it still would not solve all nanostructure problems. There remains still the whole domain of many-body problems including the atomistic basis of nanostructures, the interactions of electrons with themselves, with holes (e.g., exciton in arbitrarily shaped quantum dot), with phonons, and with the atomistic structure (e.g., desorption chemistry, such as hydrogen and deuterium desorption, defect creation, etc.). In the following several problem areas are discussed that currently defy solution and require (achievable) progress.

16.2.1 Nanostructures in Existing Semiconductor Devices

Nanostructures are embedded in many important semiconductor devices including the all-pervasive CMOS technology that features a quantum well at the silicon–silicon dioxide interface. As we have seen, this well adds a quantum capacitance to the classical oxide capacitance, an effect of increasing importance as the oxides becomes very thin. These oxides, of the order of 2 nanometer thickness, require special attention with respect to structural problems (defects) as discussed before. The nanometer dimensions of portions of conventional devices give rise to a number of electronic problems as well. The all pervasive existence and necessity of doping calls for increasingly high doping densities as the sizes are shrunk. Clearly, one cannot define a 10 nanometer structure by a doping of 10^{18} cm^{-3} . Because doping cannot be arbitrarily increased, its granularity is bound to impact device performance. This granularity needs therefore to be included into simulation and makes particle simulations necessary to understand the statistical fluctuations.

Electronic transport through the oxide by various forms of tunneling becomes a concern, although the barrier between silicon and silicon dioxide is very high. The barriers between the various semiconductor layers in III-V compound devices are much smaller and massive transport over interfaces calls for coupled solutions of the Boltzmann and Schrödinger equations. High-speed applications and optoelectronic applications of III-V compounds require the coupled solution of several equation systems. Examples include the following:

1. *Vertical cavity surface emitting lasers or more generally microcavity laser diodes.* A solution for these devices requires a combination of Maxwell-Shockley-Schrödinger and for (nonlinear gain problems) Boltzmann and the equation for lattice heating. Recent developments using parallel computation and a special Maxwell solver show that this problem can at least be solved in two dimensions.
2. *High-speed transistors such as MODFETs.* These devices are well understood in steady state (DAMOCLES, MINIMOS NT, and other software packages can be used for their simulation). However, the addition of Maxwell's equations to include transmission line and boundary problems is just currently being started. This is, from an industrial viewpoint, a relatively small but important area and needs attention.
3. *Similar considerations apply to the area of resonant tunneling devices.* Solutions of the Schrödinger equation, even using the Keldysh formalism to include extreme nonequilibrium and scattering are known; the simulator NEMO can do all of this. However, a self-consistent solution with Maxwell's equations is not included in the current approach.

16.2.2 Quantum Dots

As discussed before, the QD has emerged as the central “unit” of many ideas related to future devices ranging from already functioning QD laser diodes to the still unclear possibility of ultralarge QD memories embedded in (or surrounded by) conventional devices. Although great progress has been made and transmission of single electrons can be controlled, there are still basic problems in the simulation of quantum dots and electronic transport through them.

The connection of QDs to the environment is understood only for limiting cases. There is no approach known that describes in detail the transition from the QD to the classical reservoirs and the connected dephasing process. This problem of transition between classical and quantum transport is general and appears prominently in quantum well diode lasers and heterolayer transistors. Theories or simulation methods are available for the electron–phonon interaction in QDs; electron–electron interactions have only sporadically been treated in confined geometries.

The electronic states of QD with realistic geometries and material effects (such as compositional disorder) are beginning to be understood by the use of density functional theory. However, important many-body effects such as excitons in QDs are still lacking a numerical approach in spite of the great importance of these effects for optical transitions in devices that include QDs. Along the same lines, a detailed approach to the dielectric properties of nanostructures is lacking. The treatment of QDs at finite temperature without assuming a given energy distribution (such as a Fermi distribution) is beyond current computational possibility.

16.2.3 Structural, Atomistic, and Many-Body Effects

The area of structural and atomistic effects relates to both current and future devices and, in fact, offers great connectivity for the research on both. For example, the detailed investigation of hydrogen desorption from silicon surfaces by scanning tunneling microscopy has led to a model of transistor aging and the discovery of a giant improvement of the aging when hydrogen was replaced by deuterium. A similarly strong connectivity applies in the area of simulation. Calculations of tunneling through very thin structures assisted by impurity complexes and other defects are basic science when dealing with QDs and single electronics. At the same time tunneling through thin oxides must be understood for CMOS operation and has shown quantized behavior even for the breakdown that limits CMOS design.

Numerical approaches, based on density functional theory, are promising a very detailed understanding of structural and atomistic problems. The theoretical approaches of physicists and chemists begin to merge in this area. Nanostructures can be understood (from a theoretical view) by the nanostructure physicist just as very large molecules are by the chemist. It will be necessary to learn from

what has been developed in chemistry and to attempt to improve the methods for very large systems. For example, a QD sandwiched between two materials is beyond current simulation abilities if the atomistic basis plays a role and needs to be resolved. In this case one needs to simulate the many-particle problem of the many atoms in the QD and in addition a large number of atoms of the materials that sandwich the QD. This presents a huge numerical problem, because the numerical complexity ordinarily increases with the order of N^α , where N is the number of atoms and the exponent α is of the order of three but can be even larger. It is therefore necessary to develop so-called order N methods with $\alpha = 1$ to be able to get a detailed understanding of large systems. Possibilities to achieve order N exist by use of physical assumptions or new mathematical tools such as wavelet representations. In general, multiscale approaches are needed that combine the atomistic with the macroscopic dimensions and that permit the detailed simulation of large systems of atoms. There are various approaches that do justice to multiscale problems ranging from combinations of layers of mesh-points, with very different spacing in different regions, to novel approaches such as the use of wavelets.

Independent of all these methods, there are numerous applications beyond the QD that are highly interesting from the viewpoint of nanostructure chemistry and physics. These include the newly developing science of carbon nanotubes and Buckey balls. There are also many areas that have not yet received much attention but do open literally another dimension from a computational viewpoint. For example, the inclusion of temporal dependencies and the dynamics of nanostructures as, for example, the moving of atoms by scanning tunneling microscopes or atomic force microscopes.

In conclusion, I would like to state that nanostructure simulation poses many outstanding problems. In the arena of single particle physics, chemistry, and engineering, it is the self-consistent combined solution of several systems of differential equations. We know how to solve Maxwell's equations, but it is not known how to combine these solutions with the Schrödinger equation and link them to surrounding classical environments described by the Boltzmann's equation all having complicated boundary conditions as they are found in microcavity laser diodes and high-speed transistors. Structure, atomistic, and many-body problems provide an even larger playground for the development of more powerful future simulation tools that describe the world of semiconductors and semiconductor devices.

REFERENCES

- [1] Datta, S. *Electronic Transport in Mesoscopic Systems*, New York: Cambridge Univ. Press, 1995.
- [2] Grabert, H., and Devoret, M. H., ed. *Single Charge Tunneling*, New York: Plenum, 1992.
- [3] Lent, C. S., Tougaw, D. P., Porod, W., and Bernstein, G. H. "Quantum cellular automata," *Nanotechnology*, vol. 4, 1993, pp. 49–57.

- [4] Klimeck, G., et al. *VLSI Design*, vol. 8, 1998, p. 79.
- [5] Macucci, M., Hess, K., and Iafrate, G. J. "Simulation of electronic properties and capacitance of quantum dots," *Journal of Applied Physics*, vol. 77, 1995, p. 3267.
- [6] Sze, S. M., ed. *High-Speed Semiconductor Devices*, New York: Wiley/Interscience, 1990.
- [7] Tarucha, S., et al. "Shell filling effects in quantum dots," *Physical Review Letters*, vol. 77, 1996, p. 3613.

TUNNELING AND THE GOLDEN RULE

Consider the following model Hamiltonian

$$H_M = \begin{cases} H_L & \text{for } r \text{ in } R_L \\ H_R & \text{for } r \text{ in } R_R \end{cases} \quad (\text{A.1})$$

where R_L denotes a left-side region and R_R a right-side region. In our example of the tunneling problem in Figure 1.2, R_R would correspond to the region out of the well, that is,

$$H_R = -\frac{\hbar^2 \nabla^2}{2m} - eFz \quad (\text{A.2})$$

while H_L would be given by the Hamiltonian in the well

$$H_L = -\frac{\hbar^2 \nabla^2}{2m} + WPE \quad (\text{A.3})$$

where WPE is the well potential energy.

The total Hamiltonian is

$$H = -\frac{\hbar^2 \nabla^2}{2m} - eFz + WPE \quad (\text{A.4})$$

which is apparently not entirely equal to the sum $H_L + H_R$. Assume that we know solutions of the partial Hamiltonians H_L and H_R . We denote these solutions by

$$H_L \Psi_0 = E_0 \Psi_0 \quad (\text{A.5})$$

and

$$H_R \Psi_v = E_v \Psi_v \quad (\text{A.6})$$

where v is an integer. In other words, we assume we know the solution for the electron in a given state at the left side and in a set of states on the right side.

We now attempt to obtain an approximate solution of H by expanding the wave function that satisfies H in terms of the known wave functions

$$\psi(r, t) = \psi_0(r)e^{-iE_0 t/\hbar} + \sum_{v=1}^{\infty} a_v(t)\psi_v(r)e^{-iE_v t/\hbar} \quad (\text{A.7})$$

The expansion coefficients $a_v(t)$ have to vanish at $t = 0$; we assume that at $t = 0$ the electron wave function is ψ_0 .

The Hamiltonian needs to be brought into a form that permits perturbational solutions. In the example of the well-emitting electrons, the perturbation is given by the electric field and therefore we can write the perturbing part H' of the Hamiltonian as

$$H' = H - H_L = -eFz \quad (\text{A.8})$$

As we know from Eq. (1.41), or from insertion of Eq. (A.7) into the Schrödinger equation, all that is needed to solve quantum problems to first order is the matrix element, which we can write in shorthand notation as

$$\langle v | H - H_L | 0 \rangle \equiv \int_{V_{\text{el}}} \psi_v^*(H - H_L) \psi_0 d\mathbf{r} \quad (\text{A.9})$$

The matrix element can, of course, be easily calculated if ψ_v and ψ_0 are known, which they are for the case of the perturbed quantum well. There is, however, a possibility of rewriting Eq. (A.9) in a more general and useful form. We will outline this “rewriting” below. It is not easy to understand why things should be done just that way, but the derivation is clear and the final equations are very useful. The fact that the method is due to Bardeen says, I believe, everything.

We replace H in Eq. (A.9) by our model Hamiltonian of Eq. (A.1) and write

$$\langle v | H - H_L | 0 \rangle \approx \int_{R_R} \psi_v^*(H_M - H_L) \psi_0 d\mathbf{r} \quad (\text{A.10})$$

The integration now goes only over the right half space because $H_M - H_L$ vanishes to the left. $H_M - H_R$ is identical to zero in R_R and we therefore can write

$$\langle v | H - H_L | 0 \rangle \approx \int_{R_R} [\psi_v^*(H_M - H_L) \psi_0 - \psi_0(H_M - H_R) \psi_v^*] d\mathbf{r} \quad (\text{A.11})$$

Using Eqs. (A.5) and (A.6), we have

$$\langle v | H - H_L | 0 \rangle \approx \int_{R_R} [\psi_v^* H_M \psi_0 - \psi_0 H_M \psi_v^* - \psi_v^* E_0 \psi_0 + \psi_0 E_v \psi_v^*] d\mathbf{r} \quad (\text{A.12})$$

We are later only interested in elastic tunneling processes and therefore $\psi_0 E_v \psi_v^* = \psi_v^* E_0 \psi_0$ as demanded by the energy-conserving δ function in Eq. (1.42). Under most circumstances of interest to us, the model Hamiltonian H_M can be formally separated into two factors

$$H_M = -\frac{\hbar^2 \nabla^2}{2m} + FPE \quad (\text{A.13})$$

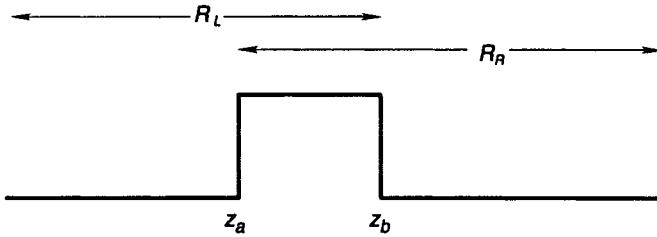


Figure A.1 Tunneling barrier with left and right space division indicated.

where FPE is a function of \mathbf{r} representing the potential energy. Therefore, $\psi_v^* FPE \psi_0 - \psi_0 FPE \psi_v^*$ also vanishes and we have

$$\langle v | H - H_L | 0 \rangle \approx \int_{R_R} \left[\Psi_v^* \left(\frac{-\hbar^2 \nabla^2}{2m} \right) \Psi_0 - \Psi_0 \left(\frac{-\hbar^2 \nabla^2}{2m} \right) \Psi_v^* \right] d\mathbf{r} \quad (\text{A.14})$$

The tunneling problem is therefore solved as soon as we know Ψ_0 and Ψ_v . The integral in Eq. (A.14) can be evaluated using Green's theorem. Because we later discuss only one-dimensional applications, we reproduce the equations in one dimension only:

$$\begin{aligned} \langle v | H - H_L | 0 \rangle &\approx \frac{-\hbar^2}{2m} \int_0^\infty \left[\Psi_v^* \frac{\partial^2}{\partial z^2} \Psi_0 - \Psi_0 \frac{\partial^2}{\partial z^2} \Psi_v^* \right] dz \\ &= \frac{-\hbar^2}{2m} \int_0^\infty \frac{\partial}{\partial z} \left[\Psi_v^* \frac{\partial}{\partial z} \Psi_0 - \Psi_0 \frac{\partial}{\partial z} \Psi_v^* \right] dz \end{aligned} \quad (\text{A.15})$$

where we have assumed the right side to be in the interval $0 < z < \infty$. This gives

$$\langle v | H - H_L | 0 \rangle \approx \frac{-\hbar^2}{2m} \left[\Psi_v^*(0) \frac{\partial \Psi_0}{\partial z} \Big|_0 - \Psi_0(0) \frac{\partial \Psi_v^*}{\partial z} \Big|_0 \right] \quad (\text{A.16})$$

$\frac{\partial \Psi_0}{\partial z} \Big|_0$ and $\frac{\partial \Psi_v^*}{\partial z} \Big|_0$ denote the derivatives at $z = 0$, that is, at the point (or surface) separating the regions R_L and R_R . Bardeen's model of tunneling¹ differs from the one described above by a slight generalization. He lets R_R and R_L overlap in the tunneling region (the rest of the derivation is similar). The functions in Eq. (A.16) can then be evaluated at any point of the overlapping region with the same result.

Below we illustrate the use of Eq. (A.16) for the case of a simple rectangular tunneling barrier, as shown in Figure A.1. We need to find Ψ_0 and Ψ_v to obtain the matrix element. Because the problem corresponding to Figure A.1 is symmetric, the left and right wave functions will be identical except for the sign of the space coordinate z and it suffices therefore to calculate one. Writing formally

¹J. Bardeen, *Physical Review Letters*, vol. 6, 1961, p. 57.

for $-E_0 + FPE$ the term $\hbar^2 \bar{k}_z^2 / 2m$ (this is the definition of \bar{k}_z), the Schrödinger equation reads

$$\frac{\partial \psi_0}{\partial z^2} + \hbar^2 \bar{k}_z^2 \psi_0 = 0 \quad (\text{A.17})$$

One has to find solutions for $z < z_a$ and for $z_a \leq z \leq z_b$ and then connect the solutions. This connection is a difficult problem and is discussed in most quantum mechanics courses. Thus we give only a simplified result here:

$$\psi_0 = \sqrt{\frac{2}{L_a}} \cos \left(\int_z^{z_a} \bar{k}_z dz - \frac{\pi}{2} \right) \quad \text{for } z < z_a \quad (\text{A.18})$$

and

$$\psi_0 = \frac{1}{\sqrt{2L_a}} \exp \left(- \int_{z_a}^z |\bar{k}_z| dz \right) \quad \text{for } z_0 \leq z \leq z_b \quad (\text{A.19})$$

where L_a is the length of the crystal left to z_a . Below we will use L_a also for the crystal length to the right of z_b . Also we have $\bar{k}_{z_a} = \bar{k}_{z_b}$ because of the symmetry of the problem. Using Eqs. (A.18) and (A.19), one obtains the square of the matrix element from Eq. (A.16), which is

$$|\langle 0 | H - H_L | 0 \rangle|^2 \approx \frac{\hbar^2}{2m} \frac{|\bar{k}_{z_a}|^2}{L_a^2} \exp \left(-2 \int_{z_a}^{z_b} |\bar{k}_z| dz \right) \quad (\text{A.20})$$

A generalization of the approach for z -dependent crystal structure (alloys!) has been given by Harrison.²

The calculation for a three-dimensional barrier proceeds along the same lines. For specular flat parallel barriers, one obtains the same results as in one dimension. In addition, one notices that the wavevector component parallel to the barrier is conserved in the tunneling process [compare with Eq. (1.28)]. We note also that tunneling treated in this form can easily be included in Monte Carlo simulations as an additional scattering mechanism.

²W. Harrison, *Physical Review*, vol. 123, 1961, p. 85.

APPENDIX B

THE ONE BAND APPROXIMATION

The proof of Eq. (3.21) rests on the following identity

$$E(-i\nabla)\psi(\mathbf{k}, \mathbf{r}) = E(\mathbf{k})\psi(\mathbf{k}, \mathbf{r}) \quad (\text{B.1})$$

which can be shown by Fourier-expanding E in terms of lattice vectors, which, according to Eq. (2.11), is always possible

$$E(\mathbf{k}) = \sum_l E_l e^{i\mathbf{R}_l \cdot \mathbf{k}} \quad (\text{B.2})$$

Therefore,

$$\begin{aligned} E(-i\nabla)\psi(\mathbf{k}, \mathbf{r}) &= \sum_l E_l e^{\mathbf{R}_l \cdot \nabla} \psi(\mathbf{k}, \mathbf{r}) \\ &= \sum_l E_l (1 + \mathbf{R}_l \cdot \nabla + 1/2(\mathbf{R}_l \cdot \nabla)^2 + \dots) \psi(\mathbf{k}, \mathbf{r}) \end{aligned} \quad (\text{B.3})$$

We notice now that the term $(1 + \mathbf{R}_l \cdot \nabla + 1/2(\mathbf{R}_l \cdot \nabla)^2 + \dots) \psi(\mathbf{k}, \mathbf{r})$ is just the Taylor expansion of the function $\psi(\mathbf{k}, \mathbf{r} + \mathbf{R}_l)$, and therefore

$$E(-i\nabla)\psi(\mathbf{k}, \mathbf{r}) = \sum_l E_l \psi(\mathbf{k}, \mathbf{r} + \mathbf{R}_l) \quad (\text{B.4})$$

which, with the help of Bloch's theorem, becomes

$$E(-i\nabla)\psi(\mathbf{k}, \mathbf{r}) = \sum_l E_l e^{i\mathbf{R}_l \cdot \mathbf{k}} \psi(\mathbf{k}, \mathbf{r}) = E(\mathbf{k})\psi(\mathbf{k}, \mathbf{r}) \quad (\text{B.5})$$

and the proof is complete.

As evident, then, from Eq. (B.1), the operator $E(-i\nabla)$ replaces the operator $-(\hbar^2/2m)\nabla^2 + V(\mathbf{r})$ in the Schrödinger equation, which can now be written without the appearance of the crystal potential $V(\mathbf{r})$.

The addition of an external potential presents no problem, as can be seen from the following arguments. Let the solution of the Schrödinger equation,

which includes V_{ext} , be a combination of all Bloch functions of all bands with index n

$$\psi(\mathbf{k}, \mathbf{r}) = \sum_n a_n \psi_n(\mathbf{k}, \mathbf{r}) \quad (\text{B.6})$$

We then have

$$\begin{aligned} \left(-\frac{\hbar^2}{2m} \nabla^2 + V(\mathbf{r}) - eV_{\text{ext}} \right) \sum_n a_n \psi_n(\mathbf{k}, \mathbf{r}) \\ = \sum_n a_n \left(-\frac{\hbar^2}{2m} \nabla^2 + V(\mathbf{r}) - eV_{\text{ext}} \right) \psi_n(\mathbf{k}, \mathbf{r}) \end{aligned}$$

and using the above identity

$$= \sum_n a_n (E_n(-i\nabla) + V_{\text{ext}}) \psi_n(\mathbf{k}, \mathbf{r}) \quad (\text{B.8})$$

If, as assumed, only the energy of one band is important and interband transition can be neglected, then $E_n(-i\nabla)$ can be replaced by a single energy $E(-i\nabla)$ and Eq. (B.8) becomes

$$(E(-i\nabla) - eV_{\text{ext}}) \sum_n a_n \psi_n(\mathbf{k}, \mathbf{r}) = (E(-i\nabla) - eV_{\text{ext}}) \psi_n(\mathbf{k}, \mathbf{r}) \quad (\text{B.9})$$

which completes the proof. The condition of one band being important only means in the semiconductors GaAs, Si, and Ge that the external electric fields are smaller than about 10^6 V/cm. The rigorous proof of this fact is involved.

APPENDIX C

TEMPERATURE DEPENDENCE OF THE BAND STRUCTURE

It is convenient to separate the temperature dependence of the energy band into contributions arising from the volume change and contributions related to the lattice vibrations.

The volume change is easily accounted for; one only has to insert the temperature-dependent lattice constant into the pseudopotential calculation, Eq. (3.15). The change in lattice constant can be calculated from the volume expansion coefficient $\partial V_{\text{ol}}/\partial T$. This coefficient and the pressure coefficient $\partial P/\partial V_{\text{ol}}$ (P is the pressure) can be found from data in the *Handbook of Chemistry and Physics*. Because the change of the energy gap with pressure is known from experiments, the change of the energy gap with temperature owing to the volume increase can be calculated from

$$\Delta E_G^{V_{\text{ol}}} = \left(\frac{\partial E_G}{\partial P} \right) \left(\frac{\partial P}{\partial V_{\text{ol}}} \right) \left(\frac{\partial V_{\text{ol}}}{\partial T} \right) \Delta T \quad (\text{C.1})$$

There is also a change in the value of the energy gap owing to the lattice vibrations that is more difficult to calculate. H. Y. Fan used second-order perturbation theory, including the potential of the phonons [as given, for example, by Eq. (7.25)]. The energy change owing to the phonons ΔE_G^{phI} is then given by

$$\Delta E_G^{\text{phI}} = \sum_{n,q} \frac{|M_q|^2}{E_m(\mathbf{k}) - E_n(\mathbf{k} + \mathbf{q}) \pm \hbar\omega_q} \quad (\text{C.2})$$

Equation (C.2) follows immediately from Eq. (1.38) with the phonon energy included in the denominator. The plus sign is appropriate for phonon absorption and the minus sign for phonon emission. In practice, only a single band needs to be considered—that is, $n = m$ (= conduction (valence) band)—and the summation over q is converted into an integration according to Eq. (5.8). Values for M_q for the various phonon coupling mechanisms are given in Chapter 7.

Another contribution to the change of the energy gap as a consequence of the lattice vibrations arises from the influence of vibrations on the structure factor

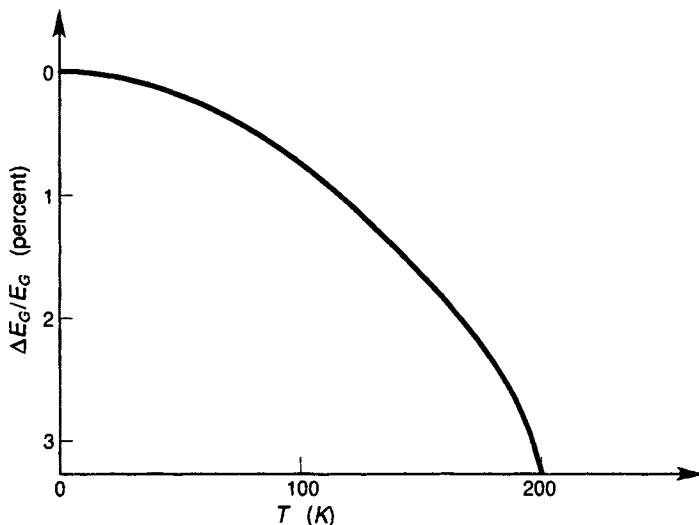


Figure C.1 Typical form of the relative change of energy gap with temperature for III-V compounds and group IV (Si, Ge) semiconductors. The energy gap decreases with temperature.

e^{-iKt} in Eq. (3.15). Because of the vibrations this structure factor is “washed out.” Antonchik (and later Brooks and Yu) proposed, therefore, to multiply the factor e^{-iKt} by $\exp(-1/6|K|^2\langle u^2 \rangle)$, where $\langle u^2 \rangle$ is the mean square displacement of the atom in question (see Eq. (1.18) for the definition of u). To obtain the change in energy gap ΔE_G^{phII} owing to this effect, one has to calculate $E(\mathbf{k})$ with the temperature-dependent structure factor according to the method described in Chapter 3.

Experimental results of changes of the energy states with temperature can be found in the *Handbook on Semiconductors* (1980).

In comparing the experimental results with the theory outlined above, one finds that ΔE_G^{phII} gives the most important contribution for III-V compounds and group IV semiconductors (which are the most important materials from a device point of view). ΔE_G^{phI} and ΔE_G^{Vol} contribute only about 20% to the total change in the energy gap. A typical result for the total change

$$\Delta E_G = \Delta E_G^{\text{Vol}} + \Delta E_G^{\text{phI}} + \Delta E_G^{\text{phII}} \quad (\text{C.3})$$

is plotted in Figure C.1.

HALL EFFECT AND MAGNETORESISTANCE FOR SMALL MAGNETIC FIELDS

The precise theory of the classical Hall effect (the quantum Hall effect will not be treated here) involves solution of the Boltzmann equation using the method of moments as described in Eqs. (11.1) through (11.8). The force term is now given by

$$-\frac{e}{\hbar}(\mathbf{F} + \mathbf{v} \times \mathbf{B})\nabla_{\mathbf{k}}f \quad (\text{D.1})$$

where $e(\mathbf{F} + \mathbf{v} \times \mathbf{B})$ replaces the force \mathbf{F}_0 in Eq. (8.12). This term complicates the method of moments significantly and we have to proceed slightly differently than we did in Chapter 11. To facilitate the notation we perform the calculation in two dimensions. This treatment applies, for example, to the two-dimensional electron gas in metal–oxide–silicon (MOS) transistors.

The treatment in three dimensions can be done in totally analogous fashion using polar coordinates instead of the cylindrical coordinates used below. For the $E(\mathbf{k})$ relation we choose a parabolic band. However, for the sake of generality, we let the surface of constant energy be an ellipse with masses m_x and m_y along the main axes [see Eq. (3.34)]. Our coordinate system for the wavevector \mathbf{k} is chosen along these main axes. In the method of moments one multiplies the Boltzmann equation by \mathbf{k} (to obtain the current) and integrates over the cylindrical coordinates $d\phi k dk$. We integrate first over angle $d\phi$ and define

$$dj_i \equiv \frac{e}{m_i^*} \int_0^{2\pi} k_i f d\phi \quad i = x, y \quad (\text{D.2})$$

and

$$dn \equiv \int_0^{2\pi} f d\phi \quad (\text{D.3})$$

The differential notation dj and dn means that integration over $k dk$ (the energy) is still necessary to obtain macroscopic current and carrier concentration. Assuming now the magnetic field in z -direction, one obtains for the differential current component in x -direction

$$\frac{dj_x}{\tau_{\text{tot}}} = E \frac{e}{m_x^* k T_c} F_x dn + \frac{e}{m_x} B_z dj_y \quad (\text{D.4})$$

and a similar equation for dj_y . τ_{tot} is the total scattering time, as given for example, in Eq. (7.20).

In vector and matrix notation we obtain, then, the following differential current

$$d\mathbf{j} = \begin{bmatrix} \bar{\mu}_x - B_z^2 \bar{\mu}_x^2 \bar{\mu}_y & \bar{\mu}_y \bar{\mu}_x B_z \\ -B_z \bar{\mu}_y \bar{\mu}_x & \bar{\mu}_y - B_z^2 \bar{\mu}_x \bar{\mu}_y^2 \end{bmatrix} \frac{F E e}{k T_c} dn \quad (\text{D.5})$$

where $\bar{\mu}_i = e\tau_{\text{tot}}/m_i^*$ and T_c is the electron temperature (which in principle also depends on the electric and magnetic fields).

The terms proportional to B_z^2 are responsible for the magnetoresistance, that is, the increase in sample resistance owing to the magnetic field. In the explanation of the classical Hall effect, as these terms are usually ignored, because B_z is small and B_z^2 therefore is negligible. If the sample is elongated in x -direction and no net current flows in y - (and z -) directions, Eq. (D.5) gives the two equations

$$j_x = \frac{2}{4\pi^2} \int_0^\infty dk k \bar{\mu}_x \frac{F_x E e}{k T_c} dn \quad (\text{D.6})$$

and

$$\int_0^\infty dk k B_z \bar{\mu}_y \bar{\mu}_x \frac{F_x E e}{k T_c} dn = \int_0^\infty dk k \bar{\mu}_y \frac{F_y E e}{k T_c} dn \quad (\text{D.7})$$

which gives the Hall field F_y

$$F_y = F_x B_z \frac{\int_0^\infty dk k \bar{\mu}_y \bar{\mu}_x E dn}{\int_0^\infty dk k \bar{\mu}_y E dn} \quad (\text{D.8})$$

With the definitions of averages in two dimensions, this is equivalent to

$$F_y = F_x B_z \mu_x \frac{\langle \tau_{\text{tot}}^2 \rangle}{\langle \tau_{\text{tot}} \rangle^2} \quad (\text{D.9})$$

We therefore obtain Eq. (7.12) modified by a “statistics factor” $\frac{\langle \tau_{\text{tot}}^2 \rangle}{\langle \tau_{\text{tot}} \rangle^2}$, which is equal to one if τ_{tot} is independent of energy (or k). This result is general and holds also in three dimensions with the averages being three-dimensional averages, as in Eq. (8.48). Notice that for large magnetic fields ($(\mu B)^2 \geq 1$), a quantum treatment is necessary.

APPENDIX E

THE POWER BALANCE EQUATION FROM THE METHOD OF MOMENTS

To derive the power balance equation, we need to take the third moment of the Boltzmann transport equation. Beginning with Eq. (11.8), we let $Q(\mathbf{k}) = E(\mathbf{k}) = \hbar^2 k^2 / 2m^*$. We will disregard for the moment the time derivative. Equation (11.8) then becomes

$$\frac{\hbar}{m^*} \nabla \{ n \langle E \mathbf{k} \rangle \} + \frac{e}{\hbar} \mathbf{F} \cdot n \langle \nabla_{\mathbf{k}} E(\mathbf{k}) \rangle = \frac{1}{4\pi^3} \int d\mathbf{k} E(\mathbf{k}) \frac{\partial f_0}{\partial t} \Big|_{\text{coll}} - \frac{1}{4\pi^3} \int d\mathbf{k} \frac{f_1}{\tau_{\text{tot}}(\mathbf{k})} E(\mathbf{k}) \quad (\text{E.1})$$

We now perform the integrals in Eq. (E.1): The first term on the left is given by

$$\langle E(\mathbf{k}) \mathbf{k} \rangle = \int \frac{\hbar^2 k^2}{2m^*} \mathbf{k} f \, d\mathbf{k} = \frac{1}{n} \int \frac{\hbar^2 k^2}{2m^*} \mathbf{k} (f_0 + f_1) \, d\mathbf{k} \quad (\text{E.2})$$

The term $k^2 \mathbf{k} f_0$ is odd and does not contribute to the integral; therefore

$$\langle E(\mathbf{k}) \mathbf{k} \rangle = \frac{1}{n} \frac{\hbar^2}{2m^*} \int k^2 \mathbf{k} f_1 \, d\mathbf{k} = \frac{1}{n} \int E(\mathbf{k}) \mathbf{k} f_1 \, d\mathbf{k} \quad (\text{E.3})$$

The second term on the right-hand side of Eq. (E.1) represents the average velocity while the first term, $(1/4\pi^3) \int d\mathbf{k} E(\mathbf{k}) (\partial f_0 / \partial t)|_{\text{coll}} = -nB(T_c)$ represents the energy loss to the lattice for a carrier temperature T_c . The last remaining integral in Eq. (E.1) vanishes if we assume that τ_{tot} is a function of energy only, because the integrand is then an odd function. Substituting these results back into (E.1), one obtains¹

$$\frac{\hbar}{m^*} \nabla \left\{ \int E(\mathbf{k}) \mathbf{k} f_1 \, d\mathbf{k} \right\} + e \mathbf{F} \cdot (n \langle \mathbf{v} \rangle) = -nB(T_c) \quad (\text{E.4})$$

which is equivalent to

$$\mathbf{j} \cdot \mathbf{F} = nB(T_c) + \nabla \cdot \mathbf{S}(T_c) \quad (\text{E.5})$$

¹R. Stratton, *Physics Review*, vol. 126, 1962, p. 2002

where we have used the definition of the current density and the energy flux: $\mathbf{S} = (\hbar/m^*) \int d\mathbf{k} E(\mathbf{k}) \mathbf{k} f_1$. To evaluate this flux term we have to assume $f_1 \ll f_0$. In order to simplify calculations, we proceed with a one-dimensional analysis. We can write the linearized Boltzmann equation as

$$f_1 = \tau(E) \left\{ \frac{eF\hbar k}{m^*} \frac{\partial f_0}{\partial E} - \frac{\hbar k}{m^*} \frac{\partial f_0}{\partial x} \right\} \quad (\text{E.6})$$

Substituting f_1 into the equation for $\mathbf{S}(T_c)$, one has

$$\mathbf{S}(T_c) = \frac{\hbar}{3m^*} \int d\mathbf{k} E(\mathbf{k}) k \tau(E) \left\{ \frac{eF\hbar k}{m^*} \frac{\partial f_0}{\partial E} - \frac{\hbar k}{m^*} \frac{\partial f_0}{\partial x} \right\} \quad (\text{E.7})$$

The integrations over \mathbf{k} are still performed in three dimensions; the one-dimensional density of states diverges as $E \rightarrow 0$ and would therefore introduce non-physical results. We further use Eqs. (11.9), (11.10), and (11.11), as well as Eq. (8.46). After some algebra, one obtains

$$\begin{aligned} S(T_c) &= \frac{-2j}{3m^* n \mu k T_c} \int_{[1]} \tau(E) E^2 f_0 d\mathbf{k} \\ &\quad + \frac{2e}{3m^* n \mu k T_c} \int_{[2]} \tau(E) E^2 f_0 \frac{\partial}{\partial x} (Dn) d\mathbf{k} - \frac{2}{3m^*} \int_{[3]} \tau(E) E^2 \frac{\partial f_0}{\partial x} d\mathbf{k} \end{aligned} \quad (\text{E.8})$$

The three integrals [1], [2], [3] are evaluated to be

$$[1] \quad \frac{-2j}{3m^* n \mu k T_c} \int \tau(E) E^2 f_0 d\mathbf{k} = \frac{-2j}{3m^* \mu k T_c} \langle \tau E^2 \rangle = \frac{-j \langle \tau E^2 \rangle}{e \langle \tau E \rangle} \quad (\text{E.9})$$

$$\begin{aligned} [2] \quad &\frac{2e}{3m^* n \mu k T_c} \int \tau(E) E^2 f_0 \frac{\partial}{\partial x} (Dn) d\mathbf{k} \\ &= \frac{1}{n \langle \tau E \rangle} \int \tau E^2 f_0 \frac{\partial}{\partial x} \left(\frac{2n}{3m^*} \langle \tau E \rangle \right) d^3 \mathbf{k} \\ &= \frac{\langle \tau E^2 \rangle}{\langle \tau E \rangle} \frac{2}{3m^*} \frac{\partial}{\partial x} (n \langle \tau E \rangle) \\ &= \frac{\langle \tau E^2 \rangle}{\langle \tau E \rangle} \frac{2}{3m^*} \frac{\partial}{\partial T_c} (n \langle \tau E \rangle) \frac{\partial T_c}{\partial x} \\ &= \frac{\langle \tau E^2 \rangle}{\langle \tau E \rangle} \frac{2n}{3m^*} \left[\frac{\langle \tau E^2 \rangle}{k T_c} - \frac{3}{2} \langle \tau E \rangle \right] \frac{1}{T_c} \frac{\partial T_c}{\partial x} \\ &= \frac{2n}{3m^*} \left[\frac{\langle \tau E^2 \rangle^2}{\langle \tau E \rangle k T_c} - \frac{3}{2} \langle \tau E^2 \rangle \right] \frac{1}{T_c} \frac{\partial T_c}{\partial x} \end{aligned} \quad (\text{E.10})$$

$$\begin{aligned}
 [3] \quad & \frac{2}{3m^*} \int \tau(E) E^2 \frac{\partial f_0}{\partial x} d\mathbf{k} = \frac{2}{3m^*} \int \tau E^2 \frac{\partial f_0}{\partial T} \frac{\partial T_c}{\partial x} d\mathbf{k} \\
 &= \frac{2}{3m^*} \int \tau E^2 f_0 \left[\frac{E}{kT_c^2} - \frac{3}{2T_c} \right] \frac{\partial T_c}{\partial x} d\mathbf{k} \\
 &= \frac{2n}{3m^*} \left[\frac{\langle \tau E^3 \rangle}{kT_c} - \frac{3}{2} \langle \tau E^2 \rangle \right] \frac{1}{T_c} \frac{\partial T_c}{\partial x}
 \end{aligned} \tag{E.11}$$

Putting all these terms together, we have

$$\begin{aligned}
 S(T_c) &= \frac{-j}{e} \frac{\langle \tau E^2 \rangle}{\langle \tau E \rangle} + \frac{2n}{3m^*} \left[\frac{\langle \tau E^2 \rangle^2}{\langle \tau E \rangle kT_c} - \frac{3}{2} \langle \tau E^2 \rangle \right] \frac{1}{T_c} \frac{\partial T_c}{\partial x} \\
 &\quad - \frac{2n}{3m^*} \left[\frac{\langle \tau E^3 \rangle}{kT_c} - \frac{3}{2} \langle \tau E^2 \rangle \right] \frac{1}{T_c} \frac{\partial T_c}{\partial x} \quad \text{and} \quad (\text{E.12}) \\
 S(T_c) &= \frac{-j}{e} \frac{\langle \tau E^2 \rangle}{\langle \tau E \rangle} + \frac{2n}{3m^* kT_c} \left[\frac{\langle \tau E^2 \rangle^2}{\langle \tau E \rangle} - \langle \tau E^3 \rangle \right] \frac{1}{T_c} \frac{\partial T_c}{\partial x}
 \end{aligned}$$

APPENDIX F

THE SELF-CONSISTENT POTENTIAL AT A HETEROJUNCTION (QUANTUM CASE)

This appendix describes a block diagram of the self-consistent calculations that are necessary to include size quantization effects at a heterolayer interface. References to important papers are also given, especially for the complications that arise in connection with many valley problems (X valleys in silicon).

To compute the energy eigenvalues of Eq. (10.33) one needs to go through an iterative procedure and therefore starts with an initial guess of the solution. A good initial guess can be obtained by using a triangular approximation for the potential $\phi(z)$. The solution of the Schrödinger equation for triangular potentials is well known and has been described, for example, by Ando, Fowler, and Stern.¹ From this treatment one obtains an initial guess of Fermi energy, subband energies, electron concentration $n(z)$, and wave functions (in the effective mass approximation). Using this initial guess, the Poisson equation is then solved in the three regions: GaAs, undoped AlGaAs (undoped space layer of 50 to 150 Å at the interface), and doped AlGaAs. The finite difference method can be used, which gives a tridiagonal matrix equation that is easily solved (even for very large matrices). The resulting potential (the contributions of exchange and correlation added) is then used to solve the Schrödinger equation, which is accomplished by the Numerov process, for example. A good review on this process and on computer solutions of the Schrödinger equation has been given by Chow.² This combined gives the block diagram in Figure F.1.

Although the formalism described above works for GaAs, silicon needs further considerations because of the six equivalent X valleys (instead of the one Γ valley in the case of GaAs). Stern and Howard³ have shown that the subbands have ellipses (circles) as lines of constant energy in the interface plane, which are

¹T. Ando, A. B. Fowler, and F. Stern, "Electronic Properties of Two-Dimensional Systems," *Review of Modern Physics*, vol. 54, 1982, p. 466.

²P. C. Chow, *American Journal of Physics*, vol. 40, 1972, pp. 780–784.

³F. Stern and W. E. Howard, *Physical Review*, vol. 163, 1967, pp. 816–835.

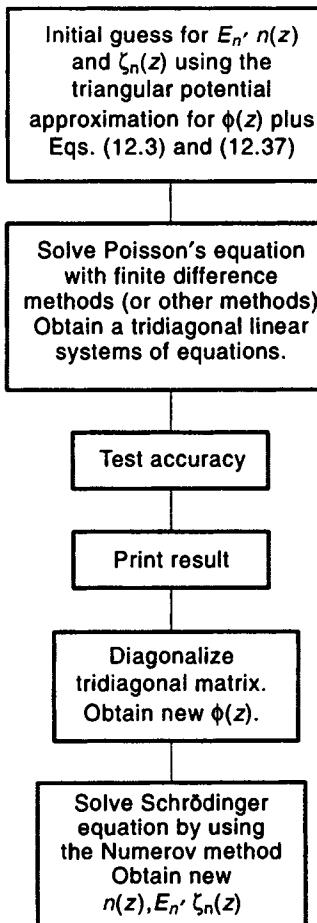


Figure F.1 Flowchart of self-consistent solution for heterolayer potentials.

obtained by the intersection of the interface plane with the ellipsoids. If the effective mass perpendicular to the interface is the same, all ellipses are equivalent and belong to the same energy eigenvalue. If the perpendicular masses differ, the ellipses with higher mass belong to lower energies than the one with smaller mass [the reason for this is evident from Eq. 1.32)].

APPENDIX G

DIFFUSIVE TRANSPORT AND THERMIONIC EMISSION IN SCHOTTKY BARRIER TRANSPORT

In this appendix we examine again the assumptions of our calculation of transport over barriers and especially the assumption of constant quasi-Fermi levels through the barrier region. We have developed the following physical picture. Below the energy of the barrier no current flows; electrons cannot propagate into neighboring layers and the distribution function is spherically symmetric. Above the barrier a thermionic current can flow without collisions like a jet stream of air. The density of this jet stream is constant everywhere because of the assumption of constant quasi-Fermi levels. In equilibrium, this assumption is indeed exact, and the jet stream is balanced by an opposing stream from the other material. If we apply an external forward voltage, however, we raise the quasi-Fermi level on one side, therefore increasing the jet stream and generating a net current. [In the reverse direction we soon reach current saturation, as can be seen from Eq. (13.12).] These assumptions oversimplify the physics of the process, as can be seen from Figure G.1, where we have divided space into a finite mesh of side length L_m where L_m is the mean free path.

The electron transport between the rectangles is dominated by certain rates. A vertical transition of electrons requires electron-electron interaction or phonon emission and absorption, which is characterized by the time constants τ_{tot}^{e-e} or τ_{tot}^{e-ph} , respectively.

A horizontal transition at the right side is characterized by the velocity of electrons and the corresponding time constant L_m/v_d , where v_d is the drift velocity. The thermionic emission over the last partition of the barrier does not involve diffusion and the corresponding time constant is determined by the z component of the electron velocity, v_z , which means it is determined by the time electrons need to propagate to the left without a collision. There will be few reflections of electrons from the left because electrons find themselves at extremely high energy in the GaAs (metal) and will spontaneously emit a phonon after $\approx 10^{-13}$ s. If the electron energy is above the L valleys in GaAs, the phonon emission is much enhanced because of intervalley deformation potential scattering and the high

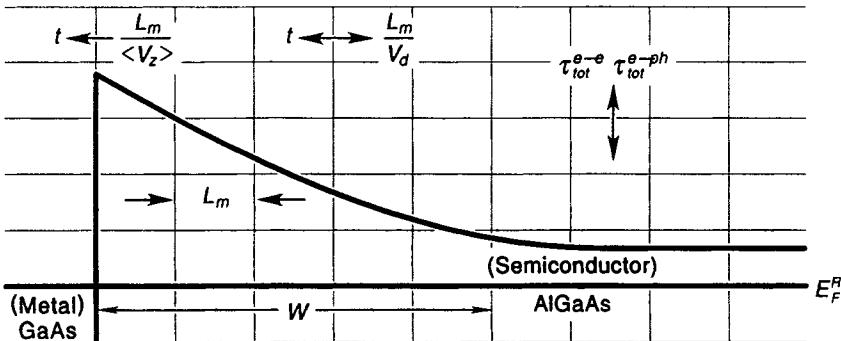


Figure G.1 Schematic of electron supply rates in heterojunctions (Schottky barriers).

density of final states, which can reduce the time constant to less than 10^{-14} s.

In the space elements away from the junction, electrons can be reflected, for example, by phonon scattering or at low energies by the barrier itself. Therefore, the time constant away from the junction will be given rather by the distance divided by the drift velocity, that is, by L_m/v_d , where v_d is the position-dependent average electron velocity (average over positive and negative values), which is much smaller than the average of the absolute value of the electron velocity or the average taken in one direction only as is done for calculating the thermionic emission current. We can see now that our thermionic jet stream theory can work only if the depletion width is very short and comparable to the mean free distance L_m of collisionless travel.

If W is longer, then electrons are lost at a rate $\langle v_z \rangle/L_m$ at the barrier, where $\langle v_z \rangle$ is the average velocity of electrons going in negative z -direction (if none returns), whereas away from the barrier they are replenished at a rate of v_d/L_m only (electrons are scattered and reflected back), which is usually much smaller (see Figure G.1). Because these two rates are very different, the assumption of a constant electron density above the barrier height breaks down and so does the assumption of a constant quasi-Fermi level. In fact, the whole concept of a quasi-Fermi levels may not be applicable because the distribution function may not even have a Fermi shape if the rates shown in Figure G.1 do not permit it. Only a very strong electron-electron interaction can help to establish at all times a Maxwellian high energy tail. However, close to the junction the electron density is low and therefore electron-electron interactions are rare.

Let us summarize the main point: The electrons need to be replenished from the bulk of the semiconductor. If the junction is rather broad, the replenishing current is of "diffusive nature." This diffusion current may be smaller than the thermionic current and therefore represents the limiting factor. Then if W is very large, a "diffusion solution" will be more appropriate than a thermionic current. The thermionic current then only presents a boundary condition, as it represents the current close to the interface (the last rectangle of Figure G.1).

Below we treat this case following Schottky as well as Crowell and Sze.

Assuming a constant mobility and diffusion constant ($T_c = T_L = T$), the current can be written as Eq. (11.10)

$$j = en\mu + eD\partial n/\partial z \quad (\text{G.1})$$

Outside the depletion region we have $F = 0$. Inside the region $F \neq 0$ and we will denote it by $-\partial\phi/\partial z$, where ϕ is the potential. Because, in steady state, the current is constant, we can integrate Eq. (G.1) using $-e\phi(z)/kT$ as integrating factor. For brevity, we denote $-e\phi(z)/kT$ by $\bar{\phi}$; then, using the Einstein relation for the diffusion constant [Eq. (11.11)], we have

$$j = eD \left(-n \frac{\partial \bar{\phi}}{\partial z} + \frac{\partial n}{\partial z} \right) \quad (\text{G.2})$$

Multiplying Eq. (G.2) by $e^{-\bar{\phi}}$, we can write

$$je^{-\bar{\phi}} = eD \frac{\partial n e^{-\bar{\phi}}}{\partial z} \quad (\text{G.3})$$

In steady state the current density j is constant and has the same value everywhere. Therefore

$$j \int_0^W e^{-\bar{\phi}} dz = eD[n(W)e^{\bar{\phi}(W)} - n(0)e^{\bar{\phi}(0)}] \quad (\text{G.4})$$

$\phi(z)$ is known from the solution of the Poisson equation in the depletion region to be

$$\phi(z) = \frac{e}{2\epsilon\epsilon_0} N_D^+ z^2 + C_1 z + C_2 \quad (\text{G.5})$$

where C_1 and C_2 are constants, which can be determined from the boundary conditions $\bar{\phi}(W) - \bar{\phi}(0) = e(V_{bi} - V_{ext})/kT$ and $\left. \frac{\partial\phi(z)}{\partial z} \right|_W = 0$. Here we have made use of the fact that all of the external potential drops in the depletion region. Denoting the function $e^{\bar{\phi}(0)} \int_0^W e^{-\bar{\phi}} dz$ by $F(N_D^+)$, we have from Eq. (G.4)

$$jF(N_D^+) = eD[n(W) \exp(-\bar{V}_{bi} - \bar{V}_{ext}) - n(0)] \quad (\text{G.6})$$

where $F(N_D^+)$ can easily be calculated from Eq. (G.5) and the boundary conditions. $n(W)$ is equal (in the spirit of the depletion approximation) to the equilibrium value of the carrier concentration at W , which we denote by $n_c(W)$. The equilibrium concentration $n_c(0)$ at the junction ($z = 0$) is then

$$n_c(W) \exp(-\bar{V}_{bi}) = n_c(0)$$

and therefore,

$$jF(N_D^+) = eD(n_c(0)e^{\bar{V}_{ext}} - n(0)) \quad (\text{G.7})$$

We have yet to determine $n(0)$, the nonequilibrium concentration, at the junction. We know that the current flowing to the left at the junction is given by the thermionic emission current, which corresponds to the carrier density $n(0)$, and zero barrier height (because we are considering now the current at the junction on the top of the barrier).

From the definition of the quasi-Fermi level, Eqs. (9.10), (5.7), and (10.19), this current is given by

$$j_{\text{th}}^0 = A^* T^2 \left(\frac{n(0)}{N_c} - \frac{n_c(0)}{N_c} \right) \quad (\text{G.8})$$

Here N_c is the so-called effective density of states $N_c = 2(2\pi m^* kT/h^2)^{3/2}$ as can be derived from Eq. (5.26) and the definition $n = N_c \exp(E_F - E_c)/kT$. [In Eq. (5.26) E_F is measured from E_c .]

The second term in Eq. (G.8) represents the current coming back from the left side (GaAs, metal), which is equal to the equilibrium current since no significant external voltage drops at the left side.

The term $A^* T^2 / N_c e$ is sometimes called the interface recombination velocity v_R (it has the correct dimension). In other words, the left side is viewed as a trap where the electrons are captured and cannot return. Now the fact is used that in steady state the current is the same everywhere and $j = j_{\text{th}}^0$. Combining Eqs. (G.7) and (G.8), we have then

$$j = \frac{eDn_c(0)(e^{\tilde{V}_{\text{ext}}} - 1)}{F(N_D^+) + \frac{D}{v_R}} \quad (\text{G.9})$$

INDEX

A

Abramo, A., 107
Abramowitz, M., 235, 245
Acoustic deformation potential, 98
Acoustic deformation-potential scattering, 103
Acoustic phonon scattering, 94, 101, 103, 104, 106, 115, 123, 173
Acoustic phonons, 252
Alam, M. A., 191
AlGaAs, 147, 159–162
 insulating qualities of, 161–162
Alloys, band structure of, 50–54
Amorphous solids, 57
Anderson, C. L., 229, 245
Ando, T., 315
Antibonding state, 33
Areal defects, 57
Ashcroft, N. W., 48, 55
Asymmetric junctions, 221–222
Atomic force microscope, 298
Atoms, coupling, 33–34
Auger recombination, 139
Avalanche breakdown, 236

B

Band diagram
 of *p*–*n* junction with external voltage applied, 206
 of tunnel diode, 242
Band diagrams
 for an abrupt *p*–*n* junction, 203
 rules for plotting, 202
Band edge discontinuity, 54
Band gap, 218, 270
 differences in the bipolar transistor, 271

energy, effect of gradients of, 125
in heterolayer transistors, 271
narrowing as a function of impurity concentration, 270, 271
shinkage owing to many body interactions, 270
temperature dependence of, 307–308
Band structure
 equation for, 45
 in QWLDS, 255
 influence of, on ionization coefficient, 236
 of alloys, 50–54
 of bipolar transistor, 266, 268
 of important semiconductors, 40
 of metal-insulator semiconductor (MIS) structure, 273
of separated semiconductor and metal, 194
parameters, 50
rise in density of states at higher energy, 233
sample region for calculation of, 41
temperature dependence of, 307–308
Band tailing, 73, 270
Baraff, G. A., 233
Bardeen transfer hamiltonian, 181
Bardeen, J., 23, 58, 97, 194, 265, 266, 272, 302, 303
 model of tunneling of, 303
Baym, G., 13, 14, 18
Beam-of-light transistor, 266
Bebb, H., 137, 146
Bergstresser, T. K., 37, 38, 40, 55
Bernstein, G. H., 298
Bethe's thermionic emission theory, 257
Bethe, H. A., 147, 230, 245

- Bipolar transistors, 266–272, 277
band structure of, 266, 268
- Bloch's theorem, 6, 27, 29, 35, 136, 305
- Body-centered cubic lattice, 20, 26, 27, 31
Brillouin zone of, 28
- Bohr radius, 60
- Boltzmann statistics, 171
- Boltzmann transport equation (BTE), 1, 96, 109–134, 139, 140, 167, 169, 170, 181, 182, 188, 189, 257, 266, 287, 295, 296, 298, 309, 311, 312
derivation of, 109–114
solutions in the relaxation time approximation of, 114–121
- Boltzmann's law, 154
- Boltzmann-Bloch equations, 256
- Bonding state, 33
- Bonding-antibonding splitting, 33
- Bose distribution, 97
- Bosons, 97
- Bragg reflection, 27, 29–30, 39, 43, 55, 176, 238
- Brattain, W., 265, 266
- Bravais lattice, 19, 20
- Brillouin zone, 7, 8, 18, 21, 27–30, 90, 103, 236, 237
- Brooks Herring formula, 281
- Brooks, H., 96
- Buckey balls, 298
- Bude, J., 229, 230, 245
- Burstein, E., 238, 245
- C**
- Caldeira, A. O., 110, 134
- Capacitance
changes in, due to dynamics of electron release, 223
depletion, 209–211, 222
diffusion, 208–210, 215, 216, 218, 221–223
quantum, 283
time-dependent, 225
total, 215
under forward bias, 210
- Capasso, F., 236, 245
- Carbon nanotubes, 298
- Carrier velocity
and Poisson equation, 176–178
- Carriers and photons
coupling of, 253–257
- Casey, H. C. Jr., 53, 55, 136–138, 146
- Charge carrier temperature T_c , 240
- Child's law, 178
- Chow, P. C., 315
- Chow, W. W., 256, 264
- Chuang, S. L., 248, 254, 255, 257, 259, 264
- Classical mechanics, equations of, 2–9
- Classical transport, 250
- CMOS devices, 273, 296
breakdown, 297
tunneling through thin oxides, 297
- Cohen, M. L., 37, 38, 40, 55
- Collector base voltage, 266
- Complementary MOS, 273, *see* CMOS devices
- Compositional disorder, 50
- Conduction band, 39, 45, 46, 50, 52, 53, 67, 70, 71
electron density in, 75–79
in Si, 43
- Conductivity matrix, 24, 122
- Connection rules, for the potential at an interface, 153–154
- Continuity, equation of, 149, 168, 169, 181, 183–186, 207, 208, 213, 243, 247, 257, 259
- Conventional devices
extensions of, 291–293
- Conwell, E. M., 101, 102, 107, 134, 172, 179
- Correlation hole, 270
- Coulomb blockade effect, 293, 294
- Coulomb gauge, 254
- Coulombic repulsion, 270
- Coupling atoms, 33–34
- Coupling of carriers and photons, 253–257
- Crowell, C. R., 229, 245, 318
- Crystal structure
energy band theory, 33–54
imperfections, 57–65
lattice vibrations, 2–9
of GaAs, 19–21
of silicon, 19–21
symmetry of lattice, 19–31
- Crystal, equations of motion in a, 42–46
- Crystal-growth techniques, 62, 292
- Current density and distribution function,

- 121–125
- Current saturation, 213, 244, 317
- Current, equations of, 183, 250
- Cyclotron orbit, 93
- D**
- DAMOCLES code, 183, 189, 285, 293, 296
- Das Sarma, S., 157, 166
- Datta, S., 94, 107, 134, 298
- de Broglie wavelength, 15, 157, 281
- Debye length, 83, 87, 151, 178, 222
- Deep impurity levels, theory of, 60–62
- Deformation-potential scattering, 97, 98, 102, 174
- acoustic, 98, 103
- optical, 100, 101, 103, 106, 115, 117, 172
- Density functional theory, 297
- Density of states, 67–73, 87, 100–102, 135, 141, 200, 233, 312, 320
- reduced, 255
- two-dimensional, 162, 164
- Depletion capacitance, 209–211, 222, 259
- Depletion voltages, estimates of, 152
- DESSIS code, 183, 188, 280
- Deuterium, 287–289, 295, 297
- Device equations, 167–179
- ideal, 279
- intermediate set of, 182, 183
- optimum set of, 182
- Shockley set of, 183
- simplest set of, 183
- Device modeling, 266–277
- Devices, scaling down, 278–279
- Devoret, M. H., 298
- Dielectric constant, 60, 81–88
- complex, 256
- Dielectric relaxation time, 223, 243
- Diffusion Capacitance, 259
- Diffusion capacitance, 208–210, 215, 216, 218, 221–223
- Diffusion current, 170, 177
- Diffusive transport and thermionic emission
- in Schottky barrier transport, 317–320
- Dingle, R., 163, 166
- Diode current, basic equations for, 207–211
- Diodes, 193–246, 265
- Esaki, 242
- laser, 247–264
- negative differential resistance and, 241–244
- resonant tunneling, 291
- Dirac's notation, 17
- Dislocations, 57, 62–64
- Disorder bowing, 53
- Dispersion relation, 7
- Distribution function, 73–74, 109–111, 114, 121, 198, 231, 232, 240
- and current density, 121–125
- nonequilibrium, 109
- DLTS (deep level transient capacitance spectroscopy), 223
- Domain formation, high field, 243
- Doped semiconductor, 76–78
- Doping
- concentration in an ideal abrupt homojunction, 202
- modulation, 163, 165, 260, 292
- Dow, J. D., 62, 65, 107
- Drain current, 275
- extremely high fields, 283
- for a submicrometer channel length transistor, 278
- MOSFETs, 282
- of MOS transistors, 277
- space-charge limited current, 276
- Drain voltage
- of MOS transistors, 277
- Drain-induced barrier lowering (DIBL), 281
- Drift velocity, 175, 284
- Drift-diffusion approximation, 257
- Drude theory, 89–94
- model of conduction, 90
- Duke, C. B., 17, 18, 238, 245
- Duncan, A., 190, 191
- Dutton, R. W., 196
- E**
- $E(k)$ relation, 30, 33–39, 43, 44, 46, 47, 50, 69, 70, 90, 229, 230, 238, 309
- calculation of with temperature-dependent structure factor, 308
- Ebers-Moll equations, 270
- Edge-emitting laser diode, 248, 249
- Effective mass, 43, 45
- approximation, 44, 94, 122, 148, 157, 183

- calculation of, 50
 in GaAs, 106
 in silicon, 104
 negative, 46
 theorem, 44, 45, 50, 58, 98, 135, 153
- E**ffective Richardson constant, 149
Ehrenfest's theorem, 42
Einstein coefficient, 255
Einstein relation, 219, 283
 for diffusion constant, 170, 226, 319
EISPACK, 12
Elastic tunneling process, 302
Electron concentration, 206
Electron density, in conduction band, 75–79
Electron spin, 1
Electron temperature approximation, 232
Electrons, equilibrium statistics for, 67–79
Emitter base voltage, 266
Emitter efficiencies, 271
Empirical pseudo-potential method, 39
Energy balance equation, 182, 183, 227, 232
Energy bands, theory of, in crystals, 33–54
Energy gap, 33, 36, 39, 51, 52
 change of
 owing to lattice vibrations, 307
 with pressure, 307
 with temperature owing to the volume increase, 307, 308
Edisorder bowing of, 52–53
 of AlAs, 51
 of AlGaAs, 161
 of AlSb, 51
 of GaAs, 51
 of Ge, 51
 of InAs, 51
 of InP, 51
 of InSb, 51
 of Si, 51
Energy transport, 183
Equilibrium statistics for electrons and holes, 67–79
Equipartition approximation, 97, 102
Esaki diode, 242
Esaki, L., 89, 107, 242
Esaki-Tsu oscillator, 90
Exchange correlation effect, 270
 potential, 157, 270
- F**
f scattering, 103
Fabry-Perot interferometer, 292
Face-centered cubic lattice, 20, 21, 26, 27, 30
 Brillouin zone of, 28
Fan, H. Y., 307
Fantner, E., 108
Fermi distribution, 73, 74, 97, 109, 115, 125, 127–129, 147, 158, 262
Fermi level E_F , 76
Fermi levels, 194
 pinning of, 194–196
Fermi, E., 14
Fermi, Golden Rule, *see* Golden Rule
Fermions, 97
Ferry, D. K., 107, 130, 134, 281, 289
Feynman, R. P., 11, 18, 90, 91
Field effect transistors, 272–278
Finite difference method, 183, 315
Forward bias, 198, 206, 209–212, 216, 219, 223, 228, 242, 262
p–*n* junction in, 205
 ac carrier concentrations and current in, 213–215
 capacitance under, 210
 extreme, 219–221, 242, 260
p–*n* junction in, 224, 259, 266, 267
 Schottky barrier under, 198
 steady-state current in, 211–213
Fourier transformation, 26, 29, 34–39, 44, 45, 50, 85–87, 94, 95, 305
Fowler, A. B., 315
Franz-Keldysh effect, 238
Free carrier concentration, *p*–*n* junction, 203, 204
Free carrier depletion, of semiconductor layers, 151–153
Frensley, W. R., 54, 55
- G**
g scattering, 103
GaAs
 approximate phonon scattering rate of, 105
 conduction band minimum of, 100
 crystal structure of, 19–21

- intrinsic carrier concentration, 78
material parameters for, 105, 106
negative resistance of, 242
scattering mechanisms in, 103–107
variation of electron energy in, after scattering events, 234
- GaP
intrinsic carrier concentration, 78
- Gate voltage, charge induced by, 159, 274
- Gauss's law, 208
- Gauss, theorem of, 159, 266
- Ge
intrinsic carrier concentration, 78
- Generation-recombination (GR) processes, 135–146, 149, 168, 205, 211, 274
rate equations, 144–145
rates, 140–144, 207, 269
- Golden Rule, 14, 16, 17, 94, 136, 229, 230, 257
and generation-recombination processes, 135
and QWLDs, 253
and tunneling, 301–304
scattering probability from the, 94–103
- Golub, G. H., 12, 18
- Gossard, H. C., 166
- Grabert, H., 298
- Green's theorem, 164, 166, 303
- Group theory, 3
elements of, 22–28
- Grupen, M., 191, 250, 252, 258, 262–264
- Gummel, H. K., 185, 186, 191, 270
- Gunn effect, 174, 244
- Gunn, J. B., 243
- H**
- Hall effect, 92, 93
and magnetic resistance for small magnetic fields, 309–310
- Hall field, 93
- Hall, R. N., 242, 245
- Halperin, B., 73, 79
- Hamilton, W. R., 2
- Hamiltonian
Bardeen transfer, 181
- Hamiltonian equations, 1–10, 17, 42, 301–303
- Harrison, W. A., 34, 54, 55, 62, 65, 100, 107, 304
- Heavy hole curve, 46
- Heil, 272
- HEMT, 292–293
- Herring, C., 96
- Hess, K., 41, 55, 100, 103, 105–108, 134, 166, 174, 176, 179, 191, 192, 220, 222, 227, 234, 237, 245, 250, 252, 258, 262–264, 287, 289, 299
- Heterojunctions, 50, 247, 291
band structure, 54
barrier, 147–166
built-in voltage in, 203
interface, conduction band edge at, 148
lattice-matched, 194
self-consistent potential at, 315–316
wave functions, 54
- Heterolayer boundaries, pronounced effects of size quantization and, 162–166
- Heterolayer transistors, 271
- High electric fields, 272, 278, 281, 283
- High electron mobility transistor (HEMT),
see HEMT
- High energy tail, 232
- High field domain formation, 243
- High field effects, in semiconductor junctions, 226–241
- Higman, J. M., 103, 105, 227, 245
- Hilsum, C., 243
- Hole concentration, 206
- Holes, 46
equilibrium statistics for, 67–79
- Hot electron, 171
desorption chemistry, 287
emission of, 240
- Hot electron degradation, 183, 189
- Hot electron effects, 281–289
and bipolar transistor operation, 271
and method of moments, 170–174
and saturation, 276
in MOS transistors, 284
in space-charge region of Schottky barriers or *p*–*n* junctions, 226, 227
- thermionic emission currents of, 226
- thermionic emission currents of from one layer, 241
- Hot phonons, 262

Howard, W. E., 158, 159, 166, 315

Hu, C., 289

Huffaker, D. L., 250, 264

Hydrodynamic equations, 183

Hydrodynamic simulations, 191

Hydrogen desorption, 297

Hydrogen passivated interface, 286

I

Iafrate, G. J., 176, 179, 299

Image force, 54, 157, 196, 197

Impact ionization, 139, 183, 189, 226, 230, 232, 233, 236, 244, 272, 284–289

a dead space for, 285

as scattering mechanism, 232

avalanche transit time (IMPATT) devices, 244

energy loss due to, in energy balance equation, 232

for electrons in silicon as function of energy, 231

in *p*–*n* junctions, 229–236

in **k** space, 229

in MOS transistors, 284, 285

in real space, 229

Imperfections, crystal structures, 57–65

Impurities, 57

Impurity concentration, band gap narrowing as function of, 270

Impurity potential, 84

Impurity scattering, 94–97, 103, 105–107, 115, 116, 123, 131, 163, 165, 270, 293

Imrefs, *see* Quasi-Fermi levels

Inkson, J. C., 287, 289

Interface, 57, 62–64

connection rules for the potential at an, 153–154

lattice-matched, 64

recombination velocity v_R , 320

roughness scattering, 107

Si-SiO₂, 64

states, 62, 273

traps, 64

Intervalley scattering, 174

Inversion channel resistance, 276

Inversion electrons, 161, 272, 273, 276

Inversion layer, 272, 273, 275, 276, 281

Ionization coefficient, influence of band structure on, 234, 236

J

Jackiw, R., 230, 245

Jakumeit, J., 190, 191

Jones, W., 53, 55

Joule's heat, 171, 253, 262

K

k space, 39, 41, 43, 49, 60, 61, 70, 96, 226, 229, 242, 244

spherical constant energy surfaces in, 68 transitions in, caused by photon emission, 137, 138

Kan, E. C., 191

Kane, E. O., 71, 72, 79, 230, 231, 238, 245

Keldysh formula, 230, 296

Keldysh, L. V., 230

Kirchhoff's law, 270

Kizilayalli, I. C., 289

Klimeck, G., 299

Ko, P. K., 282, 289

Kocevar, P., 191

Koch, S. W., 264

Kroemer, H., 54, 55, 64, 65

Kronecker delta symbol, 15

Kuchar, F., 96, 108

L

Landsberg, P. T., 23

Landsberg, P. T., 3, 9, 18, 38, 55, 73, 74, 79, 108, 146, 211, 245

LAPACK, 12

Laser diodes, 247–264

modulation response, 260 numerical solutions of the equations for, 257–263

Lattice

body-center cubic, 20

body-centered cubic, 26–28, 31

Bravais, 19, 20

face-center cubic, 20

face-centered cubic, 26–28

simple cubic, 20, 26

symmetry of a crystal, 19–31

Lattice vibrations, 2–9, 21, 90, 96

and change of value of energy gap, 307

- Lattice-matched semiconductor heterojunction, 53, 194
Laux, S. E., 220, 222, 245
Law of the junction, 219, 220
Lax, M., 73, 79
Leggett, A. J., 110, 134
Lent, C. S., 298
Lifetime of electrons, 145
Lilienfeld, 272
Line defects, 57
Loan, C. F., 12, 18
Long, M., 64, 65
Low frequency roll off, 260
Lucky electrons, 233
Lundquist, F., 238, 245
Lundstrom, M. S., 191
Lyding, J. W., 63, 289
- M**
- Macucci, M., 294, 299
Madelung, O., 31, 52, 55, 100, 101, 108
Magnetic fields, Hall effect and magnetoresistance for small, 309–310
Mahen, G. D., 84, 88
March, N., 53, 55
Maxwell equations, 1, 92, 167, 181, 182, 247, 256–258, 295, 296, 298
Maxwell-Boltzmann distribution, 73, 74, 77, 87, 109, 115, 116, 119–122, 125, 127, 129, 140, 143, 147, 170
McGill, T. C., 201
Mean value theorem, 225, 266
Mermin, N. D., 48, 55
Mertens, R., 271, 289
MESFET, 292
Metal-insulator semiconductor, *see* MIS
Metal-oxide-silicon semiconductor transistor, *see* MOS
Metal-semiconductor contact, 194
Metal-semiconductor field effect transistor, *see* MESFET
Method of moments, 167–170, 309 and hot electrons, 170–174 and mean values, 266 and space-dependent carrier distributions, 176–178 and velocity transients, 175–176
- power balance equation from, 311–313
Microcavity lasers, 263, 296
MINIMOS code, 183, 188, 280, 293, 296
Minority carrier injection, 265
MIS, 272, 273
band structure of, 273
with applied voltage, 274
Mobile (free) charge carriers, solution of Poisson's equation in the presence of classical case, 154–156 quantum mechanical case, 157–162
Mobility in small MOSFET, 281–284 of HEMT, 293
MODFET, 184, 292–293, 296
Modulation doped field effect transistor, *see* MODFET
Modulation doping, 163, 165, 260, 292
Modulation response including nonlinear gain effects, 263
Molecular beam epitaxy, 292
Momentum matrix element, 136
Momentum scattering rate, 96, 116
Monte Carlo method, 129–132
Monte Carlo simulations, 148, 150, 167, 178, 182, 189, 226, 227, 234, 236, 244, 282, 284, 287, 293 and impact ionization, 285 full-band, 183, 188–190 tunneling included in, 304
Moore's law, 288
Morgan, D. J., 23
MOS, 23, 24, 64, 160, 182, 273, 274, 277, 279
drain current versus drain voltage characteristic of, 277
emission of hot electrons from silicon into silicon dioxide on, 284
hot electron effects in, 284
impact ionization in, 284
reliability of gate oxide in, 285
value of, 273
MOSFET, 12, 276, 282, 291, 292, 295
device structure, 280
energy distribution in, 190
mobility in small, 281–284
Moss-Burstein shift, 142
Mott-Gurney law, 177

N

Nanodevice integrated circuits, *see* NIC
 Nanostructure devices, *see* ND
 Nanostructure simulation, 298
 challenges in, 295–298
 Narrow-width effect, 280
 ND, 293
 Negative differential resistance, 276
 and semiconductor diodes, 241–244
 NEMO, 292, 296
 Newton's first law, 2
 Newton's method, 187, 258
 Newton's second law, 2
 NIC, 293, 295
 Ning, T. H., 289
 NMIS, 273
 NMOS, 273
 Numerical device simulations, 181–192
 Numerov process, 315

O

Ohm's law, 23, 173, 178, 275
 Ohmic contacts, 194–201
 One band approximation, 42, 44, 111, 183,
 188, 305–306
 Openheimer, R., 17
 Optical absorption, 254
 Optical deformation-potential scattering,
 100, 101, 103, 106, 115, 117, 172
 Optical exitation, 140
 Optical matrix element, 99
 Optical phonon scattering, 173
 Optical phonons, 8, 97, 99, 101
 Optical recombination, 252, 258, 262
 Optical transitions
 in a quantum well, 138
 in QDs, 297
 Optoelectronics, 296
 Orbits in semiclassical phase space, time
 evolution of, 48
 Overall charge neutrality, 209, 266
 Overshoot effects, 176, 271
 Oxide rings, 250

P

p-n junctions, 135, 139, 144, 201–225, 248,
 267, 278
 asymmetric, 221–222

band diagram for an abrupt, 203
 band diagram with external voltage
 applied, 206
 band edges and quasi-Fermi levels of, in
 extreme forward bias, 220
 calculating dc current through, 212
 current in, 204–205, 226–228
 forward-biased, 205, 221, 266, 267
 free carrier concentration in, 203, 204
 high doped, 242
 high energy physics analogy for, 205
 impact ionization in, 229–236
 light-emitting, 267
 nonideality factor, 218
 reverse-biased, 236, 266, 267
 switching cycle of, 224
 terminal capacitance of, 217, 218, 221
 two-carrier transport in, 236
 Panish, M. B., 53, 55, 136–138, 146
 Pantelides, S. T., 64, 65
 Pauli principle, 140, 157, 270
 Penn, D. R., 87, 88
 Perturbation theory, 13, 17, 28, 83, 85, 164,
 225, 228, 302, 307
 harmonic, 15
 time-dependent, 14
 time-independent, 15
 Perturbed quantum well, 302
 Perturbed wave function, 85, 163
 Phonon coupling mechanisms, values for
 M_q , 307
 Phonon scattering, 96–107, 165, 232, 233,
 252
 acoustic, 94, 101, 103, 104, 106, 115, 123
 Phonons
 absorption, 99, 103
 acoustical, 9, 97–99, 103
 and drift velocity, 284
 displacement of sublattices, 101
 emission, 99, 103
 energy of, 307
 longitudinal, 99
 optical, 8, 9, 99, 101, 103
 polar optical, 9, 100
 scattering, 96–107
 transversal, 99
 Photon emission, transitions in k space
 caused by, 137, 138

- Photons
coupling of, carriers and, 253–257
- Photothreshold, 54
- Piezoelectric potential, 97
- Pinch-off effect, 276
- PISCES code, 183, 188, 280
- PMOS, 273
- Poetz, W., 191
- Point contact transistor, 265
- Point defects, 57
- Point group, of crystal lattice, 22–26
- Poisson equation, 81–84, 86, 147, 150, 151, 153, 154, 167, 169, 177, 181–185, 189, 209, 210, 216, 241, 247, 259–261, 280, 315, 319
and carrier velocity, 176–178
depletion approximation, 151, 152
solution of, in the presence of mobile (free) charge carriers
classical case, 154–156
quantum mechanical case, 157–162
- Polar optical phonon scattering, 100, 102, 105, 115, 170
- Polar optical phonons, 9, 252
- Polaron, 100, 101
- Polasko, K. J., 65
- Polycrystalline silicon (POLY), 279
- Pool-Frenkel effect, 228
- Poon, H. C., 270
- Porod, W., 298
- Potential at an interface, connection rules for, 153–154
- Potential $V(\mathbf{r})$, 16, 36, 37, 44, 305
- Power balance equation, from the method of moments, 241, 311–313
- Price, P. J., 134
- Pseudopotential method, 39, 50, 52, 54
- Q**
- QD, 294, 295, 297–298
- Quantum dots, *see* QD
- Quantum effects, 1, 157, 160
- Quantum field theory, 1
- Quantum Hall Effect, 94
- Quantum mechanical dephasing length, 250
- Quantum mechanics, equations of, 9–18
- Quantum number, 12
- Quantum states, 12
- Quantum transport, 250
- Quantum well, 247, 250, 255, 261
AlGaAs/GaAs, 247
coupled, 33
heterolayer structure, 137
optical transitions in, 138
perturbed, 302
states, 257
- Quantum well laser diodes, *see* QWLD
- Quasi-Brownian motion of electron, 90
- Quasi-Fermi levels (imrefs), 121, 125, 127, 128, 139–140, 142, 147, 148, 219, 220, 223, 227, 240, 241, 258, 259, 262, 273, 274, 285
band edges and, of a p – n junction in extreme forward bias, 220
constancy of, 198, 200, 205
constant, through barrier region, 317–320
depletion region, 205, 209, 213
of a barrier structure when voltage is applied, 197
of a Schottky barrier under bias, 198
- QWLD, 247, 248, 252, 256, 263, 291, 294, 297
basic geometry and equations for, 248–250
conduction and valence band edges at a dc bias, 261
energy exchange dynamics of, 253
modulation response for, 259
transport, 252
- R**
- Radiative recombination, 135–138
- Random phase approximation, 83
- Ravaoli, U., 190, 191
- Read, W. T., 244, 245
- Real space transfer (RST), 150, 183, 189–191, 226, 240–241, 244, 276, 282–284
- Reciprocal crystal lattice, 27
- Reciprocal lattice vectors, 27, 28
- Recombination in depletion region, 216–219
- Rectifiers, Schottky barriers as, 199
- Reduced density of states, 255
- Register, L. F., 132, 134, 264
- Relaxation time, 114–116, 118, 124, 165, 172

dielectric, 223, 243
 energy, 171, 172
Resonance tunneling barrier, 292
Resonant tunneling diode, 291
Reverse bias, 211, 212, 228, 242
 and energy gap, 206
 and impact ionization, 272
 effects of, 223–225
p-n junction in, 224, 229, 236, 266, 267,
 272
Schottky barrier under, 198
 simplified energy band diagrams of tunnel
 diode at, 242
Reverse saturation current, 213, 269, 272
Ridley, B. K., 138, 139, 144, 146, 230, 243,
 245
Ritzit method, 158
Robbins, V. M., 245
Rode, D. L., 93, 106, 108
Ruch, J. G., 175, 179
Rutishauser, H., 12, 18

S

Sah, C. T., 222, 223, 245
Sargent, M. III, 264
Saturation current, 213, 244, 276
 influence of hot electron phenomena on,
 271
Scaling schemes, 279
Scanning tunneling microscope, 298
Scattering mechanisms
 in gallium arsenide, 103–107
 in silicon, 103–107
Scattering probability
 from the Golden Rule, 94–103
 per unit time, 94, 96
Scattering rate
 for electrons interacting with remote
 impurities, 163–166
Scattering theory, 89–108
Scharfetter, D. L., 185, 186, 191
Scharfetter-Gummel discretization, 186
Schichijo, H., 161
Schottky barrier height, 194, 196, 197, 201
Schottky barrier transport, diffusive transport
 and therionic emission in, 317–320
Schottky barriers, 147, 148, *see* Tunneling,
 194–201, 223, 292, 293

as rectifiers, 199
 current over, 197
 lowering of, 196, 197
 quasi-Fermi levels, under bias, 198
 speed limitations of, 199
 switching cycle of, 224
 under forward bias, 198
 under reverse bias, 198
Schrieffer, J. R., 23
Schrödinger equation, 1, 10–13, 17, 35, 42,
 44, 45, 55, 157, 182, 183, 257, 261,
 295, 296, 298, 302, 304, 305, 315
 computer solutions of, 315
Schrödinger, E., 10, 12
Screening wave vector, 87
Selberherr, S., 184, 185, 187, 188, 191
Self-consistent potential, at heterojunction,
 315–316
Semiconductor junctions, high field effects
 in, 226–241
Semiconductor-metal junction, 194
Semiconductor-semiconductor junction, 197
Sfb-inductance, 221
Shenai, K., 196, 245
Shichijo, H., 41, 55, 105, 106, 108, 132, 134,
 174, 179, 192, 234, 237, 245
Shockley equations, 167–179, 189, 190, 213,
 280, 282, 283, 295, 296
 and real space transfer, 283
 and Velocity overshoot, 283
 numerical solution of, 184–191
Shockley set of device equations, 183
Shockley, W., 97, 232–234, 236, 272
Shockley-Read-Hall centers, 168, 257, 285
Short channel effect, 280
Short diodes, 215–216
Short gates and threshold voltage, 279–281
Si
 crystal structure of, 19–21
 density of states for, 71
 energy distribution in bulk, 190
 energy relaxation time of, 171
 intrinsic carrier concentration, 78
 material parameters for, 104
 scattering mechanisms in, 103–107
Simple cubic lattice, 20, 26
Size quantization effects, 157, 160, 182, 183
 and heterolayer boundaries, 162–166

- at heterolayer interface, 315–316
Sound propagation in solids, microscopic theory of, 8
Space-charge limited current, 176, 276
 drain current, 276
Space-charge region, 212
Space-dependent carrier distributions, 176–178
Space-dependent power balance equation, 241
Spectral hole, 254, 259
Spontaneous emission, 97, 255, 257
Square well trap, 228
Stegun, I. A., 235, 245
Stern, F., 23, 157–159, 166, 315
Stillman, G. E., 238, 245
Stimulated emission, 97, 255–257, 259
Störmer, H. L., 166
Stratton, R., 169, 183, 191, 311
Streetman, B. G., 193, 201, 236, 246, 270, 289
Subthreshold current, 275
Sum rules, 266
Super cell, 255
Surface states, 57, 58, 62, 63
 and pinning, 194
Surface-emitting laser diode, 248, 249
Surfaces, 57, 62–64
Switching cycle of a *p*–*n* junction, 224
Sze, S. M., 246, 272, 289, 299, 318
- T**
- Tang, J. Y., 103, 108
Tarucha, S., 294, 299
Temperature dependence of the band structure, 307–308
Temperature gradients, effect of, 125
Temperature-dependent mobility, for several scattering mechanisms, 122–125
Theory of Henry, 254
Thermionic coupling, 258
Thermionic emission, 258
 in Schottky barrier transport, diffusive transport and, 317–320
 of electrons over barriers, 147–150, 200, 226, 242
Thermionic-emission-diffusion theory, 200, 241
- Thomas-Fermi length, 87, 151
Three-dimensional tunneling barrier, 304
Threshold voltage, 275
 short gates and, 279–281
Thurmond, C. D., 78, 79
Time-dependent capacitance, 225
Total defects, 57, 58
Tougaw, D. P., 298
Transient capacitance methods, 223, 225
Transistor lifetime, 288
Transistors, 265–289
 beam-of-light, 266
 bipolar, 266–272, 277
 field effect, 272–278
 heterolayer, 271, 297
 high electron mobility, 292
 high-speed, 296
 MIS, 272–274
 MODFET, 184
 MOS, 23, 24, 147, 160, 273, 274, 277, 279, 285
 MOSFET, 276, 280–285, 291, 292
 NMIS, 273
 NMOS, 273
 PMOS, 273
 point contact, 265
 simple models of, 266–277
 single electron, 293
Transition matrix, 54
Translational invariance, 26–28
Transport, over a heterobarrier, theory of, 150
- Traps
 density of deep, 225
 time-dependent filling of, 225
- Tsu, R., 89, 107
- Tunneling, 11, 16, 17, 162, 200, 201, 242, 296
 and the Golden Rule, 301–304
 Bardeen's model of, 303
 calculations of, 297
 Zener, 226, 236–238, 241, 272
- Tyagi, M. S., 208, 246
- U**
- Uncertainty principle, 16, 60
Undoped semiconductors, current carried by, 162

V

- Vacancies, 57
Vacuum level, 53, 54
Valence band, 46, 47, 49, 50, 52, 53, 67, 75
Van Overstraaten, R., 271, 289
van De Walle, C. G., 287
Velocity of sound, 8, 90
Velocity overshoot, 174, 175, 183, 189, 190, 283, 284
Velocity saturation, 173
Velocity transients, and methods of moments, 175–176
Vertical cavity surface emitting laser (VCSEL), 249, 263, 296
Vibrations, lattice, 2–9
Virtual crystal approximation, 46, 50
Vogl, P., 100, 107, 108, 289

W

- Watkins, T. B., 243
Wavelet representations, 298
Well states, 251
Wiegmann, W., 166
Wigner-Seitz cell, 26, 27, 37, 38
Williams, E., 137, 146
Wolff, P. A., 232, 233
Woodall, J. M., 195
Wright, S. C., 65

Y

- Yale Sparse Matrix solver, 187
Yoder, P. D., 104, 108

Z

- Zener tunneling, 226, 236–238, 241, 272
Ziman, J. M., 88, 96, 97, 108

ABOUT THE AUTHOR

Karl Hess received the Ph.D. degree in applied physics from the University of Vienna, Austria, in 1970.

Dr. Hess currently holds the Swanlund Endowed Chair and is a professor of electrical and computer engineering and of physics at the University of Illinois, Urbana. He has dedicated a major portion of his research to electronic transport in semiconductors and semiconductor devices with particular emphasis on hot electron effects and effects pertinent to device miniaturization. Dr. Hess is particularly interested in problems that require large computational resources for their solution. His current research at the Beckman Institute of the University of Illinois is in the area of molecular and electronic nanostructures.

Dr. Hess has received numerous awards, including the IEEE J. J. Ebers Award of the Electron Devices Society in 1993 and the IEEE David Sarnoff Field Award for Electronics in 1995. He is a Fellow of the American Academy of Arts and Sciences.