Semiconductor Device Physics and Design

Umesh K. Mishra

Jasprit Singh

Springer

SEMICONDUCTOR DEVICE PHYSICS AND DESIGN

Semiconductor Device Physics and Design

by

UMESH K. MISHRA
University of California, Santa Barbara, CA, USA

and

JASPRIT SINGH
The University of Michigan, Ann Arbor, MI, USA

Springer

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN 978-1-4020-6480-7 (HB)
ISBN 978-1-4020-6481-4 (e-book)

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

www.springer.com

Printed on acid-free paper

This book is dedicated to our parents
Srinibas Mishra and Sushila Devi
Gurcharn Singh and Gursharan Kaur

CONTENTS

ACKNOWLEDGEMENTS

Writing a book on Semiconductor Device Physics and Design is never complete and probably never completely satisfying. The field is vast and diverse and it is difficult to decide what should be included in the book and what should not be. Of course it is always a good idea for authors to not discuss areas that they are unfamiliar with and that helped narrow the scope of this book down greatly!! In all seriousness the flow and content of this book is a consequence of the classes that we have taught at UC Santa Barbara and The University of Michigan and reflects what we believe can be taught in a manner that emphasizes physical understanding with an appropriate amount of rigor. At UCSB Prof Kroemer had developed a two-quarter sequence class on device physics which I (Umesh Mishra) took over when I arrived at UCSB in 1990. I developed the class over the past 15 years using his notes as a foundation and the new content is reflected in this book. I am grateful to Prof Kroemer for allowing me to include parts of his notes and homework problems in this book. Prof Mark Rodwell contributed to understanding that the answer to the question “How much β do we need” is application dependent. Prof Steve Long and Prof Rakesh Lal helped with the diode- applications section. Prof Tomas Palacios of MIT contributed to the AlGaN/GaN HEMT description. Dr Karthik Krishnamurthy (RFMD) allowed use of his descriptions of classes of FET amplifiers. Lastly, Dr Jeff Shealy (VP RFMD) and Dr Rama Vetury (RFMD) are thanked for their help in illustrating how the mobile phone uses multiple semiconductor technologies for optimal system performance as described in the introduction. Discussions with Profs. Lorenzo Faraone, Brett Nener and John Dell at The University of Western Australia were helpful (and fun). Drs., Lee McCarthy, Ilan Ben-Yaacov, Nicholas Fichtenbaum and Siddharth Rajan contributed significantly in helping the text layout of the book. We thank several of our colleagues who contributed figures to the book and they have been acknowledged at the appropriate places. Umesh would like to thank his wife Susan for not asking the question “Isn’t it finished?” too many times. Jasprit would like to thank his wife Teresa for drawing numerous figures and YuhRennn Wu for providing device design studies for field effect transistors. We would also like to thank the editors at Springer Verlag for their enthusiasm and support.

PREFACE

It would not be an exaggeration to say that semiconductor devices have transformed human life. From computers to communications to internet and video games these devices and the technologies they have enabled have expanded human experience in a way that is unique in history. Semiconductor devices have exploited materials, physics and imaginative applications to spawn new lifestyles. Of course for the device engineer, in spite of the advances, the challenges of reaching higher frequency, lower power consumption, higher power generation etc. provide never ending excitement. Device performances are driven by new materials, scaling, and new device concepts such as bandstructure and polarization engineering. Semiconductor devices have mostly relied on Si but increasingly GaAs, InGaAs and heterostructures made from Si/SiGe, GaAs/AlGaAs etc have become important. Over the last few years one of the most exciting new entries has been the GaN based devices that provide new possibilities for lighting, displays and wireless communications. New physics based on polar charges and polar interfaces has become important as a result of the nitrides. For students to be able to participate in this and other exciting arena, a broad understanding of physics, materials properties and device concepts need to be understood. It is important to have a textbook that teaches students and practicing engineers about all these areas in a coherent manner. While this is an immense challenge we have attempted to do so in this textbook by judiciously selecting topics which provide depth while simultaneously providing the basis for understanding the ever expanding breath of device physics.

In this book we start out with basic physics concepts including the physics behind polar heterostructures and strained heterostructures. We then discuss important devices ranging from $p - n$ diodes to bipolar and field effect devices. An important distinction users will find in this book is the discussion we have presented on how interrelated device parameters are on system function. For example, how much gain is needed in a transistor, and what kind of device characteristics are needed. Not surprisingly the needs depend upon applications. The specifications of transistors employed in A/D or D/A converter will be different from those in an amplifier in a cell phone. Similarly the diodes used in a laptop will place different requirements on the device engineer than diodes used in a mixer circuit. By relating device design to device performance and then relating device needs to system use the student can see how device design works in real world.

It is known that device dimensions and geometries are now such that one cannot solve device problems analytically. However, simulators do not allow students to see the physics of the problem and how intelligent choices on doping, geometry and heterostructures will impact devices. We have tried to provide this insight by carefully discussing and presenting analytical models and then providing simulation based advanced results. The goal is to teach the student how to approach device design from the point of view some one who wants to improve devices and can see the opportunities and challenges. The end of chapter problems chosen in this book are carefully chosen to allow students to test their knowledge by solving real life problems.

Umesh K. Mishra
University of California
Santa Barbara

Jasprit Singh
The University of Michigan
Ann Arbor

Lesson Plan for a 1 Semester course: 35-40 lectures Ⓡ : Reading Assignment

Structural Properties: 2 Lectures

Ⓡ Crystals: Lattices, Basis, and Planes

- Heteroepitaxy: Strain tensor,
- Polar Effects: Spontaneous and Piezoelectric Effects

Electronic levels in Semiconductors: 3 Lectures

Ⓡ Particles in attractive potentials, quantum wells

Ⓡ Electrons in crystalline materials

- Important bandstructures

Ⓡ Distribution functions

Ⓡ Metals and insulators

- Mobile carriers
- Doping: Dopants and polar doping
- Heterostructures: Lower dimensional systems
- Strained heterostructures

Charge Transport: 4 lectures

- Transport and scattering
- Velocity-Field relations
- Transport by diffusion
- Carrier generation and recombination
- Current continuity equation

P-N Diodes- Steady State: 3 lectures

- P-N junction under equilibrium
- Junction under bias: Ideal case
- Non-Ideal effects
- High voltage effects

Ⓡ Diode applications

Semiconductor Junctions: 3 lectures

- Ⓡ Metal interconnects
- Schottky Barrier Diode
- Ohmic contacts
- Ⓡ Insulator semiconductor junctions
- Semiconductor heterojunctions

Bipolar Junction Transistor- Steady State: 4 lectures

- Current voltage relations
- Device design and optimization
- Secondary effects in BJTs
- Heterojunction bipolar transistors

Bipolar Devices- Temporal Response: 4 lectures

- P-N diode: Small and large signal
- Schottky diode temporal response
- Bipolar transistor: Charge control model
- High frequency response of bipolar transistors
- Technology roadmap and device needs

Field Effects Transistors- MESFET and HFETs: 6 lectures

- JFET and MESFET: Charge control and current-voltage
- Charge control for MODFETs
- Polar HFETs: Charge Control
- HFET Design issues
- Temporal response and high power issues

MOSFET: 6 lectures

- MOS Capacitor
- Current-voltage characteristics
- Sub-threshold current flow
- Short channel and scaling issues in FETs

Mesoscopic Devices: 3 lectures

- Zener-Bloch oscillations
- Resonant tunneling devices
- Quantum interference effects
- Conductance fluctuations and Coulomb Blockade
- Spintronics: Overview

Lecture Plan for a two-quarter sequence of 10 weeks each with 3.5 hours of lecture per week

The basis of this lecture plan is the experience gained from teaching graduate students at UCSB. The experience has been that the class size is larger in the first quarter than in the second where a large group of graduate students from many disciplines attend the class to understand important devices at a level higher than their exposure as undergraduates. It is therefore proposed that the first quarter cover p-n junctions, heterojunctions, HBTs, FETs and MOSFETs operating under DC conditions. Here drift diffusion analysis and thermionic emission will be employed to describe current flow. In the next quarter, it is suggested that the Boltzmann transport analysis contained in the Appendix be covered and the basis for the drift-diffusion fomalism explained. Next the methodology for deriving the high frequency properties of devices such as HBTs and FETs along with their equivalent circuits is covered. Lastly, High Electron Mobility Transistors and Gallium Nitride based devices may be covered

Quarter 1

Lecture 1: Shockley-Read-Hall analysis of lifetime (this introduces the concept of lifetime essential for p-n junction analysis)

Lecture 2: P-n junction electrostatics, P-n junction transport (Forward)

Lecture 3: P-n junction transport (Reverse) and Applications

Lecture 4: Schottky barrier electrostatics and current transport

Lecture 5: Graded materials, Quasi-fields and heterojuncions

Lecture 6: HBTs, Generalized Moll-Ross relationship Early effect, Kirk effect(quick description)

Lecture 7: FETs and gradual channel analysis

Lecture 8: High Aspect Ratio design analysis

Lecture 9: MOS Capacitor and MOSFETs

Lecture 10: Non-ideal effects

Quarter 2

Lecture 1 and 2: Boltzmann Transport Equation and consequences (Drift Diffusion Equation derivation, relaxation times)

Lecutre 3: Charge Control Model (Description and application to HBTs)

Lecutre 4: Ramo-Shockley Theorem and the Kirk effect

Lecutre 5: High Frequency properties of HBTs

Lecutre 6: Equivalenbt Circuit derivation of HBTs; Figures of Merit

Lecutre 7: HEMTs; Electrostatics

Lecutre 8: Gallium Nitride; Polarization and device design

Lecutre 9: HEMTs; I-V characteristics, Design Issues

Lecutre 10: HEMTs and FETs; High Frequency Performance

INTRODUCTION

The pace of semiconductor materials and device development has been staggering, and the impact on human society monumental. Leading this advance has been the development of the silicon-based MOSFET device and its continuous high level of integration. Moore's Law (shown in figure .1), which predicts the doubling of device density every 18 months, has been the governing maxim of the industry. Sustaining Moore's Law has required:

- The development of lithography tools to achieve the 45 nm gate length MOSFETs released into production in 2006
- The continuous scaling of silicon wafers to 12 inch diameters (2005) and 15 inch in the future to enable large chip yields per wafer
- Tremendous advances in interconnect technology
- Device innovations to continuously maintain charge control and low gate leakage as the oxide thickness is scaled down along with the gate length

Though most of the chip and dollar volume of the industry has been driven by Si-based CMOS architecture, there have been critical advances made in other semiconductor technologies. The ability to grow epitaxial layers in a controlled fashion, initially by Liquid Phase Expitaxy (LPE) and Vapor Phase Epitaxy (VPE) and currently by Metalorganic Vapor Phase Epitaxy (MOVPE) and Molecular Beam Epitaxy (MBE), has enabled the compound semiconductor industry to mature into a small but critical component of the total space. The impact has been felt in both the electronics and photonics arenas. In the former, development of the Heterojunction Field Effect Transistor (HFET) and the Heterojunction Bipolar Transistor (HBT) has had a large impact on analog and mixed signal applications. In the low noise receiver area, GaAs and InP based HFETs are the preferred technology. The GaAs-based HBT is preferred for power amplifiers in cellular phones. The Si/SiGe HBT is being actively used in mixed signal applications such as A/D converters and in BiCMOS implementations.

In the optical arena, the development of Light Emitting Diodes (LEDs), lasers, and detectors has been profound. LEDs are used in prolific applications such as signage displays and remote controls as well as in communication devices. The advent of GaN-based LEDs has raised the possibility of a revolutionary advance in lighting with the emergence of solid-state lighting.

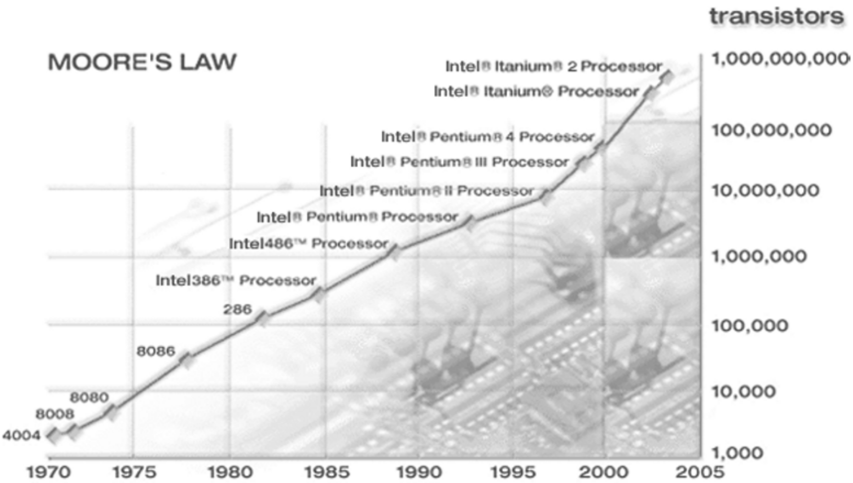


Figure .1: Illustration of Moore's Law.

Lasers and detectors have been the enabling elements in optical communications. Lasers have also enabled entertainment devices such as the DVD.

The continuous expansion of the material and device tool set has enabled system designers to choose the correct technology for the application, resulting in phenomenal advances at the system level. This is best understood by studying a commercial widespread system - the cellular phone. Consider the Motorola V551 phone, illustrated in figure .2. The key components of the transmit/receive chain in any radio architecture are the switch, filter, modulator/demodulator, LNA, mixer, gain blocks, and power amplifier. Integrating the different chips into a total radio solution places varied specifications on the different chips used to achieve the radio solution. In turn, these specifications drive the selection of the active device and process technology that is used to implement the functionality of the particular chip.

As an example of this technology selection process, consider the POLARIS total radio solution from RFMD, which is a highly integrated transciever that performs all functions of the handset radio section, operating under GSM/GPRS/EDGE standards. The POLARIS chipset consists of the following functional blocks, shown in figure .3:

1. The RF 2722 quad-band RF receiver.
2. The RF 3146 POWER STAR PA module with integrated power control.
3. The RF 6001 digital filter, fractional-N PLL, modulator, and power amplifier ramp control.

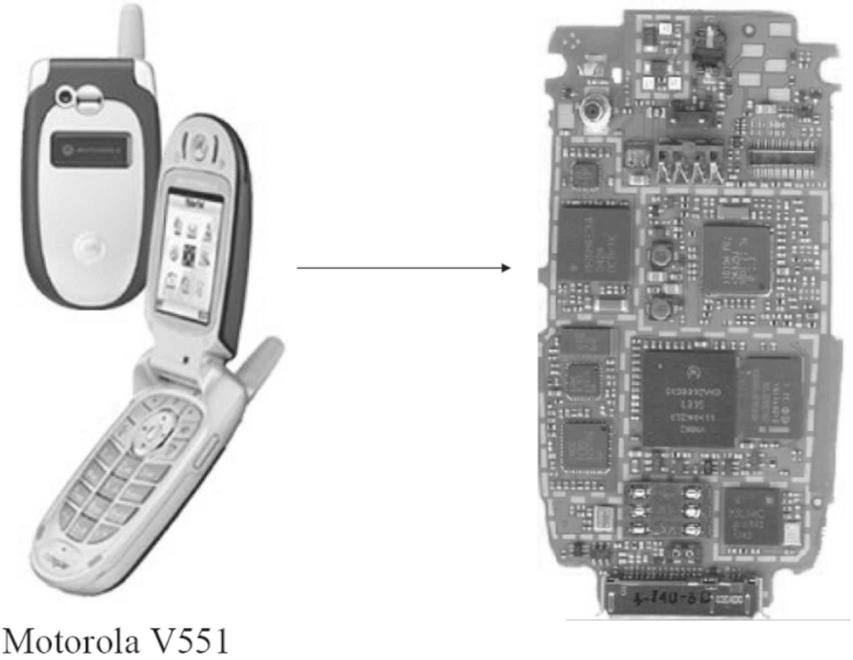


Figure .2: The Motoral V551 cellular phone. Picture courtesy of A. Upton, R. Vetury, and J. Shealy, RFMD.

The RF 2722 fulfills the functional requirement of a quad band LNA and mixer. It includes a VCO and supports very low IF (VLIF) and direct conversion receive (DCR) architectures, thus eliminating the need for IF SAW and RF interstage filters. The complexity of the circuit architecture needed and the noise and linearity requirements placed by the LNA and mixer make the technology of choice SiGe-BiCMOS.

The RF 3146 fulfills the functional requirement of a power amplifier. It includes considerable power control circuitry and can be driven from the DAC output, thus eliminating the need for directional couplers, detector diodes, and other power control circuitry. GaAs HBT technology is chosen for this component in order to achieve the optimum combination of high power, high PAE, and excellent linearity requirements at the frequency of operation.

The RF 6001 fulfills the functional requirements of a synthesizer and signal processor. To achieve the optimum combination of low cost, high levels of integration, and low power consumption, Si CMOS technology is chosen for this component.

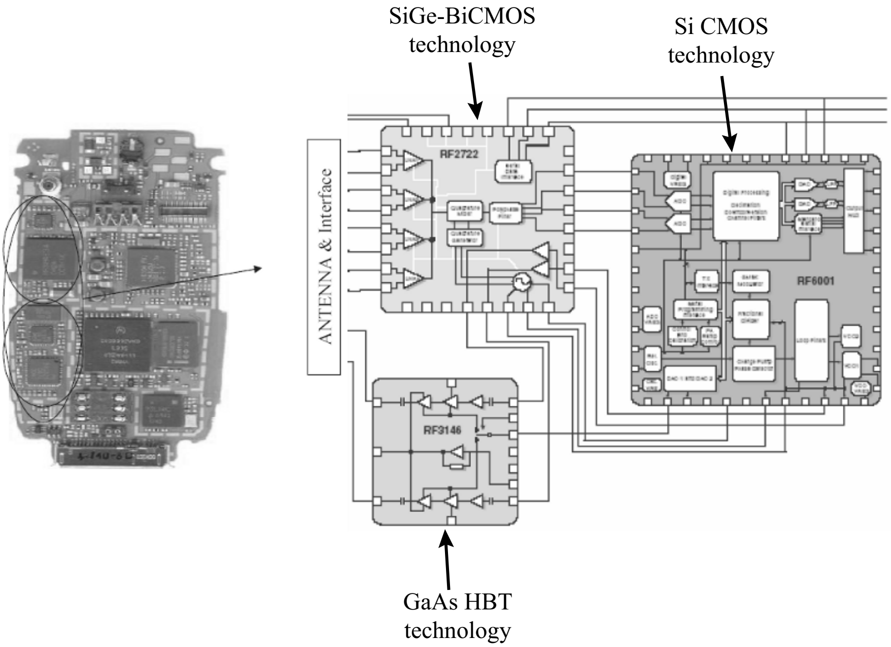


Figure .3: The POLARIS total radio solution from RFMD. Picture courtesy of A. Upton, R. Vetury, and J. Shealy, RFMD.

So what does the future hold for semiconductor based device development? There are brick walls facing the conventional scaling of CMOS circuits. Beyond the year 2012 and the 18 nm node, several of the pathways to continued scaling are not obvious. Also, the power dissipation in the chip threatens to set a thermal limit to the size and the speed of processors in the future. This is best illustrated in figure .4, where it is clear that today's chips seem hotter than a hot plate, and chips of the future in a tongue-in-cheek prediction may rival the sun's surface (obviously impossible). Hence now is the time for all of us to rethink the conventional CMOS scaling paradigm and consider what new pathways may open up. Could compound semiconductors, with their high electron mobilities and velocities, play a role in achieving high clock speeds and reduced power levels? Could large bandgap materials such as Gallium Nitride play a role in applications where the operating temperature is continuously rising? Are there completely new devices such as Carbon Nanotubes (CNTs) which operate in the ballistic regime of electron and hole transport that can emerge as the dark horse in future complementary circuits? Or is molecular electronics, the use of the electronic states of the molecule to achieve computation, the

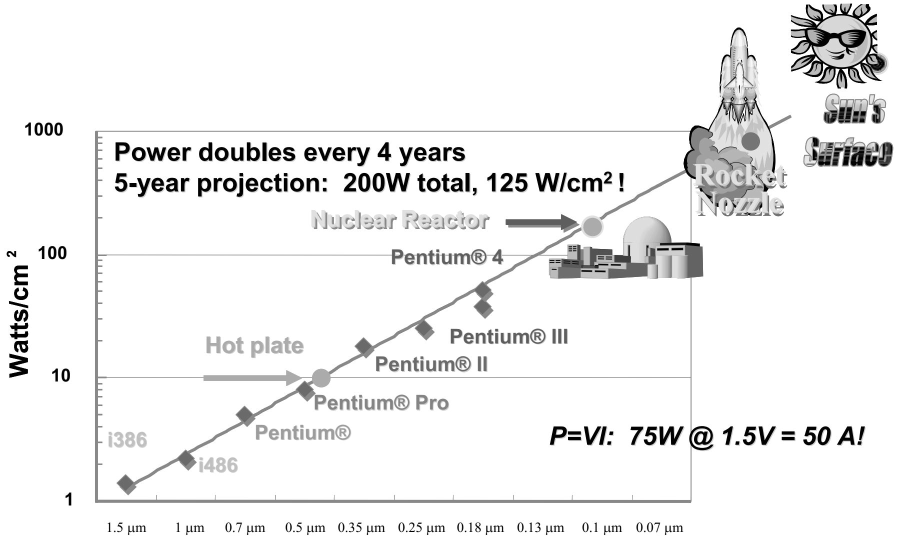


Figure .4: Chip power density is increasing exponentially with time.

answer? Is the control of electron spin rather than the total charge in the channel of the device (the emerging field of spintronics) the holy grail? Are architectures based on single electron transistors a high density, low power alternative?

The future is murky, and we as scientists and engineers have to help clarify it. This book seeks to provide an understanding of the materials, devices, and technology of the various alternatives being considerred, with detail appropriate to the maturity of the technology. A bias towards compound semiconductors is obvious, as Si-based devices have been exclusively addressed over the years in various forms. We hope that this book serves a function to academics teaching course materials, engineers and researchers in the field tackling the murky future, and today's graduate students who will be the great engineers of tomorrow.

Chapter 1

STRUCTURAL PROPERTIES OF SEMICONDUCTORS

1.1 INTRODUCTION

In this text we will be exploring state of the art electronic devices that drive modern information technology. Essentially all of these devices are based on semiconductors. Semiconductor structures have also provided the stages for exploring questions of fundamental physics. As technology advances the number of semiconductors that are used in technology steadily increases. Indeed many innovations have arisen as a result of using new materials and their heterostructures. Thus while Si, GaAs and InP have been most widely used, other materials like InAs, GaN, InN etc. are finding important uses as well. It is important to recognize that the ability to examine fundamental physics issues and to use semiconductors in state of the art device technologies depends critically on the purity and perfection of the semiconductor crystal.

Semiconductor structures can operate at their potential only if they can be grown with a high degree of crystallinity and if impurities and defects can be controlled. For high structural quality it is essential that a high quality substrate be available. This requires growth of bulk crystals which are then sliced and polished to allow epitaxial growth of thin semiconductor regions including heterostructures.

In this chapter we will discuss important semiconductor crystal structures. We also discuss strained lattice structures and the strain tensor for such crystals. Strained epitaxy and its resultant consequences are now widely exploited in semiconductor physics. High speed SiGe devices are based on strained systems as are InGaAs and AlGaN/GaN microwave devices.

We will start with some general properties of crystalline materials and then discuss some specific crystal structures important for semiconductors.

1.2 CRYSTAL STRUCTURE

As noted above high performance semiconductor devices are based on crystalline materials. Crystals are periodic structures made up of identical building blocks. While in "natural" crystals the crystalline symmetry is fixed by nature, new advances in crystal growth techniques are allowing scientists to produce artificial crystals with modified crystalline structure. These advances depend upon being able to place atomic layers with exact precision and control during growth, leading to "low dimensional systems". To define the crystal structure, two important concepts are introduced. The lattice represents a set of points in space forming a periodic structure. The lattice is by itself a mathematical abstraction. A building block of atoms called the basis is then attached to each lattice point yielding the physical crystal structure.

To define a lattice one defines three vectors $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$, such that any lattice point $\mathbf{R}'$ can be obtained from any other lattice point **R** by a translation

$$\mathbf{R}' = \mathbf{R} + m_1\mathbf{a}_1 + m_2\mathbf{a}_2 + m_3\mathbf{a}_3 \tag{1.2.1}$$

where m_1, m_2, m_3 are integers. Such a lattice is called a Bravais lattice . The crystalline structure is now produced by attaching the basis to each of these lattice points.

$$\text{lattice} + \text{basis} = \text{crystal structure} \tag{1.2.2}$$

The translation vectors $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ are called primitive if the volume of the cell formed by them is the smallest possible. There is no unique way to choose the primitive vectors. It is possible to define more than one set of primitive vectors for a given lattice, and often the choice depends upon convenience. The volume cell enclosed by the primitive vectors is called the primitive unit cell .

Because of the periodicity of a lattice, it is useful to define the symmetry of the structure. The symmetry is defined via a set of point group operations which involve a set of operations applied around a point. The operations involve rotation, reflection and inversion. The symmetry plays a very important role in the electronic properties of the crystals. For example, the inversion symmetry is extremely important and many physical properties of semiconductors are tied to the absence of this symmetry. As will be clear later, in the diamond structure (Si, Ge, C, etc.), inversion symmetry is present, while in the Zinc Blende structure (GaAs, AlAs, InAs, etc.), it is absent. Because of this lack of inversion symmetry, these semiconductors are piezoelectric, i.e., when they are strained an electric potential is developed across the opposite faces of the crystal. In crystals with inversion symmetry, where the two faces are identical, this is not possible.

1.2.1 Basic Lattice Types

The various kinds of lattice structures possible in nature are described by the symmetry group that describes their properties. Rotation is one of the important symmetry groups. Lattices can be found which have a rotation symmetry of 2π, $\frac{2\pi}{2}$, $\frac{2\pi}{3}$, $\frac{2\pi}{4}$, $\frac{2\pi}{6}$. The rotation symmetries are denoted by 1, 2, 3, 4, and 6. No other rotation axes exist; e.g., $\frac{2\pi}{5}$ or $\frac{2\pi}{7}$ are not allowed because such a structure could not fill up an infinite space.

There are 14 types of lattices in 3D. These lattice classes are defined by the relationships between the primitive vectors a_1, a_2, and a_3, and the angles α, β, and γ between them. We will focus on the cubic and hexagonal lattices which underly the structure taken by all semiconductors.

There are 3 kinds of cubic lattices: simple cubic, body centered cubic, and face centered cubic.

Simple cubic: The simple cubic lattice shown in figure 1.1is generated by the primitive vectors

$$a\mathbf{x}, a\mathbf{y}, a\mathbf{z} \tag{1.2.3}$$

where the **x**, **y**, **z** are unit vectors.

Body-centered cubic : The bcc lattice shown in figure 1.2 can be generated from the simple cubic structure by placing a lattice point at the center of the cube. If $\hat{\mathbf{x}}, \hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ are three orthogonal unit vectors, then a set of primitive vectors for the body-centered cubic lattice could be

$$a_1 = a\hat{\mathbf{x}}, a_2 = a\hat{\mathbf{y}}, a_3 = \frac{a}{2}(\hat{\mathbf{x}} + \hat{\mathbf{y}} + \hat{\mathbf{z}}) \tag{1.2.4}$$

A more symmetric set for the bcc lattice is

$$a_1 = \frac{a}{2}(\hat{\mathbf{y}} + \hat{\mathbf{z}} - \hat{\mathbf{x}}), a_2 = \frac{a}{2}(\hat{\mathbf{z}} + \hat{\mathbf{x}} - \hat{\mathbf{y}}), a_3 = \frac{a}{2}(\hat{\mathbf{x}} + \hat{\mathbf{y}} - \hat{\mathbf{z}}) \tag{1.2.5}$$

Face Centered Cubic: Another equally important lattice for semiconductors is the face-centered cubic (fcc) Bravais lattice shown in figure 1.3. To construct the face-centered cubic Bravais lattice add to the simple cubic lattice an additional point in the center of each square face. This form of packing is called close-packed.

A symmetric set of primitive vectors for the face-centered cubic lattice (see figure 1.3) is

$$a_1 = \frac{a}{2}(\hat{y} + \hat{z}), a_2 = \frac{a}{2}(\hat{z} + \hat{x}), a_3 = \frac{a}{2}(\hat{x} + \hat{y}) \tag{1.2.6}$$

The face-centered cubic and body-centered cubic Bravais lattices are of great importance, since an enormous variety of solids crystallize in these forms with an atom (or ion) at each lattice site. Essentially all semiconductors of interest for electronics and optoelectronics have a close-packed structure, either fcc or Hexagonal Close Pack(HCP) as discussed below.

1.2.2 Basic Crystal Structures

Diamond and Zinc Blende Structures

Most semiconductors of interest for electronics and optoelectronics have an underlying fcc lattice, with two atoms per basis. The coordinates of the two basis atoms are

$$(000) \text{ and } (\frac{a}{4}, \frac{a}{4}, \frac{a}{4}) \tag{1.2.7}$$

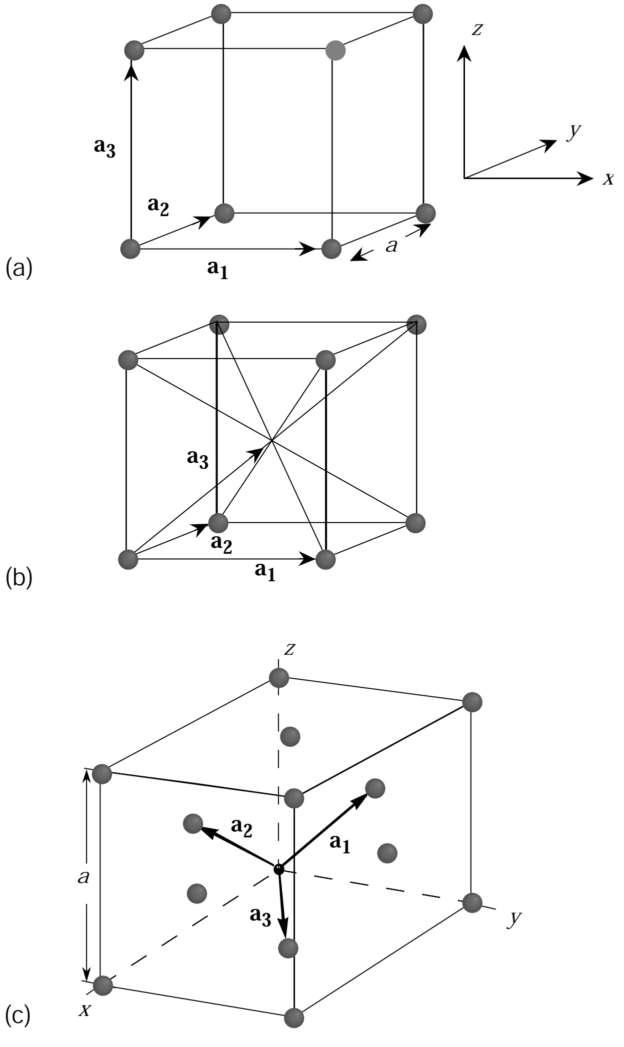
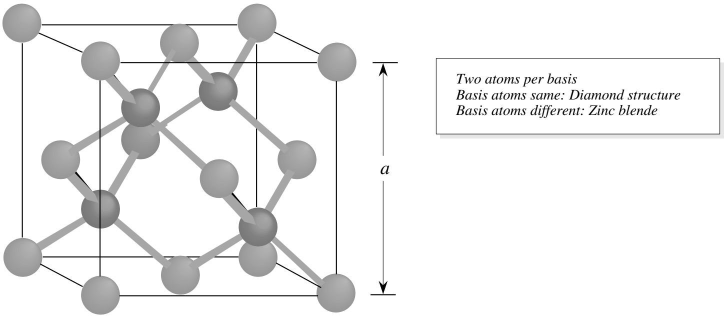


Figure 1.1: A simple cubic lattice showing the primitive vectors. The crystal is produced by repeating the cubic cell through space.



Two atoms per basis
Basis atoms same: Diamond structure
Basis atoms different: Zinc blende

Figure 1.2: The body centered cubic lattice along with a choice of primitive vectors.

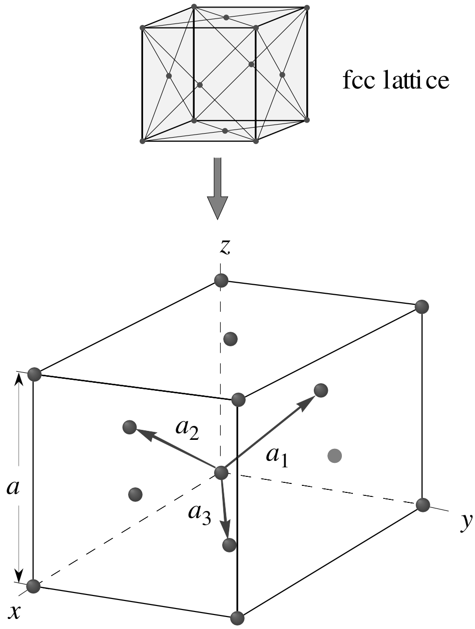


Figure 1.3: Primitive basis vectors for the face centered cubic lattice.

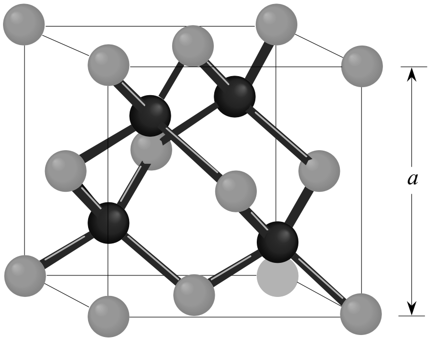


Figure 1.4: The zinc blende crystal structure. The structure consists of the interpenetrating fcc lattices, one displaced from the other by a distance $(\frac{a}{4}\frac{a}{4}\frac{a}{4})$ along the body diagonal. The underlying Bravais lattice is fcc with a two atom basis. The positions of the two atoms is (000) and $(\frac{a}{4}\frac{a}{4}\frac{a}{4})$.

Since each atom lies on its own fcc lattice, such a two atom basis structure may be thought of as two inter-penetrating fcc lattices, one displaced from the other by a translation along the body diagonal direction $(\frac{a}{4}\frac{a}{4}\frac{a}{4})$.

Figure 1.4 shows this important structure. If the two atoms of the basis are identical, the structure is called diamond. Semiconductors such as Si, Ge, C, etc., fall in this category. If the two atoms are different, the structure is called the Zinc Blende structure. Semiconductors such as GaAs, AlAs, CdS, etc., fall in this category. Semiconductors with diamond structure are often called elemental semiconductors, while the Zinc Blende semiconductors are called compound semiconductors. The compound semiconductors are also denoted by the position of the atoms in the periodic chart, e.g., GaAs, AlAs, InP are called III-V (three-five) semiconductors while CdS, HgTe, CdTe, etc., are called II-VI (two-six) semiconductors.

Hexagonal Close Pack Structure The hexagonal close pack (hcp) structure is an important lattice structure and many semiconductors such as BN, AlN, GaN, SiC, etc., also have this underlying lattice (with a two-atom basis). The hcp structure is formed as shown in figure 1.5a. Imagine that a close-packed layer of spheres is formed. Each sphere touches six other spheres, leaving cavities, as shown in figure 1.5. A second close-packed layer of spheres is placed on top of the first one so that the second layer sphere centers are in the cavities formed by the first layer. The third layer of close-packed spheres can now be placed so that center of the spheres do not fall on the center of the starting spheres (left side of figure 1.5a) or coincide with the centers of the starting spheres (right side of figure 1.5). These two sequences, when repeated, produce the fcc and hcp lattices.

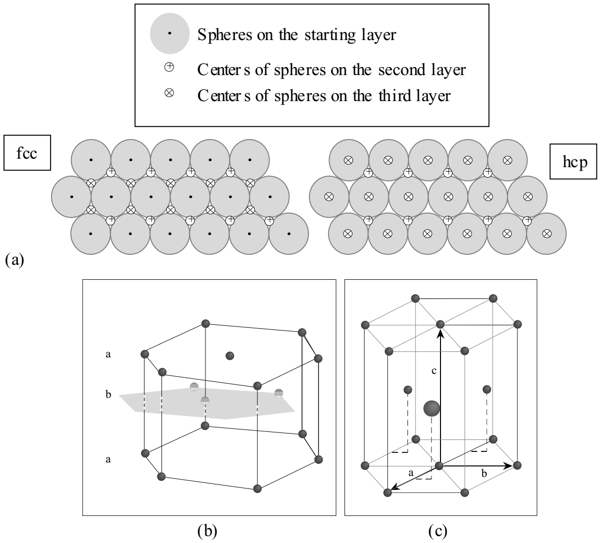

Spheres on the starting layer
Centers of spheres on the second layer
Centers of spheres on the third layer
fcc
hcp
(a)
a
b
a
c
a
b
(b)
(c)

Figure 1.5: (a) A schematic of how the fcc and hcp lattices are formed by close packing of spheres. (b) The hcp structure is produced by two interpenetrating hexagonal lattices with a displacement discussed in the text. (c) Arrangement of lattice points on an hcp lattice.

In figure 1.5b and figure 1.5c we show the detailed positions of the lattice points in the hcp lattice. The three lattice vectors $a_1,\ a_2,\ a_3$ are shown as $a,\ b,\ c$. The vector a_3 is denoted by c and the term c-axis refers to the orientation of a_3. The hexagonal planes are displaced from each other by $a_1/3 + a_2/3 + a_3/2$. In an ideal structure, if $\mid a \mid = \mid a_1 \mid = \mid a_2 \mid$,

$$\frac{c}{a} = \sqrt{\frac{8}{3}} \tag{1.2.8}$$

In table 1.1 we show the structural and some important electronic properties of some important semiconductors. Note that two or more semiconductors are randomly mixed to produce an alloy, A_xB_{1-x}, the lattice constant of the alloy is given by Vegard's law according to which the alloy lattice constant is the weighted mean of the lattice constants of the individual components

$$a_{alloy} = xa_A + (1-x)a_B \tag{1.2.9}$$

1.2.3 Notation to Denote Planes and Points in a Lattice: Miller Indices

To represent the directions and planes in a crystalline structure an agreed upon scheme is used. The planes or directions are denoted by a series of integers called the Miller indices.

Zinc Blende and Wurtzite

MATERIAL	CRYSTAL STRUCTURE	BANDGAP (EV)	STATIC DIELECTRIC CONSTANT	LATTICE CONSTANT (¯)	DENSITY (gm-cm^{3})
C	DI	5.50, I	5.570	3.56683	3.51525
Si	DI	1.1242, I	11.9	5.431073	2.329002
SiC	ZB	2.416, I	9.72	4.3596	3.166
Ge	DI	0.664, I	16.2	5.6579060	5.3234
BN	HEX	5.2, I	$\varepsilon\| = 5.06$	a = 6.6612	2.18
			$\varepsilon^{\perp} = 6.85$	c = 2.5040	
BN	ZB	6.4, I	7.1	3.6157	3.4870
BP	ZB	2.4, I	11.	4.5383	2.97
BAs	ZB			4.777	5.22
AlN	W	6.2,D	$\bar{\varepsilon} = 9.14$	a = 3.111	3.255
				c = 4.981	
AlP	ZB	2.45,I	9.8	5.4635	2.401
AlAs	ZB	2.153,I	10.06	5.660	3.760
AlSb	ZB	1.615,I	12.04	6.1355	4.26
GaN	W	3.44,D	$\varepsilon\|=10.4$	a = 3.175	6.095
			$\varepsilon_{\perp}= 9.5$	c = 5.158	
GaP	ZB	2.272,I	11.11	5.4505	4.138
GaAs	ZB	1.4241,D	13.18	5.65325	5.3176
GaSb	ZB	0.75,D	15.69	6.09593	5.6137
InN	W	1.89,D		a = 3.5446	6.81
				c = 8.7034	
InP	ZB	1.344,D	12.56	5.8687	4.81
InAs	ZB	0.354,D	15.15	6.0583	5.667
InSb	ZB	0.230,D	16.8	6.47937	5.7747
ZnO	W	3.44,D	$\varepsilon\|= 8.75$	a = 3.253	5.67526
			$\varepsilon_{\perp}= 7.8$	c = 5.213	
ZnS	ZB	3.68,D	8.9	5.4102	4.079
ZnS	W	3.9107,D	$\bar{\varepsilon} = 9.6$	a = 3.8226	4.084
				c = 6.2605	
ZnSe	ZB	2.8215,D	9.1	5.6676	5.266
ZnTe	ZB	2.3941,D	8.7	6.1037	5.636
CdO	R	0.84,I	21.9	4.689	8.15
CdS	W	2.501,D	$\bar{\varepsilon} = 9.83$	a = 4.1362	4.82
				c = 6.714	
CdS	ZB	2.50,D		5.818	
CdSe	W	1.751,D	$\varepsilon\|=10.16$	a = 4.2999	5.81
			$\varepsilon_{\perp}= 9.29$	c = 7.0109	
CdSe	ZB			6.052	
CdTe	ZB	1.475,D	10.2	6.482	5.87
PbS	R	0.41,D*	169.	5.936	7.597
PbSe	R	0.278,D*	210.	6.117	8.26
PbTe	R	0.310,D*	414.	6.462	8.219

Data are given at room temperature values (300 K).
Key: DI: diamond; HEX: hexagonal; R: rocksalt; W: wurtzite; ZB: zinc blende;
*: gap at L point; D: direct; I: indirect $\varepsilon\|$: parallel to c-axis; $\varepsilon_{\perp}$: perpendicular to c-axis

Table 1.1: Structure, lattice constant, and density of some materials at room temperature

These indicies are obtained using the following:

(1) Define the x, y, z axes (primitive vectors).
(2) Take the intercepts of the plane along the axes in units of lattice constants.
(3) Take the reciprocal of the intercepts and reduce them to the smallest integers.
The notation (hkl) denotes a family of parallel planes.
The notation (hkl) denotes a family of equivalent planes.

To denote directions, we use the smallest set of integers having the same ratio as the direction cosines of the direction.

In a cubic system the Miller indices of a plane are the same as the direction perpendicular to the plane. The notation [ ] is for a set of parallel directions; $< \ >$ is for a set of equivalent direction. Figure 1.6 shows some examples of the use of the Miller indices to define planes for a cubic system.

Example 1.1 The lattice constant of silicon is 5.43 Å. Calculate the number of silicon atoms in a Si MOSFET with dimensions of $50\mu m \times 2\mu m \times 1\mu m.$

Silicon has a diamond structure which is made up of the fcc lattice with two atoms on each lattice point. The fcc unit cube has a volume a^3. The cube has eight lattice sites at the cube edges. However, each of these points is shared with eight other cubes. In addition, there are six lattice points on the cube face centers. Each of these points is shared by two adjacent cubes. Thus the number of lattice points per cube of volume a^3 are

$$N(a^3) = \frac{8}{8} + \frac{6}{2} = 4$$

In silicon there are two silicon atoms per lattice point. The number density is, therefore,

$$N_{Si} = \frac{4 \times 2}{a^3} = \frac{4 \times 2}{(5.43 \times 10^{-8})^3} = 4.997 \times 10^{22} \text{ atoms/cm}^3$$

The number in the MOSFET are

$$N_{MOSFET} = 4.997 \times 10^{12} \text{ atoms}$$

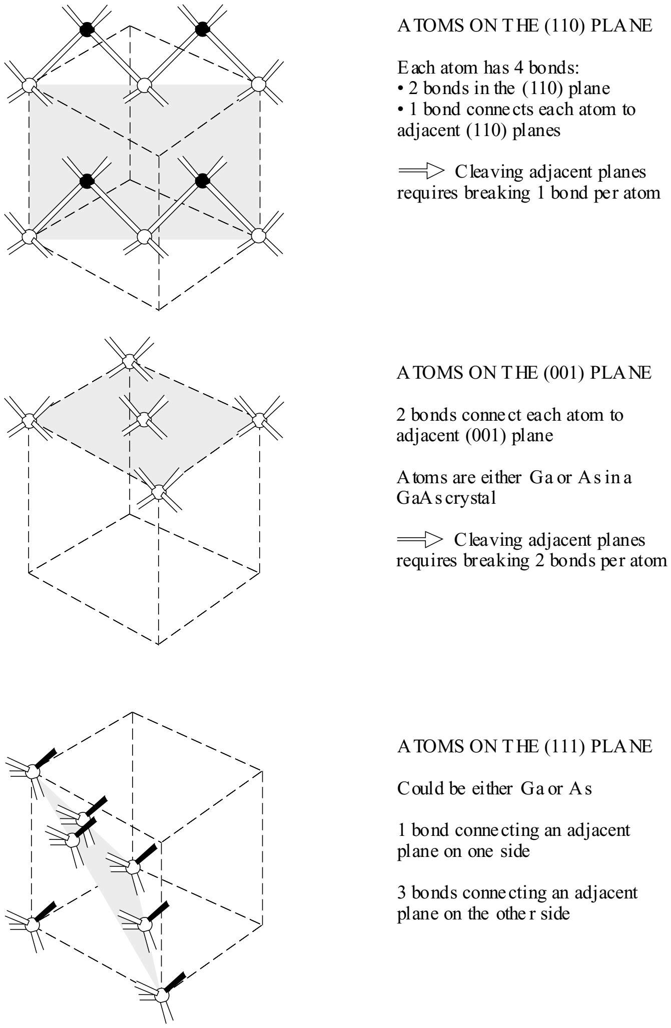
Example 1.2 Calculate the surface density of Ga atoms on a Ga terminated (001) GaAs surface.

In the (001) surfaces, the top atoms are either Ga or As leading to the terminology Ga terminated (or Ga stabilized) and As terminated (or As stabilized), respectively. A square of area a^2 has four atoms on the edges of the square and one atom at the center of the square. The atoms on the square edges are shared by a total of four squares. The total number of atoms per square is

$$N(a^2) = \frac{4}{4} + 1 = 2$$

The surface density is then

$$N_{Ga} = \frac{2}{a^2} = \frac{2}{(5.65 \times 10^{-8})^2} = 6.26 \times 10^{14} \text{ cm}^{-2}$$



ATOMS ON THE (110) PLANE

Each atom has 4 bonds:
- 2 bonds in the (110) plane
- 1 bond connects each atom to adjacent (110) planes

⇨ Cleaving adjacent planes requires breaking 1 bond per atom

ATOMS ON THE (001) PLANE

2 bonds connect each atom to adjacent (001) plane

Atoms are either Ga or As in a GaAs crystal

⇨ Cleaving adjacent planes requires breaking 2 bonds per atom

ATOMS ON THE (111) PLANE

Could be either Ga or As

1 bond connecting an adjacent plane on one side

3 bonds connecting an adjacent plane on the other side

Figure 1.6: Some important planes in the cubic system along with their Miller indices. This figure also shows how many bonds connect adjacent planes. This number determines how easy or difficult it is to cleave the crystal along these planes.

1.2.4 Artificial Structures: Superlattices and Quantum Wells

Epitaxial crystal growth techniques such as molecular beam epitaxy (MBE) and metal organic chemical vapor deposition (MOCVD) allow one to have monolayer ($\sim$3 Å) control in the chemical composition of the growing crystal. Nearly every semiconductor extending from zero bandgap (α-Sn,HgCdTe) to large bandgap materials such as ZnSe,CdS,AlN etc., has been grown by epitaxial techniques. This allows growth of quantum wells and heterostructures where electronic properties can be altered. Heteroepitaxial techniques allow one to grow heterostructures with atomic control, one can change the periodicity of the crystal in the growth direction. This leads to the concept of superlattices where two (or more) semiconductors A and B are grown alternately with thicknesses d_A and d_B respectively. The periodicity of the lattice in the growth direction is then $d_A + d_B$. A (GaAs)/(AlAs) heterostructure is illustrated in figure 1.7.

In figure 1.8, we show a cross-sectional TEM image of a structure containing InGaAs/GaAs and AlGaAs/GaAs superlattices, indicating the precision with which these structures can be produced using modern epitaxial growth techniques.

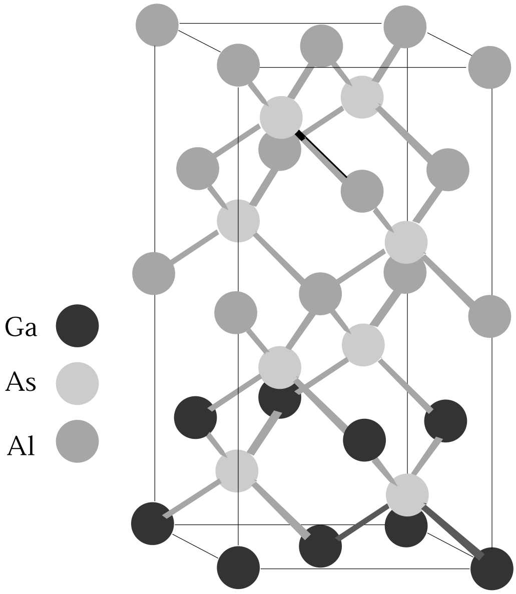


Figure 1.7: Arrangement of atoms in a (GaAs)/(AlAs) heterostructure grown along (001) direction.

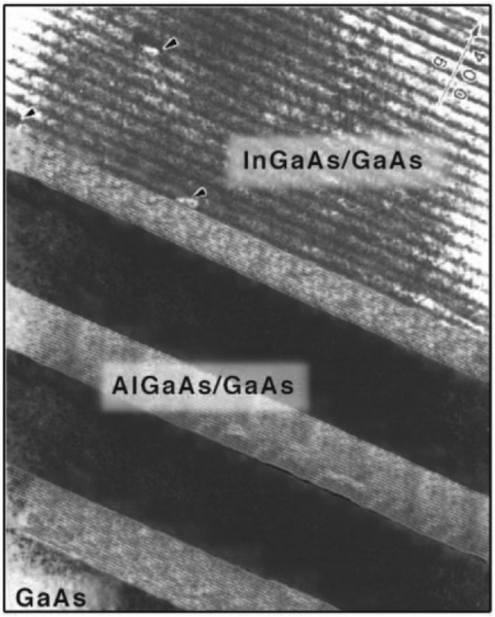


Figure 1.8: (b) This transmission electron microscope picture shows the precision with which semiconductor compositions can be altered by epitaxial growth techniques. Individual semiconductor layers as thin as 10 Å can be produced.

1.2.5 Surfaces : Ideal Versus Real

The arrangement of atoms on the surface can be quite different from that in the bulk. The bulk crystal structure is decided by the internal chemical energy of the atoms forming the crystal with a certain number of nearest neighbors, second nearest neighbors, etc. Since the surface, the number of neighbors is suddenly altered, the spatial geometries which were providing the lowest energy configuration in the bulk may not provide the lowest energy configuration at the surface. Thus, there is a readjustment or “reconstruction” of the surface bonds toward an energy minimizing configuration.

An example of such a reconstruction is shown for the GaAs surface in figure 1.9. The figure (a) shows an ideal (001) surface where the topmost atoms form a square lattice. The surface atoms have two nearest neighbor bonds (Ga-As) with the layer below, four second neighbor bonds (e.g., Ga-Ga or As-As) with the next lower layer, and four second neighbor bonds within the same layer. We could denote the ideal surface by the symbol C(1×1), representing the fact that the surface periodicity is one unit by one unit along the square lattice along [110] and [$\bar{1}$10].

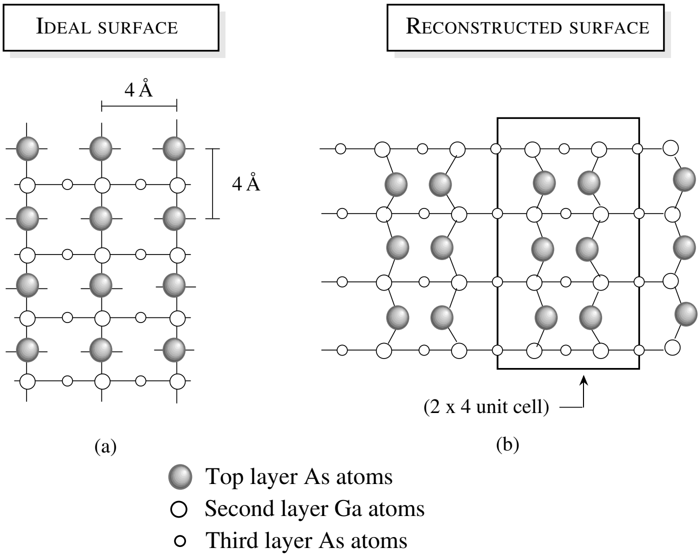


Figure 1.9: The structure (a) of the unreconstructed GaAs (001) arsenic-rich surface. The missing dimer model (b) for the GaAs (001) (2×4) surface. The As dimers are missing to create a 4 unit periodicity along one direction and a two unit periodicity along the perpendicular direction.

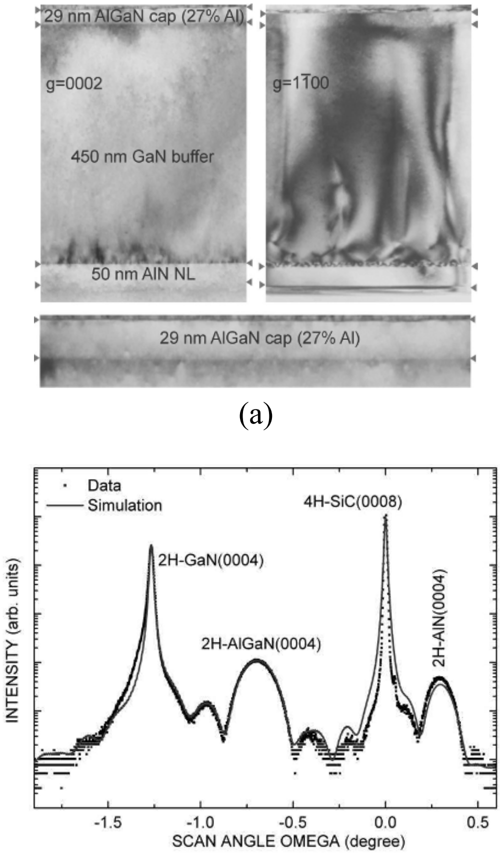
The reconstructed surfaces that occur in nature are generally classified as C(2×8) or C(2×4) etc., representing the increased periodicity along the $[\bar{1}10]$ and [110] respectively. The C(2×4) case is shown schematically in figure 1.9, for an arsenic stabilized surface (i.e., the top monolayer is As). The As atoms on the surface form dimers (along $[\bar{1}10]$ on the surface to strengthen their bonds. In addition, rows of missing dimers cause a longer range ordering as shown to increase the periodicity along the [110] direction to cause a C(2×4) unit cell. Similar reconstruction occurs for Si surfaces as well.

Example 1.1 Calculate the planar density of atoms on the (111) surface of GaAs.

As can be seen from figure 1.6, we can form a triangle on the (111) surface. There are three atoms on the tips of the triangle. These atoms are shared by six other similar triangles. There are also 3 atoms along the edges of the triangle which are shared by two adjacent triangles. Thus the number of atoms in the triangle are

$$\frac{3}{6} + \frac{3}{2} = 2$$

The area of the triangle is $\sqrt{3}a^2/2$. The density of GaAs atoms on the surface is then 7.29×10^{14} cm^{-2}.



(a)

(b)

Figure 1.10: (a) Cross-sectional TEM images of a typical AlGaN/GaN HFET structure grown on a SiC substrate. First, a 50 nm AlN nucleation layer is grown, followed by a 450 nm GaN buffer layer and a 29 nm AlGaN cap. A large number of defects are formed at the AlN/GaN interface, but many of the defects are annihilated as the GaN layer is thickened. The AlGaN cap layer is coherently strained on top of the bulk GaN. No new defects are formed at this interface. (b) High resolution X-ray diffraction scan of this structure. The close match between the data and theory indicates the high crystalline quality of the structure. Images courtesy of Prof. J. Speck, UCSB.

1.2.6 Interfaces

Like surfaces, interfaces are an integral part of semiconductor devices. We have already discussed the concept of heterostructures and superlattices which involve interfaces between two semiconductors. These interfaces are usually of high quality with essentially no broken bonds (see figure 1.10), except for dislocations in strained structures (to be discussed later). There is, nevertheless, an interface roughness of one or two monolayers which is produced because of either non-ideal growth conditions or imprecise shutter control in the switching of the semiconductor species. The general picture of such a rough interface is as shown in figure 1.11a for epitaxially grown interfaces. The crystallinity and periodicity in the underlying lattice is maintained, but the chemical species have some disorder on interfacial planes. Such a disorder can be quite important in many electronic devices. In figure 1.11b we show a TEM for a GaAs/AlAs interface.

One of the most important interfaces in electronics is the Si/SiO$_2$ interface. This interface and its quality is responsible for essentially all of the modern consumer electronic revolution. This interface represents a situation where two materials with very different lattice constants and crystal structures are brought together. However, in spite of these large differences the interface quality is quite good. In figure 1.12 we show a TEM cross-section of a Si/SiO$_2$ interface. It appears that the interface has a region of a few monolayers of amorphous or disordered Si/SiO$_2$ region creating fluctuations in the chemical species (and consequently in potential energy) across the interface. This interface roughness is responsible for reducing mobility of electrons and holes in MOS devices. It can also lead to “trap” states, which can seriously deteriorate device performance if the interface quality is poor.

Finally, we have the interfaces formed between metals and semiconductors. Structurally, these important interfaces are hardest to characterize and are usually produced in presence of high temperatures. Metal-semiconductor interfaces involve diffusion of metal elements along with complex chemical reactions.

1.2.7 Semiconductor Defects

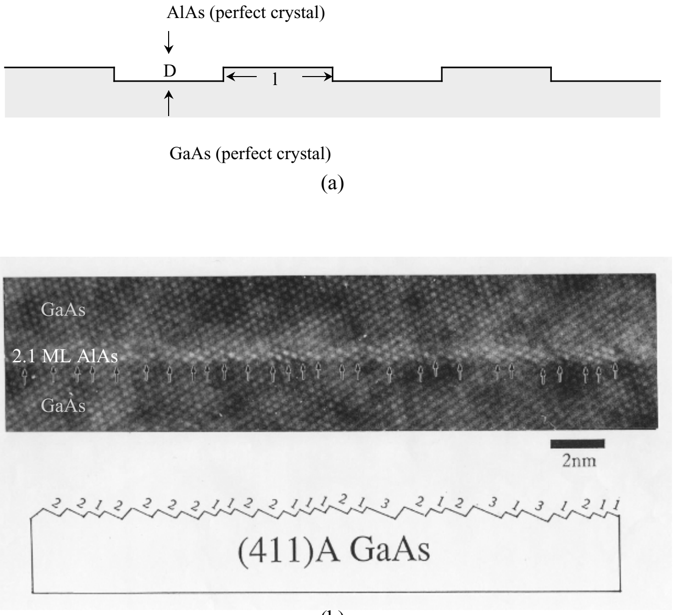
Semiconductor devices have both unintended and intentional defects. Some unintentional defects are introduced due to either thermodynamic considerations or the presence of impurities during the crystal growth process. In general, defects in crystalline semiconductors can be characterized as i) point defects; ii) line defects; iii) planar defects and iv) volume defects. These defects are detrimental to the performance of electronic and optoelectronic devices and are to be avoided as much as possible.

Localized Defects

A localized defect affects the periodicity of the crystal only in one or a few unit cells. There are a variety of point defects, as shown in figure 1.13. Defects are present in any crystal and their concentration is given roughly by the thermodynamics relation

$$\frac{N_d}{N_{Tot}} = k_d \exp\left(-\frac{E_d}{k_B T}\right) \tag{1.2.10}$$

where N_d is the vacancy density, N_{Tot} the total site density in the crystal, E_d the defect formation



(a)

(b)

Figure 1.11: (a) A schematic picture of the interfaces between materials with similar lattice constants such as GaAs/AlAs. No loss of crystalline lattice and long range order is suffered in such interfaces. The interface is characterized by islands of height Δ and lateral extent λ. (b) High resolution cross-sectional TEM image along with schematic diagram of (411A) GaAs with a very thin (2.1 monolayer) AlAs layer in the middle. A small amount of roughness can be observed at the interface. TEM image courtesy of S. Shimomura and S. Hiyamizu of Osaka University.

energy, k_d is a dimensionless parameter with values ranging from 1 to 10 in semiconductors, and T, the crystal growth temperature. Defect formation energy is in the range of an eV for most semiconductors.

Dislocations

In contrast to point defects, line defects (called dislocations) involve a large number of atomic sites that can be connected by a line. Dislocations are produced if, for example, an extra half plane of atoms are inserted (or taken out) of the crystal as shown in figure 1.14. Such dislocations are called edge dislocations. Dislocations can also be created if there is a slip in the crystal so that part of the crystal bonds are broken and reconnected with atoms after the slip. In the nitride technology where alternate substrates are used, dislocation densities can be quite large.

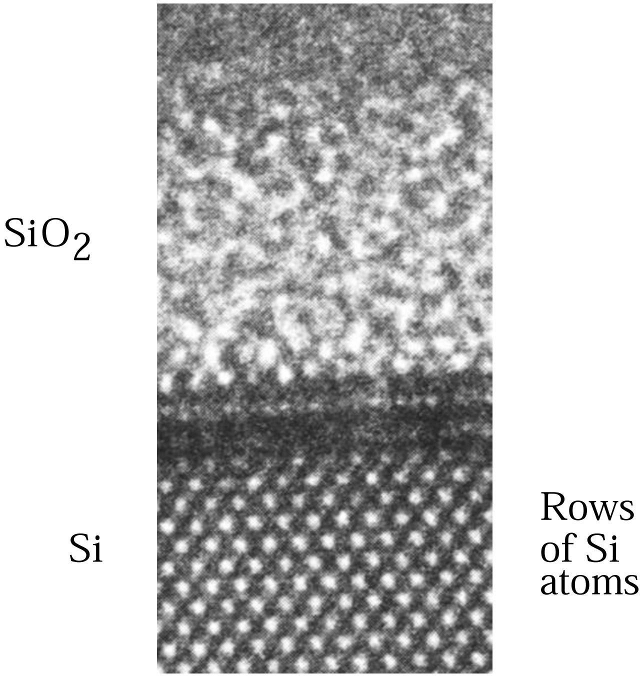


Figure 1.12: The tremendous success of Si technology is due to the Si/SiO_2 interface. In spite of the very different crystal structure of Si and SiO_2, the interface is extremely sharp, as shown in the TEM picture in this figure. TEM image courtesy of Bell Labs.

1.3 LATTICE MISMATCHED STRUCTURES

It is relatively easy to grow heterostructures where the overlayer lattice constant is the same or similar to that of the substrate. In such lattice matched epitaxy the interface quality can be very high with essentially negligible interface defects and atomically abrupt interface. However one often needs structures where there is lattice mismatch between the overlayer and the substrate. The motivation for lattice mismatched epitaxy is two fold:

i) Incorporation of built-in strain: When a lattice mismatched semiconductor is grown on a substrate and the thickness of the overlayer is very thin, the overlayer has a built-in strain. This built-in strain has important effects on the electronic and optoelectronic properties of the material and can be exploited for high performance devices. It can be exploited in nitride heterostructures to effectively dope structures. It can also be exploited in Si/SiGe systems.

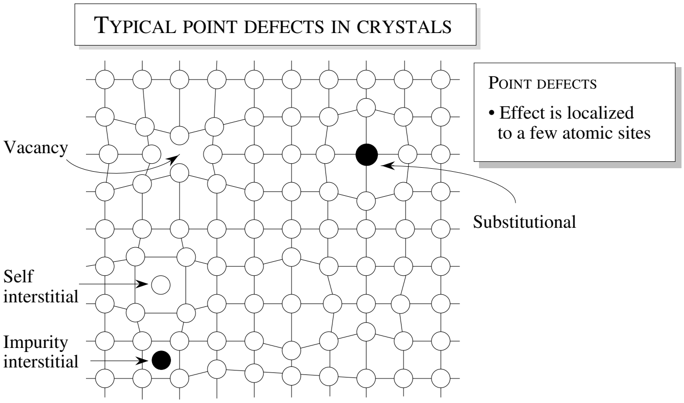

TYPICAL POINT DEFECTS IN CRYSTALS
POINT DEFECTS
• Effect is localized to a few atomic sites
Vacancy
Substitutional
Self interstitial
Impurity interstitial

Figure 1.13: A schematic showing some important point defects in a crystal.

ii) New effective substrate: High quality substrates are only available for Si, GaAs and InP (sapphire, SiC and quartz substrates are also available and used for some applications). Since most semiconductors are not lattice-matched to these substrates a solution that has emerged is to grow a thick overlayer on a mismatched substrate. If the conditions are right, dislocations are generated and eventually the overlayer forms its own substrate. This process allows a tremendous flexibility in semiconductor technology. Not only can it, in principle, resolve the substrate availability problem, it also allows the possibility of growing GaAs on Si, CdTe on GaAs, GaN on SiC etc. Thus different semiconductor technologies can be integrated on the same wafer.

In figure 1.15 we show a TEM image of an InP/InAs double-barrier resonant tunneling device (DBRT). The InP barriers are 5 nm wide, enclosing a 15 nm InAs quantum dot. The InP is coherently strained, with no dislocations created at the interfaces. The sharpness of the interfaces was determined to be 1-3 lattice spacings.

Coherent and Incoherent Structures

Consider situation shown schematically in figure 1.16 where an overlayer with lattice constant a_L is grown on a substrate with lattice constant a_S. The strain between the two materials is defined as

$$\epsilon = \frac{a_S - a_L}{a_L} \tag{1.3.1}$$

If the lattice constant of the overlayer is maintained to be a_L, it is easy to see that after every $1/\epsilon$ bonds between the overlayer and the substrate, either a bond is missing or an extra bond appears as shown in figure 1.16b. In fact, there would be a row of missing or extra bonds since we have a 2-dimensional plane. These defects are the dislocations discussed earlier.

An alternative to the incoherent case is shown in figure 1.16c. Here all the atoms at the interface of the substrate and the overlayer are properly bonded by adjusting the in-plane lattice

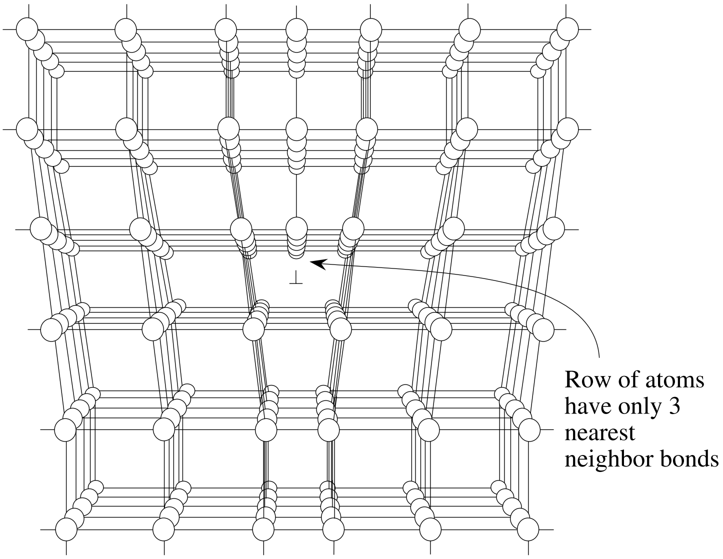


Figure 1.14: A schematic showing the presence of a dislocation. This line defect is produced by adding an extra half plane of atoms.

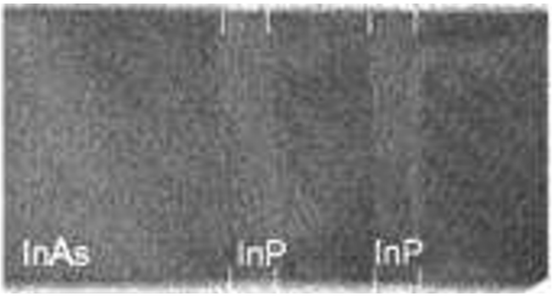


Figure 1.15: TEM image of an InP/InAs double-barrier resonant tunneling device (DBRT) consisting of 5 nm InP barriers surrounding a 15 nm InAs quantum dot. The InP is coherently strained, with no dislocations created at the interfaces. Image courtesy of M. Bjork, Lund University.

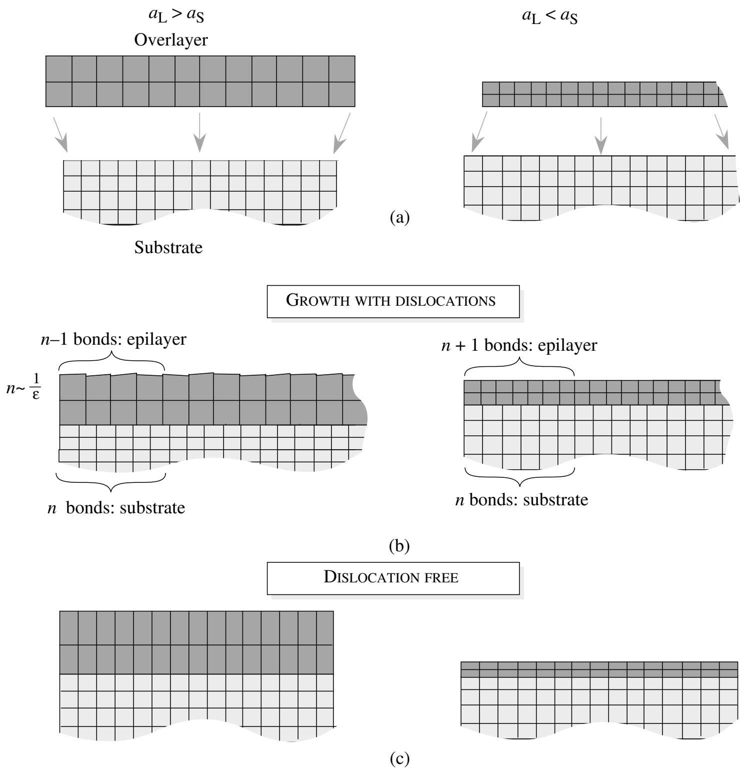


Figure 1.16: (a) An overlayer with one lattice constant is placed without distortion on a substrate with a different lattice constant. (b) Dislocations are generated at positions where the interface bonding is lost. (c) The case is shown where the overlayer is distorted so that no dislocation is free and coherent with the substrate.

constant of the overlayer to that of the substrate. This causes the overlayer to be under strain and the system has a certain amount of strain energy. This strain energy grows as the overlayer thickness increases. In the strained epitaxy, the choice between the state of the structure shown in figure 1.16b and the state shown in figure 1.16c is decided by free energy minimization considerations. The general observations can be summarized as follows:

For small lattice mismatch ($\epsilon < 0.03$), the overlayer initially grows in perfect registry with the substrate, as shown in figure 1.16c. However, as noted before, the strain energy will grow as the overlayer thickness increases. As a result, it will eventually be favorable for the overlayer to generate dislocations. In simplistic theories this occurs at an overlayer thickness called the critical thickness , d_c, which is approximately given by

$$d_c \cong \frac{a_S}{2|\epsilon|} \tag{1.3.2}$$

where a_S is the lattice constant of the substrate and ϵ the lattice mismatch. In reality, the point in growth where dislocations are generated is not so clear cut and depends upon growth conditions, surface conditions, dislocation kinetics, etc. However, one may use the criteria given by equation 1.3.2 for approximately characterizing two regions of overlayer thickness for a given lattice mismatch. Below critical thickness, the overlayer grows without dislocations and the film is under strain. Under ideal conditions above critical thickness, the film has a dislocation array, and after the dislocation arrays are generated, the overlayer grows without strain with its free lattice constant.

If the strain value is greater than 0.03 one can still have strained epitaxy but the growth occurs in the island mode where islands of the over-layer are formed. Such self-assembled islands are being used for quantum dot structures.

Epitaxy beyond the critical thickness is important to provide new effective substrates for new material growth. For these applications the key issues center around ensuring that the dislocations generated stay near the overlayer-substrate interface and do not propagate into the overlayer as shown in figure 1.17. A great deal of work has been done to study this problem. Often thin superlattices in which the individual layers have alternate signs of strain are grown to “trap” or “bend” the dislocations. It is also useful to build the strain up gradually.

In recent years, the GaN material system has seen much progress in electronic and optoelectronic applications. Since GaN substrates are still not readily available, it is typically grown on Al_2O_3 (sapphire) or SiC , neither of which are closely lattice matched to GaN. The resulting material is therefore highly dislocated. Many of the dislocations propagate upwards and are terminated at the surface. In figure 1.18a, we show a cross-sectional transmission electron microscope image of GaN grown on sapphire. The vertical lines propagating upwards from the highly defective interface are dislocations. Figure 1.18b is an atomic force microscope (AFM) image of the GaN surface. The black pits are dislocations that have propagated upwards. Also evident are the atomic steps that are typical of GaN surfaces. Such surface reconstructions were described in section 1.2.5. Note that these atomic steps are always terminated by a dislocation.

In figure 1.18c, we show an AFM image of the surface of dislocation-free GaN. In contrast to the dislocated material in figure 1.18b, there are no pits visible on the surface, and the surface step structure is smooth and continuous.

1.4 STRAINED EPITAXY: STRAIN TENSOR

Growth of an epitaxial layer whose lattice constant is close, but not equal, to the lattice constant of the substrate can result in a coherent strain. What is the strain tensor in such epitaxy? The strain tensor determines how the electronic properties are altered. If the strain is small one can have monolayer-by-monolayer. In this case the lattice constant of the epitaxial layer in the directions parallel to the interface is forced to be equal to the lattice constant of the substrate.

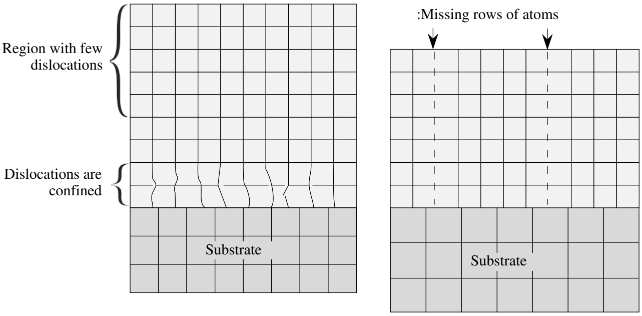


Figure 1.17: Strained epitaxy above critical thickness . The left hand side figure shows a desirable structure in which the dislocations are confined near the overlayer-substrate interface. On the right hand side, the dislocations are penetrating the overlayer.

The lattice constant of the epitaxial perpendicular to the substrate will be changed by the Poisson effect . These two cases are depicted in figure 1.16c. This type of coherently strained crystal is called pseudomorphic .

For layer-by-layer growth, the epitaxial semiconductor layer is biaxially strained in the plane of the substrate, by an amount $\epsilon_{||}$, and uniaxially strained in the perpendicular direction, by an amount $\epsilon_{\perp}$. For a thick substrate, the in-plane strain of the layer is determined from the bulk lattice constants of the substrate material, a_S, and the layer material, a_L:

$$\begin{aligned} e_{||} &= \frac{a_S}{a_L} - 1 \\ &= \epsilon \end{aligned} \tag{1.4.1}$$

Since the layer is subjected to no stress in the perpendicular direction, the perpendicular strain, $\epsilon_{\perp}$, is simply proportional to $\epsilon_{||}$:

$$\epsilon_{\perp} = \frac{-\epsilon_{||}}{\sigma} \tag{1.4.2}$$

where the constant σ is known as Poisson's ratio .

Noting that there is no stress in the direction of growth it can be simply shown that for the strained layer grown on a (001) substrate (for an fcc lattice)

$$\begin{aligned} \sigma &= \frac{c_{11}}{2c_{12}} \qquad (1.4.3) \\ \epsilon_{xx} &= \epsilon_{||} \\ \epsilon_{yy} &= \epsilon_{xx} \\ \epsilon_{zz} &= \frac{-2c_{12}}{c_{11}}\epsilon_{||} \\ \epsilon_{xy} &= 0 \\ \epsilon_{yz} &= 0 \\ \epsilon_{zx} &= 0 \end{aligned}$$

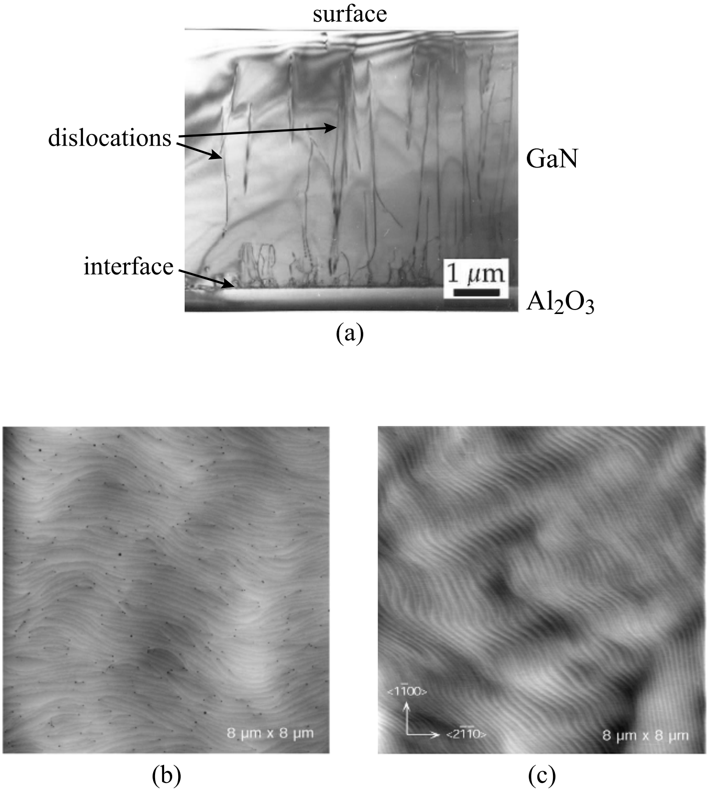


Figure 1.18: (a) Cross-sectional TEM image of GaN grown heteroepitaxially on sapphire, indicating the highly defective interface and the dislocations that propagate upwards. (b) AFM surface image of the dislocated GaN , showing the atomic step structure which is typical of GaN surfaces. The black dots are dislocations that have propagated upwards to the surface. (c) AFM surface image of non-dislocated GaN, exhibiting a smooth and continuous step structure. Images courtesy of P. Fini and H. Marchand of UCSB.

while in the case of strained layer grown on a (111) substrate

$$\begin{aligned}
\sigma &= \frac{c_{11}+2c_{12}+4c_{44}}{2c_{11}+4c_{12}-4c_{44}} \\
\epsilon_{xx} &= \left[\frac{2}{3}-\frac{1}{3}\left(\frac{2c_{11}+4c_{12}-4c_{44}}{c_{11}+2c_{12}+4c_{44}}\right)\right]\epsilon_{\|} \\
\epsilon_{yy} &= \epsilon_{xx} \\
\epsilon_{zz} &= \epsilon_{xx} \\
\epsilon_{xy} &= \left[\frac{-1}{3}-\frac{1}{3}\left(\frac{2c_{11}+4c_{12}-4c_{44}}{c_{11}+2c_{12}+4c_{44}}\right)\right]\epsilon_{\|} \\
\epsilon_{yz} &= \epsilon_{xy} \\
\epsilon_{zx} &= \epsilon_{yz} \qquad (1.4.4)
\end{aligned}$$

In general, the strained epitaxy causes a distortion of the lattice and, depending upon the growth orientation, the distortions produce a new reduced crystal symmetry. It is important to note that for (001) growth, the strain tensor is diagonal while for (111), and several other directions, the strain tensor has nondiagonal terms. The nondiagonal terms can be exploited to produce built-in electric fields in certain heterostructures as will be discussed in the next section.

An important heterostructure system involves growth of hcp lattice-based AlGaN or InGaN on a GaN substrate along the c-axis. In this case the strain tensor is given by (a_L is the substrate lattice constant, a_S is overlayer lattice constant)

$$\begin{aligned}
\epsilon_{xx} &= \epsilon_{yy} = \frac{a_S}{a_L} - 1 \\
\epsilon_{zz} &= -2\frac{c_{13}}{c_{33}}\epsilon_{xx} \qquad (1.4.5)
\end{aligned}$$

This strain is exploited to generate piezoelectric effect based interface charge as discussed in the next chapter. Such a charge can cause effective doping in heterostructures as discussed in Chapter 2. In table 1.1 we provide values of elastic constant of several important semiconductors.

1.5 TECHNOLOGY CHALLENGES

Metal and insulator (glass) technologies have been around for thousands of years. Compared to these semiconductor technology is relatively new. Semiconductors need to be extremely "pure" if they are to be useful. Defect densities of a percent may have minimal effect on metals and insulators, but will ruin a semiconductor device. For most high performance devices, defect densities of less than one part in 100 million are needed

Semiconductor substrate technology is available (i.e., bulk crystals can be grown in sufficient size/purity) for a handful of materials. These include Si, GaAs, InP, and Ge, which are widely available and SiC, Al_2O_3, and GaSb, etc., which are available only in small pieces (a few square centimeters) and are very expensive. Since most semiconductors do not have a substrate available from either bulk crystal growth or another lattice matched substrate, this severely restricts the

Material	C_{11}(N/m^2)	C_{12}(N/m^2)	C_{41}(N/m^2)
Si	1.66 x 10^{11}	0.64 x 10^{11}	0.8 x 10^{11}
Ge	1.29 x 10^{11}	0.48 x 10^{11}	0.67 x 10^{11}
GaAs	1.2 x 10^{11}	0.54 x 10^{11}	0.59 x 10^{11}
C	10.76 x 10^{11}	1.25 x 10^{11}	5.76 x 10^{11}

Material	C_{13}(N/m^2)	C_{33}(N/m^2)
GaN	10.9 x 10^{11}	35.5 x 10^{11}
AlN	12 x 10^{11}	39.5 x 10^{11}

Table 1.2: Elastic constant for some fcc and hcp based semiconductors. (For Si, Ge, GaAs see H. J. McSkimin and P. Andreatch, J. Appl. Phys., **35**, 2161 (1964) and D. I. Bolef and M. Meres, J. Appl. Phys., **31**, 1010 (1960). For nitrides see J. H. Edgar, Properties of III-V Nitrides, INSPEC, London (1994) and R. B. Schwarz, K. Khachaturyan, and E. R. Weber, Appl. Phys. Lett., **74**, 1122 (1997).)

use of a wide range of semiconductors. In table 1.3 we show an overview of some important substrates and issues in semiconductor technology.

1.6 PROBLEMS

Problem 1.1 A 10.0 μm Si epitaxial layer is to be grown. The Si flux is 10^{14} cm^{-2} s^{-1}. How long will it take to grow the film if the sticking coefficient for Si atoms is 0.95?

Problem 1.2 A Si wafer is nominally oriented along the (001) direction, but is found to be cut 2° off, toward the (110) axis. This off axis cut produces “steps” on the surface which are 2 monolayers high. What is the lateral spacing between the steps of the 2° off-axis wafer?

Problem 1.3 Conduct a literature search to find out what the lattice mismatch is between GaAs and AlAs at 300 K and 800 K. Calculate the mismatch between GaAs and Si at the same temperatures.

Problem 1.4 In high purity Si crystals, defect densities can be reduced to levels of 10^{13} cm^{-3}. On an average, what is the spacing between defects in such crystals? In heavily doped Si, the dopant density can approach 10^{19} cm^{-3}. What is the spacing between defects for such heavily doped semiconductors?

IMPORTANT SUBSTRATES	Statue / ISSUES
1. Silicon (Si)	Mature, 12-inch diameter. Next generation 15-inch diameter.
2. Gallium Arsenide (GaAs)	Mature, 6-inch diameter.
3. Indium Phosphide (InP)	Mature, brittle, maximum diameter 4 inches.
4. Silicon Carbide (SiC)	Developing technology, 3-inch diameter in production. Micropipe density 1 cm^{-2} for n-type and 100 cm^{-2} for semi-insulating.
5. Germanium (Ge)	6-inch diameter. Limited supply. Water-soluble oxide.
6. Sapphire (Al_2O_3)	Hydrothermal growth. 4-inch diameter available. Low thermal conductivity.
7. Aluminum Nitride (AlN)	1-inch diameter. Early stages of development, sublimation growth technique.
8. Gallium Nitride (GaN)	2-inch diameter substrates by HVPE. Dislocation density 10^{6} cm^{-2}.
9. Indium Antimonide (InSb)	2-inch diameter, early stages of development.
10. Zinc Oxide (ZnO)	Hydrothermal growth. 2-inch diameter available. Dislocation density < 100 cm^{-2} for n-type.

Table 1.3: A brief overview of important substrates available in semiconductor technology.

Problem 1.5 Assume that a Ga-As bond in GaAs has a bond energy of 1.0 eV. Calculate the energy needed to cleave GaAs in the (001) and (110) planes.

Problem 1.6 Consider a hcp structure shown in the text. Prove the relation given by $c/a = \sqrt{8/3} = 1.633$.

Problem 1.7 Why are entropy considerations unimportant in dislocation generation?

Problem 1.8 A coherently strained quantum well laser has to made from $\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{As}$ on a GaAs substrate. If the minimum thickness of the region is 50 Å, calculate the maximum composition of In that can be tolerated. Assume that the lattice constant of the alloy can be linearly interpolated from its components.

Problem 1.9 Assume that in a semiconductor alloy, the lattice constant scales as a linear weighted average. Find the composition of the $In_xGa_{1-x}As$ alloy that lattice matches with an InP substrate.

Problem 1.10 Calculate the critical thickness for the growth of AlAs on a GaAs substrate.

Problem 1.11 A 100 Å $In_{0.2}Ga_{0.8}As$ film is grown on a GaAs substrate. The film is coherent. Calculate the strain energy per cm^2 in the film.

Problem 1.12 Consider a coherently grown film of $Si_{0.8}Ge_{0.2}$ grown on a Si substrate. Calculate the thickness of the film at which the strain energy density (eV cm^{-2}) becomes equal to the energy density arising from a square array of dislocations in the film.

Assume that the dislocations are on a planar square grid with one broken bond per spacing of a/ϵ where a is the film lattice constant and ϵ is the strain. The energy per broken bond is 1.0 eV.

1.7 FURTHER READING

- **Crystal Structure**
  - M. M. Woolfson, An Introduction to Crystallography, Cambridge University Press (1997).
  - McGraw-Hill Encyclopedia of Science and Technology, Volume 4, McGraw-Hill (1997).
  - A. C. Gossard (ed.), Epitaxial Microstructures in Semiconductors and Semi metals, Volume 40, Academic Press (1994).
  - G. Benedek (ed.), Point and Extended Defects in Semiconductors, Plenum Publishing Press (1989).
  - Landolt-Bornstein, Numerical Data and Functional Relationships in Science and Technology, (O. Madelung, M. Schultz, and H. Weiss, eds.), Springer (1985).
- **Strained Structures**
  - J. F. Nye, Physical Properties of Crystals: Their Representation by Tensors and Matrices, Oxford University Press (1987).
  - T. Ikeda, Fundamentals of Piezoelectricity, Oxford University Press (1990).
  - E. Bernardini, V. Fiorentini, and D. Vanderbilt, Spontaneous Polarization, and Piezoelectric Constant of III-V Nitrides, Physical Review B, vol. 56, p. R10024 (1997).
  - J. H. Edgar, Properties of Group III Nitrides, INSPEC, London (1994).

Chapter 2

ELECTRONIC LEVELS IN SEMICONDUCTORS

2.1 INTRODUCTION

Semiconductor electronic and optoelectronic devices depend upon how electrons inside materials behave and how they are influenced by external perturbations which may be electrical, electromagnetic, mechanical, or magnetic, etc. The simplest approach to understanding such properties would be to use classical physics. Based on classical physics the general problem could be solved by using Newton's equation

$$\frac{d\mathbf{p}}{dt} = e\left(\mathcal{E} + \mathbf{v} \times \mathbf{B}\right)$$

where $\mathbf{p}$ is the electron momentum, $\mathbf{v}$ the velocity, and $\mathcal{E}$ and $\mathbf{B}$ are the electrical and magnetic fields, respectively. Additional forces, if present, can be added on the right-hand side of the equation. Although classical physics has been successful in describing many of nature's phenomena, it fails completely when it is used to describe electrons in solids. To understand the underlying physical properties that form the basis of modern intelligent information devices, we need to use quantum mechanics.

According to quantum mechanics particles such as electrons behave as waves while waves such as electromagnetic waves behave as particles. The wave nature of particles is manifested for electrons in solids. To the level needed in device physics, the electronic properties are described by the Schrödinger equation . However, It turns out we can develop effective descriptions for the behavior of electrons and then use simple classical physics. Of course to develop this effective description we have to solve the Schrödinger equation. But once this description is developed we can use Newton's equation to understand how electrons respond to external forces. This allows us to use simple models to describe electronic devices.

In this chapter we will review a few important outcomes of quantum mechanics. In particular we will discuss the following:

• Electronic properties in an atom: All solids are made up of atoms and the properties of electrons in atoms allows us to develop insight into the electronic properties of solids. We will discuss the hydrogen atom problem since it is the simplest atom and captures the useful physics needed to understand the theory of doping.

• Electrons in a quantum well: Quantum wells, both naturally occurring and artificially created in semiconductor structures are very important in modern technology. In devices such as MOSFETs, lasers, modulators etc. electrons are in quantum wells of various sizes and shapes.

• Electrons in free space and in crystalline materials: Most high performance semiconductor devices are based on high quality crystals. In these periodic structures electrons have allowed energies that form bands separated from each other (in energy) by gaps. Almost every semiconductor property depends upon these bands. Once we understand the band theory, i.e properties of electrons in crystalline solids we can develop the effective description mentioned above and use simple classical concepts.

• Occupation of electronic states: Quantum mechanics has very specific rules on the actual occupation of energies allowed by Schrödinger equation. This occupation theory is central to understanding solid state physics and device behavior.

Once we have developed the basic quantum theory structure we will discuss properties of various semiconductors and their heterostructures.

2.2 PARTICLES IN AN ATTRACTIVE POTENTIAL: BOUND STATES

We will now examine several important quantum problems that have impact on materials and physical phenomena useful for device applications. The Schrödinger equation for electrons can be written in as

$$\left[-\frac{\hbar^2}{2m_0}\nabla^2 + V(r,t)\right]\Psi(r,t) = E\Psi(r,t)$$

where m_0 is the mass of the electron and $V(r,t)$ is the potential energy. This is a differential equation with solutions Ψ. Once the equation is solved we get a series of allowed energies and wavefunctions. Energies are allowed while others not consistent with the equation are forbidden. The band theory that forms the basis of all semiconductor devices is based on energy bands and gaps.

2.2.1 Electronic levels in a hydrogen atom

The hydrogen atom problem is of great relevance in understanding dopants in semiconductors. We will briefly summarize these findings. The hydrogen atom consists of an electron and a proton interacting with the Coulombic interaction. The problem can be solved exactly and provides insight into how electrons behave inside atoms.

Wavefunctions in the H-atom problem have the following term:

$$\psi_{n\ell m}(r,\theta,\phi) = R_{n\ell}(r)F_{\ell m}(\theta)G_m(\phi)$$

The symbols n, ℓ, m are the three quantum numbers describing the solution. The three quantum numbers have the following allowed values:

$$\begin{aligned}
\text{principle number}, n &: \text{Takes values } 1, 2, 3, \ldots \\
\text{angular momentum number}, \ell &: \text{Takes values } 0, 1, 2, \ldots n-1 \\
\text{magnetic number}, m &: \text{Takes values } -\ell, -\ell+1, \ldots \ell
\end{aligned}$$

The principle quantum number specifies the energy of the allowed electronic levels. The energy eigenvalues are given by

$$E_n = -\frac{\mu e^4}{2\left(4\pi\epsilon_0\right)^2 \hbar^2 n^2} \tag{2.2.1}$$

The spectrum is shown schematically in figure 2.1. Due to the much larger mass of the nucleus as compared with the mass of the electron, the reduced mass μ is essentially the same as the electron mass m_0. The ground state of the hydrogen atom is given by

$$\psi_{100} = \frac{1}{\sqrt{\pi a_0^3}} e^{-r/a_0} \tag{2.2.2}$$

The parameter a_0 appearing in the functions is called the Bohr radius and is given by

$$a_0 = \frac{4\pi\epsilon_0\hbar^2}{m_0 e^2} = 0.53 \text{ Å} \tag{2.2.3}$$

It roughly represents the spread of the ground state.

As noted earlier the dopant problem is addressed by using the potential of the H-atom

2.2.2 Electrons in a quantum well

As noted in the previous chapter, using semiconductor heterostructures it is possible to fabricate quantum well systems. These systems are used for high-performance devices, such as transistors, lasers and modulators. The quantum well problem can also be used to understand how defects create trap levels.

A quantum well potential profile is shown in figure 2.2. The well (i.e., region where potential energy is lower) is described by a well size $W = 2a$ as shown and a barrier height V_0. In general the potential could be confining in one dimension with uniform potential in the other two directions (quantum well), or it could be confining in two dimensions (quantum wire) or in all three dimensions (quantum dot). As discussed later in this chapter such quantum wells are formed in semiconductor structures and we can use the results discussed in this section to understand these problems.

We assume that the potential has a form

$$V(r) = V(x) + V(y) + V(z)$$

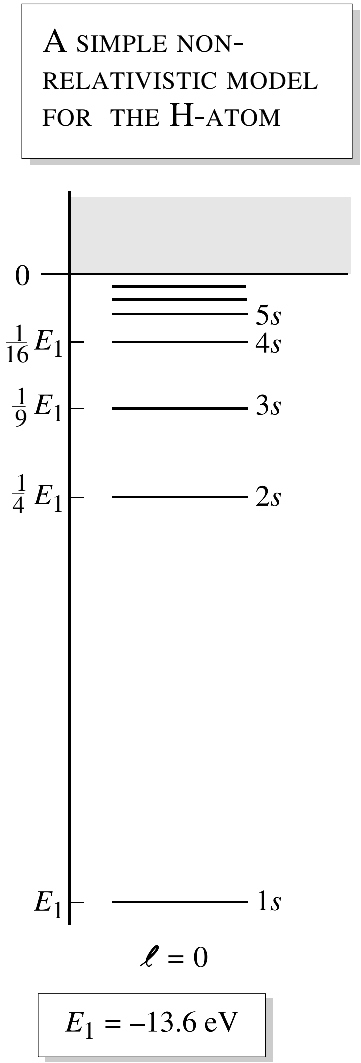


Figure 2.1: Allowed energy levels of electrons in a hydrogen atom .

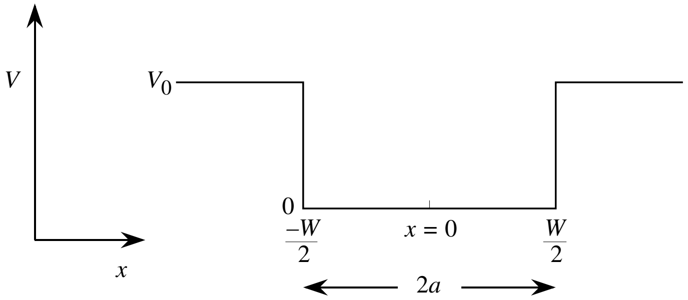


Figure 2.2: Schematic of a quantum well of width $2a$ and infinite barrier height or barrier height V_0 .

so that the wavefunction is separable and of the form

$$\psi(r) = \psi(x)\psi(y)\psi(z)$$

We will briefly discuss the problem of the square potential well, and in section 2.10 we will use the quantum well physics to discuss semiconductor quantum wells of importance in devices.

The simplest form of the quantum well is one where the potential is zero in the well and infinite outside. The equation to solve then is (the wave function is non-zero only in the well region)

$$-\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2} = E\psi \tag{2.2.4}$$

which has the general solutions

$$\begin{aligned} \psi(x) &= B\cos\frac{n\pi x}{2a}, \quad n \text{ odd} \\ &= A\sin\frac{n\pi x}{2a}, \quad n \text{ even} \end{aligned} \tag{2.2.5}$$

The energy is

$$E = \frac{\pi^2\hbar^2 n^2}{8ma^2} \tag{2.2.6}$$

Note that the well size is $2a$.

The normalized particle wavefunctions are

$$\begin{aligned} \psi(x) &= \sqrt{\frac{2}{W}}\cos\frac{n\pi x}{W}, \quad n \text{ odd} \\ &= \sqrt{\frac{2}{W}}\sin\frac{n\pi x}{W}, \quad n \text{ even} \end{aligned} \tag{2.2.7}$$

If the potential barrier is not infinite, we cannot assume that the wavefunction goes to zero at the boundaries of the well. Let us define two parameters.

$$\begin{aligned} \alpha &= \sqrt{\frac{2mE}{\hbar^2}} \\ \beta &= \sqrt{\frac{2m(V_0 - E)}{\hbar^2}} \end{aligned} \tag{2.2.8}$$

The conditions for the allowed energy levels are given by the transcendental equations

$$\frac{\alpha W}{2} \tan \frac{\alpha W}{2} = \frac{\beta W}{2} \tag{2.2.9}$$

and

$$\frac{\alpha W}{2} \cot \frac{\alpha W}{2} = -\frac{\beta W}{2} \tag{2.2.10}$$

An important outcome of these solutions is that as in the H-atom case, only some energies are allowed for the electron. This result is of importance in electronic devices as will be discussed in section 2.10.

2.3 ELECTRONS IN CRYSTALLINE SOLIDS

The devices discussed in this text are made from crystalline materials. It is, therefore, important to understand the electronic properties of these materials. Let us first examine the simpler problem of electrons in free space. It turns out that electrons in crystals can be considered to behave as if they are in free space except they have a different "effective properties". In the free space problem the background potential energy is uniform in space. The time-independent equation for the background potential in a solid equal to V_0 is

$$\frac{-\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right)\psi(r) = (E - V_0)\psi(r) \tag{2.3.1}$$

A general solution of this equation is

$$\psi(\mathbf{r}) = \frac{1}{\sqrt{V}} e^{\pm i\mathbf{k}\cdot\mathbf{r}} \tag{2.3.2}$$

and the corresponding energy is

$$E = \frac{\hbar^2 k^2}{2m} + V_0 \tag{2.3.3}$$

where the factor $\frac{1}{\sqrt{V}}$ in the wavefunction occurs because we wish to have one particle per volume V or

$$\int_V d^3r \mid \psi(r) \mid^2 = 1 \tag{2.3.4}$$

We assume that the volume V is a cube of side L. Note that if we assign the momentum of the electron as $\hbar k$ the energy-momentum relation of free electrons is the same as that in classical physics. Later we will see that in crystalline material one can use a similar relationship except the mass of the electron is modified by an effective mass.

Density of states for a three-dimensional system

We will now discuss the extremely important concept of density of states. The concept of density of states is extremely powerful, and important physical properties in materials, such as optical absorption, transport, etc., are intimately dependent upon this concept. Density of states is the number of available electronic states per unit volume per unit energy around an energy E. If we denote the density of states by $N(E)$, the number of states in a unit volume in an energy interval dE around an energy E is $N(E)dE$.

Accounting for spin, the density of states can be shown to be (see Appendix C)

$$N(E) = \frac{\sqrt{2}m_0^{3/2}(E - V_0)^{1/2}}{\pi^2\hbar^3} \tag{2.3.5}$$

In figure 2.4a we show the form of the three-dimensional density of states.

Density of states in sub-three-dimensional systems

The use of heterostructures has allowed one to make sub-three-dimensional-systems. In these systems the electron can be confined in two-dimensions (forming a quantum well) or in one-dimensional (quantum wire) and zero-dimensional (quantum dot) space. The two-dimensional density of states is defined as the number of available electronic states per unit area per unit energy around an energy E. It can be shown that the density of states for a parabolic band (for energies greater than V_0) is (see figure 2.3b)

$$N(E) = \frac{m_0}{\pi\hbar^2} \tag{2.3.6}$$

Finally, we can consider a one-dimensional system often called a "quantum wire." The one-dimensional density of states is defined as the number of available electronic states per unit length per unit energy around an energy E. In a 1D system or a "quantum wire" the density of states is (including spin) (see figure 2.3c)

$$N(E) = \frac{\sqrt{2}m_0^{1/2}}{\pi\hbar}(E - V_0)^{-1/2} \tag{2.3.7}$$

Notice that as the dimensionality of the system changes, the energy dependence of the density of states also changes. As shown in figure 2.3, for a three-dimensional system we have $(E - V_0)^{1/2}$ dependence, for a two-dimensional system we have no energy dependence, and for a one-dimensional system we have $(E - V_0)^{-1/2}$ dependence.

We will see later in the next section that when a particle is in a periodic potential, its wavefunction is quite similar to the free particle wavefunction. Also, the particle responds to external forces as if it is a free particle except that its energy-momentum relation is modified by the presence of the periodic potential. In some cases it is possible to describe the particle energy by the relation

$$E = \frac{\hbar^2k^2}{2m^*} + E_{edge} \tag{2.3.8}$$

where m^* is called the effective mass in the material and E_{edge} is the bandedge energy. The effective mass in general summarizes the appropriate way to modify the free electron mass based

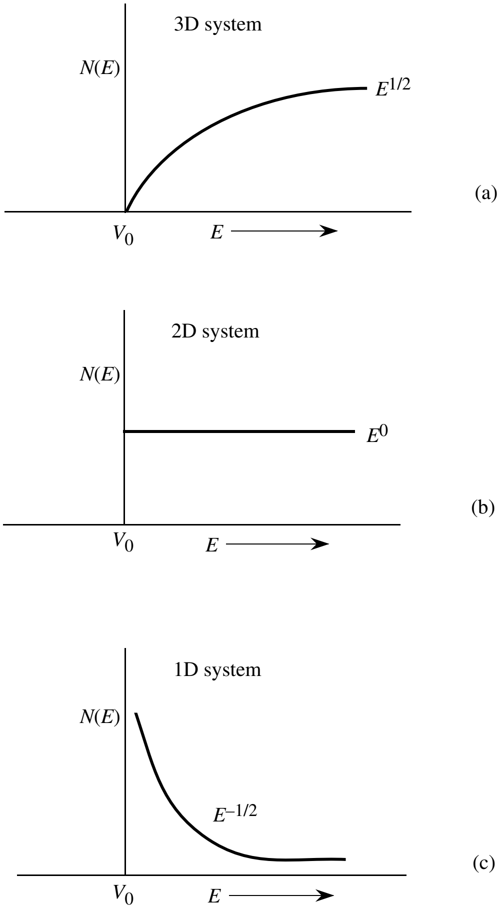


Figure 2.3: Energy dependence of the density of states in: (a) three-dimensional, (b) two-dimensional, and (c) one-dimensional systems.

on the physical property being characterized. Appendix C describes the various forms of effective mass in detail. The expressions derived for the free electron density of states can then be carried over to describe the density of states for a particle in a crystalline material (which has a periodic potential) by simply replacing m_0 by m^*.

EXAMPLE 2.1 Calculate the density of states of electrons in a 3D system and a 2D system at an energy of 1.0 eV. Assume that the background potential is zero.

The density of states in a 3D system (including the spin of the electron) is given by (E is the energy in Joules)

$$\begin{aligned} N(E) &= \frac{\sqrt{2}(m_0)^{3/2}E^{1/2}}{\pi^2\hbar^3} \\ &= \frac{\sqrt{2}(0.91 \times 10^{-30}\ \text{kg})(E^{1/2})}{\pi^2(1.05 \times 10^{-34}\ \text{J}\cdot\text{s})^3} \\ &= 1.07 \times 10^{56} E^{1/2}\ \text{J}^{-1}\ \text{m}^{-3} \end{aligned}$$

Expressing E in eV and the density of states in the commonly used units of eV^{-1} cm^{-3}, we get

$$\begin{aligned} N(E) &= 1.07 \times 10^{56} \times (1.6 \times 10^{-19})^{3/2}(1.0 \times 10^{-6})E^{1/2} \\ &= 6.8 \times 10^{21} E^{1/2}\ \text{eV}^{-1}\ \text{cm}^{-3} \end{aligned}$$

At $E = 1.0$ eV we get

$$N(E) = 6.8 \times 10^{21}\ \text{eV}^{-1}\ \text{cm}^{-3}$$

For a 2D system the density of states is independent of energy and is

$$N(E) = \frac{m_0}{\pi\hbar^2} = 4.21 \times 10^{14}\ \text{eV}^{-1}\ \text{cm}^{-2}$$

2.3.1 Particle in a periodic potential: Bloch theorem

Band theory, which describes the properties of electrons in a periodic potential arising from the periodic arrangement of atoms in a crystal, is the basis for semiconductor technology.

The Schrödinger equation in the crystal

$$\left[\frac{-\hbar^2}{2m_0}\nabla^2 + U(\mathbf{r})\right]\psi(\mathbf{r}) = \mathbf{E}\psi(\mathbf{r}) \tag{2.3.9}$$

where $U(\mathbf{r})$ is the background potential seen by the electrons. Due to the crystalline nature of the material, the potential $U(\mathbf{r})$ has the same periodicity

$$U(\mathbf{r}) = U(\mathbf{r} + \mathbf{R})$$

We have noted earlier that if the background potential is V_0, the electronic function in a volume V is

$$\psi(\mathbf{r}) = \frac{e^{i\mathbf{k}\cdot\mathbf{r}}}{\sqrt{V}}$$

and the electron momentum and energy are

$$\begin{aligned} \mathbf{p} &= \hbar\mathbf{k} \\ E &= \frac{\hbar^2k^2}{2m_0} + V_0 \end{aligned}$$

The wavefunction is spread in the entire sample and has equal probability ($\psi^*\psi$) at every point in space. In the periodic crystal electron probability is the same in all unit cells of the crystal because each cell is identical. This is shown schematically in figure 2.4.

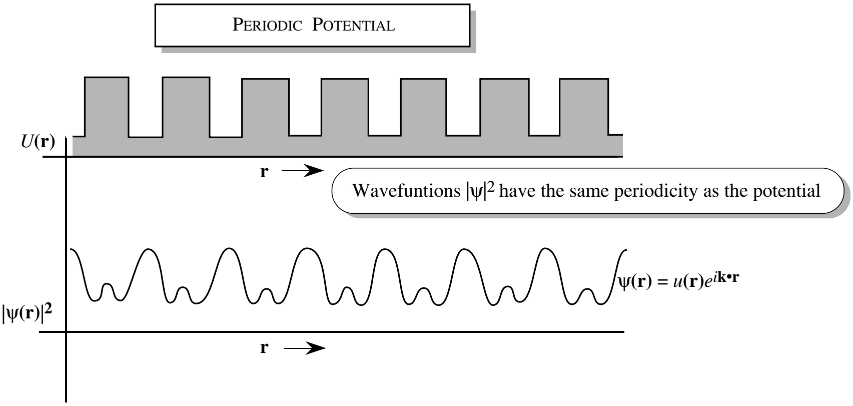

PERIODIC POTENTIAL
U(r)
r →
Wavefuntions |ψ|² have the same periodicity as the potential
ψ(r) = u(r)e^{ik•r}
|ψ(r)|²
r →

Figure 2.4: A periodic potential, $|\psi|^2$ has the same spatial periodicity as the potential.

Bloch's theorem states the eigenfunctions of the Schrödinger equation for a periodic potential are the product of a plane wave $e^{i\mathbf{k}\cdot\mathbf{r}}$ and a function $u_{\mathbf{k}}(\mathbf{r})$, which has the same periodicity as the periodic potential. Thus

$$\psi_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} u_{\mathbf{k}}(\mathbf{r}) \tag{2.3.10}$$

is the form of the electronic function. The periodic part $u_{\mathbf{k}}(\mathbf{r})$ has the same periodicity as the crystal, i.e.

$$u_{\mathbf{k}}(\mathbf{r}) = u_{\mathbf{k}}(\mathbf{r}+\mathbf{R}) \tag{2.3.11}$$

The wavefunction has the property

$$\begin{aligned} \psi_{\mathbf{k}}(\mathbf{r}+\mathbf{R}) &= e^{i\mathbf{k}\cdot(\mathbf{r}+\mathbf{R})} u_{\mathbf{k}}(\mathbf{r}+\mathbf{R}) = e^{i\mathbf{k}\cdot\mathbf{r}} u_{\mathbf{k}}(\mathbf{r}) e^{i\mathbf{k}\cdot\mathbf{R}} \\ &= e^{i\mathbf{k}\cdot\mathbf{R}} \psi_{\mathbf{k}}(\mathbf{r}) \end{aligned} \tag{2.3.12}$$

To obtain the allowed energies, i.e. the band structure, computer techniques are used to solve the Schrödinger equation. One obtains a series of allowed energy bands separated by bandgaps as shown schematically in figure 2.5. Each band has an E vs. k relation Examples of such relations called bandstructure will be shown later in section 2.6. The product of $\hbar$ and the k-vector behaves like an effective momentum for the electron inside the crystal.

The smallest k-values lie in a k-space called the Brillouin zone (see figure 2.6). If the k-value is chosen beyond the Brillouin zone values, the energy values are simply repeated. The concept

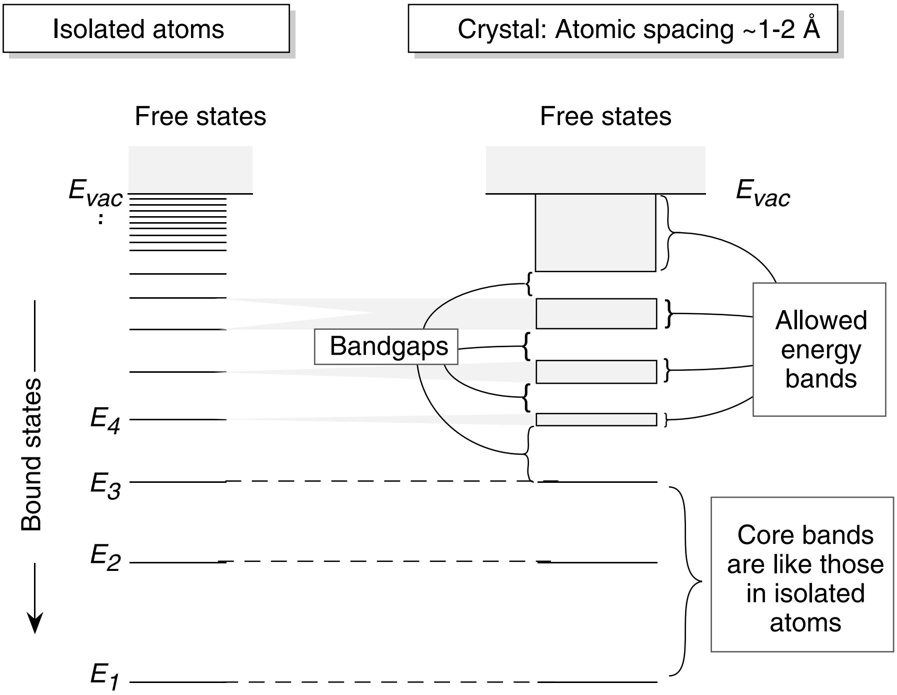


Figure 2.5: A schematic description of allowed energy levels and energy bands in an atom and in crystalline materials.

of allowed bands of energy separated by bandgaps is central to the understanding of crystalline materials. Near the bandedges it is usually possible to define the electron E–k relation as

$$E = \frac{\hbar^2(k - k_o)^2}{2m^*}$$

where k_o is the k-value at the bandedge and m^* is the effective mass. The concept of an effective mass is extremely useful, since it represents the response of the electron–crystal system to the outside world.

k-vector
According to the Bloch theorem, in the perfectly periodic background potential that the crystal presents, the electron propagates without scattering. The electronic state $\sim \exp(i\mathbf{k}\cdot\mathbf{r})$ is an extended wave which occupies the entire crystal. To describe the response of the electron waves to external forces one uses the wavepacket description. The equation of motion for electrons in

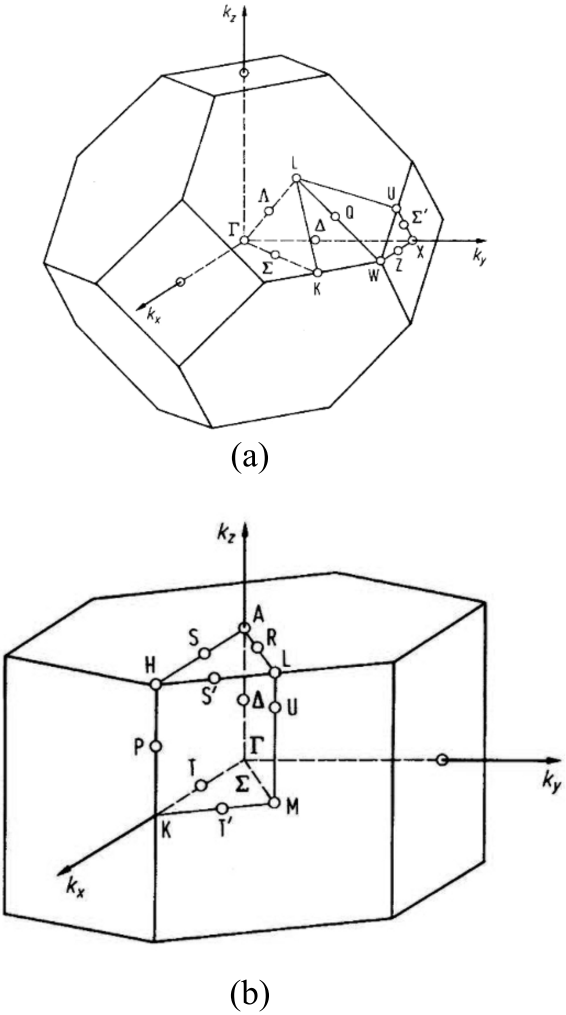


Figure 2.6: Brillouin zone of (a) the face centered cubic lattice and (b) the hexagonal lattice.

general is

$$\frac{d\mathbf{p}}{dt} = \mathbf{F}_{\text{ext}} + \mathbf{F}_{\text{int}}$$

However this is not very useful for a meaningful description of the electron because it includes the internal forces on the electron. We need a description which does not include the evaluation of the internal forces. Using a wavepacket description of electrons as with any wave phenomena it is the wave group velocity that represents the propagation of wave energy. In the case of a particle wave the group velocity represents the particle velocity. The group velocity of this wavepacket is

$$\mathbf{v}_g = \frac{d\omega}{d\mathbf{k}} \tag{2.3.13}$$

where ω is the frequency associated with the electron of energy E; in quantum mechanics, $\omega = E/\hbar$:

$$\begin{aligned} \mathbf{v}_g &= \frac{1}{\hbar}\frac{dE}{d\mathbf{k}} \\ &= \frac{1}{\hbar}\nabla_{\mathbf{k}}E(\mathbf{k}) \end{aligned}$$

If we have an electric field $\mathcal{E}$ present, the work done on the electron during a time interval δt is

$$\delta E = -e\mathcal{E} \cdot \mathbf{v}_g \delta t \tag{2.3.14}$$

We may also write, in general

$$\begin{aligned} \delta E &= \left(\frac{dE}{d\mathbf{k}}\right)\delta\mathbf{k} \\ &= \hbar\mathbf{v}_g \cdot \delta\mathbf{k} \end{aligned} \tag{2.3.15}$$

Comparing the two equations for δE, we get

$$\delta\mathbf{k} = -\frac{e\mathcal{E}}{\hbar}\delta t$$

giving us the relation

$$\hbar\frac{d\mathbf{k}}{dt} = -e\mathcal{E} \tag{2.3.16}$$

In general, we may write

$$\hbar\frac{d\mathbf{k}}{dt} = \mathbf{F}_{\text{ext}} \tag{2.3.17}$$

The term $\hbar\mathbf{k}$ responds to the external forces as if it is the momentum of the electron, although, as can be seen by comparing the true Newtons equation of motion, it is clear that $\hbar\mathbf{k}$ contains the effects of the internal crystal potentials and is therefore not the true electron momentum. The quantity $\hbar\mathbf{k}$ is called the crystal momentum . We can, for all practical purposes, treat the electrons as if they are free and obey the effective Newtons equation of motion. This physical picture is summarized in figure 2.7.

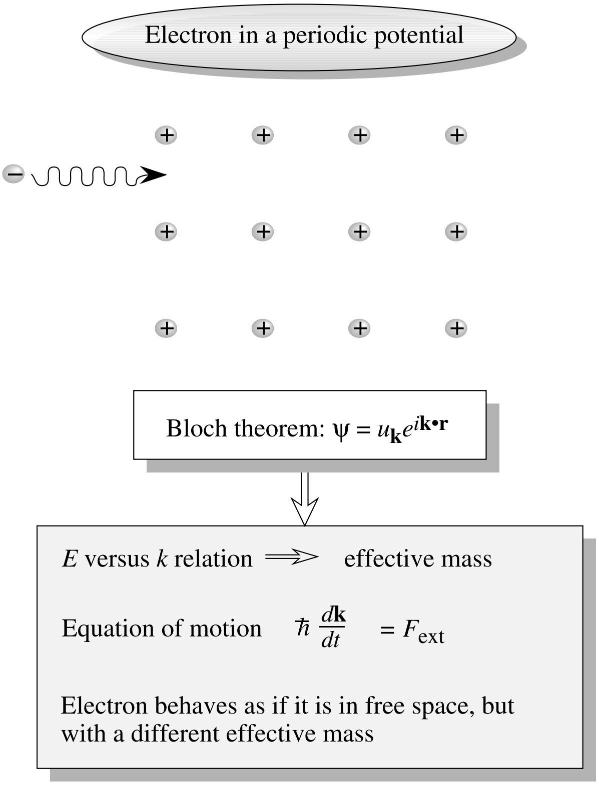

Electron in a periodic potential
Bloch theorem: $\psi = u_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{r}}$
E versus k relation ⟹ effective mass
Equation of motion $\hbar \frac{d\mathbf{k}}{dt} = F_{\text{ext}}$
Electron behaves as if it is in free space, but with a different effective mass

Figure 2.7: Electrons in a periodic potential can be treated as if they are in free space except that their energy–momentum relation is modified because of the potential. Near the bandedges the electrons respond to the outside world as if they have an effective mass m^*. The effective mass can have a positive or negative value.

2.4 OCCUPATION OF STATES: DISTRIBUTION FUNCTION

Bandstructure calculations give us the allowed energies for the electron. How will the particles distribute among the allowed states? To answer this question we need to use quantum statistical physics. According to quantum mechanics particles (this term includes classical particles and classical waves which are represented by particles) have an intrinsic angular momentum called spin. The spin of particles can take a value of $0, 1/2\hbar, \hbar, 3/2\hbar$, etc. Particles which have integral spins (in units of $\hbar$) are called bosons, while those that have half-integral spins are called fermions.

According to thermodynamics, a system with a large number of particles can be described by macroscopic properties such as temperature, pressure, volume, etc. Under equilibrium conditions (no exchange of net energy with other systems) the system is described by a distribution function, which gives us the occupation number for any energy level. To find this occupation we have to minimize the free energy F of the system subject to any constraints from quantum mechanics (such as the Pauli exclusion principle). The following distribution functions are obtained for equilibrium:

- For fermions such as electrons

$$f(E) = \frac{1}{\exp\left[\frac{E-E_F}{k_BT}\right]+1}$$

Here $f(E)$ is the occupation function; E_F is the Fermi energy and its value depends upon particle density).

In classical physics the occupation function for electrons is

$$f(E) = \frac{1}{\exp\left(\frac{E-E_F}{k_BT}\right)} \tag{2.4.1}$$

Note that if $E - E_F \gg k_BT$; i.e., $f(E) \ll 1$, the classical function approaches the quantum Fermi distribution function.

For completeness we note the distributed function for bosons as well.

- Massless bosons (like photons)

$$f(E) = \frac{1}{\exp\left(\frac{E}{k_BT}\right)-1} \tag{2.4.2}$$

- Bosons with mass (this applies to electron pairs that occur in superconductors)

$$f(E) = \frac{1}{\exp\left(\frac{E-\mu}{k_BT}\right)-1} \tag{2.4.3}$$

where μ is an energy determined from the particle density.

- There is a distribution function that proves to be useful in solid state devices. When solving the Schrödinger equation we can get more than one solution with the same energy. This is the degeneracy g_d of a state. Consider a case where a state has a degeneracy g_i and can, in principle, be occupied by g_d electrons. However, for dopants and defect levels, when one electron is placed in the allowed state, the next one cannot be placed because of the Coulombic repulsion. This happens for some states, such as those states associated with donors or acceptors, traps, etc.

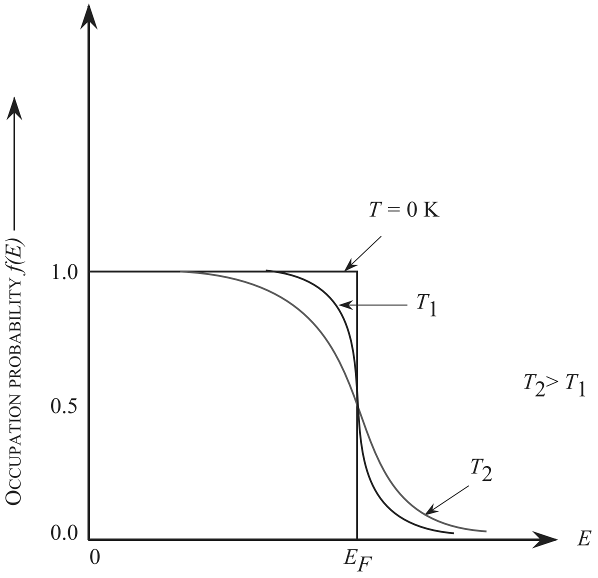


Figure 2.8: Schematic of the Fermi function for electrons and other fermions. In general the position of E_F is dependent on temperature. The occupation probability is at 0.5 at the Fermi energy.

Thus, even though Pauli exclusion principle would allow two (or more) electrons to reside on the state, the repulsion would not. In such cases the occupation function can be shown to be

$$f(E) = \frac{1}{\frac{1}{g_d}\exp\left(\frac{E-E_F}{k_BT}\right)+1} \tag{2.4.4}$$

In figure 2.8 we show a schematic of the Fermi function for electrons and its dependence on temperature. It is important to note that at $E = E_F, f(E) = 0.5$ regardless of the temperature. At zero temperature, the Fermi function becomes a step function with $f(E < E_F) = 1.0$ and $f(E) > E_F = 0.0$.

2.5 METALS AND INSULATORS

Band theory shows that the allowed energy states of electrons in a crystalline material are described by a series of allowed bands separated by forbidden bandgaps. Two important situations

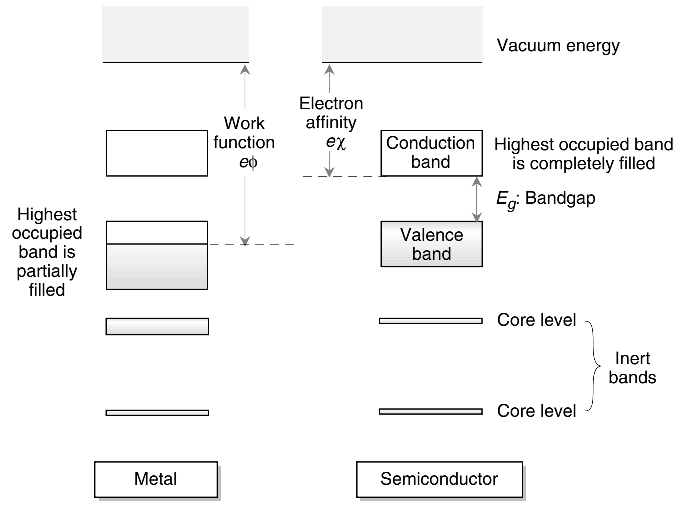


Figure 2.9: Electron occupation of the bands in a metal and semiconductor (or insulator). In a metal, the highest occupied band at 0 K is partially filled with electrons. In a semiconductor at 0 K, the highest occupied band is completely filled with electrons and the next band is completely empty. The separation between the two bands is the bandgap E_g.

arise when we examine the electron occupation of allowed bands. As shown in figure 2.9 we can have a situation where an allowed band is completely filled with electrons, while the next allowed band is separated in energy by a gap E_g and is completely empty at 0 K. In a second case, the highest occupied band is only half full (or partially full).

When an allowed band is completely filled with electrons, the electrons in the band cannot conduct any current. Since electrons are fermions they cannot carry any net current in a filled band since an electron can only move into an empty state. Because of this effect, when we have a material in which a band is completely filled, while the next allowed band is separated in energy and empty, the material has, in principle, infinite resistivity and is an insulator or a semiconductor. The material in which a band is only half full with electrons has a very low resistivity and is a metal.

The band that is normally filled with electrons at 0 K in semiconductors is called the valence band, while the upper unfilled band is called the conduction band . The energy difference between the vacuum level and the highest occupied electronic state in a metal is called the metal work function . The energy between the vacuum level and the bottom of the conduction band is called the electron affinity.

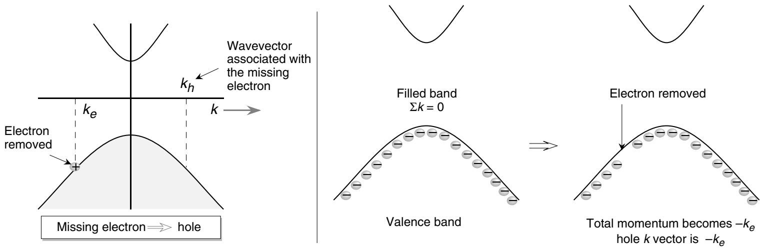


Figure 2.10: Illustration of the wavevector of a filled valence band with a missing electron k_e. The wavevector is $-\mathbf{k}_e$, which is associated with the hole.

Semiconductors have zero conductivity at 0 K and quite low conductivity at finite temperatures, but it is possible to alter their conductivity by orders of magnitude through doping or applied electric potentials. This makes semiconductors useful for active devices.

2.5.1 Electrons and Holes

In semiconductors the valence band is full of electrons and the conduction band is empty at 0 K. At finite temperatures some of the electrons leave the valence band and occupy the conduction band. Electrons in the conduction band can carry current. When electrons leave the valence band there are unoccupied states. Consider the situation as shown in figure 2.10, where an electron with momentum $\mathbf{k}_e$ is missing from the valence band. When all of the valence band states are occupied, the sum of the total momentum is zero; i.e.

$$\sum \mathbf{k}_i = 0 = \sum_{\mathbf{k}_i \neq \mathbf{k}_e} \mathbf{k}_i + \mathbf{k}_e \tag{2.5.1}$$

This result is just an indication that there are as many positive k states occupied as there are negative ones. Now, in the situation where the electron at wavevector $\mathbf{k}_e$ is missing, the total wavevector is

$$\sum_{\mathbf{k}_i \neq \mathbf{k}_e} \mathbf{k}_i = -\mathbf{k}_e \tag{2.5.2}$$

The missing state is called a hole and the wavevector of the system $-\mathbf{k}_e$ is attributed to it. It is important to note that the electron is missing from the state $\mathbf{k}_e$ and the momentum associated with the hole is at $-\mathbf{k}_e$. The position of the hole is depicted as that of the missing electron. But in reality the hole wavevector $\mathbf{k}_h$ is $-\mathbf{k}_e$, as shown in figure 2.10 and we have

$$\mathbf{k}_h = -\mathbf{k}_e \tag{2.5.3}$$

If an electric field is applied, all the electrons move in the direction opposite to the electric field. This results in the unoccupied state moving in the field direction. The hole thus responds as if it has a positive charge. It therefore responds to external electric and magnetic fields $\mathcal{E}$ and $\mathbf{B}$, respectively, according to the equation of motion

$$\hbar\frac{d\mathbf{k}_h}{dt} = e\left[\mathcal{E} + \mathbf{v}_h \times \mathbf{B}\right] \tag{2.5.4}$$

where $\hbar\mathbf{k}_h$ and $\mathbf{v}_h$ are the momentum and velocity of the hole.

Thus the equation of motion of holes is that of particles with a positive charge e. The mass of the hole has a positive value, although the electron mass in its valence band is negative. When we discuss the valence band properties, we refer to holes. This is because in the valence band only the missing electrons or holes lead to charge transport and current flow.

2.6 BANDSTRUCTURE OF SOME IMPORTANT SEMICONDUCTORS

In this section we will examine the band structure near the band edges for several important materials. To represent the bandstructure on a figure that is two-dimensional, we draw the E-k diagram in several panels where k goes from zero to its maximum value along the (100) direction or the (111) direction, etc within the Brillouin zone. As shown in figure 2.6 for the fcc lattice, the maximum k-value along the (100) direction is $2\pi/a(1,0,0)$. This point is called the X-point and there are five other equivalent points, due to the cubic symmetry of the lattice. Similarly, along the (111) direction, the maximum k-point is $\pi/a(1,1,1)$ and seven other similar points. This point is called the L-point. Thus we commonly display the E-k diagram with k going from the origin (called the Γ-point) to the X-point and from the origin to the L-point.

2.6.1 Direct and indirect semiconductors

Two types of band structures arise in semiconductors- direct and indirect. The top of the valence band of most semiconductors occurs at effective momentum equal to zero. A typical bandstructure of a semiconductor near the top of the valence band is shown in figure 2.11. We notice the presence of three bands near the valence bandedge. These curves or bands are labeled I, II, and III in the figure and are called the heavy hole (HH), light hole (LH), and the split off hole bands.

The bottom of the conduction band in some semiconductors occurs at $k = 0$. Such semiconductors are called direct bandgap materials. Semiconductors, such as GaAs, InP, GaN, InN, etc., are direct bandgap semiconductors. In other semiconductors, the bottom of the conduction band does not occur at the $k = 0$ point, but at certain other points. Such semiconductors are called indirect semiconductors. Examples are Si, Ge, AlAs, etc.

Due to the law of momentum conservation, direct gap materials have a strong interaction with light. Indirect gap materials have a relatively weak interaction with electrons.

When the bandedges are at $k = 0$ it is possible to represent the bandstructure by a simple relation of the form

$$E(\mathbf{k}) = E_c + \frac{\hbar^2 k^2}{2m^*} \tag{2.6.1}$$

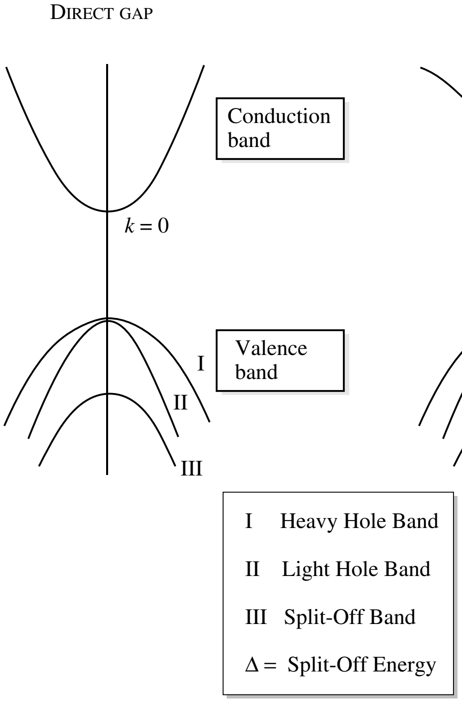


Figure 2.11: Schematic of the valence band, direct bandgap, and indirect bandgap conduction bands. The curves I, II, III in the valence band are called heavy hole, light hole, and split-off hole states, respectively.

where E_c is the conduction bandedge, and the bandstructure is a simple parabola. The equation for the E–k relation looks very much like that of an electron in free space as noted in the previous section.

Silicon

The most important semiconductor is silicon. Silicon has an indirect bandgap as shown in figure 2.12. The bottom of the conduction band in Si is at point ($\sim$ $(2\pi/a)(0.85, 0.0)$; i.e., close to the X-point. There are six degenerate X-points and, consequently, six conduction bandedge valleys. The near bandedge bandstructure can be represented by ellipsoids of energy with simple E vs. k relations of the form (for examples for the [100] valley)

$$E(\mathbf{k}) = \frac{\hbar^2 k_x^2}{2m_l^*} + \frac{\hbar^2 \left(k_y^2 + k_z^2\right)}{2m_t^*} \tag{2.6.2}$$

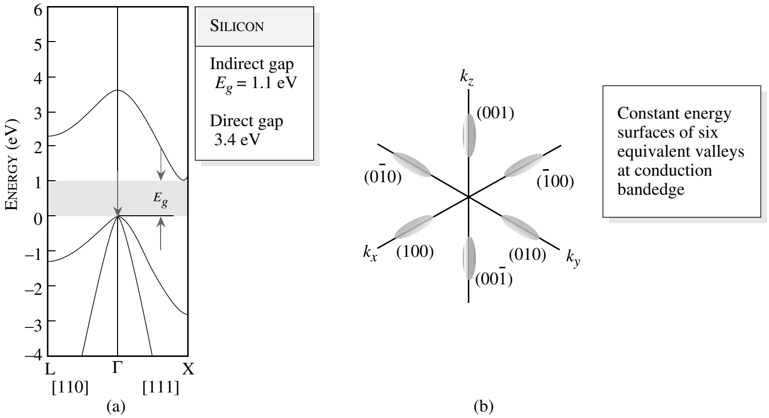


Figure 2.12: (a) Bandstructure of Si. (b) Constant energy ellipsoids for the Si conduction band. There are six equivalent valley in Si at the bandedge.

where we have two masses, the longitudinal and transverse. The constant energy surfaces of Si are ellipsoids according to Eq. 2.6.2. The six surfaces are shown in figure 2.12

The longitudinal electron mass m_l^* is approximately 0.98 m_0, while the transverse mass is approximately 0.19 m_0.

The next valley in the conduction band is the L-point valley, which is about 1.1 eV above the bandedge. Above this is the Γ-point edge. Due to the six-fold degeneracy of the conduction bandedge, the electron transport in Si is quite poor because of the very large density of states near the bandedge, leading to a high scattering rate in transport.

GaAs

GaAs is a direct gap material with small electron effective mass. The near bandedge bandstructure of GaAs is shown in figure 2.13. The bandstructure can be represented by the relation (referenced to E_c)

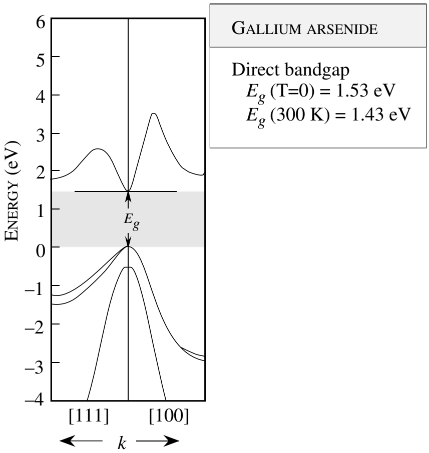
$$E = \frac{\hbar^2 k^2}{2m^*} \tag{2.6.3}$$

with $m^* = 0.067 m_0$. A better relationship is the non-parabolic approximation

$$E(1 + \alpha E) = \frac{\hbar^2 k^2}{2m^*} \tag{2.6.4}$$

with $\alpha = 0.67$ eV^{-1}.

For high electric field transport, it is important to note that the valleys above Γ-point are the L-valleys. There are eight L-points, but, since half of them are connected by a reciprocal lattice vector, there are four valleys. The separation $\Delta E_{\Gamma L}$ between the Γ- and L- minima is 0.29 eV.



GALLIUM ARSENIDE

Direct bandgap
E_g (T=0) = 1.53 eV
E_g (300 K) = 1.43 eV

Figure 2.13: Bandstructure of GaAs. The bandgap at 0 K is 1.51 eV and at 300 K it is 1.43 eV. The bottom of the conduction band is at $k = (0, 0, 0)$, i.e., the Γ-point. The upper conduction band valleys are at the L-point.

The L-valley has a much larger effective mass than the Γ-valley. For GaAs, $m_L^* \sim 0.25m_0$. This difference in masses is extremely important for high electric field transport as will be discussed in the next chapter.

The valence band of GaAs has the standard HH, LH, and SO bands. Due to the large spin–orbit splitting, for most purposes the SO band does not play any role in electronic properties.

The bandstructures of Ge and AlAs, two other important semiconductors, are shown in figure 2.14, along with brief comments about their important properties.

InN, GaN, and AlN

The III–V nitride family of GaN, InN, and AlN have become quite important due to progress in the ability to grow the semiconductor. These materials are typically grown with a wurtzite structure, and have bandgaps ranging from $\sim$1.0 eV to over 6.0 eV. This large bangap is very useful for short wavelength light emitters and high power electronics. In figure 2.15 we show the bandstructure for nitrides.

It is important to note is that the bandgap of semiconductors generally decreases as temperature increases. The bandgap of GaAs, for example, is 1.51 eV at $T = 0$K and 1.43 eV at room temperature. In table 2.1 we show the temperature dependence of bandgaps of several semiconductors.

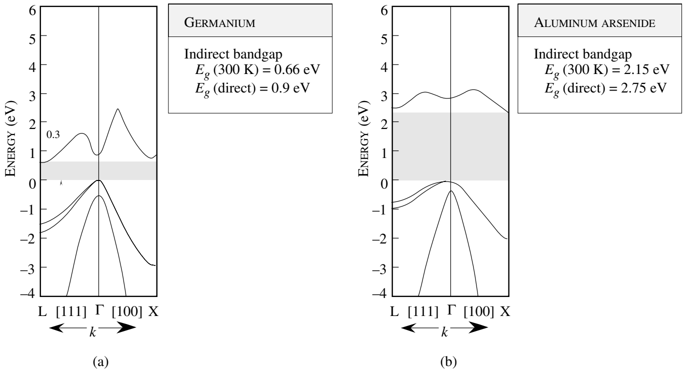


Figure 2.14: (a) Bandstructure of Ge and AlAs.

2.7 MOBILE CARRIERS

From our brief discussion of metals and semiconductors in Section 2.5, we see that in a metal current flows because of the electrons present in the highest (partially) filled band. As shown schematically in figure 2.16a. The density of such electrons is very high ($\sim 10^{23}$ cm^{-3}). In a semiconductor, in contrast, no current flows if the valence band is filled with electrons and the conduction band is empty of electrons. However, if somehow empty states or holes are created in the valence band by removing electrons, current can flow through the holes. Similarly, if electrons are placed in the conduction band, these electrons can carry current. This is shown schematically in figure 2.16b. If the density of electrons in the conduction band is n and that of holes in the valence band is p, the total mobile carrier density is $n + p$.

2.7.1 Mobile electrons in metals

In a metal, we have a series of filled bands and a partially filled band called the conduction band. The filled bands are inert as far as electrical and optical properties of metals are concerned. The conduction band of metals can be assumed to be described by the parabolic energy–momentum relation

$$E(k) = E_c + \frac{\hbar^2 k^2}{2m_0} \tag{2.7.1}$$

Note that we have used an effective mass equal to the free electrons mass. This is a reasonable approximation for metals. The large electron density in the band "screens" out the background potential and the electron effective mass is quite close to the free space value.

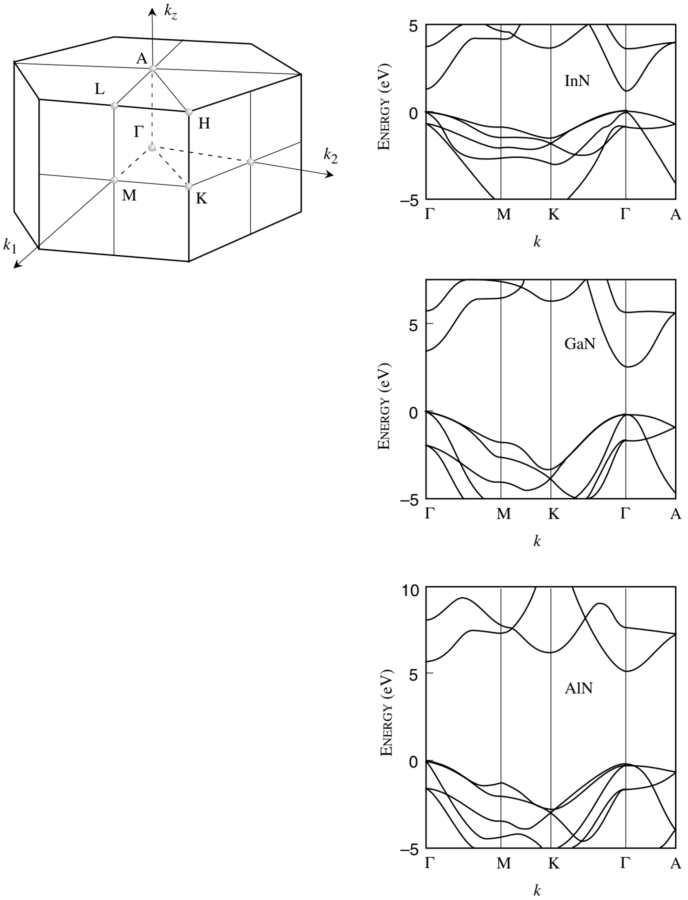


Figure 2.15: Bandstructure of InN, GaN and AlN.

The electron density in the conduction band of a metal is related to the Fermi level by the relation

$$n = \int_{E_c}^{\infty} \frac{\sqrt{2} m_0^{3/2}}{\pi^2 \hbar^3} \frac{E^{1/2} dE}{\exp\left(\frac{E - E_F}{k_B T}\right) + 1} \tag{2.7.2}$$

		Experimental bandgap E_G (eV)		
	Type of			Temperature dependence
Compound	bandgap	0 K	300 K	of bandgap $E_G(T)$ (eV)
AlP	Indirect	2.52	2.45	2.52 – 3.18 $\times 10^{-4} T^2/(T+588)$
AlAs	Indirect	2.239	2.163	2.239 – 6.0 $\times 10^{-4} T^2/(T+408)$
AlSb	Indirect	1.687	1.58	1.687 – 4.97 $\times 10^{-4} T^2/(T+213)$
GaP	Indirect	2.338	2.261	2.338 – 5.771 $\times 10^{-4} T^2/(T+372)$
GaAs	Direct	1.519	1.424	1.519 – 5.405 $\times 10^{-4} T^2/(T+204)$
GaSb	Direct	0.810	0.726	0.810 – 3.78 $\times 10^{-4} T^2/(T+94)$
InP	Direct	1.421	1.351	1.421 – 3.63 $\times 10^{-4} T^2/(T+162)$
InAs	Direct	0.420	0.360	0.420 – 2.50 $\times 10^{-4} T^2/(T+75)$
InSb	Direct	0.236	0.172	0.236 – 2.99 $\times 10^{-4} T^2/(T+140)$

Table 2.1: Bandgaps of binary III–V compounds (From Casey and Panish, 1978).

This integral is particularly simple to evaluate as 0 K, since, at this temperature

$$\begin{aligned} \frac{1}{\exp\left(\frac{E-E_F}{k_BT}\right)+1} &= 1 \text{ if } E \leq E_F \\ &= 0 \text{ otherwise} \end{aligned}$$

this gives

$$n = \int_{E_C}^{E_F} N(E)dE$$

We then have

$$\begin{aligned} n &= \frac{\sqrt{2}m_0^{3/2}}{\pi^2\hbar^3} \int_{E_C}^{E_F} (E-E_C)^{1/2}\, dE \\ &= \frac{2\sqrt{2}m_0^{3/2}}{3\pi^2\hbar^3} (E_F - E_C)^{3/2} \end{aligned}$$

or

$$E_F - E_C = \frac{\hbar^2}{2m_0}\left(3\pi^2 n\right)^{2/3} \tag{2.7.3}$$

The expression is applicable to metals such as copper, gold, etc. In Table 2.2 we show the conduction band electron densities for several metals. The quantity E_F, which is the highest occupied energy state at 0 K, is called the Fermi energy. We can define a corresponding wavevector

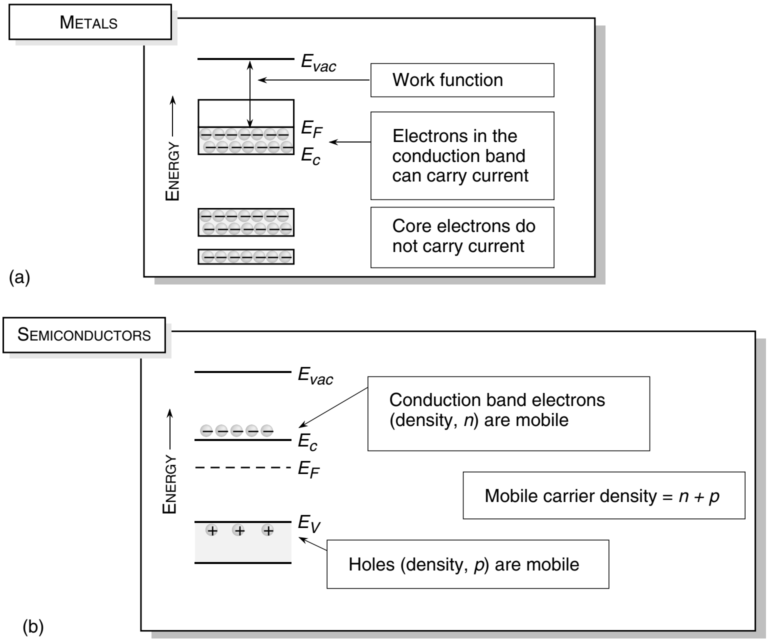

METALS
E_{vac}
Work function
E_F
E_c
Electrons in the conduction band can carry current
ENERGY
Core electrons do not carry current
(a)
SEMICONDUCTORS
E_{vac}
Conduction band electrons (density, n) are mobile
E_c
E_F
ENERGY
Mobile carrier density = $n + p$
E_V
Holes (density, p) are mobile
(b)

Figure 2.16: (a) In metals the highest occupied band is partially filled and electrons can carry current. (b) A schematic showing the valence band and conduction band in a typical semiconductor. Only electrons in the conduction band and holes in the valence band can carry current.

k_F, called the Fermi vector, and a velocity v_F, called the Fermi velocity as

$$\begin{aligned} k_F &= \left(3\pi^2 n\right)^{1/3} \\ v_F &= \left(\frac{\hbar}{m_0}\right)\left(3\pi^2 n\right)^{1/3} \end{aligned} \qquad (2.7.4)$$

It is important to note that even at 0 K, the velocity of the highest occupied state is v_F and not zero, as would be the case if we used classical statistics. At finite temperatures, the Fermi level is approximately given by

$$E_F(T) = E_F(0)\left[1 - \frac{\pi^2}{12}\frac{(k_B T)^2}{(E_F(0))^2}\right] \qquad (2.7.5)$$

where $E_F(T)$ and $E_F(0)$ are the Fermi levels at temperatures T and 0 K, respectively. In Metals there is very little change in the Fermi level with temperature.

ELEMENT	VALENCE	DENSITY (gm/cm^3)	CONDUCTION ELECTRON DENSITY (10^{22} cm^{-3})
Al	3	2.7	18.1
Ag	1	10.5	5.86
Au	1	19.3	5.90
Na	1	0.97	2.65
Fe	2	7.86	17.0
Zn	2	7.14	13.2
Mg	2	1.74	8.61
Ca	2	1.54	4.61
Cu	1	8.96	8.47
Cs	1	1.9	0.91
Sn	4	7.3	14.8

Table 2.2: Properties of some metals. In the case of elements that display several values of chemical valence, one of the values has been chosen arbitrarily.

EXAMPLE 2.1 A particular metal has 10^{22} electrons per cubic centimeter. Calculate the Fermi energy and the Fermi velocity (at 0 K).

The Fermi energy is the highest occupied energy state at 0 K and is given by (measured from the conduction bandedge)

$$\begin{aligned} E_F &= \frac{\hbar^2}{2m_0}\left(3\pi^2 n\right)^{2/3} \\ &= \frac{(1.05\times 10^{-34})^2[3\pi^2(10^{28})]^{2/3}}{2(0.91\times 10^{-30})} = 2.75\times 10^{-19}\text{ J} \\ &= 1.72\text{ eV} \end{aligned}$$

The Fermi velocity is

$$\begin{aligned} v_F &= \frac{\hbar}{m_0}\left(3\pi^2 n\right)^{1/3} \\ &= \frac{(1.05\times 10^{-34}\text{ J.s})(3\pi^2\times 10^{28}\text{ m}^{-3})^{1/3}}{0.91\times 10^{-30}\text{ kg}} = 7.52\times 10^5\text{ m/s} \\ &= 7.52\times 10^7\text{ cm/s} \end{aligned}$$

Thus, the highest energy electron has a large energy and is moving with a very large speed.

2.7.2 Electrons and holes in semiconductors

In pure semiconductors there are no mobile carriers at zero temperature. As temperature is raised, electrons from the valence band are thermally excited into the conduction band, and in equilibrium there is an electron density n and an equal hole density p, as shown in figure 2.17a Note that the density of allowed states has the form

$$N(E) = \frac{\sqrt{2}\left(m^*_{\text{dos}}\right)^{3/2}\left(E-E_c\right)^{1/2}}{\pi^2\hbar^3} \tag{2.7.6}$$

where m^*_{dos} is the density of states mass and E_c is the conduction bandedge. A similar expression exists for the valence band except the energy term is replaced by $\left(E_v - E\right)^{1/2}$ and the density of states exist below the valence bandedge E_v. Figure 2.17 shows a schematic view of the density of states.

It is important to note that the density of states mass has a special term in indirect gap materials. In direct gap semiconductors m^*_{dos} is just the effective mass for the conduction band. In indirect gap materials it is given by (see Appendix C)

$$m^*_{dos} = \left(m^*_1 m^*_2 m^*_3\right)^{1/3}$$

where $m^*_1 m^*_2 m^*_3$ are the effective masses along the three principle axes. For Si counting the six degenerate X-valleys we have

$$m^*_{dos} = 6^{2/3}\left(m_\ell m_t^2\right)^{1/3}$$

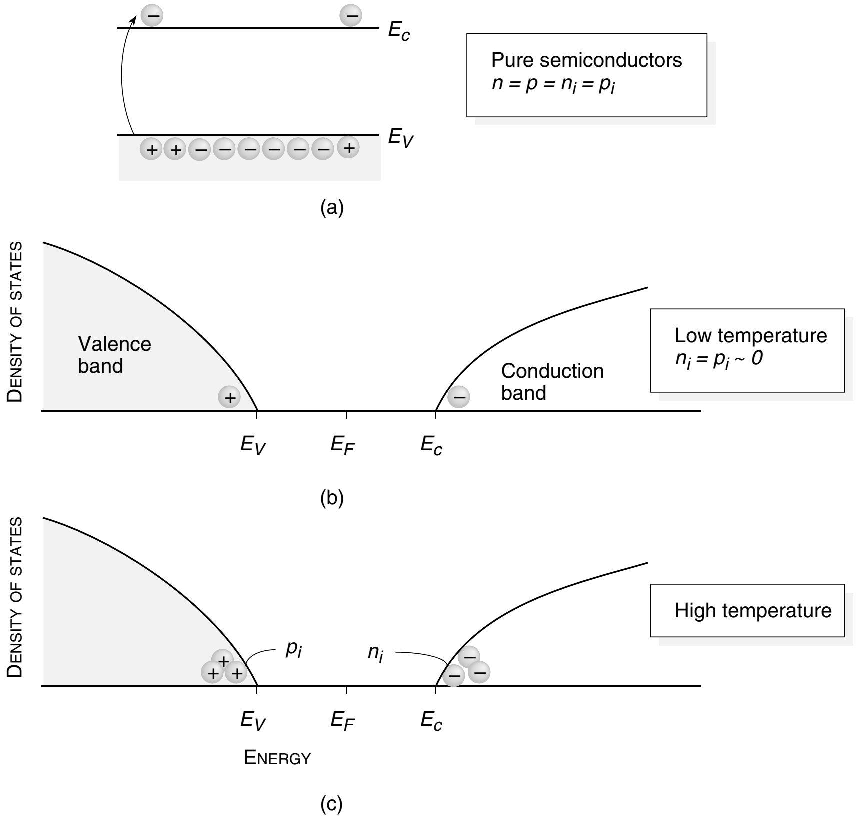


Figure 2.17: (a) A schematic showing that electron and hole densities are equal in a pure semiconductor. (b) Density of states and Fermi occupation function at low temperatures. (c) Density of states and Fermi function at high temperatures when n_i and p_i become large.

For the valence band we can write a simple expression for a density of states masses , which includes the HH and LH bands

$$m^*_{dos} = \left(m^{*3/2}_{hh} + m^{*3/2}_{\ell h}\right)^{2/3}$$

In calculating the position of the Fermi energy, charge density, etc. we need to use the density of states mass. In pure semiconductors, electrons in the conduction come from the valence band and $n = p = n_i = p_i$, where n_i and p_i are the intrinsic carrier concentrations. In general the electron density in the conduction band is

$$\begin{aligned} n &= \int_{E_c}^{\infty} N_e(E) f(E) dE \\ n &= \frac{1}{2\pi^2}\left(\frac{2m_e^*}{\hbar^2}\right)^{3/2} \int_{E_c}^{\infty} \frac{(E-E_c)^{1/2} dE}{\exp\,(\frac{E-E_F}{k_BT})+1} \end{aligned} \qquad (2.7.7)$$

For small values of n (non-degenerate statistics where we can ignore the unity in the Fermi function) we get

$$n = N_c \exp\left[\left(E_F - E_c\right)/k_BT\right] \qquad (2.7.8)$$

where the effective density of states N_c is given by

$$N_c = 2\left(\frac{m_e^* k_BT}{2\pi\hbar^2}\right)^{3/2}$$

A similar derivation for hole density gives

$$p = N_v \exp\left[\left(E_v - E_F\right)/k_BT\right] \qquad (2.7.9)$$

where the effective density of states N_v is given by

$$N_v = 2\left(\frac{m_h^* k_BT}{2\pi\hbar^2}\right)^{3/2}$$

We also obtain

$$np = 4\left(\frac{k_BT}{2\pi\hbar^2}\right)^3 (m_e^* m_h^*)^{3/2}\ \exp\ (-E_g/k_BT) \qquad (2.7.10)$$

Notice that within our low carrier density approximation, the product np is independent of the position of the Fermi level and is dependent only on the temperature and intrinsic properties of the semiconductor. This is the law of mass action. If n increases, p must decrease, and vice versa. For the intrinsic case $n = n_i = p = p_i$, we have from the square root of the equation above

$$\begin{aligned} n_i = p_i &= 2\left(\frac{k_BT}{2\pi\hbar^2}\right)^{3/2} (m_e^* m_h^*)^{3/4}\ \exp\ (-E_g/2k_BT) \\ E_{Fi} &= \frac{E_c + E_v}{2} + \frac{3}{4} k_BT \ln\left(m_h^*/m_e^*\right) \end{aligned} \qquad (2.7.11)$$

Thus the Fermi level of an intrinsic material lies close to the midgap.

In Table 2.3 we show the effective densities and intrinsic carrier concentrations in Si, Ge, and GaAs The values given are those accepted from experiments. These values are lower than the ones we get by using the equations derived in this section. The reason for this difference is due to inaccuracies in carrier masses and the approximate nature of the analytical expressions.

We note that the carrier concentration increases exponentially as the bandgap decreases. Results for the intrinsic carrier concentrations for Si, Ge, GaAs, and GaN are shown in figure 2.18.

MATERIAL	CONDUCTION BAND EFFECTIVE DENSITY (N_c)	VALENCE BAND EFFECTIVE DENSITY (N_v)	INTRINSIC CARRIER CONCENTRATION ($n_i = p_i$)
Si (300 K)	2.78 x 10^{19} cm^{-3}	9.84 x 10^{18} cm^{-3}	1.5 x 10^{10} cm^{-3}
Ge (300 K)	1.04 x 10^{19} cm^{-3}	6.0 x 10^{18} cm^{-3}	2.33 x 10^{13} cm^{-3}
GaAs (300 K)	4.45 x 10^{17} cm^{-3}	7.72 x 10^{18} cm^{-3}	1.84 x 10^{6} cm^{-3}

Table 2.3: Effective densities and intrinsic carrier concentrations of Si, Ge, and GaAs. The numbers for intrinsic carrier densities are the accepted values even though they are smaller than the values obtained by using the equations derived in the text.

In electronic devices where current has to be modulated by some means, the concentration of intrinsic carriers is fixed by the temperature and therefore is detrimental to device performance. Once the intrinsic carrier concentration increases to $\sim 10^{15}$ cm^{-3}, the material becomes unsuitable for electronic devices, due to the high leakage current arising from the intrinsic carriers. A growing interest in high-bandgap semiconductors, such as diamond (C), SiC, etc., is partly due to the potential applications of these materials for high-temperature devices where, due to their larger gap, the intrinsic carrier concentration remains low up to very high temperatures. For GaN the background defect density usually does not allow one to reach theoretical intrinsic carrier densities.

EXAMPLE 2.2 Calculate the effective density of states for the conduction and valence bands of GaAs and Si at 300 K. Let us start with the GaAs conduction-band case. The effective density of states is

$$N_c = 2\left(\frac{m_e^* k_B T}{2\pi\hbar^2}\right)^{3/2}$$

Note that at 300 K, $k_B T$ = 26 meV = 4×10^{-21} J.

$$\begin{aligned} N_c &= 2\left(\frac{0.067 \times 0.91 \times 10^{-30}\ (\text{kg}) \times 4.16 \times 10^{-21}\ (\text{J})}{2 \times 3.1416 \times (1.05 \times 10^{-34}\ (\text{Js}))^2}\right)^{3/2} \quad \text{m}^{-3} \\ &= 4.45 \times 10^{23}\ \text{m}^{-3} = 4.45 \times 10^{17}\ \text{cm}^{-3} \end{aligned}$$

In silicon, the density of states mass is to be used in the effective density of states. This is given by

$$m^*_{\text{dos}} = 6^{2/3}(0.98 \times 0.19 \times 0.19)^{1/3}\, m_0 = 1.08\, m_0$$

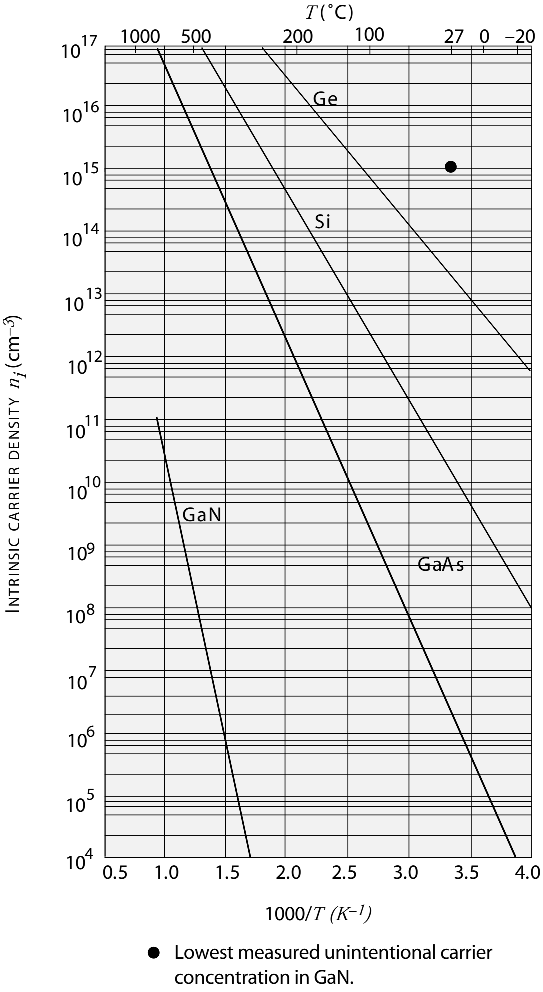


Figure 2.18: Intrinsic carrier densities of Ge, Si, GaAs, and GaN as a function of reciprocal temperature. Currently, the lowest measured unintentional background density in GaN at room temperature is around 1×10^{15} cm^{-3}, indicating that the electronic properties are dominated by defects (either extrinsic or intrinsic point defects).

The effective density of states becomes

$$\begin{aligned} N_c &= 2\left(\frac{m^*_{\rm dos} k_B T}{2\pi\hbar^2}\right)^{3/2} \\ &= 2\left(\frac{1.06 \times 0.91 \times 10^{-30}\ ({\rm kg}) \times 4.16 \times 10^{-21}\ ({\rm J})}{2 \times 3.1416 \times (1.05 \times 10^{-34}\ ({\rm Js}))^2}\right)^{3/2} \ {\rm m}^{-3} \\ &= 2.78 \times 10^{25}\ {\rm m}^{-3} = 2.78 \times 10^{19}\ {\rm cm}^{-3} \end{aligned}$$

We can see the large difference in the effective density between Si and GaAs.

In the case of the valence band, we have the heavy hole and light hole bands, both of which contribute to the effective density. The effective density is

$$N_v = 2\left(m_{hh}^{3/2} + m_{\ell h}^{3/2}\right)\left(\frac{k_B T}{2\pi\hbar^2}\right)^{3/2}$$

For GaAs we use $m_{hh} = 0.45 m_0, m_{\ell h} = 0.08 m_0$ and for Si we use $m_{hh} = 0.5 m_0, m_{\ell h} = 0.15 m_0$, to get

$$\begin{aligned} N_v({\rm GaAs}) &= 7.72 \times 10^{18} {\rm cm}^{-3} \\ N_v({\rm Si}) &= 9.84 \times 10^{18} {\rm cm}^{-3} \end{aligned}$$

2.8 DOPING OF SEMICONDUCTORS

To avoid leakage current in the ‘OFF’ state, semiconductor devices operate at temperatures where the intrinsic carrier density is small ($\lesssim 10^{15}$ cm^{-3}). To introduce electrons and holes in a semiconductor the material is doped with dopants. The electrons (holes) created by the dopants are used in device design.

Donors are dopants which can donate an electron to the conduction band and acceptors are dopants which can accept an electron from the valence band and thus create a hole. The donor atom replaces a host atom in the crystal and contains one (or more) extra electrons in its outer shell. The donor atom could be a pentavalent atom in Si or a Si atom on a Ga site in GaAs. Focusing on the pentavalent atom in Si, four of the valence electrons of the donor atom behave as they would in a Si atom; the remaining fifth electron now sees a positively charged ion to which it is attracted, as shown in figure 2.19. The ion has a charge of unity and the attraction is simply Coulombic suppressed by the dielectric constant of the material. The problem is now that of the hydrogen atom case, except that the electron mass is the effective mass at the bandedge. The attractive potential is

$$U(r) = \frac{-e^2}{4\pi\epsilon r} \tag{2.8.1}$$

where ϵ is the dielectric constant of the semiconductor; i.e., the product of ϵ_0 and the relative dielectric constant. In this simplification the properties of the dopant atom can be described by a simple hydrogen-like model, where the electron mass is simply the effective mass at the bandedge.

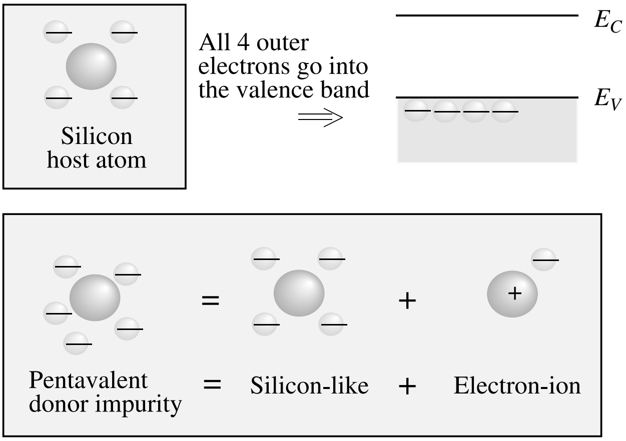

Silicon host atom
All 4 outer electrons go into the valence band
E_C
E_V
Pentavalent donor impurity
=
Silicon-like
+
Electron-ion

Figure 2.19: A schematic showing the approach we can take to understand donors in semiconductors. The donor problem is treated as the host atom problem, together with a Coulombic interaction term.

We have seen that electrons in the crystal can be represented by an effective mass near the bandedge. We get the effective mass equation for the donor level which has an energy for E_d of

$$\left[\frac{-\hbar^2}{2m_e^*}\nabla^2 - \frac{e^2}{4\pi\epsilon r}\right] F_c(r) = (E_d - E_c)F_c(r) \tag{2.8.2}$$

where m_e^* is the conduction bandedge mass and $E_d - E_c$ is the impurity energy with respect to the conduction bandedge E_c levels.

This equation is now essentially the same as that of an electron in the hydrogen atom problem. The only difference is that the electron mass is m^* and the Coulombic potential is reduced by ϵ_0/ϵ.

The energy solutions for this problem are

$$E_d = E_c - \frac{e^4 m_e^*}{2(4\pi\epsilon)^2\hbar^2}\,\frac{1}{n^2}, \quad n = 1, 2, ... \tag{2.8.3}$$

A series of energy levels are produced, with the ground state energy level being at

$$\begin{aligned} E_d &= E_c - \frac{e^4 m_e^*}{2(4\pi\epsilon)^2\hbar^2} \\ &= E_c - 13.6\left(\frac{m^*}{m_o}\right)\left(\frac{\epsilon_o}{\epsilon}\right)^2 \text{ eV} \end{aligned} \tag{2.8.4}$$

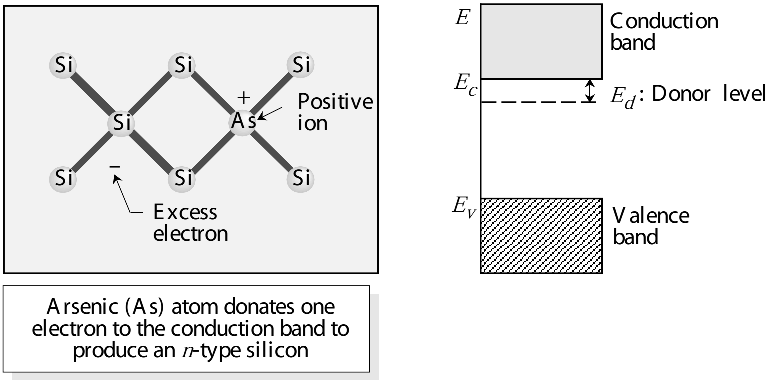


Figure 2.20: A schematic of doping of Si with arsenic (or other group V dopant). A donor level is produced below the conduction bandedge.

Note that in the hydrogen atom problem the electron levels are measured from the vacuum energy level which is taken as $E = 0$. In the donor problem, the energy level is measured from the bandedge. Figure 2.20 shows the energy level associated with a donor impurity.

The wavefunction of the ground state is as in the hydrogen atom problem

$$F_c(r) = \frac{1}{\sqrt{\pi a^3}} e^{-r/a} \tag{2.8.5}$$

where a is the donor Bohr radius and is given by

$$a = \frac{(4\pi\epsilon)\hbar^2}{m_e^* e^2} = 0.53 \left(\frac{\epsilon/\epsilon_0}{m_e^*/m_0} \right) \text{Å} \tag{2.8.6}$$

For most semiconductors the donor energies are a few meVs below the conduction bandedge and the Bohr radius is ∼100 Å.

Note that donors are defect levels, which are neutral when an electron occupies the defect level and positively charged when unoccupied. Acceptors are neutral when empty and negatively charged when occupied by an electron. The acceptor levels are produced when impurities, which have a similar core potential as the atoms in the host lattice, but have one less electron in the outermost shell, are introduced into the crystal.

As shown in figure 2.21 the acceptor impurity potential could now be considered to be equivalent to a host atom potential, together with the Coulombic potential of a negatively charged particle. The "hole" (i.e., the absence of an electron in the valence band) can then bind to the acceptor potential. The effective mass equation can again be used, since only the top of the valence band contributes to the acceptor level. The valence band problem is considerably more complex and requires the solution of multiband effective mass theory. However, the acceptor level can be reasonably predicted by using the heavy hole mass. Due to the larger hole mass, acceptor levels are usually deeper in the bandgap than donor levels.

Population of dopant levels

The presence of a dopant impurity creates a bound level E_d (or E_a) near the conduction (or valence) bandedge. If the extra electron associated with the donor occupies the donor level, it does not contribute to the mobile carrier density. The purpose of doping is to create a mobile electron or hole. When the electron associated with a donor (or a hole associated with an acceptor) is in the conduction (or valence) band, the dopant is said to be ionized. To calculate densities of electrons and holes at finite temperatures in doped semiconductors we note that carrier densities the electrons will be redistributed, but their numbers will be conserved and will satisfy the following equality resulting from charge neutrality

$$(n - n_i) + n_d = N_d \tag{2.8.7}$$

$$(p - p_i) + p_a = N_a \tag{2.8.8}$$

or

$$n + n_d = N_d - N_a + p + p_a \tag{2.8.9}$$

where

$$\begin{aligned} n &= \text{total free electrons in the conduction band} \\ n_d &= \text{electrons bound to the donors} \\ p &= \text{total free holes in the valence band} \\ p_a &= \text{holes bound to the acceptors} \end{aligned}$$

The number density of electrons attached to the donors has been derived in equation 2.4.4 and is given by

$$\frac{n_d}{N_d} = \frac{1}{\frac{1}{2}\ \exp\ \left(\frac{E_d - E_F}{k_B T}\right) + 1} \tag{2.8.10}$$

The factor $\frac{1}{2}$ essentially arises from the fact that there are two states an electron can occupy at a donor site corresponding to the two spin-states.

The probability of a hole being trapped to an acceptor level is given by

$$\frac{p_a}{N_a} = \frac{1}{\frac{1}{4}\ \exp\ \left(\frac{E_F - E_a}{k_B T}\right) + 1} \tag{2.8.11}$$

The factor of $\frac{1}{4}$ comes about because of the presence of the two bands, light hole, heavy hole, and the two spin-states.

To find the fraction of donors or acceptors that are ionized, we have to use a computer program in which the position of the Fermi level is adjusted so that the charge neutrality condition given Eq. 2.8.9 is satisfied. Once E_F is known, we can calculate the electron or hole densities in the conduction and valence bands. For doped systems, it is useful to use the Joyce–Dixon

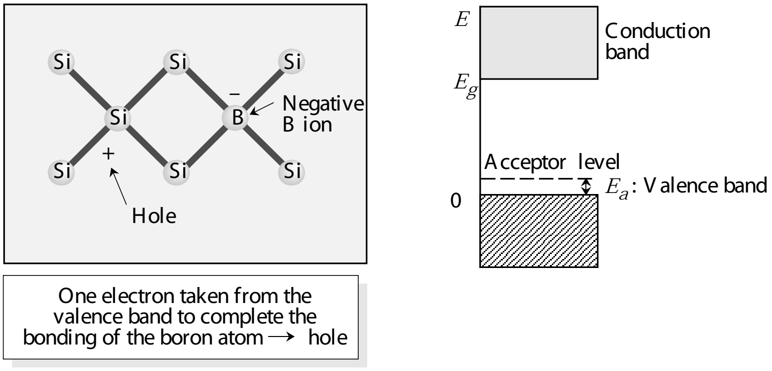


Figure 2.21: Boron has only three valence electrons. It can complete its four fold tetrahedral bonds only by taking an electron from an Si–Si bond, leaving behind a hole in the silicon valence band. The positive hole is then available for conduction.

approximation, which gives the relation between the Fermi level and the free carrier concentration. This approximation is more accurate than the Boltzmann approximation. According to the Joyce–Dixon approximation , we have

$$E_F = E_c + k_BT \left[\ln \frac{n}{N_c} + \frac{1}{\sqrt{8}} \frac{n}{N_c} \right] = E_v - k_BT \left[\ln \frac{p}{N_v} + \frac{1}{\sqrt{8}} \frac{p}{N_v} \right] \tag{2.8.12}$$

This relation can be used to obtain the Fermi level if n is specified. Or else it can be used to obtain n if E_F is known by solving for n iteratively. If the term $(n/\sqrt{8}\, N_c)$ is ignored, the result corresponds to the Boltzmann approximation.

If we examine the mobile carrier density dependence upon temperature, there are three regimes, as shown in figure 2.22 for an $n-$type material. At low temperatures, the electrons coming from the donors are attached to the donors and occupy the impurity levels E_d. Thus there is no contribution to the mobile carrier density from the dopants. This regime is called the carrier freeze out regime. At higher temperatures, the dopants ionize until most of them are ionized out over a temperature regime, the mobile carrier is essentially equal to the dopant density and independent of temperature. This is the saturation regime and semiconductor devices are operated in this regime. At very high temperatures, the intrinsic carrier density overwhelms the dopant density and the material acts as an intrinsic material.

In figure 2.23 we show experimentally measured properties of Mg in GaN (Mg acts as a deep acceptor in GaN). When the temperature is not extremely high, the hole concentration is much less than the effective acceptor concentration $N_A - N_D$, since deep acceptors are not fully ionized at lower temperatures.

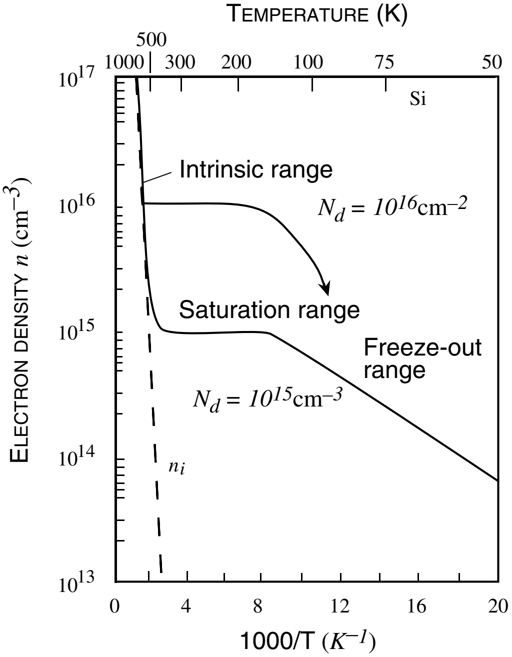


Figure 2.22: Electron density as a function of temperature for a Si sample with donor impurity concentration of 10^{15} cm^{-3}.

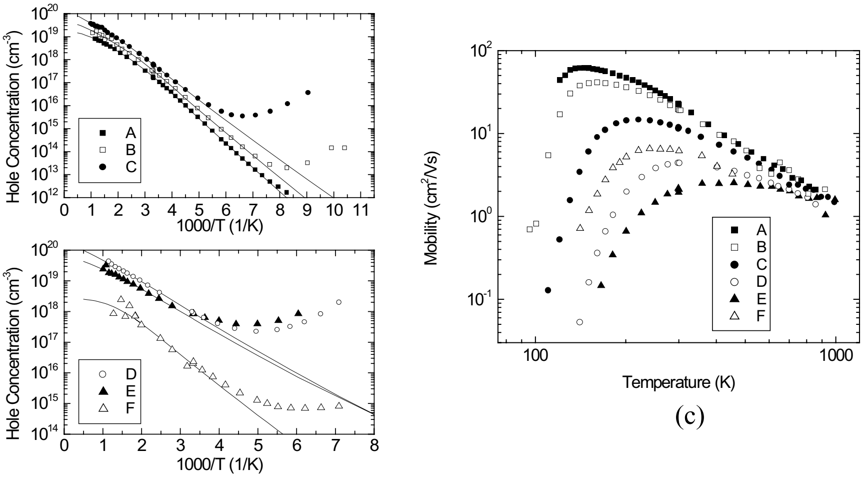
2.9 DOPING IN POLAR MATERIALS

Semiconductors such as GaN, In, and AlN are called polar materials since they can have net polarization due to a shift in the cation and anion sublattices. In unstrained zinc-blende structures the cation and anion sublattices are arranged in such a way that there is no net polarization in the material. However, in the wurtzite crystal (like InN, GaN, AlN) the arrangement of the cation and anion sublattices can be such that there is a relative movement from the ideal wurtzite position to produce a "spontaneous polarization" in the crystal which becomes very important for heterostructures. This effect is illustrated in figure 2.24. Also given in table 2.4 are the values of the spontaneous polarization which is aligned along the c-axis of the crystal.

In addition to spontaneous polarization is another phenomena which can lead to polarization in the material. Strain can cause a relative shift between the cation and anion sublattices and create net polarization in the material. This is the piezoelectric effect. In figure 2.25 we show how the movement of rows can cause polarization effect by looking at the structural arrangements of atoms in barium titanate.

Sample	Growth	Cp_2Mg flow		ΔE_A (meV)	N_A (cm^{-3})	N_D (cm^{-3})
		sccm	nmol/ min			
A	990206PB	15	24.4	190	1.8×10^{19}	1.1×10^{18}
B	990206PA	26	42.3	174	4.6×10^{19}	3.0×10^{18}
C	990205PD	47	76.5	152	1.4×10^{20}	1.2×10^{19}
D	980901PA	83	135	118	2.2×10^{20}	4.1×10^{19}
E	990205PC	140	228	112	8.6×10^{19}	2.7×10^{19}
F	980901PE	263	428	165	7.6×10^{18}	5.0×10^{18}

(a)



(c)

(b)

Figure 2.23: Measured properties of a deep acceptor: Mg doping of GaN. (a) Doping parameters for six different samples. (b) Hole concentration as a function of temperature. Notice that for all these samples, when the temperature is not extremely high, the hole concentration is much less than the effective acceptor concentration $N_A - N_D$, since deep acceptors are not fully ionized at lower temperatures. (c) Hole mobility as a function of temperature. Figures are from the PhD dissertation of Peter Kozodoy, UCSB.

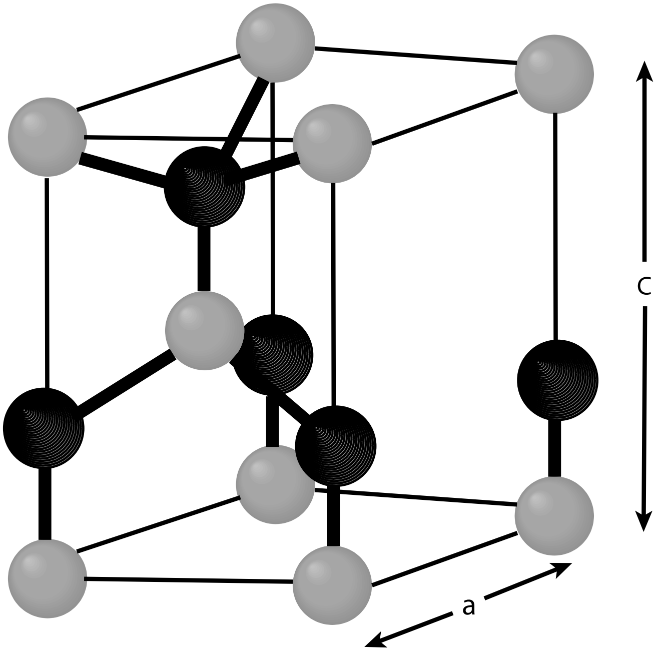


Figure 2.24: The wurtzite crystal structure unit cell. In the ideal wurtzite structure the c lattice constant is related to the a lattice constant by the relation $c = 2\sqrt{\frac{2}{3}}a$. However, when the cation-anion bond lengths cause a deviation from this relationship a net spontaneous polarization is created.

Polar Charge at Heterointerfaces

If there is a net movement of one sublattice against each other, a polarization field is set up. This results in a positive and negative polar charge. Under most conditions the polar charge on the free surfaces is neutralized by charges present in the atmosphere. This causes depolarization of the material. If, however, a heterostructure is synthesized and the two materials forming the structure have different values for the polarization, there is a net polar charge (and polarization) at the interface as shown in figure 2.30. In semiconductors this polar charge can cause a built-in electric field

$$\mathcal{E} = \frac{P}{\epsilon} \tag{2.9.1}$$

The interface charge $P_A - P_B$ and the built-in interface field (see figure 2.26) can be exploited in device design since for most applications this fixed polar charge can act as dopant (see figure 2.27 and figure 2.28).

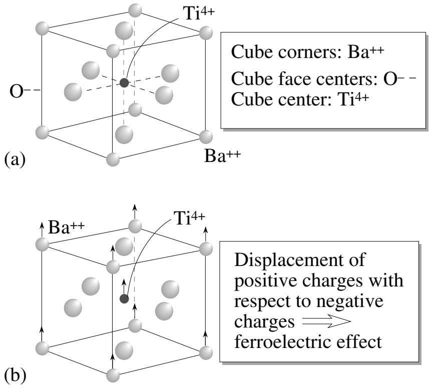

Ti4+
O−−
Ba++
(a)
Cube corners: Ba++
Cube face centers: O−−
Cube center: Ti4+
Ba++
Ti4+
(b)
Displacement of positive charges with respect to negative charges ⟹ ferroelectric effect

Figure 2.25: The structure of a typical perovskite crystal illustrated by examining barium titanate. (b) The movement of the ions leads to a ferroelectric effect.

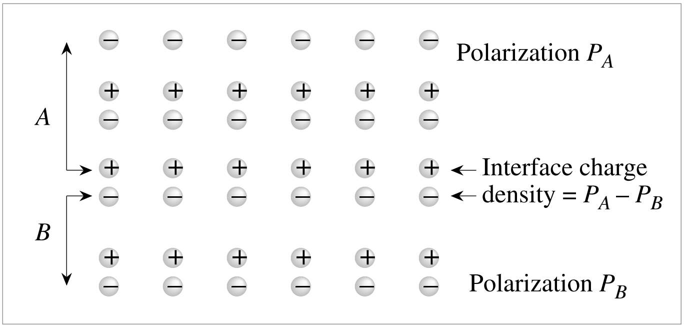


Figure 2.26: A schematic showing how interface charge density can be produced at heterointerfaces of two polar materials.

For example, in AlGaN/GaN HFETs , a fixed sheet charge is formed at the heterointerface due to the the piezoelectric polarization in the strained AlGaN, and the discontinuity in the spontaneous polarization at the interface (see figure 2.27). To screen the net positive charge at the AlGaN/GaN junction, a 2DEG is formed. The same effect can also be used to create a bulk three-dimensional electron slab, as shown in figure 2.28. This is achieved by grading from GaN to AlGaN, thus spreading the polarization-induced charge over the graded region. The polarization-induced carrier density, ρ_π, is given by the equation $\rho_\pi = \nabla \cdot \mathbf{P}$; here $\mathbf{P}$ is the total polarization in the material. Since the AlGaN composition and polarization are shown to be well-approximated by Vegard's law, any desired channel charge profile can be obtained by choosing

Zinc blende

Material	e_{14} (C/m^2)
AlAs	–0.23
GaAs	–0.16
GaSb	–0.13
GaP	–0.10
InAs	–0.05
InP	–0.04

Wurtzite (c-axis growth)

Material	e_{31} (C/m^2)	e_{33} (C/m^2)	$P_{\rm sp}$ (C/m^2)
AlN	–0.6	1.46	–0.081
GaN	–0.49	0.73	–0.029
InN	–0.57	0.97	–0.032

Table 2.4: Piezoelectric constants in some important semiconductors. For the nitrides the spontaneous polarization values are also given. (Data for zinc-blende material from S. Adachi, J. Appl. Phys. vol. 58, **R1** (1985). For nitrides see E. Bernardini, V. Fiorentini, and D. Vanderbilt, Phys. Rev. B vol. 56, **R10024** (1997).)

the appropriate grading scheme. This polarization induced channel charge can be modulated by a gate in a structure called a polarization-doped FET or PolFET can be used to tailor the g$_m$-V$_{gs}$ profile of the PolFET. This is analogous to impurity doped MESFETs, where the g$_m$-V$_{gs}$ profile is modified by dopant profile design. In figure 2.29, we show experimentally measured electrical characteristics of doped GaN, GaN 2DEG structures, and GaN 3DEG structures.

Piezoelectric Effect

As noted above, when a structure is under strain a net polarization can arise—a phenomenon called piezoelectric effect. The value of the polar charge induced by strain depends upon the strain tensor. In the previous section we have discussed the nature of the strain tensor in strained epitaxy (i.e., in the coherent growth regime).

Nitride heterostructures have polarization charges at interfaces because of strain related piezoelectric effect as well as from spontaneous polarization. For growth along (0001) orientation the strain tensor for coherently strained wurtzite crystals is given in Chapter 1. The piezoelectric polarization is related to the strain tensor by the following relation

$$P_{pz} = e_{33}\epsilon_{zz} + e_{31}(\epsilon_{xx} + \epsilon_{yy}) \tag{2.9.2}$$

Piezoelectric effect is also present in zinc blende structures. However, the piezoelectric effect only occurs when the strain tensor has off-diagonal components. The polarization values are given by

$$\begin{aligned} P_x &= e_{14}\epsilon_{yz} \\ P_y &= e_{14}\epsilon_{xz} \\ P_z &= e_{14}\epsilon_{xy} \end{aligned} \tag{2.9.3}$$

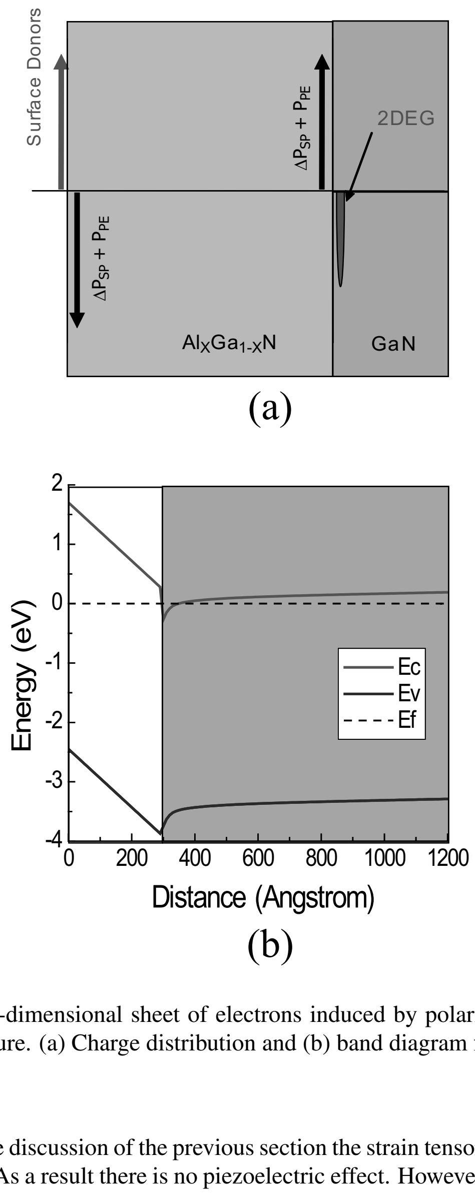
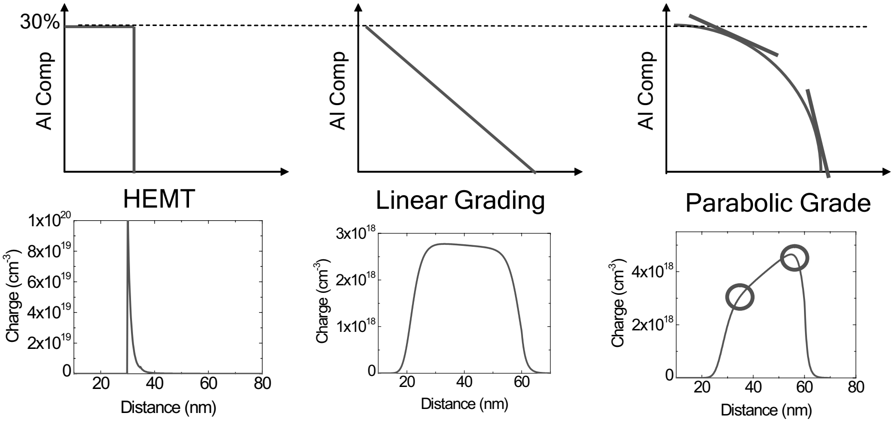


Figure 2.27: Mobile 2-dimensional sheet of electrons induced by polarization fields in an AlGaN/GaN heterostructure. (a) Charge distribution and (b) band diagram for the structure.

As can be seen from the discussion of the previous section the strain tensor is diagonal for growth along (001) direction. As a result there is no piezoelectric effect. However for other orientations, notably for (111) growth there is a strong piezoelectric effect.



(a) (b) (c)

Figure 2.28: A 3-dimensional charge distribution can be induced in polar materials via bandgap grading. (a) 2-dimensional charge distribution induced via an abrupt interface. (b) Linear grade and (c) parabolic grade result in the displayed 3-dimensional charge distributions. (Figure courtesy S. Rajan, UCSB)

Piezoelectric effect can be exploited to create interface charge densities as high as 10^{13} cm^{-2} in materials. In Table 2.4 we provide the values of piezoelectric constants for some semiconductors. In addition to the polarization induced by strain, the cation and anion sublattices are spontaneously displaced with respect to each other producing an additional polarization. For heterostructures the difference of the spontaneous polarization appears at the interfaces, as noted earlier. In Chapter 1 we have provided the values of spontaneous polarization for AlN, GaN, and InN.

EXAMPLE 2.5 A thin film of $Al_{0.3}Ga_{0.7}N$ is grown coherently on a GaN substrate. Calculate the polar charge density and electric field at the interface.

The lattice constant of $Al_{0.3}Ga_{0.7}N$ is given by Vegard's law

$$a_{all} = 0.3a_{AlN} + 0.7a_{GaN} = 3.111 \text{ Å}$$

The strain tensor is

$$\epsilon_{xx} = 0.006$$

Using the elastic constant values from Chapter 1

$$\epsilon_{zz} = -0.6 \times 0.006 = 0.0036$$

The piezoelectric effect induced polar charge then becomes

$$P_{pz} = 0.0097 \text{ C/m}^2$$

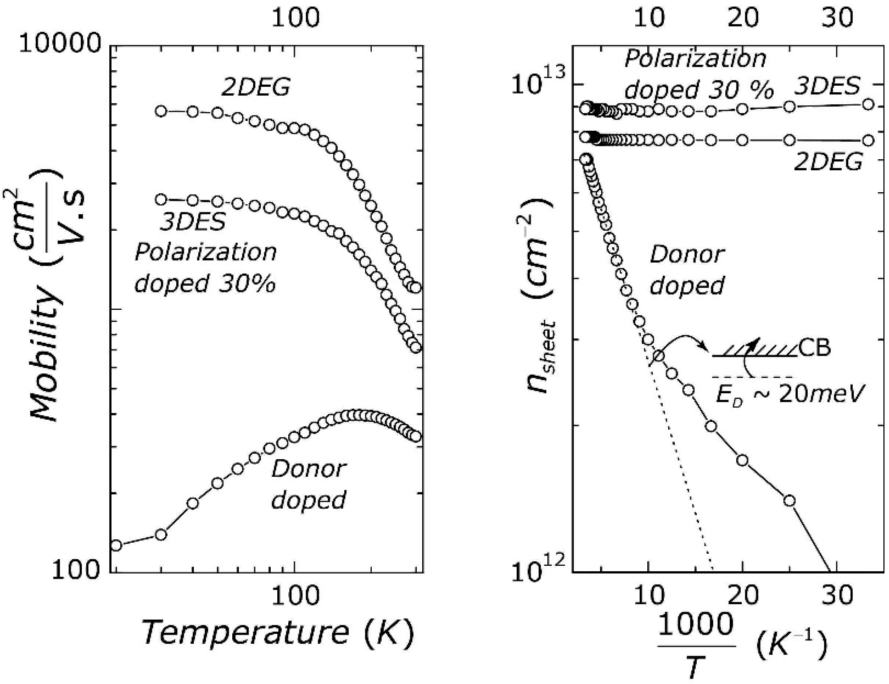


Figure 2.29: Measured electrical characteristics as a function of temperature for three different GaN samples. The sample with the lowest sheet charge is doped with Si (a shallow donor in GaN) to generate mobile electrons. In the sample with the highest sheet charge, carriers are generated by grading AlGaN from 0% Al to 30% Al, resulting in a 3-dimensional electron distribution (as in figure 2.28b and c). In the third sample, a 2DEG is generated in an AlGaN/GaN heterostructure (as in figure 2.28a). While the charge in the Si-doped sample decreases as temperature is decrease (carrier freeze-out), the charge in the other two samples remains constant. Figures courtesy of D. Jena, University of Notre Dame.

This corresponds to a density of 6.06×10^{12} cm^{-2} electronic charges.

In addition to the piezoelectric charge the spontaneous polarization charge is

$$P_{sp} = 0.3(0.089) + 0.7(0.029) - 0.029 = 0.018 \text{ C/m}^2$$

which corresponds to a density of 1.125×10^{13} cm^{-2} charges. The total charge (fixed) arising at the interface is the sum of the two charges.

2.10 TAILORING ELECTRONIC PROPERTIES

In many applications we need bandgaps or carrier properties that are not available in naturally occurring materials. It is possible to tailor electronic properties by using alloys and quantum wells.

2.10.1 Electronic properties of alloys

Alloys are made from combinations of two or more materials and can be exploited to create new bandgaps or lattice constants. In Chapter 1 we have discussed how the lattice constant of alloys changes with composition. To the first order the electronic properties are also given by a similar relation. Consider an alloy A_xB_{1-x} made from materials A with bandstructure given by $E_A(k)$ and B with bandstructure given by $E_B(k)$. The bandstructure of the alloy is then given by

$$E_{all}(k) = xE_A(k) + (1-x)E_B(k) \tag{2.10.1}$$

Note that the energy averaging is done at the same k value. If we make an alloy from a direct and an indirect material, one does not simply average the bandgaps to get the alloy bandgap. Instead the bandgaps at the same k values are averaged and the bandgap is then given by the lowest energy difference between the conduction and valence energies.

Based on the equation above the effective mass of the alloy is to be averaged as

$$\frac{1}{m^*_{all}} = \frac{x}{m^*_A} + \frac{(1-x)}{m^*_B} \tag{2.10.2}$$

It is important to note that alloys have inherent disorder since they have random arrangements of atoms. This leads to disorder related scattering discussed in the next chapter.

2.10.2 Electronic properties of quantum wells

Quantum wells offer a very useful approach to bandstructure tailoring. In Section 2.2 we have discussed electronic properties in quantum wells. In quantum wells electrons behave as if they are in a 2-dimensional space and acquire properties that are especially useful for many electronic and optoelectronic applications.

When two semiconductors with different bandgaps (and chemical compositions) form an interface, We need to know how does the conduction band (valence band) on one material line up with the other materials bands? This information is usually obtained through experiments. There are three possible scenarios as shown in figure 2.30. In type I structures the layer bandgap material “surrounds” the bandgap of the small gap material. In quantum wells made from such materials, both electrons and holes are confined in the same physical quantum well. Most electronic and optoelectronic devices are based on type I lineup. In type II lineup the conduction band of material A is below that of the material B, but the valence band of A is above that of B as shown. In quantum wells made from such materials the electrons and holes are confined in spatially different quantum wells. These structures are useful for applications in the long wavelength regime, since their “effective” bandgap can be very small. Finally, in type III

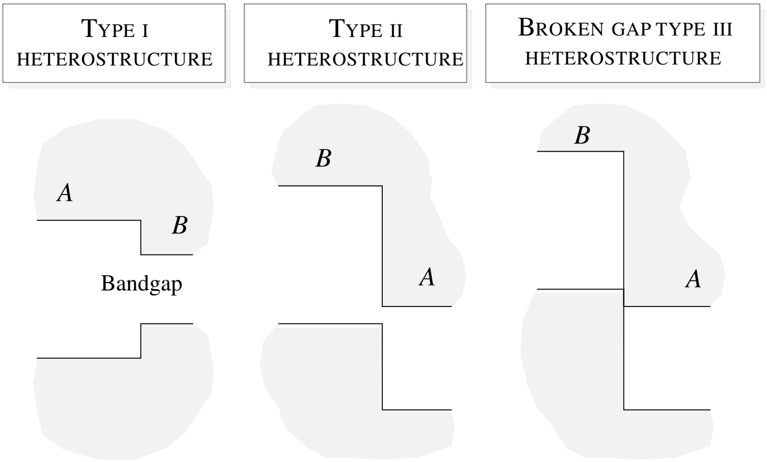


Figure 2.30: Various possible bandedge lineups in semiconductors A and B.

heterostructures, both the conduction and valence band edges of material A are above the conduction band edge of material B. In figure 2.31 we show bandlineups for a number of different material systems.

In figure 2.32 we show a schematic of a type I quantum well made from a smaller bandgap material B sandwiched between a large bandgap material A. To understand the electronic properties of the quantum well we use the effective mass approach and the discussion of Section 2.2. The key difference in semiconductor quantum wells is that we need to use the effective mass instead of the free electron mass.

The confinement of electrons and holes by quantum wells alters the electronic properties of the system. This has important consequences for optical properties and optoelectronic devices. In an infinite quantum well the confined energies are

$$E_n = \frac{\pi^2 \hbar^2 n^2}{2m^* W^2} \tag{2.10.3}$$

The energy of the electron bands are then

$$E = E_n + \frac{\hbar^2 k_{\|}^2}{2m^*} \tag{2.10.4}$$

The two-dimensional quantum well structure thus creates electron energies that can be described by subbands $(n = 1, 2, 3 \cdots)$. The subbands for the conduction band and valence band are shown schematically in figure 2.33.

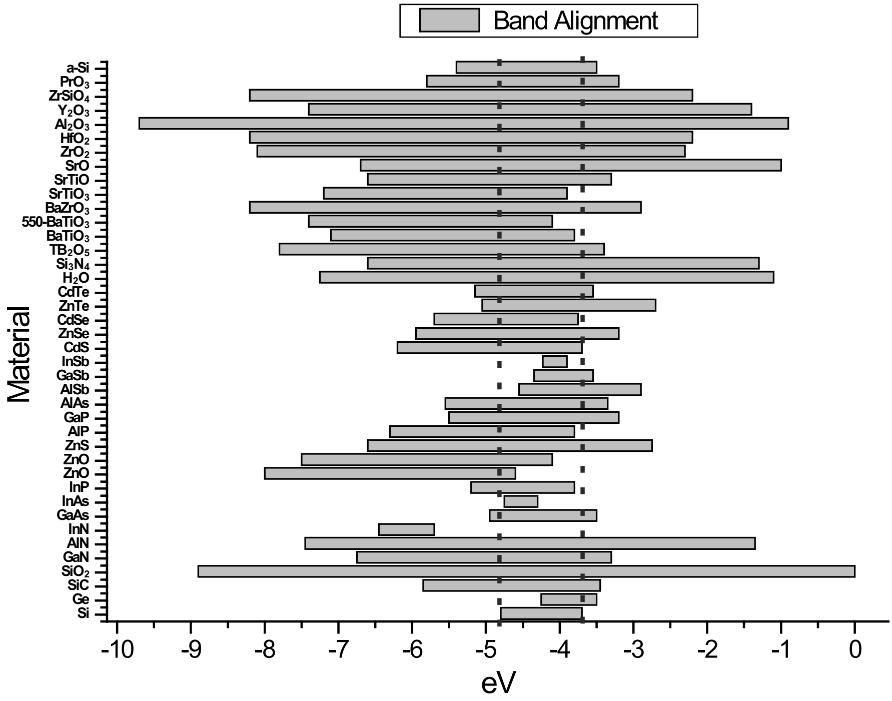


Figure 2.31: Bandedge lineups in a variety of materials.

If the barrier potential V_c is not infinite, the wavefunction decays exponentially into the barrier region, and is a sine or cosine function in the well. By matching the wavefunction and its derivative at the boundaries one can show that the energy and the wavefunctions are given by the solution to the transcendental equations (see Section 2.2)

$$\begin{aligned} \alpha \tan \frac{\alpha W}{2} &= \beta \\ \alpha \cot \frac{\alpha W}{2} &= -\beta \end{aligned} \qquad (2.10.5)$$

where

$$\begin{aligned} \alpha &= \sqrt{\frac{2m^* E}{\hbar^2}} \\ \beta &= \sqrt{\frac{2m^*(V_c - E)}{\hbar^2}} \end{aligned}$$

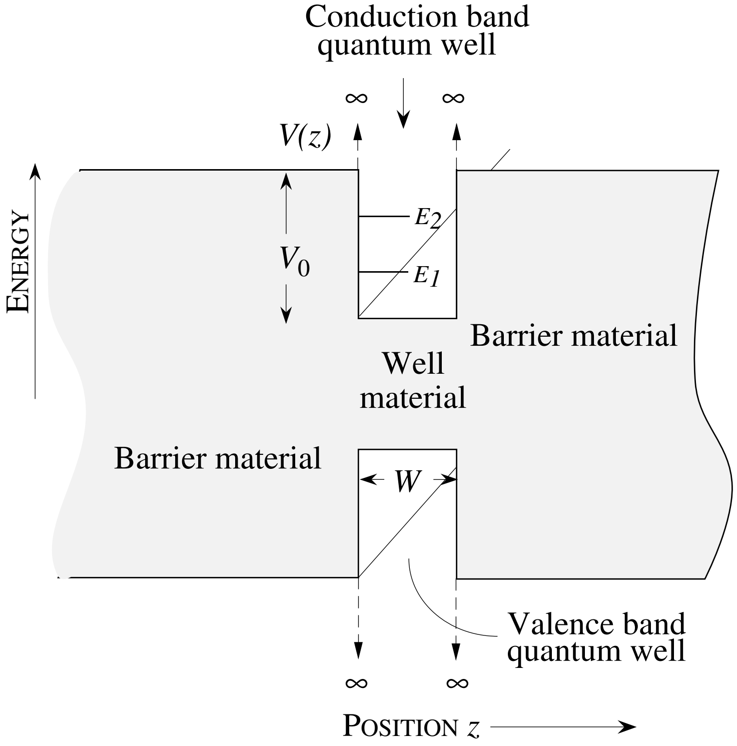


Figure 2.32: A schematic of a quantum well formed for the electron and holes in a heterostructure.

These equations can be solved numerically. The solutions give the energy levels E_1, E_2, E_3 . . . and the wavefunctions, $f_1(z), f_2(z), f_3(z), \cdots$.

Each level E_1, E_2, etc., is actually a subband due to the electron energy in the x–y plane. As shown in figure 2.33 we have a series of subbands in the conduction and valence band. In the valence band we have a subband series originating from heavy holes and another one originating from light holes.

The subband structure has important consequences for the optical and transport properties of heterostructures. An important manifestation of this subband structure is the density of states of the electronic bands. The density of states figures importantly in both electrical and optical properties of any system. In Section 2.3 we have discussed how dimensionality alters the density of states.

The density of states in a quantum well is

- Conduction band

$$N(E) = \sum_i \frac{m^*}{\pi\hbar^2}\theta(E - E_i) \tag{2.10.6}$$

where θ is the heavyside step function (unity if $E > E_i$; zero otherwise) and E_i are the subband energy levels.

- Valence band

$$N(E) = \sum_i \sum_{j=1}^{2} \frac{m_j^*}{\pi\hbar^2}\theta(E_{ij} - E) \tag{2.10.7}$$

where i represents the subbands for the heavy hole $(j = 1)$ and light holes $(j = 2)$. The density of states is shown in figure 2.33 and has a staircase-like shape.

The differences between the density of states in a quantum well and a three-dimensional semiconductor is one of the important reasons why quantum wells are useful for optoelectronic devices. The key difference is that the density of states in a quantum well is large and finite at the effective bandedges (lowest conduction subband and highest valence subband). As a result the carrier distribution is highest at the bandedges.

The relationship between the electron or hole density (areal density for 2D systems) and the Fermi level is different from that in three-dimensional systems because the density of states function is different. The 2D electron density in a single subband starting at energy E_1^e is

$$\begin{aligned} n &= \frac{m_e^*}{\pi\hbar^2}\int_{E_1^e}^{\infty} \frac{dE}{\exp\left(\frac{E-E_F}{k_BT}\right)+1} \\ &= \frac{m_e^* k_BT}{\pi\hbar^2}\left[\ln\left\{1+\exp\left(\frac{E_F - E_1^e}{k_BT}\right)\right\}\right] \\ \text{or} \qquad E_F &= E_1^e + k_BT\ln\left[\exp\left(\frac{n\pi\hbar^2}{m_e^* k_BT}\right) - 1\right] \end{aligned} \tag{2.10.8}$$

If more than one subband is occupied we can add their contribution similarly. For the hole density we have (considering both the HH and LH ground state subbands)

$$p = \frac{m_{hh}^*}{\pi\hbar^2}\int_{E_1^{hh}}^{-\infty} \frac{dE}{\exp\left(\frac{E_F-E}{k_BT}\right)+1} + \frac{m_{\ell h}^*}{\pi\hbar^2}\int_{E_1^{\ell h}}^{-\infty} \frac{dE}{\exp\left(\frac{E_F-E}{k_BT}\right)+1} \tag{2.10.9}$$

where m_{hh}^* and $m_{\ell h}^*$ are the in-plane density of states masses of the HH and LH subbands. We then have

$$\begin{aligned} p &= \frac{m_{hh}^* k_BT}{\pi\hbar^2}\left[\ln\left\{1+\exp\frac{(E_1^{hh} - E_{Fp})}{k_BT}\right\}\right] \\ &+ \frac{m_{\ell h}^* k_BT}{\pi\hbar^2}\left[\ln\left\{1+\exp\frac{(E_1^{\ell h} - E_{Fp})}{k_BT}\right\}\right] \end{aligned} \tag{2.10.10}$$

$$\text{If} \qquad E_1^{hh} - E_1^{\ell h} > k_BT$$

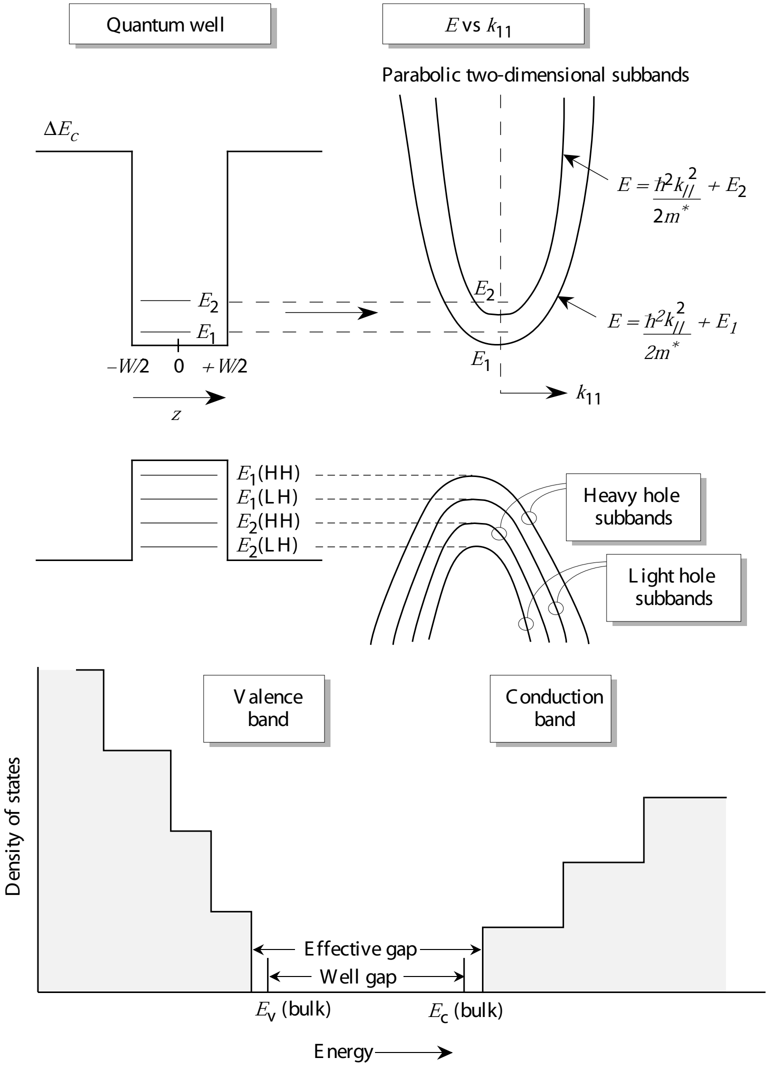


Figure 2.33: Schematic of density of states in a 3–, 2– and 1–dimensional system . The subband structure is shown.

The occupation of the light hole subband can be ignored.

In many electronic devices used for information processing, a quantum well with a "triangular" shape is produced. The potential for electrons may be written in the form

$$\begin{aligned} V(x) &= \infty \quad x < 0 \\ &= e\tilde{E}x \quad x > 0 \end{aligned} \tag{2.10.11}$$

The potential energy of the particle is of the form

$$V(x) = Fx + \text{constant} \tag{2.10.12}$$

where F is the force on the particle (say, an electron) and has a value $e\mathcal{E}$. We choose the constant in the potential energy to be such that at $x = 0, V(x) = \mathcal{E}x$ as shown in figure 2.34. The solutions to this problem are the Airy functions

$$\Phi(\xi) = \frac{1}{\sqrt{\pi}} \int_0^\infty \cos\left(\frac{u^3}{3} + u\xi\right) du \tag{2.10.13}$$

with a normalized solution

$$\psi(\xi) = A\Phi(\xi) \tag{2.10.14}$$

The normalization constant can be shown to have the value

$$A = \frac{(2m)^{1/3}}{\pi^{1/2}\mathcal{E}^{\frac{1}{6}}\hbar^{2/3}} \tag{2.10.15}$$

The Airy functions have the following asymptotic behavior:

$$\begin{aligned} \Phi(\xi) &\sim \frac{1}{2}(\xi)^{-1/4}\exp\left(-\frac{2\xi^{3/2}}{3}\right), \quad \xi > 0 \\ \Phi(\xi) &\sim \mid \xi \mid^{-1/4} \sin\left(\frac{2 \mid \xi \mid^{3/2}}{3} + \frac{\pi}{4}\right), \quad \xi < 0 \end{aligned} \tag{2.10.16}$$

Note that at $x = 0$ the second form is to be used, since $\xi < 0$.

The solutions for the energy levels turn out to be:

$$E_n = \left(\frac{\hbar^2}{2m}\right)^{1/3} \left(\frac{3}{2}\pi\mathcal{E}\right)^{2/3} \left(n - \frac{1}{4}\right)^{2/3}, \quad n = 1, 2, \ldots \tag{2.10.17}$$

As shown in figure 2.35, in electronic devices such as a MOSFET or a MODFET the device consists of an insulator-semiconductor junction. Electrons are injected at the interface on the semiconductor side by a controlling electrode (the gate). The free charge causes a bending of the semiconductor band to produce an approximately triangular quantum well, as shown. The triangular quantum well is defined by an electric field $\mathcal{E}_s$ which is related to the areal charge density by Gauss's law

$$\mathcal{E}_s = \frac{en_s}{\epsilon_s} \tag{2.10.18}$$

As a result of the confinement, quantized energy levels are formed in the triangular well. Approximate positions of these levels can be obtained from the results given above.

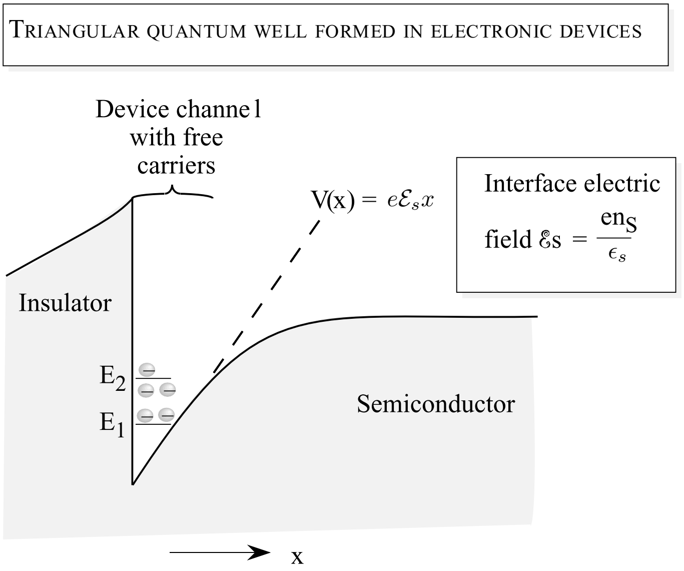

TRIANGULAR QUANTUM WELL FORMED IN ELECTRONIC DEVICES
Device channe l
with free
carriers
V(x) = $e\mathcal{E}_s x$
Interface electric
field $\mathcal{E}$s = $\frac{en_S}{\epsilon_s}$
Insulator
E_2
E_1
Semiconductor
x

Figure 2.34: A schematic of free electrons (conduction electrons) in a semiconductor device confined to an approximately triangular quantum well.

2.11 STRAINED HETEROSTRUCTURES

As noted in chapter 1 it is now possible to incorporate strain into an epitaxial film. In fact, strain of a few percent can be built-in simply by growing a film on a mismatched substrate. ne of the most important strained heterostructure is the SiGe/Si structure. This system is compatible with Si based technology since it uses Si substrates. Due the modifications in the bandstructure high performance SiGe electronic devices can be made. Other important strained structures are InGaAs grown on GaAs or InP substrates and the AlGaN/GaN structure.

Once the strain tensor is known, we are ready to apply the deformation potential theory to calculate the effects of strain on various eigenstates in the Brillouin zone. The strain perturbation Hamiltonian is defined and its effects are calculated in the simple first order perturbation theory. In general we have

$$H_\epsilon^{\alpha\beta} = \sum_{ij} D_{ij}^{\alpha\beta} \epsilon_{ij} \tag{2.11.1}$$

where D_{ij} is the deformation potential operator which transforms under symmetry operations as a second rank tensor. $D_{ij}^{\alpha\beta}$ are the matrix elements of D_{ij}.

The built in strain causes several different effects on electronic properties: i) It can lift the degeneracies or band edges; ii) it can change the bandgap; iii) it can alter effective masses. To calculate the effect of strain one uses perturbation theory using equation 2.11.1. we will summarize the relevant equations for a direct gap conduction band, an indirect gap X-valley conduction bandedge and for the valence bands.

Case 1: Let us first examine how strain influences the bottom of the non degenerate $\Gamma_2^{'}$ state which represents the conduction bandedge of direct bandgap semiconductors. This state is an s-type state and has the full cubic symmetry associated with it. The effect of the strain is to produce a shift in energy.

$$\begin{aligned} \delta E^{(000)} &= H_\epsilon \\ &= D_{xx}(\epsilon_{xx} + \epsilon_{yy} + \epsilon_{zz}) \end{aligned} \tag{2.11.2}$$

Conventionally we write

$$D_{xx} = \Xi_d^{(000)} \tag{2.11.3}$$

where $\Xi_d^{(000)}$ represents the dilation deformation potential for the conduction band (000) valley.

Case 2: In this next case we will examine indirect gap materials like Si which have the conduction bandedge along the (100) and equivalent directions. The bandedges are shifted according to the following equations.

$$\delta E^{(100)} = \Xi_d^{(100)}(\epsilon_{xx} + \epsilon_{yy} + \epsilon_{zz}) + \Xi_u^{(100)}\epsilon_{xx} \tag{2.11.4}$$

By symmetry we can write

$$\delta E^{(010)} = \Xi_d^{(100)}(\epsilon_{xx} + \epsilon_{yy} + \epsilon_{zz}) + \Xi_u^{(100)}\epsilon_{yy} \tag{2.11.5}$$

$$\delta E^{(001)} = \Xi_d^{(100)}(\epsilon_{xx} + \epsilon_{yy} + \epsilon_{zz}) + \Xi_u^{(100)}\epsilon_{zz} \tag{2.11.6}$$

We note that if the strain tensor is such that the diagonal elements are unequal (as is the case in strained epitaxy), the strain will split the degeneracy of the six valleys in Si. This occurs in the SiGe/Si structures so that the 6-fold degenerate valleys split into 2-fold and 4-fold valleys. The amount of splitting will be given later in this section.

Case 3: The triple degenerate states describing the valence bandedge.

The valence band states are defined (near the bandedge) by primarily p_x, p_y, p_z (denoted by x,y,z) basis states. We have already discussed the strain tensor in epitaxial growth. For (001) growth which has been the main growth direction studied because of its compatibility with technology of processing we have

$$\begin{aligned} \epsilon_{xx} = \epsilon_{yy} &= \epsilon \\ \epsilon_{zz} &= -\frac{2c_{12}}{c_{11}}\epsilon \end{aligned} \tag{2.11.7}$$

The effects of the strain can be shown to be like heavy hole and light hole degeneracy at the valence bandedge. This also causes the hole mass to become smaller. For the In$_y$Ga$_{1-y}$As system the separation between the HH and LH state is given by $\delta = -5.966\epsilon$ eV. The effect of strain on bandstructure for both conduction band and valence band states is illustrated by examining the direct bandgap material $\text{In}_x\text{Ga}_{1-x}\text{As}$ grown on GaAs and the indirect bandgap material $\text{Ge}_x\text{Si}_{1-x}$ alloy grown on Si. For direct bandgap materials conduction bands, the strain tensor only moves the position of the bandedge and has a rather small effect on the carrier mass.

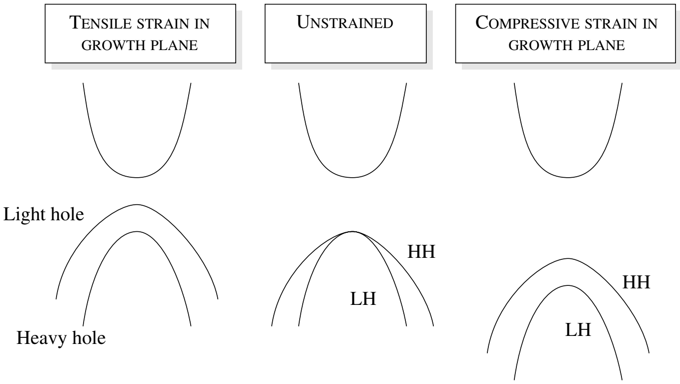

TENSILE STRAIN IN GROWTH PLANE
UNSTRAINED
COMPRESSIVE STRAIN IN GROWTH PLANE
Light hole
Heavy hole
HH
LH
HH
LH

Figure 2.35: Effect of strain on bandedges of a direct bandgap material. Due to the epitaxial strain, the valence band degeneracy is lifted.

In figure 2.35 we show a schematic of how strain in a layer grown along the (001) direction influences the bandedges in a direct gap semiconductor. The conduction bandedge moves up or down with respect to its unstrained position as discussed earlier, but since it is a non-degenerate state there is no splitting. The valence bandedge is degenerate in the unstrained system. This degeneracy is lifted by quantum confinement even in an unstrained quantum well, but the splitting produced by quantum confinement is usually small ($\sim$ 10–15 meV). Under biaxial compressive strain the bandgap of the material increases and the HH and LH degeneracy is lifted. The splitting can easily approach 100 meV making strain an important resource to alter valence band density of states. Under biaxial compressive strain the HH state is above the LH state, while under biaxial tensile strain the LH state is above the HH state, as shown in figure 2.35.

In the case of the indirect bandgap $\text{Si}_{1-x}\text{Ge}_x$ alloy grown on Si, the conduction band also is significantly affected according to equation 2.11.4 through equation 2.11.6. For (001) growth there is splitting in the 6 equivalent valleys.The results on the bandedge states are shown in figure 2.36. Note that the biaxial compressive strain causes a lowering of the four-fold in-plane valleys below the 2 two-fold out of plane valleys. We see that the bandgap of SiGe falls rapidly as Ge is added to Si. This makes the SiGe very useful for Si/SiGe heterostructure devices such as heterojunction bipolar transistors. the splitting of the HH, LH and SO bands also cause a sharp reduction in the density of states mass near the bandedge. The splitting of the conduction bandedge valleys also reduces the conduction band density of states in SiGe.

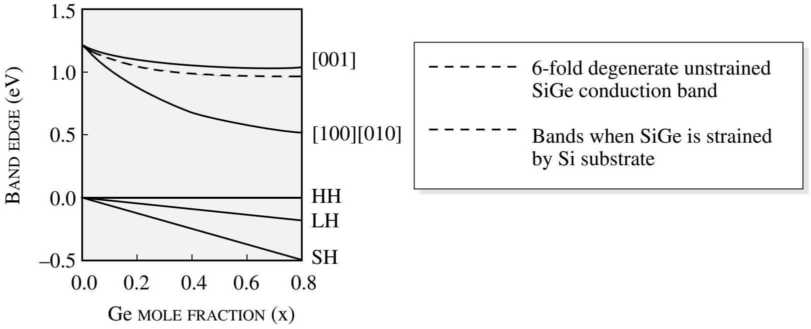


Figure 2.36: Epitaxial strain induced splittings of the conduction band and valence band as a function of alloy composition for $\mathrm{Si}_{1-x}\mathrm{Ge}_x$ grown on (001) Si. UCB: unstrained conduction band, HH: heavy hole, LH: light hole, SH: split-off hole.

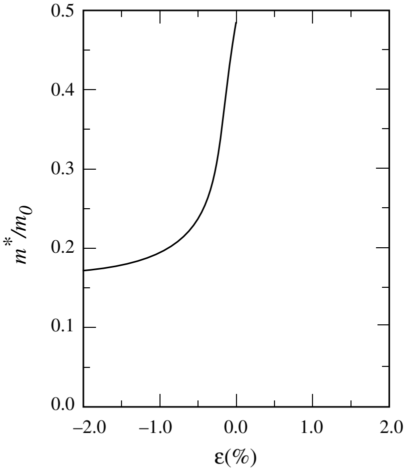


Figure 2.37: Change in the density of states mass at the valence bandedges as a function of strain for the Al$_0$.3Ga$_0$.7As/InGaAs system.

2.12 DEFECT STATES IN SOLIDS

The band theory discussed in this chapter is valid only for perfect crystals. Even in good-quality crystals there are defects, which break the periodicity of the structure. Typical defects in crystalline materials are: (i) defects in the structure arise from missing atoms (vacancies), atoms at the wrong sites, unintended impurities, etc. (ii) We may also have dislocations at surfaces of a crystal the arrangement of atoms does not have the same periodicity as in the bulk. (iii) We could also have absorbed atoms or molecules at the surface; disordered solids such as amorphous or polycrystalline materials.

Defects and surface states

In figure 2.38 we show a schematic of a perfectly periodic material and one with a defect. A deep potential region indicates the region of defect. In the case of the periodic system we have seen the electrons see a bandedge and are described by simple a effective mass equations near the bandedge. There are no allowed states in the bandgap region. In the case of a defect the deep level causes new electronic states, which can have energies in the bandgap.

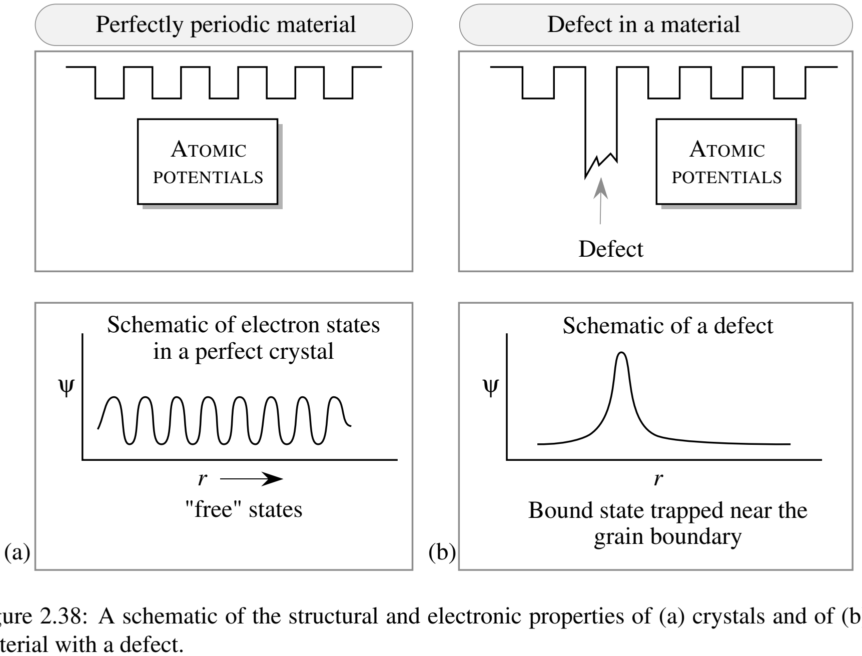
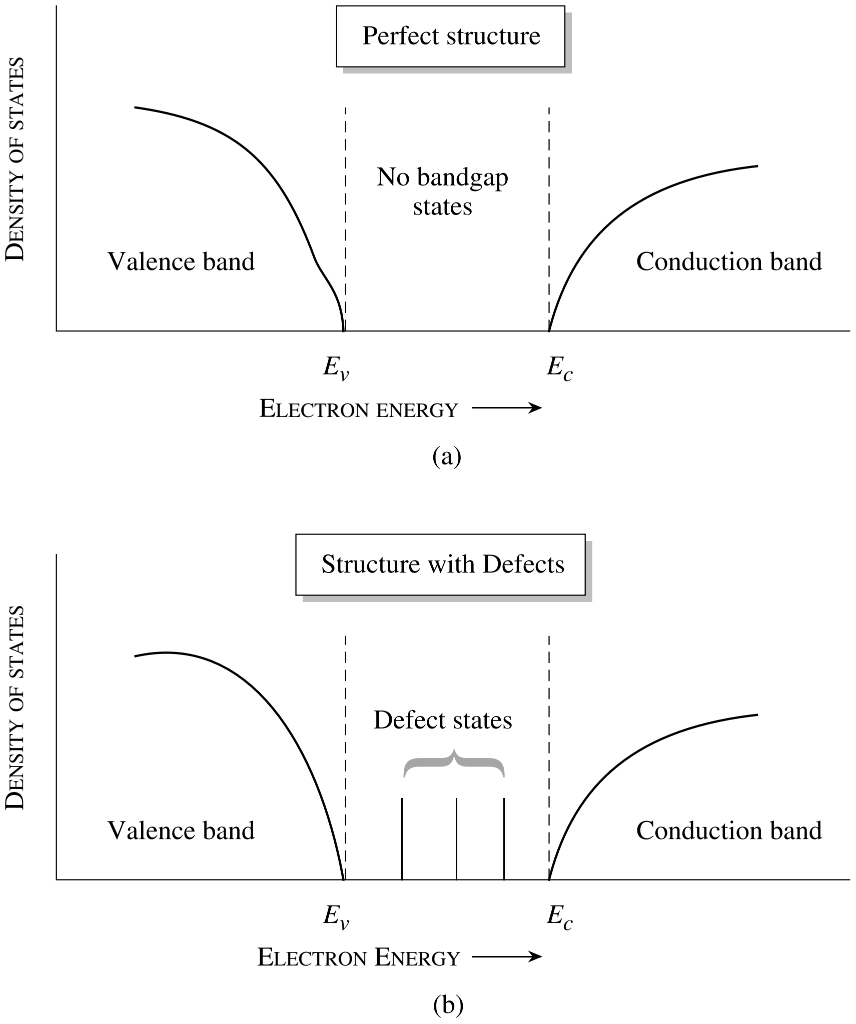

Perfectly periodic material
Defect in a material
ATOMIC POTENTIALS
ATOMIC POTENTIALS
Defect
Schematic of electron states in a perfect crystal
ψ
r
"free" states
(a)
Schematic of a defect
ψ
r
Bound state trapped near the grain boundary
(b)

Figure 2.38: A schematic of the structural and electronic properties of (a) crystals and of (b) a material with a defect.



(a)

(b)

Figure 2.39: Schematic of density of states (a) in a perfectly periodic solid and (b) in a material with defects. The presence of bangap states influences semiconductor devices.

The key difference between electronic states in the perfect crystal and a non-perfect crystal is related to the wavefunction. In the periodic state, the electron state is extended over the entire system, as shown in figure 2.38a. This reflects the fact that the electron can propagate from one region to another. In the case of a defect a bandgap state may be created with an associated wavefunction that is spatially localized near the defect region, as shown in figure 2.38b. When an electron is occupying such a localized state its transport (mobility, diffusion) properties are seriously affected. Localized electrons cannot move across the material as easily.

In figure 2.39 we show a comparison of the density of states in a perfectly periodic and of a defect-containing material. In the case of the perfect material we have a well-defined bandgap, while in the presence of defects we have bandgap states. Electrons can be trapped into the bandgap states (hence these states are also called traps).

2.13 TECHNOLOGY ISSUES

We have examined some of the driving forces behind some of the technologies. The use of alloys and heterostructures adds a tremendous versatility to the available parameter space to exploit. Semiconductor alloys are already an integral part of many advanced technology systems. Consider the following examples.

- The HgCdTe alloy is the most important high-performance imaging material for long wavelength applications (10 – 14 μm). These applications include night vision, seeing through fog, thermal imaging of the human body parts for medical applications, and a host of special purpose applications involving thermal tracking.

- The AlGaAs alloy is an important ingredient in GaAs/AlGaAs heterostructure devices which drive a multitude of technologies including microwave circuits operating up to 100 GHz, lasers for local area networks, and compact disc players.

- InGaAs and InGaAsP alloy systems are active ingredients of MMICs operating above 100 GHz and long-haul optical communication lasers.

While alloys are important ingredients of many technologies, it must be emphasized again that alloys are not perfectly periodic structures. This results in random potential fluctuations which leads to an important scattering mechanism that limits certain performances. For example, the low temperature low field mobility is severely affected by alloy scattering as is the exciton line width of optical modulators. The growth and fabrication issues in alloy systems are also sometimes serious due to miscibility gaps that may be present.

2.14 PROBLEMS

Problem 2.1 Plot the conduction band and valence band density of states in Si and GaAs from the bandedges to 0.5 eV into the bands. Use the units eV^{-1} cm^{-3}. Use the following data:

$$\begin{aligned}
\text{Si}\ :\ m_1^* &= m_\ell^* = 0.98\,m_0\\
m_2^* &= m_3^* = m_t^* = 0.19\,m_0\\
m_{hh}^* &= 0.49\,m_0\\
m_{\ell h}^* &= 0.16\,m_0\\
\text{GaAs}\ :\ m_e^* &= m_{dos}^* = 0.067\,m_0\\
m_{hh}^* &= 0.45\,m_0\\
m_{\ell h}^* &= 0.08\,m_0
\end{aligned}$$

The wavevector of a conduction band electron in GaAs is $k = (0.1, 0.1, 0.0)$ Å^{-1}. Calculate the energy of the electron measured from the conduction bandedge.

Problem 2.2 Calculate the lattice constant, bandgap, and electron effective mass of the alloy In$_x$Ga$_{1-x}$As as a function of composition from $x = 0$ to $x = 1$.

Problem 2.3 Calculate the effective density of states at the conduction and valence bands of Si GaAs, and GaN at 77 K, 300 K, and 500 K.

Problem 2.4 Estimate the intrinsic carrier concentration of diamond at 700 K. You can assume that the carrier masses are similar to those in Si. Compare the results with those for GaAs, Si, SiC and GaN.

Problem 2.5 Estimate the change in intrinsic carrier concentration per K change in temperature for InAs, Si, and GaAs at near room temperature.

Problem 2.6 Calculate the Fermi energy and Fermi velocity for the following metals: Ag, Au, Ca, Cs, Cu, Na.

Problem 2.7 Calculate the change in the Fermi level as temperature changes from 0 to 1000 K for Al and Cu.

Problem 2.8 Consider a donor an energy $\mathrm{E_D}$ from the conduction band as shown in figure 2.40. If the density of the donor device is $\mathrm{N_D}$(cm^{-3}) derive a relationship for the position of the fermi level as a function of temperature in terms of $\mathrm{N_C}$ and $\mathrm{N_V}$. Plot the fermi level as a function of temperature for the case $\mathrm{N_D} = \mathrm{N_A} = \mathrm{N_V}$. Physically explain your result. Repeat for the case of a donor and an acceptor of densities $\mathrm{N_D}$ and $\mathrm{N_A}$ respectively. What will be the dependence of the fermi level on temperature if (i) $\mathrm{N_D} = \mathrm{N_A}$, (ii) $\mathrm{N_D} > \mathrm{N_A}$, and (iii) $\mathrm{N_D} < \mathrm{N_A}$. Explain.

Problem 2.9 Consider a slab of GaAs that is doped n-type with $10^{17}cm^{-3}$.

1. Consider the case where there is a surface donor state 0.5 eV from the conduction band. What is the fermi level at the surface as the density of this level is increased from $10^{10}cm^{-2}$ to $10^{14}cm^{-2}$?

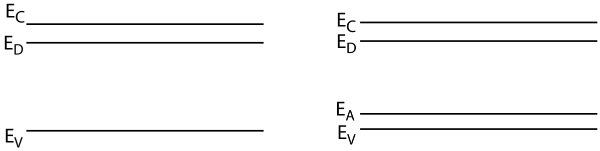


Figure 2.40: Figure for problem 2.8.

2. Solve the previous part for the case with only an acceptor state 0.5 eV from the conduction band.
3. Assume now that there are two defect levels of equal density, one donor-like and the other acceptor-like, at the surface. The acceptor state is 0.3 eV from the conduction band edge and the donor state is 0.5 eV from the conduction band edge. How does the fermi level pinning at the surface change as the areal density of each of these states is kept equal and increased from $10^{10} cm^{-2}$ to $10^{14} cm^{-2}$?
4. Now the positions of the defect levels are changed. The acceptor state is 0.5 eV from the conduction band edge and the donor state is 0.3 eV from the conduction band edge. How does the fermi level pinning at the surface change as the density of each of these states is kept equal and increased from $10^{10} cm^{-2}$ to $10^{14} cm^{-2}$?
5. Metals X and Y are now evaporated on the surface with $10^{13} cm^{-2}$ donor states at 0.5 eV from the conduction band. Find the position of the fermi level at the surface for metal X($\Phi_{ms} = 0.3eV$) and metal Y($\Phi_{ms} = 0.7eV$).
6. Repeat part 5 but with acceptor states this time, assuming they have the same energy level and areal density.

Draw band diagrams to explain your solutions.

Problem 2.10 Assume a pn junction with an acceptor close to the valence band edge, so that the acceptors are fully ionized at 300K. Assume $N_A = N_D = 10^{18} cm^{-3}$. What is the built-in voltage of the junction? Now, the choice of acceptor is changed such that only 1/10th of the acceptors are ionized.

1. What is the acceptor level relative to the valence band?
2. What is the new built-in voltage of the diode. Make reasonable approximations which *should* be justified.
3. Draw a band diagram of the system showing the acceptor level and the Fermi level.

Problem 2.11 Using Vegard's law for the lattice constant of an alloy (i.e., the lattice constant is the weighted average) find the bandgaps of alloys made in InAs, InP, GaAs, GaP which can be lattice matched to InP.

Problem 2.12 For long-haul optical communication, the optical transmission losses in a fiber dictate that the optical beam must have a wavelength of either 1.3 μm or 1.55 μm. Which alloy combinations lattice matched to InP have a bandgap corresponding to these wavelengths?

Problem 2.13 Calculate the composition of Hg$_x$Cd$_{1-x}$Te which can be used for a night vision detector with bandgap corresponding to a photon energy of 0.1 eV. Bandgap of CdTe is 1.6 eV and that of HgTe is -0.3 eV at low temperatures around 4 K.

Problem 2.14 In the In$_{0.53}$Ga$_{0.47}$As/InP system, 40% of the bandgap discontinuity is in the conduction band. Calculate the conduction and valence band discontinuities. Calculate the effective bandgap of a 100 Å quantum well. Use the infinite potential approximation and the finite potential approximation and compare the results.

Problem 2.15 In an n-type Si crystal the doping changes abruptly from $N_D = 10^{15}$ to $N_D = 10^{17}$. Make a qualitative sketch of the band diagram. Calculate

1. the built-in potential at the n^+/n^- interface, in eV. Also calculate how much of the band-bending occurs on each side of the junction,
2. the electric field at the n^+/n^- interface and
3. the electron concentration at the n^+/n^- interface.
   Assume $T = 300K$.

Problem 2.16 Calculate the first and second subband energy levels for the conduction band in a GaAs/Al$_{0.3}$Ga$_{0.7}$As quantum well as a function of well size. Assume that the barrier height is 0.18 eV.

Problem 2.17 Calculate the width of a GaAs/AlGaAs quantum well structure in which the effective bandgap is 1.6 eV. The effective bandgap is given by

$$E_g^{eff} = E_g(\text{GaAs}) + E_1^e + E_1^h$$

where E_g (GaAs) is the bandgap of GaAs (= 1.5 eV) and E_1^e and E_1^h are the ground state energies in the conduction and valence band quantum wells. Assume that $m_e^* = 0.067\, m_0, m_{hh}^* = 0.45\, m_0$. The barrier heights for the conduction and valence band well is 0.2 eV and 0.13 eV, respectively.

Problem 2.18 Assume that a particular defect in silicon can be represented by a three-dimensional quantum well of depth 1.5 eV (with reference to the conduction bandedge). Calculate the position of the ground state of the trap level if the defect dimensions are 5 Å× 5 Å× 5 Å. The electron effective mass is 0.26 m_0.

Problem 2.19 A defect level in silicon produces a level at 0.5 eV below the conduction band. Estimate the potential depth of the defect if the defect dimension is 5 Å× 5 Å×5 Å. The electron mass is 0.25 m_0.

Problem 2.20 In an n-type Si crystal the doping changes abruptly from $N_D = 10^{15}$ to $N_D = 10^{17}$. Make a qualitative sketch of the band diagram. Calculate
(a) the built-in potential at the n^+/n^- interface, in eV. Also calculate how much of the band-bending occurs on each side of the junction,
(b) the electric field at the n^+/n^- interface and
(c) the electron concentration at the n^+/n^- interface.
Assume $T = 300K$.

Problem 2.21 Consider a schottky barrier formed between Al($s\phi_M = 4.1eV$) and GaAs($q\chi = 4.04eV$). Consider that the surface has both acceptor and donor states in equal concentration, 1.0 eV and 0.6 eV from the conduction band respectively. Assume that the concentrations are equal (measured in cm^{-2}). Assume a thin insulator ($\delta \mathring{A}$ thick) between the metal and the semiconductor to help set-up the problem. Calculate the barrier height as a function of the density of states $D_s cm^{-2}$. Solve the problem for both n and p type semiconductors doped at $10^{17} cm^{-3}$. Plot.
Note: The problem is solve by balancing charges in the system.

2.15 FURTHER READING

- **General bandstructure**
  - H.C. Casey Jr. and M.B. Panish, Heterostructure Lasers, Part A, "Fundamental Principles," Part B, "Materials and Operating Characteristics," Academic Press, New York (1978).
  - R.E. Hummel, Electronic Properties of Materials–An Introduction for Engineers, Springer Verlag, New York (1985).
  - Landolt-Bornstein, Numerical Date and Functional Relationship in Science and Technology, Vol. 22, Eds. O. Madelung, M. Schulz, and H. Weiss, Springer-Verlog, New York (1987).
  - K. Seeger, Semiconductor Physics: An Introduction, Springer, Berlin (1985).
  - H.F. Wolf, Semiconductors, Wiley-Interscience, New York (1971).
- **Bandstructure modification**
  - A.G. Milnes and D.L. Feucht, Heterojunctions and Metal Semiconductor Junctions, Academic Press, New York (1972).
  - For a simple discussion of electrons in quantum wells any book on basic quantum mechanics is adequate. An example is L. Schiff, Quantum Mechanics, McGraw-Hill, New York (1968).
  - J. Singh, Electronic and Optoelectronic Properties of Semiconductor Structures, Cambridge University Press (2003).

- **Intrinsic and extrinsic carriers**
  - J.S. Blakemore, Electron. Commun., 29, 131 (1952).
  - J.S. Blakemore, Semiconductor Statistics, Pergamon Press, New York (1962) reprinted by Dover, New York (1988).
  - K. Seeger, Semiconductor Physics: An Introduction, Springer Verlag, Berlin (1985)

Chapter 3

CHARGE TRANSPORT IN MATERIALS

3.1 INTRODUCTION

Electronic devices rely on transport of electrons (holes) in materials. This transport occurs either under the influence of an electric field or carrier concentration gradients. In this chapter we will examine how electrical current flows occur in materials. The charges in a solid can be loosely classified as fixed and mobile. When an external perturbation is applied (e.g., an electric field) the mobile charges can move from one point in space to another. In particular they can move from one contact on a device to another. The fixed charge, however, can only be disturbed slightly from its equilibrium position, but cannot move over the length of a device. As shown in figure 3.1 both fixed charges and mobile charges play an important role in the physics of semiconductors. Essentially all electronic devices such as field effect transistors, bipolar transistors, diodes, as well as optoelectronic devices, such as lasers and detectors depend upon free or mobile charges. Mobile charges are the electrons in the conduction band and holes in the valence band for semiconductors and insulators. As we have discussed in the previous chapter, in metals the mobile charges are the electrons in the conduction band.

Fixed charges in materials also play an important role in devices, even though they cannot participate in current flow. Small movements in the position of the fixed charges are responsible for the dielectric response of solids. The fixed charges are also responsible for polarization effects, which are exploited for devices, such as sensors and detectors.

Mobile carriers respond to electric fields and carrier concentration gradients. Electrons and holes also combine with each other. In this chapter we will examine the physical processes that form the basis of electronic devices

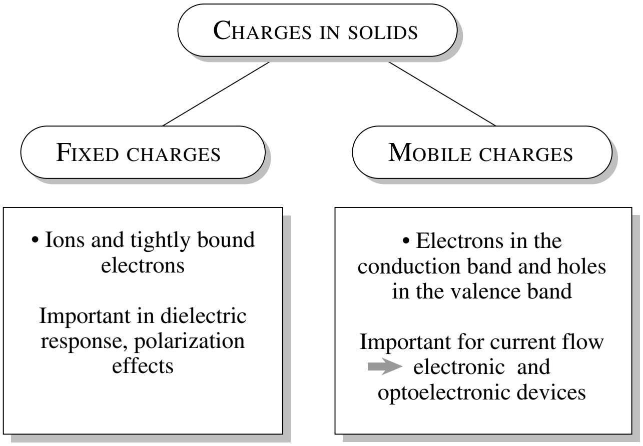

CHARGES IN SOLIDS
FIXED CHARGES
MOBILE CHARGES
• Ions and tightly bound electrons
Important in dielectric response, polarization effects
• Electrons in the conduction band and holes in the valence band
Important for current flow
electronic and optoelectronic devices

Figure 3.1: An overview of fixed and mobile charges in solids and their impact on physical phenomena. Semiconductor devices are dependent upon mobile electrons and holes.

3.2 CHARGE TRANSPORT: AN OVERVIEW

Before discussing issues in free carrier (or mobile carrier) transport we remind the reader of the nature of electronic states in solids in figure 3.2. As noted in chapter 2, in the case of the perfect crystal we see that in the conduction and valence bands the electronic states are "free,". There are no allowed energy levels in the bandgap (density of states is zero in the bandgap, as shown). In the case of a crystal with defects we still have the free states in the conduction and the valence bands, but we also have defect-related allowed states in the bandgap region, as shown in figure 3.2b. In these states (trap states) electrons are not free to move.

We will first provide a simple overview of how electrons respond to applied electric fields. In figure 3.3 we show a schematic of how electrons (holes) move through a sample when an electric field is applied. In figure 3.3a we show the situation in a good-quality crystalline material. The electron moves under the electric field force, but suffers a number of scattering processes. The scattering occurs due to various imperfections, such as defects and vibrations of atoms (due to thermal energy). The relation between the electron velocity or distance traveled and applied field is complex. However at low fields the relation can be described by a simple relation. If we examine the distance versus time trajectory of a typical electron we observe that the electron shows a path as shown in figure 3.3. On average the electron trajectory is described by

$$\begin{aligned} d &= vt \\ v &= \mu E \end{aligned} \qquad (3.2.1)$$

where d is the distance traveled in time t. The velocity v is proportional to the electric field

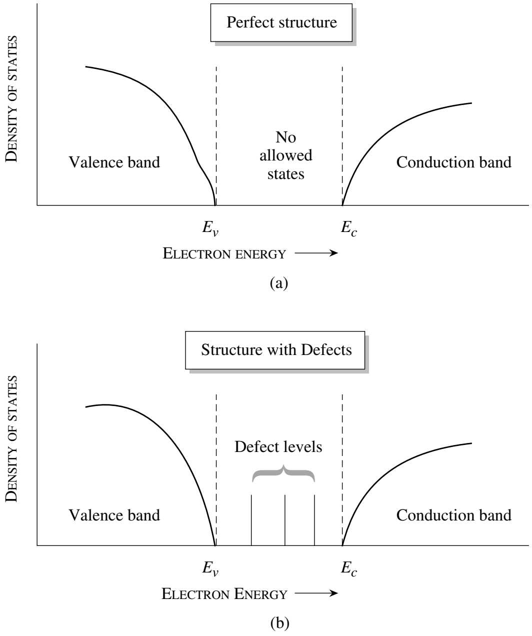


Figure 3.2: A schematic of the nature of electronic states in solids: (a) for a perfect crystal, (b) for a crystal with defects.

applied through μ, the mobility. When the electric field in large the relationship between velocity and applied field is not so simple and will be discussed later.

3.3 TRANSPORT AND SCATTERING

The problem of transport involves non-equilibrium physics. We need to find the distribution function for electrons in energy and momentum space under an applied field or under carrier concentration gradients. We know that under equilibrium the electron (hole) distribution in energy (or momentum) is given by the Fermi–Dirac distribution

$$\begin{aligned} f(E) = f^\circ(E) &= \frac{1}{\exp\left(\frac{E-E_F}{k_BT}\right)+1} \\ E &= E_i + \frac{\hbar^2k^2}{2m^*} \end{aligned}$$

where E_i is the bandedge.

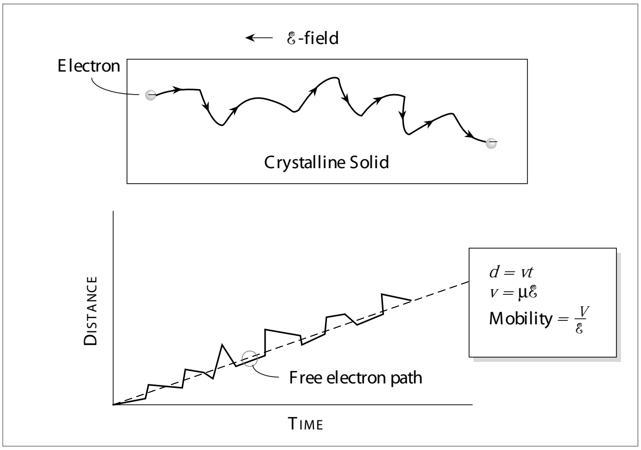


Figure 3.3: A typical electron trajectory in a sample and the distance versus time profile.

We can see that in the absence of any applied electric field, the occupation of a state with momentum $+\hbar\mathbf{k}$ is the same as that of a $-\hbar\mathbf{k}$ state. Thus there is net cancellation of momenta and there is no net current flow. The distribution function in momentum space is shown schematically in figure 3.4a. The question we would like to answer is the following: If an electric field is applied, what happens to the free electrons (holes)? When a field is applied the electron distribution will shift, as shown schematically in figure 3.4b, and there will be a net momentum of the electrons. This will cause current to flow. If the crystal is rigid and perfect, according to the Bloch theorem the electron states are described by

$$\psi_k(r,t) = u_k \ \exp \ i(k \cdot r - \omega t) \tag{3.3.1}$$

where $\omega = E/\hbar$ is the electron wave frequency. There is no scattering of the electron in the perfect system. Also, if an electric field $\mathcal{E}$ is applied, the electron behaves as a "free" space electron would, obeying the equation of the motion

$$\frac{\hbar d\mathbf{k}}{dt} = F_{\text{ext}} = -e\mathcal{E} \tag{3.3.2}$$

According to this equation the electron will behave just as in classical physics (in absence of scattering) except the electron will gain energy according to the appropriate bandstructure relation.

In a real material, there are always imperfections which cause scattering of electrons so that the equation of motion of electrons is not given by equation 3.3.2. A conceptual picture of electron transport can be developed where the electron moves in space for some time, then scatters and

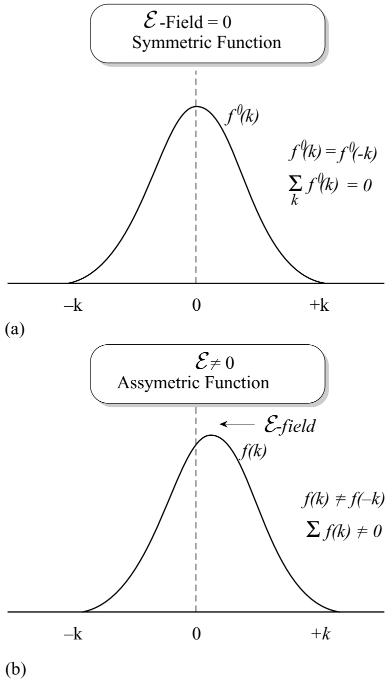


Figure 3.4: A schematic of the electron momentum distribution function in (a) equilibrium where $f(k) = f(-k)$ and (b) in the presence of an electric field.

then again moves in space and again scatters. The process is shown schematically in figure 3.5. The average behavior of the ensemble of electrons will then represent the transport properties of the electron.

3.3.1 Quantum Mechanics and Scattering of electrons

As noted above in absence of scattering the electron transport is very simple to understand. However, scattering dominates transport in semiconductor devices. The scattering problem in solids is treated by using the perturbation theory in quantum mechanics. The electron problem is formally represented by

$$H\Phi = E\Phi \tag{3.3.3}$$

where H is the full hamiltonian (potential energy + kinetic energy operator) of the problem and the electron states are denoted by Φ. This hamiltonian is, in our case, the sum of the hamiltonian

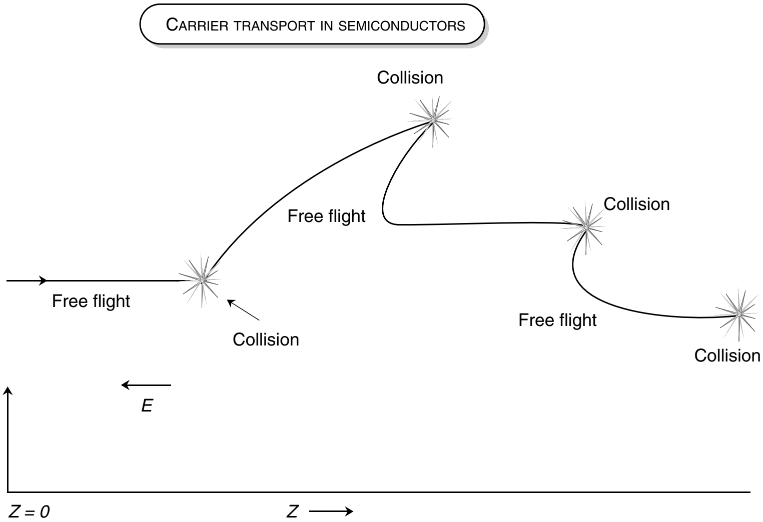


Figure 3.5: Schematic view of an electron as it moves under an electric field in a semiconductor. The electron suffers a scattering as it moves. In between scattering the electron moves according to the "free" electron equation of motion.

of the perfect crystal H_o and the energy V corresponding to the imperfection causing scattering. Thus

$$H = H_o + V \tag{3.3.4}$$

The problem

$$H_o\psi = E\psi \tag{3.3.5}$$

just gives us the bandstructure of the semiconductor which has been discussed in chapter 2. In the perturbation theory, we use the approach that the effect of the perturbation V is to cause scattering of the electron from one perfect crystalline state to another. This theory works well if the perturbation is small. The effect of the scattering is shown schematically in figure 3.6. The rate of scattering for an electron initially in state i to a state f in the presence of a perturbation of the form

$$V(r,t) = V(r)\ \exp\ (i\omega t) \tag{3.3.6}$$

is given by the Fermi golden rule

$$W_{if} = \frac{2\pi}{\hbar} \mid M_{ij} \mid^2 \delta(E_i \pm \hbar\omega - E_f) \tag{3.3.7}$$

where the various quantities in the equation represent the following:

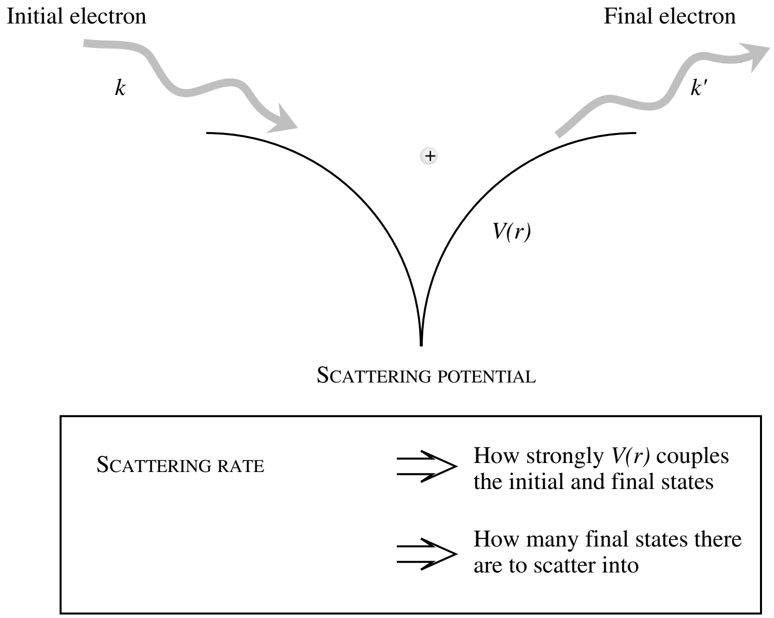

Initial electron
Final electron
k
k′
+
V(r)
SCATTERING POTENTIAL
SCATTERING RATE
How strongly V(r) couples the initial and final states
How many final states there are to scatter into

Figure 3.6: Scattering of an electron initially with momentum $\hbar\mathbf{k}$ from a scattering potential $V(r)$. The final momentum is $\hbar k'$. The scattering process is assumed to be instantaneous.

- $\mid M_{ij} \mid^2$: The quantity is called the matrix element of the scattering and is given by

$$M_{ij} = \int \psi_f^* V(r) \psi_i d^3r \tag{3.3.8}$$

The matrix element tells us how the potential couples the initial and the final state. A stronger coupling causes a higher rate of scattering.

- $\delta(E_i \pm \hbar\omega - E_f)$:This δ-function is simply a representative of energy conservation. The process where

$$E_f = E_i + \hbar\omega \tag{3.3.9}$$

is called absorption, while the process

$$E_f = E_i - \hbar\omega \tag{3.3.10}$$

is called emission. Thus, both absorption or emission of energy can occur if the perturbation has a time dependence exp($i\omega t$). If the potential is time independent (defects of various kinds), the scattering is elastic ($E_i = E_f$).

The dominant scattering of carriers involves lattice vibrations resulting from thermal energy. Carriers may scatter from various crystal imperfections including dopants and other point defects, alloy disorder, and interface imperfections.

Phonon scattering

In chapter 1, we discussed the crystalline structure in which atoms were at fixed periodic positions. In reality, the atoms in the crystal are vibrating around their mean positions. These lattice vibrations are represented by "particles" in quantum mechanics and are called phonons. The properties of the lattice vibrations are represented by the relation between the vibration amplitude, u, frequency, ω, and the wavevector q. The vibration of a particular atom, i, is given by

$$u_i(q) = u_{oi} \ \exp \ i(q \cdot r - \omega t) \tag{3.3.11}$$

which represents an oscillation with quantum energy $\hbar\omega$. In a semiconductor there are two kinds of atoms in a basis. This results in a typical ω vs. k relation shown in figure 3.7. Although the results are for GaAs, they are typical of all compound semiconductors. We notice two kinds of lattice vibrations, denoted by acoustic and optical. Additionally, there are two transverse and one longitudinal modes of vibration for each kind of vibration. The acoustic branch can be characterized by vibrations where the two atoms in the basis of a unit cell vibrate with the same sign of the amplitude as shown in figure 3.7b. In optical vibrations, the two atoms with opposing amplitudes are shown.

As noted above, in quantum mechanics lattice vibrations are treated as particles carrying energy $\hbar\omega$. According to the discussion on Bose-Einstein statistics in chapter 2, the phonon occupation is given by

$$n_\omega = \frac{1}{\exp\left(\frac{\hbar\omega}{k_B T}\right) - 1} \tag{3.3.12}$$

According to quantum mechanics, the total energy contained in the vibration is given by

$$E_\omega = (n_\omega + \frac{1}{2})\hbar\omega \tag{3.3.13}$$

Note that even if there are no phonons in a particular mode, there is a finite "zero point" energy $\frac{1}{2}\hbar\omega$ in the mode. This is important since even if $n = 0$ one can have scattering processes.

The vibrations of the atoms produce three kinds of potential disturbances that result in the scattering of electrons. A schematic of the potential disturbance created by the vibrating atoms is shown in figure 3.8. In a simple physical picture, we can imagine the lattice vibrations causing spatial and temporal fluctuations in the conduction and valence band energies. The electrons (holes) then scatter from these disturbances. The acoustic phonons produce a strain field in the crystal and the electrons see a disturbance which produces a potential of the form

$$V_{AP} = D\frac{\partial u}{\partial x} \tag{3.3.14}$$

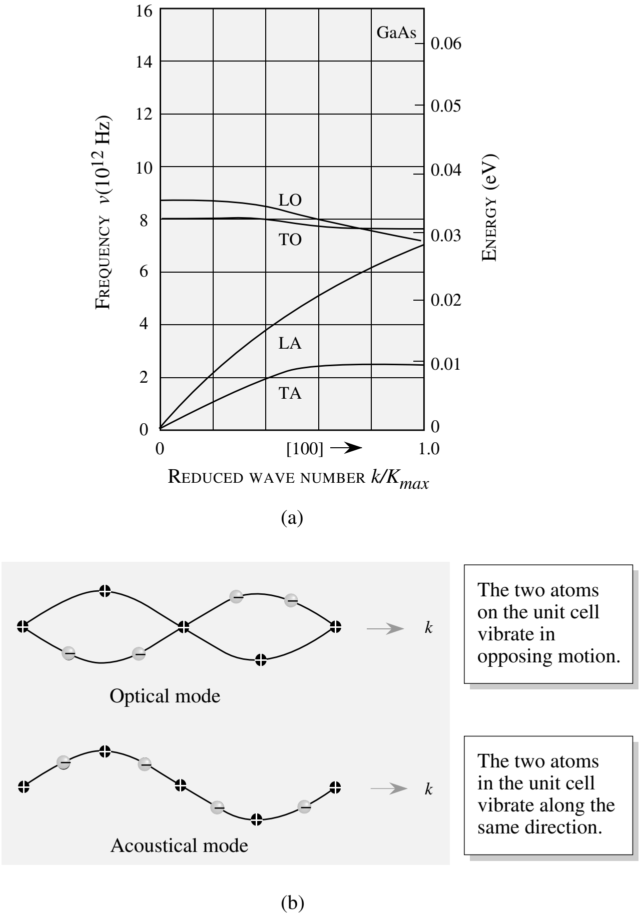


Figure 3.7: (a) Typical frequency-wavenumber relations of a semiconductor (GaAs in this case). (b) The displacement of atoms in the optical and acoustic branches of the vibrations is shown. The motion of the atoms is shown for small k vibrations.

where D is called a deformation potential (units are eV) and $\frac{\partial u}{\partial x}$ is the amplitude gradient of the atomic vibrations.

The optical phonons produce a potential disturbance, which is proportional to the atomic vibration amplitude, since in the optical vibrations the two atoms in the basis vibrate opposing

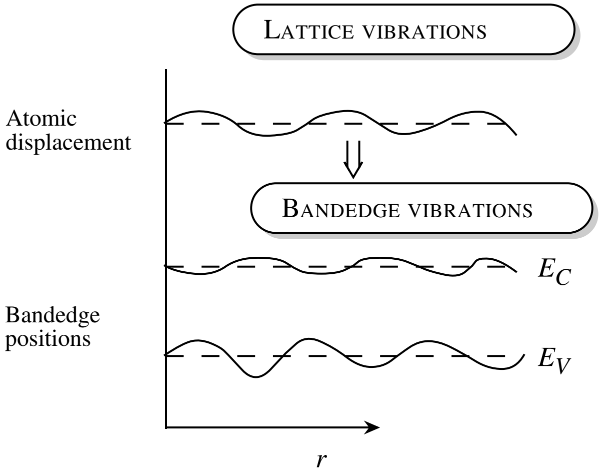


Figure 3.8: A schematic showing the effect of atomic displacement due to lattice vibrations on bandedge energy levels in real space.

each other

$$V_{op} = D_o u \tag{3.3.15}$$

where D_o (units are eV/cm) is the optical deformation potential.

In compound semiconductors the two atoms on the basis are different and there is an effective positive and negative charge e^* on each atom. When optical vibrations take place, the effective dipole in the unit cell vibrates, causing polarization fields from which the electron scatters. This scattering, called polar optical phonon scattering, has a scattering potential of the form

$$V_{po} \sim e^* u \tag{3.3.16}$$

Each material has its own effective charge which is related to the ionicity of the material. By using the Fermi golden rule we can calculate the scattering rates of electrons due to lattice vibrations. The acoustic acoustic phonon scattering rate for an electron with energy E_k to any other state is given by

$$W_{\text{ac}}(E_k) = \frac{2\pi D^2 k_B T N(E_k)}{\hbar \rho v_s^2} \tag{3.3.17}$$

where $N(E_k)$ is the electron density of states, ρ is the density of the semiconductor, v_s is the sound velocity and T is the temperature.

In materials like GaAs, the dominant optical phonon scattering is polar optical phonon scattering, and the scattering rate is given by (assuming the bandstructure is defined by a non-parabolic band; ϵ_∞ and ϵ_s are the high frequency and static dielectric constants of the semiconductor, while ϵ_o is the free space dielectric constant)

$$W(k) = \frac{e^2 m^{*1/2} \omega_o}{4\pi\sqrt{2}\hbar} \left(\frac{\epsilon_o}{\epsilon_\infty} - \frac{\epsilon_o}{\epsilon_s} \right) \frac{1 + 2\alpha E'}{\gamma^{1/2}(E)} F_o(E, E')$$

$$\times \left\{ \begin{array}{ll} n(\omega_o) & \text{absorption} \\ n(\omega_o) + 1 & \text{emission} \end{array} \right\}. \tag{3.3.18}$$

where

$$\begin{aligned}
E' &= E + \hbar\omega_o \text{ for absorption} \\
&= E - \hbar\omega_o \text{ for emission} \\
\gamma(E) &= E(1 + \alpha E) \\
F_o(E, E') &= C^{-1} \left(A \ln \left| \frac{\gamma^{1/2}(E) + \gamma^{1/2}(E')}{\gamma^{1/2}(E) - \gamma^{1/2}(E')} \right| + B \right) \\
A &= \left[2(1 + \alpha E)(1 + \alpha E') + \alpha \left\{ \gamma(E) + \gamma(E') \right\} \right]^2 \\
B &= -2\alpha\gamma^{1/2}(E)\gamma^{1/2}(E') \\
&= \times \left[4(1 + \alpha E)(1 + \alpha E') + \alpha \left\{ \gamma(E) + \gamma(E') \right\} \right] a \\
C &= 4(1 + \alpha E)(1 + \alpha E')(1 + 2\alpha E)(1 + 2\alpha E')
\end{aligned}$$

It is important to examine typical values of scattering rates from these processes. The values for GaAs are shown in figure 3.9. Note that the phonon emission process can start only after the electron has energy equal to the phonon energy. Optical phonon scattering is the most important scattering mechanism for high-field or high-temperature transport of electrons. The emission rate is stronger than the absorption rate by the rate $n(\omega_{0p} + 1)$ to $n(\omega_0)$. Optical phonon emission is the dominant mechanism for electrons to lose energy they gain from the electric field.

Ionized impurity scattering

An important scattering mechanism is due to ionized dopants. The scattering potential is Coulombic in nature, except that the potential is suppressed by screening effects due to free carriers. The screening is due to the presence of the other free electrons or holes, which form a cloud around the ion. There are several models for the ionized impurity scattering potential. A good approximation for the potential seen by electrons in a semiconductor is given by the screened Coulombic potential

$$V(r) = \frac{e^2}{\epsilon} \frac{e^{-\lambda r}}{r} \tag{3.3.19}$$

where

$$\lambda^2 = \frac{ne^2}{\epsilon k_B T} \tag{3.3.20}$$

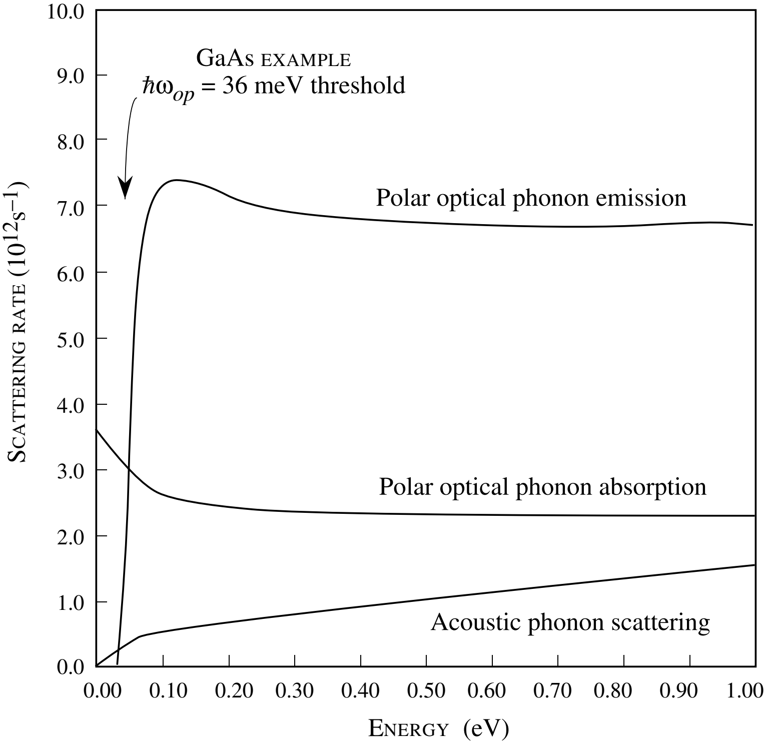


Figure 3.9: Scattering rates due to acoustic and optical phonons for GaAs electrons at room temperature.

with n the free electron density. The scattering rate for an electron with energy E_k and momentum $\hbar k$ can be shown to be

$$\begin{aligned} W(k) &= 4\pi F \left(\frac{2k}{\lambda}\right)^2 \left[\frac{1}{1+(\lambda/2k)^2}\right] \\ F &= \frac{1}{\hbar}\left(\frac{e^2}{\epsilon}\right)^2 \frac{N(E_k)}{32k^4} N_I \end{aligned} \qquad (3.3.21)$$

where N_I is the ionized impurity density. Note that ionized impurities (and the scattering processes discussed here) do not alter the spin of the electron. Thus $N(E)$ is the density of states without counting the spin degeneracy i.e it is half the usual density of states.

Alloy scattering

Alloys are made from combinations of two or more materials. Since atoms on the lattice are arranged randomly there is random potential fluctuation which causes scattering. The scattering rate for an alloy A_xB_{1-x} is found to be

$$\begin{aligned} W_{\rm tot} &= \frac{2\pi}{\hbar}\left(\frac{3\pi^2}{16}V_0\right) U_{all}^2 \; N(E_{\bf k}) \left[x\,(1-x)^2 + (1-x)\,x^2\right] \\ a = & \frac{3\pi^3}{8\hbar} V_0 \, U_{all}^2 \; N(E_{\bf k}) \; x\,(1-x) \end{aligned} \tag{3.3.22}$$

Here U_{all} is the potential difference between A type and B type potentials (see Appendix B), V_0 is the volume of the unit cell in the lattice and $N(E)$ is the density of states without counting spin degeneracy.

While the phonon and impurity scattering are the dominant scattering processes for most transport problems, electron–electron scattering, electron–hole scattering, and alloy potential scattering, etc., can also play an important role.

> **Example 3.1** Calculate the ratio of the polar optical phonon emission rate to the absorption rate for GaAs and GaN at 300K.
>
> The optical phonon energies in GaAs and GaN are 36 meV and 90 meV respectively. If the electron energies are below these values, there is no phonon emission. The phonon occupation number in GaAs at 300 K is 0.33 and in GaN is 0.032. Thus above threshold, the emission to absorption ratios are approximately 4:1 and 32:1 respectively.

3.4 TRANSPORT UNDER AN ELECTRIC FIELD

The problem of finding the distribution function of electrons under an electric field is quite complicated. Two important approaches to understanding transport in semiconductors are the solution of the transport equation using numerical methods and the Monte Carlo method using computer simulations. We will summarize the results of such theories by examining the drift velocity versus electric field relations in semiconductors.

3.4.1 Velocity–electric field relations in semiconductors

When an electron distribution is subjected to an electric field, the electrons tend to move in the field direction (opposite to the field $\mathcal{E}$ and gain velocity from the field. However, because of imperfections in the crystal potential, they suffer scattering. A steady state is established in which the electrons have some net drift velocity in the field direction. The response of the electrons to the field can be represented by a velocity–field relation. We will briefly discuss the velocity-field relationships at low electric fields and moderately high electric fields.

Low field response: mobility

At low electric fields, the macroscopic transport properties of the material (mobility, conductivity) can be related to the microscopic properties (scattering rate or relaxation time) by simple arguments. We will not solve the Boltzmann transport equation, but we will use simple conceptual arguments to understand this relationship. In this approach we make the following assumptions:

(i) The electrons in the semiconductor do not interact with each other. This approximation is called the independent electron approximation.
(ii) Electrons suffer collisions from various scattering sources and the time τ_{sc} describes the mean time between successive collisions.
(iii) The electrons move according to the free electron equation

$$\frac{\hbar dk}{dt} = e\mathcal{E} \tag{3.4.1}$$

in between collisions. After a collision, the electrons lose all their excess energy (on the average) so that the electron gas is essentially at thermal equilibrium. This assumption is really valid only at very low electric fields.

According to these assumptions, immediately after a collision the electron velocity is the same as that given by the thermal equilibrium conditions. This average velocity is thus zero after collisions. The electron gains a velocity in between collisions; i.e., only for the time τ_{sc}.

This average velocity gain is then that of an electron with mass m^*, traveling in a field $\mathcal{E}$, for a time τ_{sc}

$$\mathbf{v}_{\text{avg}} = -\frac{e\mathcal{E}\tau_{sc}}{m^*} = \mathbf{v}_d \tag{3.4.2}$$

where v_d is the drift velocity . The current density is now

$$\mathbf{J} = -nee\mathbf{v}_d = \frac{ne^2\tau_{sc}}{m^*}\mathcal{E} \tag{3.4.3}$$

Comparing this with the Ohm's law result for conductivity σ

$$\mathbf{J} = \sigma\mathcal{E} \tag{3.4.4}$$

we have

$$\sigma = \frac{ne^2\tau_{sc}}{m^*} \tag{3.4.5}$$

The resistivity of the semiconductor is simply the inverse of the conductivity. From the definition of mobility μ, for electrons

$$\mathbf{v}_d = \mu\mathcal{E} \tag{3.4.6}$$

we have

$$\mu = \frac{e\tau_{sc}}{m^*} \tag{3.4.7}$$

If both electrons and holes are present, the conductivity of the material becomes

$$\sigma = ne\mu_n + pe\mu_p \tag{3.4.8}$$

	Bandgap		Mobility at 300 K	
	(eV)		(cm^2/V-s)	
Semiconductor	300 K	0 K	Elec.	Holes
C	5.47	5.48	1800	1200
GaN	3.4	3.5	1400	350
Ge	0.66	0.74	3900	1900
Si	1.12	1.17	1500	450
α-SiC	3.00	3.30	400	50
GaSb	0.72	0.81	5000	850
GaAs	1.42	1.52	8500	400
GaP	2.26	2.34	110	75
InSb	0.17	0.23	80000	1250
InAs	0.36	0.42	33000	460
InP	1.35	1.42	4600	150
CdTe	1.48	1.61	1050	100
PbTe	0.31	0.19	6000	4000
In$_{0.53}$Ga$_{0.47}$As	0.8	0.88	11000	400

Table 3.1: Bandgaps along with electron and hole mobilities in several semiconductors. Properties of large bandgap materials (C, GaN, SiC) are continuously changing (mobility is improving), due to progress in crystal growth. Zero temperature bandgap is extrapolated.

where μ_n and μ_p are the electron and hole mobilities and n and p are their densities.

Notice that the mobility has an explicit $\frac{1}{m^*}$ dependence in it. Additionally τ_{sc} also decreases with m^*. Thus the mobility has a strong dependence on the carrier mass. In table 3.1 we show the mobilities of several important semiconductors at room temperature. The results are shown for pure materials. If the semiconductors are doped, the mobility decreases. Note that Ge has the best hole mobility among all semiconductors.

The scattering rate (or inverse of scattering time) due to ionized impurity scattering is

$$\begin{aligned}\frac{1}{\langle\langle\tau\rangle\rangle} &= N_i \frac{1}{128\sqrt{2}\pi}\left(\frac{Ze^2}{\epsilon}\right)^2 \frac{1}{m^{*1/2}\left(k_BT\right)^{3/2}} \\ &\times \left[\ln\left(1+\left(\frac{24m^*k_BT}{\hbar^2\lambda^2}\right)^2\right) - \frac{1}{1+\left(\frac{\hbar^2\lambda^2}{8m^*k_BT}\right)^2}\right] \end{aligned} \quad (3.4.9)$$

The mobility limited from ionized impurity scattering is

$$\mu = \frac{e\langle\langle\tau\rangle\rangle}{m^*}$$

The mobility limited by ionized dopant has the special feature that it decreases with temperature ($\mu \sim T^{3/2}$). This temperature dependence is quite unique to ionized impurity scattering.

One can understand this behavior physically by saying that at higher temperatures, the electrons are traveling faster and are less affected by the ionized impurities.

After doing the proper ensemble averaging the relaxation time for the alloy scattering is

$$\frac{1}{\langle\langle\tau\rangle\rangle} = \frac{3\pi^3}{8\hbar} V_0 U_{all}^2 x(1-x) \frac{m^{*3/2}(k_BT)^{1/2}}{\sqrt{2}\pi^2\hbar^3} \frac{1}{0.75} \tag{3.4.10}$$

according to which the mobility due to alloy scattering is

$$\mu_0 \propto T^{-1/2}$$

The temperature dependence of mobility is in contrast to the situation for the ionized impurity scattering. The value of U_{all} is usually in the range of 1.0 eV.

Example 3.2 Consider a semiconductor with effective mass $m^* = 0.26\, m_0$. The optical phonon energy is 50 meV. The carrier scattering relaxation time is 10^{-13} sec at 300 K. Calculate the electric field at which the electron can emit optical phonons on the average.

In this problem we have to remember that an electron can emit an optical phonon only if its energy is equal to (or greater than) the phonon energy. According to the transport theory, the average energy of the electrons is (v_d is the drift velocity)

$$E = \frac{3}{2}k_BT + \frac{1}{2}m^*v_d^2$$

In our case, this has to be 50 meV at 300 K. Since $k_BT \sim 26$ meV at 300 K, we have

$$\frac{1}{2}m^*v_d^2 = 50 - 39 = 11 \text{ meV}$$

or

$$\begin{aligned} v_d^2 &= \frac{2 \times (11 \times 10^{-3} \times 1.6 \times 10^{-19} \text{ J})}{(0.91 \times 10^{-30} \times 0.26 \text{ kg})} \\ v_d &= 1.22 \times 10^5 \text{ m/s} \end{aligned}$$

$$v_d = \frac{e\tau\mathcal{E}}{m^*}$$

Substituting for v_d, we get (for the average electrons) for the electric field

$$\begin{aligned} \mathcal{E} &= \frac{(0.26 \times 0.91 \times 10^{-30} \text{ kg})(1.22 \times 10^5 \text{ m/s})}{(4.8 \times 10^{-10} \text{ esu})(10^{13} \text{ s})} \\ &= 18.04 \text{ kV/cm} \end{aligned}$$

The results discussed correspond approximately to silicon. Of course, since the distribution function has a spread, electrons start emitting optical phonons at a field lower than the one calculated above for the average electron.

Example 3.3 The mobility of electrons in pure GaAs at 300 K is 8500 cm^2/V·s. Calculate the relaxation time. If the GaAs sample is doped at N$_d$ = 10^{17} cm^{-3}, the mobility decreases to 5000 cm^2/V·s. Calculate the relaxation time due to ionized impurity scattering.

The relaxation time is related to the mobility by

$$\begin{aligned} \tau_{sc}^{(1)} &= \frac{m^*\mu}{e} = \frac{(0.067 \times 0.91 \times 10^{-30}\ \text{kg})(8500 \times 10^{-4}\ \text{m}^2/\text{V} \cdot \text{s})}{1.6 \times 10^{-19}\ \text{C}} \\ &= 3.24 \times 10^{-13}\ \text{s} \end{aligned}$$

If the ionized impurities are present, the time is

$$\tau_{sc}^{(2)} = \frac{m^*\mu}{e} = 1.9 \times 10^{-13}\ \text{s}$$

The total scattering rate is the sum of individual scattering rates. Since the scattering rate is inverse of scattering time we find that (this is called Mathieson's rule) the impurity-related time $\tau_{sc}^{(imp)}$ is given by

$$\frac{1}{\tau_{sc}^{(2)}} = \frac{1}{\tau_{sc}^{(1)}} + \frac{1}{\tau_{sc}^{(imp)}}$$

which gives

$$\tau_{sc}^{(\text{imp})} = 4.6 \times 10^{-13} s$$

Example 3.4 The mobility of electrons in pure silicon at 300 K is 1500 cm^2/Vs. Calculate the time between scattering events using the conductivity effective mass.

The conductivity mass for indirect semiconductors, such as Si, is given by (see Appendix C)

$$\begin{aligned} m_\sigma^* &= 3\left(\frac{2}{m_t^*} + \frac{1}{m_\ell^*}\right)^{-1} \\ &= 3\left(\frac{2}{0.19m_o} + \frac{1}{0.98m_o}\right)^{-1} = 0.26m_o \end{aligned}$$

The scattering time is then

$$\begin{aligned} \tau_{sc} &= \frac{\mu m_\sigma^*}{e} = \frac{(0.26 \times 0.91 \times 10^{-30})(1500 \times 10^{-4})}{1.6 \times 10^{-19}} \\ &= 2.2 \times 10^{-13}\ \text{s} \end{aligned}$$

Example 3.5 Consider two semiconductor samples, one Si and one GaAs. Both materials are doped n-type at $N_d = 10^{17}$ cm^{-3}. Assume 50 % of the donors are ionized at 300 K. Calculate the conductivity of the samples. Compare this conductivity to the conductivity of undoped samples.

You may assume the following values:

$$\begin{aligned}\mu_n(\mathrm{Si}) &= 1000 \quad \mathrm{cm}^2/\mathrm{V}\cdot\mathrm{s}\\ \mu_p(\mathrm{Si}) &= 350 \quad \mathrm{cm}^2/\mathrm{V}\cdot\mathrm{s}\\ \mu_n(\mathrm{GaAs}) &= 8000 \quad \mathrm{cm}^2/\mathrm{V}\cdot\mathrm{s}\\ \mu_p(\mathrm{GaAs}) &= 400 \quad \mathrm{cm}^2/\mathrm{V}\cdot\mathrm{s}\end{aligned}$$

In the doped semiconductors, the electron density is (50 % of 10^{17} cm^{-3})

$$n_{n0} = 5 \times 10^{16} \text{ cm}^{-3}$$

and hole density can be found from

$$p_{n0} = \frac{n_i^2}{n_{n0}}$$

For silicon we have

$$p_{n0} = \frac{2.25 \times 10^{20}}{5 \times 10^{16}} = 4.5 \times 10^3 \text{ cm}^{-3}$$

which is negligible for the conductivity calculation.

The conductivity is

$$\sigma_n = n_{n0} e \mu_n + p_{n0} e \mu_p = 8 \ \ (\Omega \text{ cm})^{-1}$$

In the case of undoped silicon we get $(n = n_i = p = 1.5 \times 10^{10} \text{ cm}^{-3})$

$$\sigma_{\text{undoped}} = n_i e \mu_n + p_i e \mu_p = 3.24 \times 10^{-6} \ \ (\Omega \text{ cm})^{-1}$$

For GaAs we get

$$\sigma_n = 5 \times 10^{16} \times 1.6 \times 10^{-19} \times 8000 = 64 \ \ (\Omega \text{ cm})^{-1}$$

For undoped GaAs we get $(n_i = 1.84 \times 10^6 \text{ cm}^{-3})$

$$\sigma_{\text{undoped}} = n_i e \mu_n + p_i e \mu_p = 2.47 \times 10^{-9} \ \ (\Omega \text{ cm})^{-1}$$

You can see the very large difference in the conductivities of the doped and undoped samples. Also there is a large difference between GaAs and Si.

Example 3.6 Consider a semiconductor in equilibrium in which the position of the Fermi level can be placed anywhere within the bandgap.

What is the maximum and minimum conductivity for Si and GaAs at 300 K? You can use the data given in the problem above.

The maximum carrier density occurs when the Fermi level coincides with the conduction bandedge if $N_c > N_v$ or with the valence bandedge if $N_v > N_c$. If $N_c > N_v$; the Boltzmann approximation gives

$$n_{\max} = N_c$$

while if $N_v > N_c$ we get

$$p_{\max} = N_v$$

This gives us for the maximum density: i) for Si, 2.78×10^{19} cm^{-3} ii) for GaAs, 7.72×10^{18} cm^{-3}. Based on these numbers we can calculate the maximum conductivity:

For Si

$$\sigma_{\max} = 2.78 \times 10^{19} \times 1.6 \times 10^{-19} \times 1000 = 4.45 \times 10^{3}\ (\Omega\ \mathrm{cm})^{-1}$$

For GaAs

$$\sigma_{\max} = 7.72 \times 10^{18} \times 1.6 \times 10^{-19} \times 400 = 4.9 \times 10^{2}\ (\Omega\ \mathrm{cm})^{-1}$$

To find the minimum conductivity we need to find the minima of the expression

$$\begin{aligned} \sigma &= ne\mu_n + pe\mu_p \\ &= \frac{n_i^2}{p} e\mu_n + pe\mu_p \end{aligned}$$

To find the minimum we take the derivative with respect to p and equate the result to zero. This gives

$$p = n_i \sqrt{\frac{\mu_n}{\mu_p}}$$

This then gives for the minimum conductivity

$$\sigma_{\min} = n_i e[\mu_n \sqrt{\frac{\mu_p}{\mu_n}} + \mu_p \sqrt{\frac{\mu_n}{\mu_p}}]$$

For Si this gives upon plugging in numbers

$$\sigma_{\min} = 2.8 \times 10^{-6}\ (\Omega\ \mathrm{cm})^{-1}$$

and for GaAs

$$\sigma_{\min} = 1.05 \times 10^{-9}\ (\Omega\ \mathrm{cm})^{-1}$$

Note that these values are lower than the values we get in the the previous problem for the undoped cases. This example shows the tremendous variation in conductivity that can be obtained in a semiconductor.

High field transport: velocity–field relations

In most electronic devices a significant portion of the electronic transport occurs under strong electric fields. This is especially true of field effect transistors. At such high fields ($\sim$ 1$-$500 kV/cm) the electrons get "hot" and acquire a high average energy. The extra energy comes due to the strong electric fields. The drift velocities are also quite high. The description of electrons at such high electric fields is quite complex and requires either numerical techniques or computer simulations. We will only summarize the results.

At high electric field as the carriers gain energy from the field they suffer greater rates of scattering, i.e., τ_{sc} decreases. The mobility thus starts to decrease. It is usual to represent the response of the carriers to the electric field by velocity–field relations. There are several important regimes in the velocity-field relation. At lower fields the relation is linear as discussed above. As electrons (holes) gain enough energy to emit optical phonons the scattering rates increase and the differential mobility starts to decrease as shown in figure 3.10. The relation is no longer linear.

In the case of direct gap materials an interesting phenomena occurs that leads to negative differential relation as shown in figure 3.10. As carriers gain energy comparable to the inter-valley separation in the conduction band they get scattered out of the low mass lower energy valley to higher mass upper valley. As a result the velocity drops as can be seen for GaAs and InP in Figure 3. 10. The negative differential mobility (resistance) is exploited by microwave devices such as Gunn diodes to generate microwave power.

At very high fields the drift velocity becomes saturated; i.e., becomes independent of the electric field. This occurs because the scattering rates increase as the field increases so that the electrons gain energy from the field but their net velocity does not change. The drift velocity for carriers in most materials saturates to a value of $\sim 10^7$ cm/s. The fact that the velocity saturates is very important in understanding current flow in semiconductor devices.

It is important to note that the concept of velocity-field relation is valid if the fields are changing slowly over distances comparable the electron mean free path. This is the case in devices that are longer than a micron or so. For sub-micron devices electrons can move without scattering for a some distance. In this case the transport is called ballistic transport and is described by the Newton's equation without scattering,

$$m^* \frac{dx}{dt} = eF \qquad (3.4.11)$$

For short distances electrons can display overshoot effects i.e they can have velocities larger than what may be expected from a steady state velocity-field relation. For light mass semiconductors such as GaAs and InGaAs velocity overshoot effects dominate modern devices.

Example 3.7 The mobility of electrons in a semiconductor decreases as the electric field is increased. This is because the scattering rate increases as electrons become hotter due to the applied field. Calculate the relaxation time of electrons in silicon at 1 kV/cm and 100 kV/cm at 300 K.

The velocity of the silicon electrons at 1 kV/cm and 100 kV/cm is approximately 1.4×10^6 cm s and 1.0×10^7 cm/s, respectively, from the v-F curves given in figure 3.10. The mobilities are then

$$\begin{aligned} \mu(1\,\mathrm{kV/cm}) &= \frac{v}{\mathcal{E}} = 1400\,\mathrm{cm^2/V\cdot s} \\ \mu(100\,\mathrm{kV/cm}) &= 100\,\mathrm{cm^2/V\cdot s} \end{aligned}$$

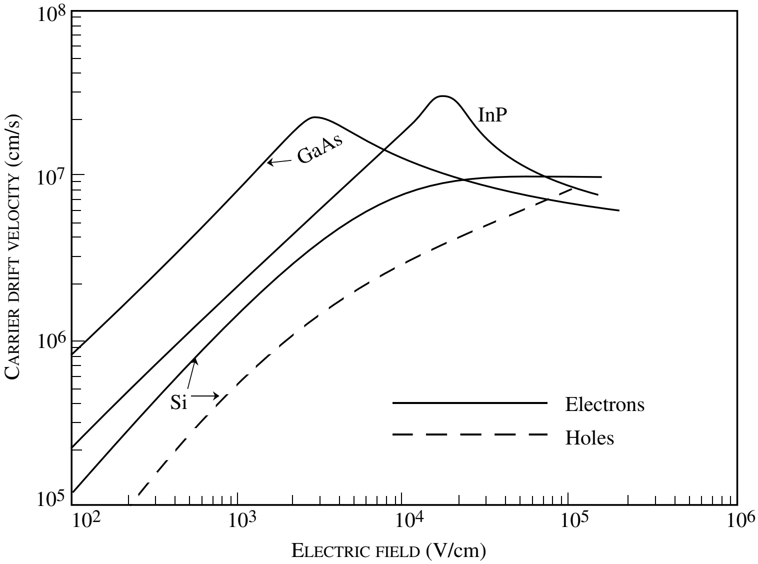


Figure 3.10: Velocity–field relations for several semiconductors at 300 K.

The corresponding relaxation times are

$$\begin{aligned}\tau_{sc}(1\,\text{kV/cm}) &= \frac{(0.26\times 0.91\times 10^{-30}\,\text{kg})(1400\times 10^{-4}\,\text{m}^2/\text{V})}{1.6\times 10^{-19}\,\text{C}} = 2.1\times 10^{-13}\,\text{s}\\ \tau_{sc}(100\,\text{kV/cm}) &= \frac{(0.26\times 0.91\times 10^{-30})(100\times 10^{-4})}{1.6\times 10^{-19}} = 1.48\times 10^{-14}\,\text{s}\end{aligned}$$

Thus the scattering rate has dramatically increased at the higher field.

Example 3.8 The average electric field in a particular 0.1 μm GaAs device is 50 kV/cm. Calculate the transit time of an electron through the device (a) if the transport is ballistic; (b) if the saturation velocity value of 10^7 cm/s is used.

For ballistic transport the transit time is

$$\tau_{tr} = \sqrt{\frac{2L}{a}}$$

with the acceleration, a given by

$$a = \frac{e\mathcal{E}}{m^*}$$

This gives a transit time of 0.123 ps.

The transit time, if the saturation velocity (which is the correct velocity value) is used, is

$$\tau_{\rm tr} = \frac{L}{v} = \frac{1 \times 10^{-5}}{10^7} = 1 \text{ ps}$$

This example shows that in short channel devices, ballistic effects can be very strong.

Very high field transport: breakdown phenomena

When the electric field becomes extremely high ($\sim$ 100 kV cm^{-1}), the semiconductor suffers a "breakdown" in which the current has a "runaway" behavior. The breakdown occurs due to carrier multiplication, which arises from the two sources discussed below. By carrier multiplication we mean that the number of electrons and holes that can participate in current flow increases. Of course, the total number of electrons is always conserved.

Avalanche breakdown

In the transport considered in the previous subsections, the electron (hole) remains in the same band during the transport. At very high electric fields, this does not hold true. In the impact ionization process shown schematically in figure 3.11, an electron, which is "very hot" (i.e., has a very high energy due to the applied field) scatters with an electron in the valence band via Coulombic interaction, and knocks it into the conduction band. The initial electron must provide enough energy to bring the valence-band electron up into the conduction band. Thus the initial electron should have energy slightly larger than the bandgap (measured from the conduction-band minimum). In the final state we now have two electrons in the conduction band and one hole in the valence band. Thus the number of current carrying charges have multiplied, and the process is often called avalanching. Note that the same could happen to "hot holes" and thus could then trigger the avalanche.

Once avalanching starts, the carrier density in a device changes as

$$\frac{dn(z)}{dz} = \alpha_{imp} n \tag{3.4.12}$$

where n is the carrier density and $\alpha_{\rm imp}$ represents the average rate of ionization per unit distance.

The coefficients $\alpha_{\rm imp}$ for electrons and $\beta_{\rm imp}$ for holes depend upon the bandgap of the material in a very strong manner. This is because, as discussed above, the process can start only if the initial electron has a kinetic energy equal to a certain threshold (roughly equal to the bandgap). This is achieved for lower electric fields in narrow gap materials.

If the electric field is constant so that α_{imp} is constant, the number of times an initial electron will suffer impact ionization after traveling a distance x is

$$n(x) = \exp\ (\alpha_{\rm imp} z) \tag{3.4.13}$$

A critical breakdown field $\mathcal{E}_{crit}$ is defined where $\alpha_{\rm imp}$ or $\beta_{\rm imp}$ approaches 10^4 cm^{-1}. When $\alpha_{\rm imp}$ ($\beta_{\rm imp}$) approaches 10^4 cm^{-1}, there is about one impact ionization when a carrier travels

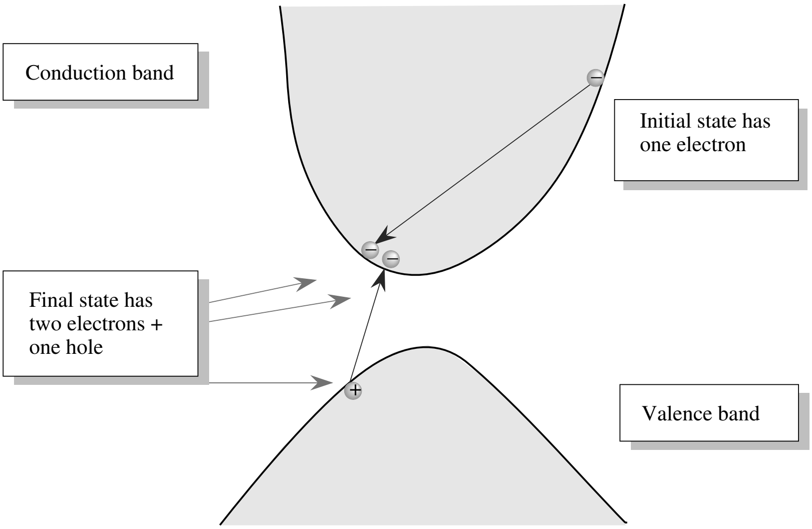

Conduction band
Initial state has one electron
Final state has two electrons + one hole
Valence band

Figure 3.11: How carriers multiply. The impact ionization process where a high energy conduction-band electron scatters from a valence-band electron, producing two conduction-band electrons and a hole.

a distance of one micron. Values of the critical field are given for several semiconductors in table 3.2. The avalanche process places an important limitation on the power output of devices. Once the process starts, the current rapidly increases due to carrier multiplication and the control over the device is lost.[1] The push for high-power devices is one of the reasons for research in large gap semiconductor devices. It must be noted that in certain devices, such as avalanche photodetectors, the process is exploited for high gain detection. The process is also exploited in special microwave devices.

Band-to-band tunneling breakdown

In quantum mechanics electrons behave as waves and one of the outcomes of this is that electrons can tunnel through regions where classically they are forbidden. Thus they can penetrate regions where the potential energy is larger than their total energy. This process is described by the tunneling theory. This theory is invoked to understand another phenomenon responsible for high field breakdown. Consider a semiconductor under a strong field, as shown in figure 3.12a. At strong electric fields, the electrons in the valence band can tunnel into an unoccupied state in the conduction band. As the electron tunnels, it sees the potential profile shown in figure 3.12b.

[1] An analytical treatment of the avalanche breakdown process of a $p - n$ junction is presented in section 4.7

Material	Bandgap (eV)	Breakdown electric field (V/cm)
GaAs	1.43	4 x 10^5
Ge	0.664	10^5
InP	1.34	
Si	1.1	3 x 10^5
$In_{0.53}Ga_{0.47}As$	0.8	2 x 10^5
C	5.5	10^7
SiC	2.9	2-3 x 10^6
SiO_2	9	-10^7
Si_3N_4	5	-10^7
GaN	3.4	2 x 10^6

Table 3.2: Breakdown electric fields in some materials.

The tunneling probability through the triangular barrier is given by

$$T = \exp\left(\frac{-4\sqrt{2m^*}E_g^{3/2}}{3e\hbar\mathcal{E}}\right) \tag{3.4.14}$$

where $\mathcal{E}$ is the electric field in the semiconductor.

In narrow bandgap materials this band-to-band tunneling or Zener tunneling can be very important. It is the basis of the Zener diode, where the current is essentially zero until the band-to-band tunneling starts and the current increases very sharply. A tunneling probability of $\sim 10^{-6}$ is necessary to start the breakdown process.

Example 3.9 Calculate the band-to-band tunneling probability in GaAs and InAs at an applied electric field of 2×10^5 V/cm.

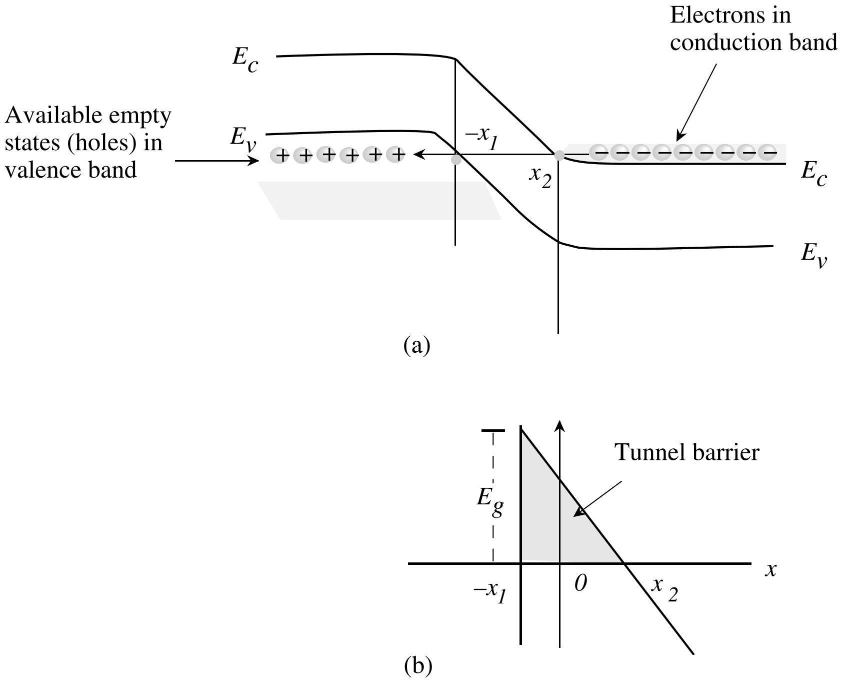

Electrons in conduction band
E_c
Available empty states (holes) in valence band
E_v
$-x_1$
x_2
E_c
E_v
(a)
Tunnel barrier
E_g
x
$-x_1$
0
x_2
(b)

Figure 3.12: (a) A schematic showing the band profile for a p–n junction. An electron in the conduction band can tunnel into an unoccupied state in the valence band or vice versa. (b) The potential profile seen by the electron during the tunneling process.

The exponent for the tunneling probability is (m^*(GaAs) = 0.065 m$_0$; m^*(InAs) $\sim$ 0.02 m$_0$; E_g(GaAs) = 1.5 eV; E_g(InAs) = 0.4 eV) for GaAs

$$\begin{aligned} & - \frac{4\times(2\times0.065\times0.91\times10^{-30}\text{ kg})^{1/2}(1.5\times1.6\times10^{-19}\text{ J})^{3/2}}{3\times(1.6\times10^{-19}\text{ C})(1.05\times10^{-34}\text{ Js})(2\times10^{7}\text{ V/m})} \\ = & -160 \end{aligned}$$

The tunneling probability is exp(-160) $\cong$ 0. For InAs the exponent turns out to be -12.5 and the tunneling probability is

$$T = \exp\ (-12.5) = 3.7\times10^{-6}$$

In InAs the band-to-band tunneling will start becoming very important if the field is $\sim 2\times10^{5}$ V/cm.

3.5 SOME IMPORTANT ISSUES IN TRANSPORT

We will discuss some important issues in transport and how bandgap, carrier masses, device length, etc. influence transport. We note that in absence of collisions, electron transport is given by the modified Newton's expression

$$\hbar \frac{dk}{dt} = e\mathcal{E} \tag{3.5.1}$$

which (for the simple parabolic band)

$$E(k) = \frac{\hbar^2 k^2}{2m^*} \tag{3.5.2}$$

Of course, in reality, as we have discussed earlier, scattering modifies this simple picture. In figure 3.13(a) we show a schematic of carrier velocity as a function of electric field in steady state for electrons in a direct bandgap material (solid line) and electrons in indirect bandgap materials (dashed line) or holes (dashed line), the negative resistance region arises due to electrons transferring from a low mass direct gap valley to high mass indirect valley.
As indicated on the figure, at low fields the important scattering mechanisms are acoustic phonon scattering, ionized impurity scattering, and optical phonon absorption. There is not much optical phonon emission since electron energies are small compared to optical phonon energy. At high fields, the optical phonon emission dominates. As a result of the different mechanisms dominating scattering at low and high fields, when temperature is lowered, low field mobility is greatly enhanced (since phonon occupation is lower) but there is not much change in high field velocity.

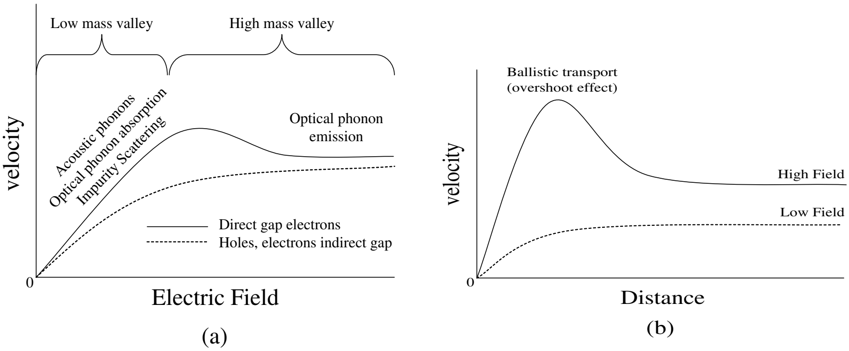


Figure 3.13: a) A schematic of how different scattering mechanisms dominate in various regions of electric field; under steady state field conditions. b) non-steady state transport velocity versus distance profile. The electron sees a step in field profile at the origin. At high fields, velocity overshoot effects occur.

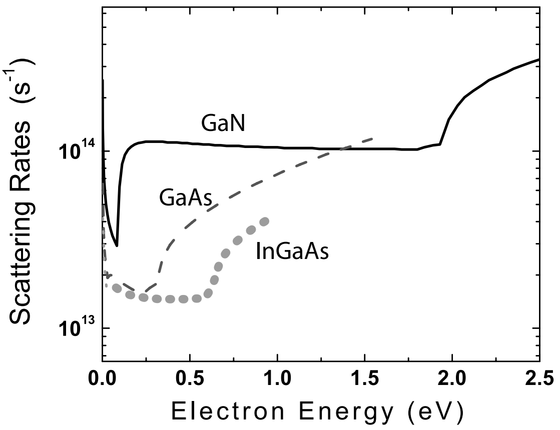
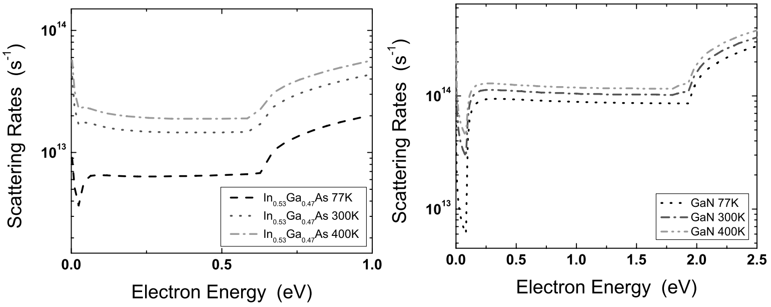


Figure 3.14: Scattering rates in InGaAs, GaAs, and GaN in 2-dimensional HFET channels

In figure 3.13b we show how an electron evolves with distance (or time) when electrons come into a high field region. The important point to note is that electrons take time to scatter and during that initial time ($\sim$ picoseconds or smaller) travel ballistically according to equation equation 3.5.1 As a result of ballistic transport, electrons can exhibit overshoot effect of high fields where electron velocity can be larger than what is expected from steady state velocity. This effect is quite dominant in materials such as InGaAs and GaAs where scattering times are long. To illustrate some of the points mentioned above, we examine electron transport in $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$, GaAs, and GaN. Transport in Si falls in between GaAs and GaN in terms of scattering rates. In figure 3.14 we show scattering rates in these three materials in 2-dimensional HFET channels (not in bulk). We note that for low electron energies there is a great difference in the scattering rates between the materials. At higher energies the relative difference is smaller. In table 3.3 we show some of the important scattering mechanisms. The rates are given for low electron energies and higher energies. In figure 3.15 we show the temperature dependence of scattering rate versus energy for InGaAs and GaN. materials. The rates drop quite dramatically at small electron energies due to phonon occupation number becoming small. Later when we examine device properties in chapter 8 we will see how the issues disscus in long and short channel devices.

3.6 CARRIER TRANSPORT BY DIFFUSION

Semiconductor devices fall into two broad categories: majority carrier devices and minority carrier devices. In the majority carrier devices, current flow is dominated by electric field driven current. In minority carrier devices current flow is dominated by diffusion effects. Whenever



a) b)

Figure 3.15: Temperature dependence of scattering rates in 2DEGs for (a) InGaAs and (b) GaN

Table 1: Scattering Comparison

Point	A			B		
	GaN	GaAs	InGaAs	GaN	GaAs	InGaAs
Energy (eV)	0.2	0.2	0.2	0.8	0.8	0.8
Optical phonon emission	8.85x 10^{13}	6.60x 10^{12}	6.33x 10^{12}	8.95x 10^{13}	6.35x 10^{12}	7.00x 10^{12}
Optical phonon absorption	3.48x 10^{12}	1.79x 10^{12}	2.08x 10^{12}	2.88x 10^{12}	1.61x 10^{12}	2.11x 10^{12}
acoustic Phonon	1.50x 10^{12}	3.44x 10^{11}	1.64x 10^{11}	3.23x 10^{12}	9.06x 10^{11}	5.69x 10^{11}
Alloy scattering	0.0	0.0	9.56x 10^{10}	0.0	0.0	4.23x 10^{11}
ionized impurity	1.47x 10^{13}	7.41x 10^{12}	6.41x 10^{12}	8.47x 10^{12}	5.11x 10^{12}	5.08x 10^{12}
dislocation	2.33x 10^{12}	7.35x 10^{8}	4.90x 10^{8}	2.33x 10^{12}	7.35x 10^{8}	4.90x 10^{8}
Nonequivalent intervalley emission(Γ–L)	0.0	0.0	0.0	0.0	3.42x 10^{13}	1.23x 10^{13}
Nonequivalent intervalley absorption(Γ–L)	0.0	0.0	0.0	0.0	1.15x 10^{13}	5.46x 10^{12}
Total (s^{-1})	1.11x 10^{14}	1.61x 10^{13}	1.51x 10^{13}	1.06x 10^{14}	5.97x 10^{13}	3.30x 10^{13}

Table 3.3: Scattering rate mechanisms in InGaAs, GaAs, and GaN 2-DEG channels

there is a gradient in the concentration of a species of mobile particles, the particles diffuse from the regions of high concentration to the regions of low concentration. As the mobile charges move they suffer random collisions, as discussed in the previous section. The collision process can be described by the mean free path ℓ and the mean collision time τ_{sc}. The mean free path is the average distance the electron (hole) travels between successive collisions. In between the collisions the electrons move randomly, with equal probability of moving in any direction (there is no electric field). We are interested in finding out how the electrons move (diffuse) when there is a concentration gradient in space.

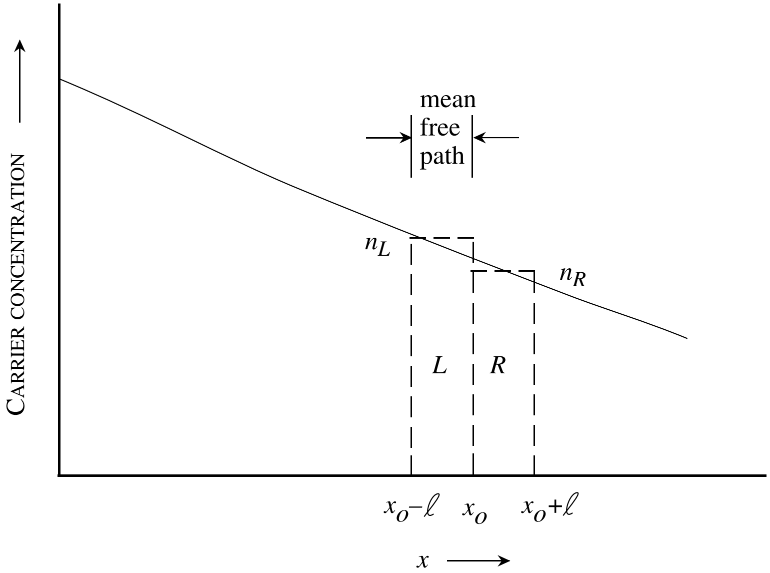


Figure 3.16: The concentration profile of electrons as a function of space. The terms n_L, n_R, L, and R are defined in the text. The distance ℓ is the mean free path for electrons; i.e., the distance they travel between collisions.

In figure 3.16 is shown a concentration profile $n(x,t)$ of electrons at time t,. We are going to calculate the electron flux $\phi(x,t)$ across a plane $x = x_o$ at any instant of time. Consider a region of space a mean free path ℓ to each side of x_o, from which electrons can come across the $x = x_o$ boundary in time τ_{sc}. Electrons from regions further away will suffer collisions that will randomly change their direction. Since in the two regions labeled L and R in figure 3.16, the electrons move randomly, half of the electrons in region L will go across $x = x_o$ to the right and half in the region R will go across $x = x_o$ to the left in time τ_{sc}. The flux to the right is

$$\phi_n(x,t) = \frac{(n_L - n_R)\ell}{2\tau_{\text{sc}}} \tag{3.6.1}$$

where n_L and n_R are the average carrier densities in the two regions. Since the two regions L and R are separated by the distance ℓ, we can write

$$n_L - n_R \cong -\frac{dn}{dx} \cdot \ell \tag{3.6.2}$$

The total flux is

$$\phi_n(x,t) = -\frac{\ell^2}{2\tau_{\text{sc}}}\,\frac{dn(x,t)}{dx} = -D_n\frac{dn(x,t)}{dx} \tag{3.6.3}$$

where D_n is called the diffusion coefficient of the electron system and depends upon the scattering processes that control ℓ and the τ_{sc}. Since the mean free path is essentially $v_{th}\tau_{sc}$, where v_{th} is the mean thermal speed, the diffusion coefficient depends upon the temperature as well. In a similar manner, the hole diffusion coefficient gives the hole flux due to a hole density gradient

$$\phi_p(x,t) = -D_p \frac{dp(x,t)}{dx} \tag{3.6.4}$$

The electron and hole flux causes current to flow in the structure This current is given by

$$\begin{aligned} J_{\text{tot}}(diff) &= J_n(diff) + J_p(diff) \\ &= eD_n \frac{dn(x,t)}{dx} - eD_p \frac{dp(x,t)}{dx} \end{aligned} \tag{3.6.5}$$

Note that the electron charge is $-e$ while the hole charge is e. While both electrons and holes move in the direction of lower concentration of electrons and holes respectively, the currents they carry are opposite, since electrons are negatively charged, while holes are positively charged.

3.6.1 Drift and diffusion transport: Einstein's relation

In case both electric field and carrier concentration gradients are present, the current is given by

$$\begin{aligned} J_n(x) &= e\mu_n n(x)\mathcal{E}(x) + eD_n \frac{dn(x)}{dx} \\ J_p(x) &= e\mu_p p(x)\mathcal{E}(x) - eD_p \frac{dp(x)}{dx} \end{aligned} \tag{3.6.6}$$

The diffusion and drift processes are linked by scattering processes. We will now establish an important relationship between mobility and diffusion coefficients. Consider a case where a uniform electric field is applied, as shown in figure 3.17a. The potential energy associated with the field is shown in figure 3.17b. There is a positive potential on the left-hand side in relation to the right-hand side. For a uniform electric field the potential energy is

$$U(x) = U(0) - e\mathcal{E}x \tag{3.6.7}$$

The applied force is related to the potential energy by

$$\text{Force} = -\nabla U(x) \tag{3.6.8}$$

Thus, since the electron charge $-e$ is negative, the bands bend as shown in figure 3.17c according to the relation

$$E_c(x) = E_c(0) + e\mathcal{E}x \tag{3.6.9}$$

Thus, if a positive potential is applied to the left of the material and a negative to the right, the energy bands will be lower on the left-hand side, as shown in figure 3.17c. The electrons drift downhill in the energy band picture and thus opposite to the field.

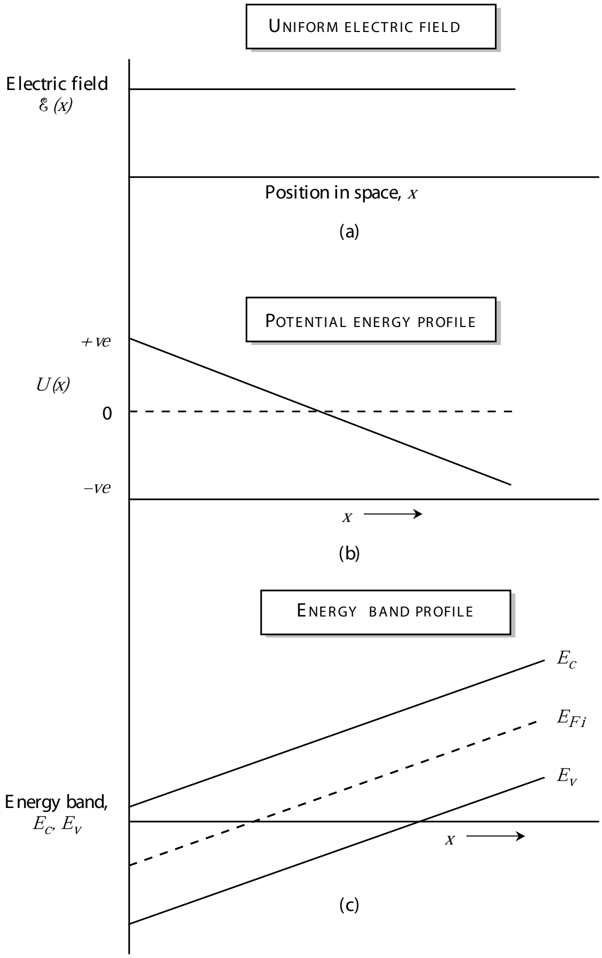


Figure 3.17: (a) Electric field profile in a semiconductor. (b) Plot of the potential energy associated with the electric field. (c) Electron energy band profile. The negative charge of the electron causes the energy band profile to have the opposite sign to the potential energy profile.

To find the relation between diffusion parameters and drift parameters (i.e. between D and μ) we assume that the system is in equilibrium and the total electron and hole currents are individually zero and we have from equation 3.6.6 for the electrons

$$\mathcal{E}(x) = -\frac{D_n}{\mu_n}\frac{1}{n(x)}\frac{dn(x)}{dx} \tag{3.6.10}$$

To obtain the derivative of carrier concentration, we write $n(x)$ in terms of the intrinsic Fermi level, E_{Fi}, which serves as a reference level, and the Fermi level in the semiconductor, $E_F(x)$. If we assume that the electron distribution is given by the Boltzmann distribution we have

$$n(x) = n_i \exp\left\{-\left(\frac{E_{Fi} - E_F(x)}{k_BT}\right)\right\} \tag{3.6.11}$$

This gives

$$\frac{dn(x)}{dx} = \frac{n(x)}{k_BT}\left(-\frac{dE_{Fi}}{dx} + \frac{dE_F}{dx}\right) \tag{3.6.12}$$

At equilibrium, the Fermi level cannot vary spatially, otherwise the probability of finding electrons along a constant energy position will vary along the semiconductor. This would cause electrons at a given energy in a region where the probability is low to move to the same energy in a region where the probability is high. Since this is not allowed by definition of equilibrium conditions, i.e. no current is flowing, the Fermi level has to be uniform in space at equilibrium, or

$$\frac{dE_F}{dx} = 0 \tag{3.6.13}$$

We then have from equation 3.6.10 and equation 3.6.12

$$\frac{D_n}{\mu_n} = \frac{k_BT}{e}$$

using

$$\mathcal{E}(x) = \frac{1}{e}\frac{dE_{Fi}}{dx}$$

This relation is known as the Einstein relation with an analogous relation for the holes. As we can see from table 3.4 which lists the mobilities and diffusion coefficients for a few semiconductors at room temperature, the Einstein relation is quite accurate.

> **Example 3.10** Use the velocity–field relations for electrons in silicon to obtain the diffusion coefficient at an electric field of 1 kV/cm and 10 kV/cm at 300 K.
>
> According to the v-$\mathcal{E}$ relations given in figure 3.10, the velocity of electrons in silicon is $\sim 1.4 \times 10^6$ cm/s and $\sim 7 \times 10^6$ cm/s at 1 kV/cm and 10 kV/cm. Using the Einstein relation, we have for the diffusion coefficient
>
> $$D = \frac{\mu k_BT}{e} = \frac{vk_BT}{e\mathcal{E}}$$

	D_n (cm^2/s)	D_p (cm^2/s)	μ_n (cm^2/V • s)	μ_p (cm^2/V • s)
Ge	100	50	3900	1900
Si	35	12.5	1350	480
GaAs	220	10	8500	400

Table 3.4: Low field mobility and diffusion coefficients for several semiconductors at room temperature. The Einstein relation is satisfied quite well.

This gives

$$\begin{aligned} D(1kV/\,\mathrm{cm}^{-1}) &= \frac{(1.4\times 10^4 \mathrm{m/s})(0.026\times 1.6\times 10^{-19}\,\mathrm{J})}{(1.6\times 10^{-19}\,\mathrm{C})(10^5\,\mathrm{V/m}^{-1})} \\ &= 3.64\times 10^{-3}\,\mathrm{m}^2/\mathrm{s} = 36.4\,\mathrm{cm}^2/\mathrm{s} \\ D(10kV/\,\mathrm{cm}^{-1}) &= \frac{(7\times 10^4\,\mathrm{m/s})(0.026\times 1.6\times 10^{-19}\,\mathrm{J})}{(1.6\times 10^{-19}\,\mathrm{C})(10^6\,\mathrm{Vm}^{-1})} \\ &= 1.82\times 10^{-3}\,\mathrm{m}^2/\mathrm{s} = 18.2\,\mathrm{cm}^2/\mathrm{s} \end{aligned}$$

The diffusion coefficient decreases with the field because of the higher scattering rate at higher fields.

3.7 CHARGE INJECTION AND QUASI-FERMI LEVELS

In semiconductor devices the electron and hole distributions are usually not under equilibrium. electric fields and optical energy causes electron densities and velocities to be different from the equilibrium values. If electrons and holes are injected into a semiconductor, either by external contacts or by optical excitation, the question arises: What kind of distribution function describes the electron and hole occupation? We know that in equilibrium the electron and hole occupation is represented by the Fermi function. It is possible to describe the non-equilibrium distribution by using the concept of quasi-equilibrium

3.7.1 Non-equilibrium Distributions

Under equilibrium conditions, electrons in the conduction band and holes in the valence band are in equilibrium with each other. Under non-equilibrium conditions it is often reasonable to assume that electrons are in equilibrium in the conduction band, while holes are in equilibrium in the valence band. In this case, the quasi-equilibrium electron and holes can be represented by an electron Fermi function f^e (with electron Fermi level) and a hole Fermi function f^h (with a different hole Fermi level). We then have

$$n = \int_{E_c}^{\infty} N_e(E) f^e(E) dE \tag{3.7.1}$$

$$p = \int_{-\infty}^{E_v} N_h(E) f^h(E) dE \tag{3.7.2}$$

where

$$f^e(E) = \frac{1}{\exp\ (\frac{E - E_{Fn}}{k_B T}) + 1} \tag{3.7.3}$$

and

$$\begin{aligned} f^h(E) = 1 - f^v(E) &= 1 - \frac{1}{\exp\ (\frac{E - E_{Fp}}{k_B T}) + 1} \\ &= \frac{1}{\exp\ (\frac{E_{Fp} - E}{k_B T}) + 1} \end{aligned} \tag{3.7.4}$$

Each band is described by its own Fermi level, E_{Fn} and E_{Fp}. At equilibrium $E_{Fn} = E_{Fp}$. If excess electrons and holes are injected into the semiconductor, the electron Fermi level E_{Fn} moves toward the conduction band, while the hole Fermi level E_{Fp} moves toward the valence band. This is shown schematically in figure 3.18. By defining separate Fermi levels for the electrons and holes, one can study the properties of excess carriers using the same relationship between Fermi level and carrier density as we developed for the equilibrium problem. Thus, in the Boltzmann approximation we have

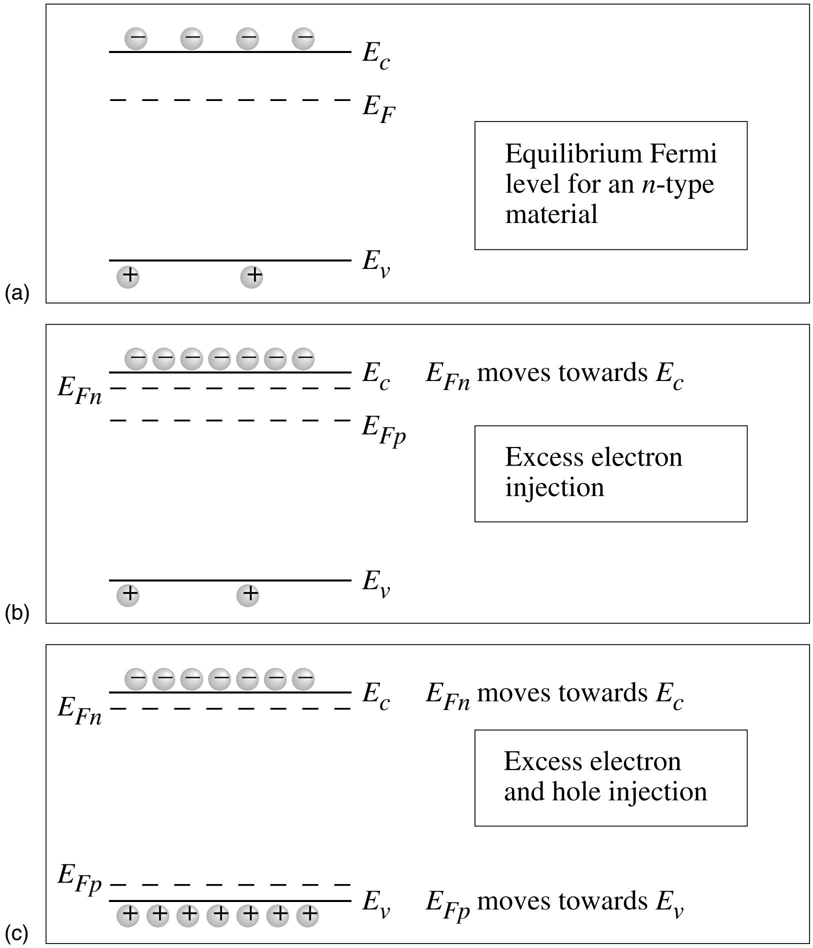
$$\begin{aligned} n &= N_c\ \exp\ \left[\frac{(E_{Fn} - E_c)}{k_B T}\right] \\ p &= N_v\ \exp\ \left[\frac{(E_v - E_{Fp})}{k_B T}\right] \end{aligned} \tag{3.7.5}$$

For high carrier densities, we have the more accurate Joyce-Dixon approximation:

$$\begin{aligned} E_{Fn} - E_c &= k_B T \left[\ell n\ \frac{n}{N_c} + \frac{n}{\sqrt{8} N_c}\right] \\ E_v - E_{Fp} &= k_B T \left[\ell n \frac{p}{N_v} + \frac{p}{\sqrt{8} N_v}\right] \end{aligned} \tag{3.7.6}$$

3.8 CARRIER GENERATION AND RECOMBINATION

In this section we will examine how mobile carrier densities change when temperature is changed or light shines on a semiconductor: The electron may start out in the valence band, then



E_c
E_F
Equilibrium Fermi level for an n-type material
E_v
(a)
E_{Fn}
E_c
E_{Fn} moves towards E_c
E_{Fp}
Excess electron injection
E_v
(b)
E_{Fn}
E_c
E_{Fn} moves towards E_c
Excess electron and hole injection
E_{Fp}
E_v
E_{Fp} moves towards E_v
(c)

Figure 3.18: (a) Schematic of an equilibrium Fermi level position in an $n-$type semiconductor. (b) The positions of the quasi-Fermi levels for the case where excess electrons are injected in the conduction band. (c) The position of the quasi-Fermi levels when excess electrons and holes are injected.

jump to the conduction band, then fall into a trap, etc. On a microscopic level there are generation recombination processes occurring in a material which cause electrons to jump between valence band, conduction band and trap states, as shown in figure 3.19.

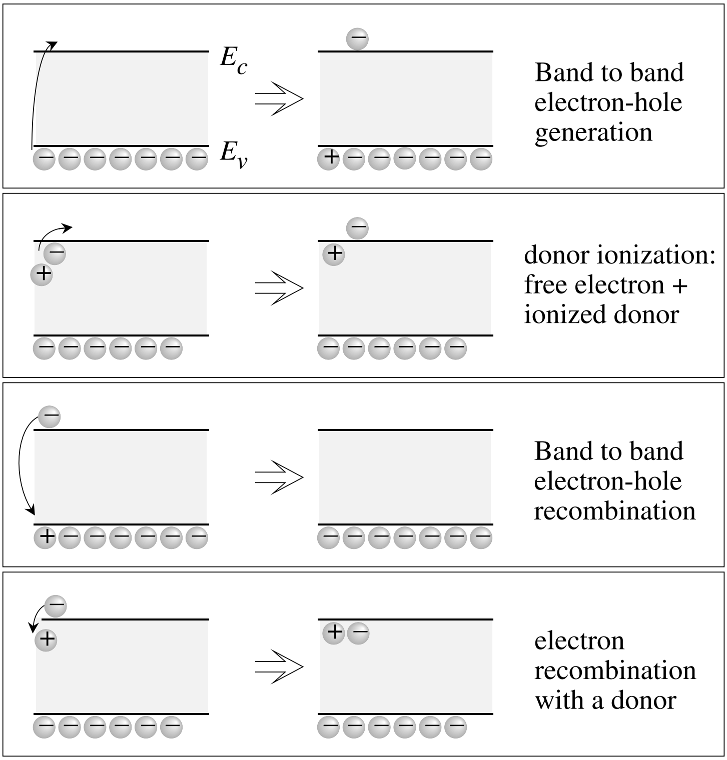

E_c
E_v
Band to band electron-hole generation
donor ionization: free electron + ionized donor
Band to band electron-hole recombination
electron recombination with a donor

Figure 3.19: A schematic of carrier generation and recombination. Processes involving band to band transitions are shown along with processes involving dopant or other impurity levels.

At equilibrium, thermal energy is responsible for exciting electrons from the valence band to the conduction band. Such a generation process is called thermal generation. We can also see that if electrons are continuously excited up from the valence band into the conduction band, there will be a build-up of free carriers. In order to reach an equilibrium concentration there has to be carrier recombination as well. Under steady state conditions we have

$$G = R \tag{3.8.1}$$

where G is the generation rate and R is the carrier recombination rate.

In figure 3.19 we show a schematic description of carrier generation and recombination. Free carriers can be generated if an electron leaves the valence-band and goes to the conduction-band. They can also be generated if electrons leave a donor and go into the conduction-band.

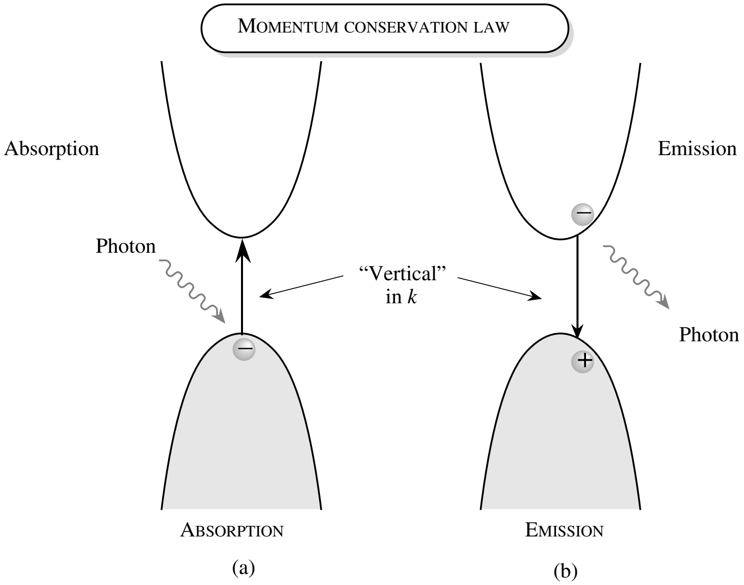


Figure 3.20: Band-to-band absorption in semiconductors. Momentum conservation ensures that only vertical transitions are allowed during absorption and emission.

An electron from the valence-band going to an acceptor causes a hole to be generated. Reverse processes can also occur.

One of the most important mechanisms for carrier generation and recombination is absorption of light and emission of light.

3.8.1 Optical Absorption and Emission in Semiconductors

According to quantum mechanics, electromagnetic radiation is made up of particles called photons, each carrying an energy $\hbar\omega$. The particle nature of $\mathcal{E}$-M waves is manifested in semiconductor devices. When light shines on a semiconductor it can cause an electron in the valence band to go into the conduction band. This process generates electron-hole pairs. It is also possible for an electron and a hole to recombine and emit light. The most important optoelectronic interaction in semiconductors as far as devices are concerned is the band-to-band transition shown in figure 3.20. In the photon absorption process, a photon scatters an electron in the valence band, causing the electron to go into the conduction band. In the reverse process the electron in the conduction band recombines with a hole in the valence band to generate a photon. These two processes are of obvious importance for light-detection and light-emission devices.

These processes are controlled by the conservation laws.

- **Conservation of energy**: In the absorption and emission process we have for the initial and final energies of the electrons E_i and E_f

$$\begin{aligned} \text{absorption}: \quad E_f &= E_i + \hbar\omega & (3.8.2) \\ \text{emission}: \quad E_f &= E_i - \hbar\omega & (3.8.3) \end{aligned}$$

where $\hbar\omega$ is the photon energy. Since the minimum energy difference between the conduction and valence band states is the bandgap E_g, the photon energy must be larger than the bandgap.

- **Conservation of momentum**: In addition to the energy conservation, one also needs to conserve the effective momentum $\hbar k$ for the electrons and the photon system. The photon k_{ph} value is given by

$$k_{ph} = \frac{2\pi}{\lambda} \tag{3.8.4}$$

The k-value of photons with energies equal to the bandgaps of typical semiconductors $\sim 10^{-4}$ Å, which is essentially zero compared to the k-values for electrons. Thus k-conservation ensures that the initial and final electrons have the same k-value. Thus for optical processes only transitions which are“vertical” in k are allowed in the bandstructure picture, as shown in figure 3.20.

Because of k-conservation, in semiconductors where the valence band and conduction bandedges are at the same $k = 0$ value (the direct semiconductors), the optical transitions are quite strong. In indirect materials like Si, Ge, etc. the optical transitions are very weak near the bandedges because they require the help of lattice vibrations to satisfy k-conservation.

Electromagnetic waves traveling through a medium like a semiconductor are described by Maxwell's equations which show that the waves have a form given by the electric field vector dependence

$$\mathcal{E} = \mathcal{E}_o \; \exp \; \left\{ i\omega \left(\frac{n_r z}{c} - t \right) \right\} \; \exp \; \left(-\frac{\alpha z}{2} \right) \tag{3.8.5}$$

Here z is the propagation direction, ω the frequency, n_r the refractive index, and α the absorption coefficient of the medium. As the $\mathcal{E}$-M wave propagates through a material, its intensity decays as

$$I(z) = I(0) \; \exp \; \{-\alpha z\} \tag{3.8.6}$$

In figure 3.21 we show the absorption coefficient of some direct and indirect bandgap semiconductors. Note that for indirect gap semiconductors the absorption coefficient is weak near the bandedge but once the photon energy is large enough to cause direct (vertical in k) transitions, the absorption coefficient increases.

When a photon is absorbed it creates an electron and a hole. If $\tilde{P}_{op}$ is the optical power density of light impinging on a semiconductor, the photon flux is

$$\Phi = \frac{\tilde{P}_{op}}{\hbar\omega}$$

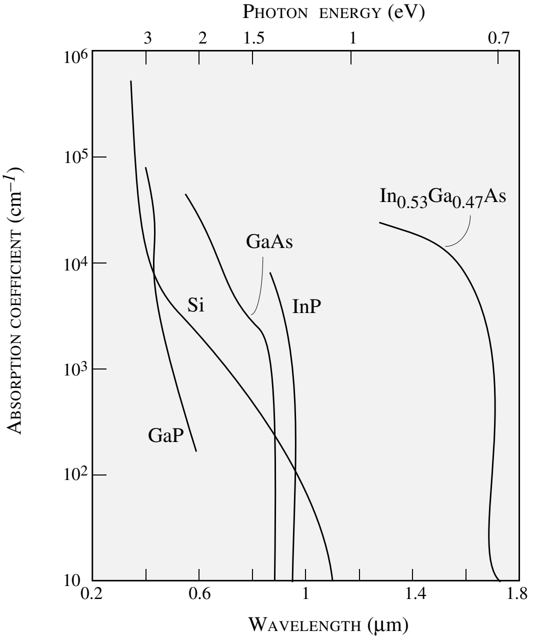


Figure 3.21: Absorption coefficient of some direct and indirect gap semiconductors. For the direct gap material, the absorption coefficient is very strong once the photon energy exceeds the bandgap. For indirect materials the absorption coefficient is small near the bandedge, but once the photon energy is more than the direct gap, the absorption coefficient increases rapidly.

and the electron-hole pair generation rate is

$$R_G = \alpha\Phi = \frac{\alpha\tilde{P}_{op}}{\hbar\omega} \tag{3.8.7}$$

Under equilibrium conditions, electron occupation in the valence band is close to unity while the occupation in the conduction band is close to zero. Assuming this is the case the absorption coefficient for direct gap materials is

$$\alpha(\hbar\omega) = \frac{\pi e^2\hbar}{2n_r c\epsilon_o m_0}\left(\frac{2p_{cv}^2}{m_0}\right)\frac{N_{cv}(\hbar\omega)}{\hbar\omega}\cdot\frac{2}{3} \tag{3.8.8}$$

Here n_r is the refractive index of the material, p_{cv} is the momentum matrix element for the scattering process, c is the speed of light in vacuumand N_{cv} is the joint density of states for the electron-hole system and is

$$N_{cv}(E) = \frac{\sqrt{2}(m_r^*)^{3/2}(E - E_g)^{1/2}}{\pi^2 \hbar^3} \tag{3.8.9}$$

If we express the energy in eV, and the absorption coefficient in $\mathrm{cm}^{-1}$for most direct gap semiconductors the absorption coefficient is approximately

$$\alpha(\hbar\omega) \sim 5.6 \times 10^4 \frac{(\hbar\omega - E_g)^{1/2}}{\hbar\omega} \mathrm{cm}^{-1} \tag{3.8.10}$$

For indirect gap materials the absorption coefficient is an order of magnitude smaller than the result given above since in first order transitions momentum is not conserved. Thus for materials like Si and Ge near bandedge absorption is weak.If there are electrons in the conduction band and holes in the valence band they can recombine to emit photons. If the occupation of an electron state is unity and the occupation of the corresponding hole state is also unity the recombination rate is given by

$$W_{em} = \frac{1}{\tau_0} = \frac{e^2 n_r}{6\pi\epsilon_o m_0 c^3 \hbar^2} \left(\frac{2p_{cv}^2}{m_0} \right) \hbar\omega \tag{3.8.11}$$

Using typical values of the momentum matrix element p_{cv} for direct gap materials the result is

$$W_{em} = \frac{1}{\tau_0} = 10^9 E_g \, s^{-1} \tag{3.8.12}$$

When electrons and holes are injected into the conduction and valence bands of a semiconductor, they recombine with each other. In general the occupation of electrons and holes is given by the quasi-Fermi levels. Theemission rateor the electron-hole recombination rate is (units are $\mathrm{cm}^{-3}\mathrm{s}^{-1}$)

$$R_{spon} = \frac{1}{\tau_o} \int d(\hbar\omega) N_{cv} \{f^e(E^e)\}\{f^h(E^h)\} \tag{3.8.13}$$

The spontaneous recombination rate is quite important for both electronic and optoelectronic devices. It is important to examine the rate for several important cases. We will give results for the electron hole recombination for the following cases: i) **Minority carrier injection:** If $n \gg p$ and the sample is heavily doped, we can assume that $f^e(E^e)$ is close to unity. We then have for the rate at which holes will recombine with electrons,

$$\begin{aligned} R_{spon} &\cong \frac{1}{\tau_o} \int d(\hbar\omega) N_{cv} f^h(E^h) \cong \frac{1}{\tau_o} \int d(\hbar\omega) N_h f^h(E^h) \left(\frac{m_r^*}{m_h^*} \right)^{3/2} \\ &\cong \frac{1}{\tau_o} \left(\frac{m_r^*}{m_h^*} \right)^{3/2} p \end{aligned} \tag{3.8.14}$$

Thus the recombination rate is proportional to the minority carrier density (holes in this case). ii) **Strong injection:** This case is important when a high density of both electrons and holes is injected and we can assume that both f^e and f^h are step functions with values 1 or zero. We get for this case

$$R_{spon} = \frac{n}{\tau_o} = \frac{p}{\tau_o} \tag{3.8.15}$$

iii) **Weak injection:** In this case we can use the Boltzmann distribution to describe the Fermi functions. We have

$$f^e \cdot f^h \cong \exp\left\{-\frac{(E_c - E_{Fn})}{k_BT}\right\} \exp\left\{-\frac{(E_{Fp} - E_v)}{k_BT}\right\} \cdot \exp\left\{-\frac{(\hbar\omega - E_g)}{k_BT}\right\} \quad (3.8.16)$$

The spontaneous emission rate now becomes

$$R_{spon} = \frac{1}{2\tau_o}\left(\frac{2\pi\hbar^2 m_r^*}{k_BTm_e^*m_h^*}\right)^{3/2} np \quad (3.8.17)$$

If we write the total charge as equilibrium charge plus excess charge,

$$n = n_o + \Delta n; p = p_o + \Delta n \quad (3.8.18)$$

we have for the excess carrier recombination (note that at equilibrium the rates ofrecombination and generation are equal)

$$R_{spon} \cong \frac{1}{2\tau_o}\left(\frac{2\pi\hbar^2 m_r^*}{k_BTm_e^*m_h^*}\right)^{3/2} (\Delta np_o + \Delta pn_o) \quad (3.8.19)$$

If $\Delta n = \Delta p$, we can define the rate of a single excess carrier recombination as

$$\frac{1}{\tau_r} = \frac{R_{spon}}{\Delta n} = \frac{1}{2\tau_o}\left(\frac{2\pi\hbar^2 m_r^*}{k_BTm_e^*m_h^*}\right)(n_o + p_o) \quad (3.8.20)$$

At low injection τ_r is much larger than τ_o, since at low injection, electrons have a low probability to find a hole with which to recombine. iv) **Inversion condition:** Another useful approximation occurs when the electron and hole densities are such that $f^e + f^h = 1$. This is the condition for inversion when the emission and absorption coefficients become equal. If we assume in this case $f^e \sim f^h = 1/2$, we get the approximate relation

$$R_{spon} \cong \frac{n}{4\tau_o} \cong \frac{p}{4\tau_o} \quad (3.8.21)$$

The recombination lifetime is approximately $4\tau_o$ in this case. This is a useful result to estimate the threshold current of semiconductor lasers.

Example 3.11 Optical radiation with a power density of 1.0 kW/cm^2 impinges on GaAs. The photon energy is 1.5 eV and the absorption coefficient is 3×10^3 cm^{-1}. Calculate the carrier generation rate at the surface of the sample. If the $e - h$ recombination time is 1 ns, calculate the steady state excess carrier density.

At the surface the carrier generation rate is

$$\begin{aligned} G(0) &= \frac{(3 \times 10^3 \text{ cm}^{-1})(10^3 \text{ W cm}^{-2})}{(1.5 \times 1.6 \times 10^{-19} \text{ J})} \\ &= 1.25 \times 10^{25} \text{ cm}^{-3}\text{s}^{-1} \end{aligned}$$

The excess carrier density is

$$\delta_n = \delta_p = 1.25 \times 10^{25} \times 10^{-9} = 1.25 \times 10^{16} \text{ cm}^{-3}$$

3.8.2 Schockley Read Hall Statistics

Semiconductor behavior is determined primarily by controlled impurities. Shallow impurities give rise to dopants, while deep impurities give rise to traps. In either case, the occupancy of all states, whether in the bands or the gap, is determined by the occupancy function.

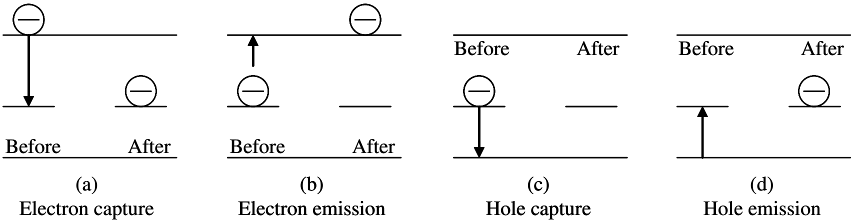


Figure 3.22: Exchanges with the conduction band are dealt as electron capture and emission, whereas exchanges with the valence band are considered hole capture and emission. The arrows indicate electron transitions

In equilibrium, the occupancy function for these states may be written:

$$f = \frac{1}{1 + \exp\left(\left(E_t - E_f\right)/k_B T\right)} \tag{3.8.22}$$

where E_t is the trap energy and E_f is the Fermi energy. For non-equilibrium a quasi-fermi level should be used, which in general applies to each set of states separately e.g. the conduction band, valence band, and each group of traps separately. Each process shown in figure 3.22 has a rate, r.

$$r_a \propto n \cdot N_t \left(1 - f\right) \tag{3.8.23}$$

where n is the concentration of available electrons and the $N_t \left(1 - f\right)$ term represents the concentration of empty traps. To calculate the proportionality constant, we recognize that electrons must be in the vicinity of the trap to be captured. We call this region σ_n cm^{-2}, a capture cross section as shown in figure 3.23.

The numbers of electrons that sweep past a trap in every second are contained in the volume defined by:

$$V = \sigma_n \cdot v_{th} \tag{3.8.24}$$

with units of cm^3/s where v_{th} is the thermal velocity of the electron. Those electrons contained in the volume described by this product in a given unit of time will be captured by the trap.

Consider an electron as shown in figure 3.24, v_{th} cms away from the trap position, x_0. After 1 second the electron will be at x_0, and therefore in the capture cross section of the trap. Any electron $v_{th} + \Delta L_2$ cms away will, after 1 second still be ΔL_2 away (case 2) from x_0 and hence not be captured. All electrons closer than v_{th} cms away (as for case 3 of the electron ΔL_3 cms

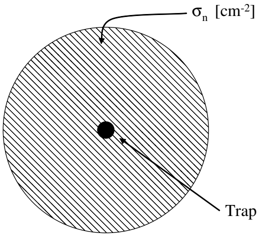


Figure 3.23: Picture of the capture cross section

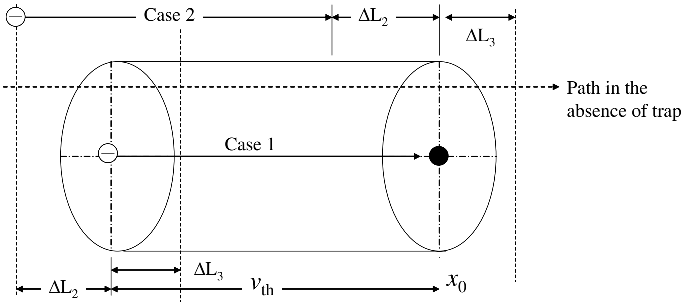


Figure 3.24: Electrons within the “volume” above will be captured by the trap

closer) would have intersected the capture cross section and be captured. Hence all electrons in the volume $V = \sigma_n \cdot v_{th}$ will be captured each second by available empty traps. Thus the number of electrons available to be captured per second is

$$n\sigma_n v_{th} \tag{3.8.25}$$

and recalling the concentration of available empty traps is $N_t\,(1-f)$, then the rate, r_a can be written:

$$r_a = v_{th}\sigma_n n N_t\,(1-f) \tag{3.8.26}$$

OR, the proportionality constant is $\sigma_n v_{th}$ (for the rate of electron capture)

$$v_{th} = \sqrt{\frac{2E}{m^*}} = \sqrt{2\cdot\frac{3k_BT}{2m^*}} \tag{3.8.27}$$

where E is thermal energy, $3/2k_BT$ in three dimensions. Thus

$$v_{th} = \sqrt{\frac{3k_BT}{m^*}} \simeq 10^7 \mathrm{cm/sec} \tag{3.8.28}$$

For the electron emission process, b,

$$r_b = e_n N_t f \tag{3.8.29}$$

where e_n is the emission rate from the trap and N_{tf} is the concentration of occupied traps. The capture rate for holes, process c, will be analogous to process a with the difference that holes are captured by occupied traps.

$$r_c = v_{th}\sigma_p p N_t f$$

Finally, the emission of holes has a rate:

$$r_d = e_p N_t \left(1 - f\right) \tag{3.8.30}$$

where e_p is the emission probability for holes. The next step is to determine the emission probabilities e_n and e_p. In general this is a very difficult problem since, f is known only in equilibrium. So first consider the equilibrium values of e_n, and e_p. In equilibrium transition rates into and out of the conduction band must be equal, or $r_a = r_b$. Inserting

$$n = N_c \ \exp\left(-\left(E_C - E_F\right)/k_BT\right) = n_i \ \exp\left(\left(E_F - E_i\right)/k_BT\right) \tag{3.8.31}$$

into $r_a = r_b$ leads to:

$$\mathrm{e}_n = v_{th}\sigma_n n_i \ \exp\left(\left(E_t - E_i\right)/k_BT\right) \tag{3.8.32}$$

or

$$e_n = v_{th}\sigma_n N_C \ \exp\left(-\left(E_C - E_t\right)/k_BT\right) \tag{3.8.33}$$

Thus the emission probability of electrons into the conduction band rises exponentially as the trap gets closer to E_C which we expect intuitively. From $r_c = r_d$ and

$$p = N_V \exp\left[-\left(E_f - E_V\right)/k_BT\right]$$
$$e_p = v_{th}\sigma_p N_V \ \exp\left(-\left(E_V - E_t\right)/k_BT\right)$$
$$= v_{th}\sigma_p n_i \ \exp\left(+\left(E_i - E_t\right)/k_BT\right)$$

In non-equilibrium (the case of most interest) f is unknown and has to be calculated. To do so, rate equations are solved. Assume that non-equilibrium is generated by optical excitation resulting in a generation rate of G_L electron-hole pairs/second. We also assume that the emission rates, e_n, and e_p are not a function of illumination and the same as that calculated at equilibrium.

In steady state, the concentration of electrons, n_n and holes, p_n in an $n-$type semiconductor is not a function of time and from figure 3.25 we get:

$$\frac{dn_n}{dt} = G_L - \left(r_a - r_b\right) = 0 \tag{3.8.34}$$

$$\frac{dP_n}{dt} = G_L - \left(r_c - r_d\right) = 0 \tag{3.8.35}$$

$$\therefore r_a - r_b = r_c - r_d \tag{3.8.36}$$

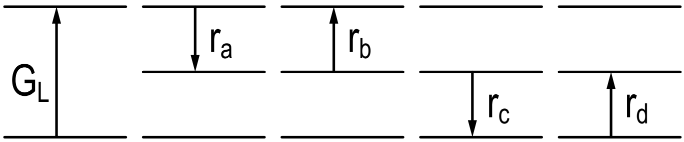


Figure 3.25: Possible Recombination processes

or the net capture rate of electrons = net capture rate of holes. This leads us to:

$$v_{th}\sigma_n n N_t \left[1 - f(E_t)\right] - v_{th} n_i \ \exp\left(E_{t-i}/k_B T\right) N_t f(E_t) =$$
$$v_{th}\sigma_p p N_t f - v_{th}\sigma_p n_i \exp\left[\left(E_i - E_t\right)/k_B T\right] N_t \left[1 - f(E_t)\right]$$

Since we are in non-equilibrium, $f\left(E_t\right)$, the distribution function for the traps has to be calculated from the above equation, where we have substituted for r_a through r_d,

$$f\left(E_t\right) = \frac{\sigma_n n + \sigma_p n_i \ \exp\left(E_{i-t}/k_B T\right)}{\sigma_n \left[n + n_i \ \exp\left(E_{t-i}/k_B T\right)\right] + \sigma_p \left[p + n \cdot \ \exp\left(E_{i-t}/k_B T\right)\right]} \tag{3.8.37}$$

where for compactness we have used the notation: $E_{i-t} = E_i - E_t$ and vice versa. Resubstituting to find a net rate of recombination:

$$U = r_a - r_b = r_c - r_d \tag{3.8.38}$$

leads to:

$$U = \frac{\sigma_p \sigma_n v_{th} N_t \left(pn - n_i^2\right)}{\sigma_n \left[n + n_i \ \exp\left(E_{t-i}/k_B T\right)\right] + \sigma_p \left[p + n_i \ \exp\left(E_{i-t}/k_B T\right)\right]} \tag{3.8.39}$$

Let us now consider some special cases:

1. for $\sigma_n = \sigma_p = \sigma$
$$U = \sigma v_{th} N_t \ \frac{pn - n_i^2}{n + p + 2n_i \cosh\left(E_{t-i}/k_B T\right)} \tag{3.8.40}$$

2. for $\sigma_n = \sigma_p = \sigma_p$ and when $E_t = E_i$
$$U = \frac{1}{\tau}\frac{pn - n_i^2}{n + p + 2n_i} \tag{3.8.41}$$

We see clearly that $pn - n_i^2$ is the driving force for recombination. We can also see that $n + p + 2n_i$ is a resistance to recombination term, which is minimized when $n + p$ is minimized. For low level injection, we assume that $n_n \gg p_n$ and

$$n_n \gg n_i \ \exp\left(\left(E_t - E_i\right)/k_B T\right) \tag{3.8.42}$$

as $E_T \sim E_i$ for efficient recombination. Then the recombination rate becomes:

$$U = \frac{\sigma_p \sigma_n v_{th} N_t}{\sigma_n n_n}\left[n_n p_n - n_i^2\right] \quad (3.8.43)$$

$$= \sigma_p v_{th} N_t \left[p_n - n_i^2/n_n\right] \quad (3.8.44)$$

$$= \sigma_p v_{th} N_t \left[p_n - p_{n0}\right] \quad (3.8.45)$$

$$U = \frac{\Delta p_n}{\tau_p} \quad (3.8.46)$$

where the minority carrier lifetime, τ_p is defined as

$$\frac{1}{\tau_p} = \sigma_p v_{th} N_t \quad (3.8.47)$$

Here the rate limiting step is the capture of the minority carrier. This is also achieved by recognizing the hole capture rate, r_c is the dominant step. In an $n-$type semiconductor, since E_F is close to the conduction band and $f(E_T) \rightarrow\ 1$ which makes r_a and r_d both negligible. Typical values of σ are $10^{-15} - 10^{-16}\text{cm}^{-2}$.

Generation occurs when $n_i^2 \gg pn$. From equation 3.8.39

$$U = -\frac{\sigma_p \sigma_n v_{th} N_t n_i^2}{\sigma_n\left[n + n_i\ \exp\left(E_{t-i}/k_BT\right)\right] + \sigma_p\left[p + n_i\ \exp\left(E_{i-t}/k_BT\right)\right]}$$

$$= -\frac{\sigma_p \sigma_n v_{th} N_t n_i}{\sigma_n\ \exp\left(E_{t-i}/k_BT\right) + \sigma_p n_i\ \exp\left(E_{i-t}/k_BT\right)}$$

For the case $\sigma_n = \sigma_p = \sigma$

$$U = -\frac{\sigma v_{th} N_t n_i}{2\cosh\left(E_{t-i}/k_BT\right)} \quad (3.8.48)$$

Thus, generation rate peaks when the trap energy is at mid-gap:

$$U = -\frac{n_i}{2\tau} \quad (3.8.49)$$

when $E_i \rightarrow E_t$ the lifetime

$$\tau = \frac{1}{\sigma v_{th} N_T}$$

3.9 CURRENT CONTINUITY(The law of conservation of electrons and holes separately)

In the previous sections we have considered several elements of non-equilibrium phenomena in semiconductors. These include drift and diffusion, carrier generation and recombination.

If we consider a volume of space in which charge transport and recombination is taking place, we have the simple equality (see figure 3.26a) As a result of consideration of particle current,

Net Rate of particle flow = Particle flow rate due to current −
Particle loss rate due to recombination + Particle gain due to generation.

Let us now collect the various terms in this continuity equation. If δn is the excess carrier density in the region, the recombination rate R in the volume $A \cdot \Delta x$ shown in figure 3.26 may be written approximately as

$$R = \frac{\delta n}{\tau_n} \cdot A \cdot \Delta x \tag{3.9.1}$$

where τ_n is the electron recombination time per excess particle due to both the radiative and the nonradiative components. The particle flow rate into the same volume due to the current J_n is given by the difference of particle current coming into the region and the particle current leaving the region,

$$\left[\frac{J_n(x)}{(-e)} - \frac{J_n(x+\Delta x)}{(-e)}\right] A \cong \frac{1}{e}\,\frac{\partial J_n(x)}{\partial x}\Delta x \cdot A \tag{3.9.2}$$

If G is the generation rate per unit volume, the generation rate in the volume $A \cdot \Delta x$ is $GA\Delta x$. The rate of electron build up in the volume $A \cdot \Delta x$ is then

$$A \cdot \Delta x \left[\frac{\partial n(x,t)}{\partial t} \equiv \frac{\partial \delta n}{\partial t} = \frac{1}{e}\,\frac{\partial J_n(x)}{\partial x} - \frac{\delta n}{\tau_n}\right] \tag{3.9.3}$$

where $\delta n/\tau_n$ is $U = G - R$, the net recombination rate of electrons. We have similar terms for holes, collecting the various terms we have, for the electrons and holes, the continuity equations (note the sign difference in the particle current density for electrons and holes)

$$\begin{aligned} \frac{\partial \delta n}{\partial t} &= \frac{1}{e}\,\frac{\partial J_n(x)}{\partial x} - \frac{\delta n}{\tau_n} && (3.9.4)\\ \frac{\partial \delta p}{\partial t} &= -\frac{1}{e}\,\frac{\partial J_p(x)}{\partial x} - \frac{\delta p}{\tau_p} && (3.9.5) \end{aligned}$$

Using these expressions, the the diffusion currents are

$$\begin{aligned} J_n(diff) &= eD_n\frac{\partial \delta n}{\partial x} && (3.9.6)\\ J_p(diff) &= -eD_p\frac{\partial \delta p}{\partial x} && (3.9.7) \end{aligned}$$

We get

$$\begin{aligned} \frac{\partial \delta n}{\partial t} &= D_n\frac{\partial^2 \delta n}{\partial x^2} - \frac{\delta n}{\tau_n} && (3.9.8)\\ \frac{\partial \delta p}{\partial t} &= D_p\frac{\partial^2 \delta p}{\partial x^2} - \frac{\delta p}{\tau_p} && (3.9.9) \end{aligned}$$

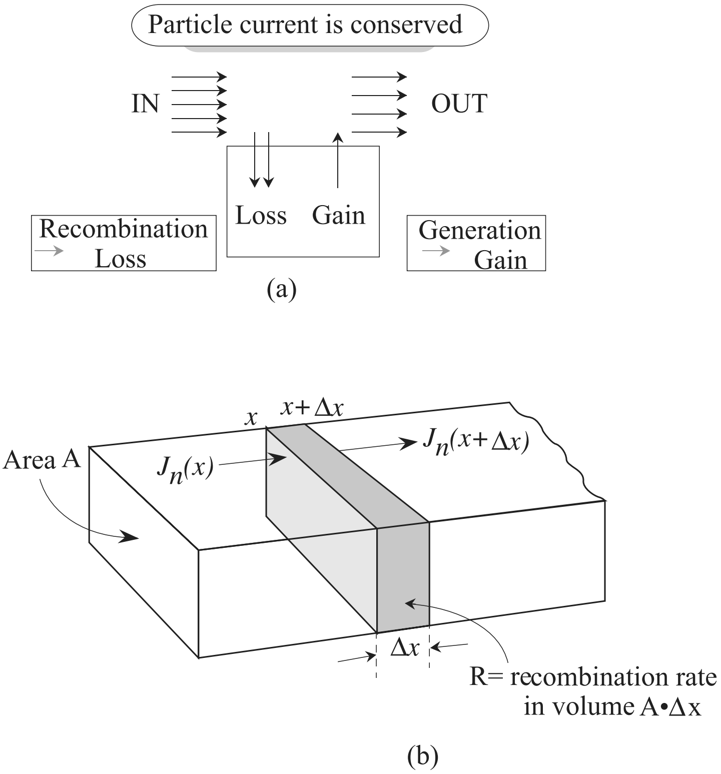


Figure 3.26: (a) A conceptual description of the continuity equation. (b) Geometry used to develop the current continuity equation.

the time dependent continuity equation for electrons and holes, valid separately. These equations will be used when we discuss the transient time responses of the p-n diodes and bipolar transistors. These equations are also used to study the steady-state charge profile in these devices. In steady state we have (the time derivative is zero)

$$\begin{aligned} \frac{d^2\delta n}{dx^2} &= \frac{\delta n}{D_n\tau_n} = \frac{\delta n}{L_n^2} && (3.9.10) \\ \frac{d^2\delta p}{dx^2} &= \frac{\delta p}{D_p\tau_p} = \frac{\delta p}{L_p^2} && (3.9.11) \end{aligned}$$

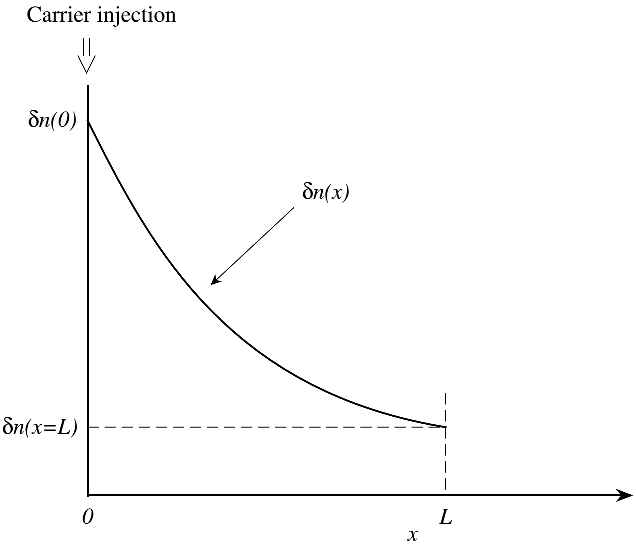


Figure 3.27: Electrons are injected at $x = 0$ into a sample. At $x = 0$, a fixed carrier concentration is maintained. The figure shows how the excess carriers decay into the semiconductor.

where $L_n(L_p)$ defined as $D_n\tau_n(D_p\tau_p)$ are called the diffusion lengths We will see below that the diffusion length represents the distance an electron (hole) will travel before it recombines with a hole (electron). Let us examine the schematic of the equation derived above. Consider the case where an excess electron density $\delta n(0)$ is maintained at the semiconductor at $x = 0$, as shown in figure 3.27. At some point L in the semiconductor the excess carrier density is fixed at $\delta(L)$. We are interested in finding out how the excess density varies with position. The general solution of the second-order differential equation 3.9.11 is

$$\delta n(x) = A_1 e^{x/L_n} + A_2 e^{-x/L_n}$$

Using the boundary conditions at $x = 0$ and $x = L$, we find that the coefficients A_1 and A_2 are

$$\begin{aligned} A_1 &= \frac{\delta n(L) - \delta n(0)e^{-L/L_n}}{e^{L/L_n} - e^{-L/L_n}} \\ A_2 &= \frac{\delta n(0)e^{L/L_n} - \delta n(L)}{e^{L/L_n} - e^{-L/L_n}} \end{aligned} \tag{3.9.12}$$

This gives for the excess carrier concentration

$$\delta n(x) = \frac{\delta n(0)\sinh\left(\frac{L-x}{L_n}\right) + \delta n(L)\sinh\left(\frac{x}{L_n}\right)}{\sinh\left(\frac{L}{L_n}\right)} \tag{3.9.13}$$

There are two important cases that occur in bipolar devices, we will examine them here:
(i) $L \gg L_n$ and $\delta n(L) = 0$: In this case the semiconductor sample is much longer than L_n. This happens in the case of the long p-n diode , which will be discussed in chapter 4 For this case we have

$$\delta n_p(x) = \delta n_p(0)e^{-x/L_n} \tag{3.9.14}$$

Thus the carrier density simply decays exponentially into the semiconductor.
(ii) $L \ll L_n$: This case is very important in discussing the operation of bipolar transistors and narrow p-n diodes . Using the small x expansion for $\sinh(x)$

$$\sinh(x) = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \ldots$$

and retaining only the first-order terms we get

$$\delta n_p(x) = \delta n_p(0) - \frac{x\left[\delta n_p(0) - \delta n_p(L)\right]}{L} \tag{3.9.15}$$

i.e., in this case the carrier density goes linearly from one boundary value to the other.

Note that once the carrier density is known the diffusion current can be simply obtained by taking its derivative.

Let us examine the case where excess carriers are injected into a thick semiconductor sample. As the excess carriers diffuse away into the semiconductor they recombine. The diffusion length L_n represents the distance over which the injected carrier density falls to $1/e$ of its original value. It also represents the average distance an electron will diffuse before it recombines with a hole. This can be seen as follows.

The probability that an electron survives up to point x without recombination is, from equation 3.9.15,

$$\boxed{\frac{\delta n_p(x)}{\delta n_p(0)} = e^{-x/L_n}} \tag{3.9.16}$$

The probability that it recombines in a distance Δx is

$$\frac{\delta n_p(x) - \delta n_p(x+\Delta x)}{\delta n_p(x)} = -\frac{\Delta x}{\delta n_p(x)}\,\frac{d\delta n_p(x)}{dx} = \frac{1}{L_n}\Delta x \tag{3.9.17}$$

where we have expanded $\delta n_p(x + \Delta x)$ in terms of $\delta n_p(x)$ and the first derivative of δn_p. Thus the probability that the electron survives up to a point x and then recombines is

$$P(x)\Delta x = \frac{1}{L_n}e^{-x/L_n}\Delta x \tag{3.9.18}$$

Thus the average distance an electron can move and then recombine is

$$\begin{aligned} < x > &= \int_o^\infty x P(x) dx = \int_0^\infty \frac{x e^{-x/L_n}}{L_n} dx \\ &= L_n \end{aligned} \tag{3.9.19}$$

This average distance ($= \sqrt{D_n \tau_n}$) depends upon the recombination time and the diffusion constant in the material. In the derivations of this section, we used a simple form of recombination rate

$$R = \frac{\delta n_p}{\tau_n} \tag{3.9.20}$$

where τ_n is given in terms of the radiative and nonradiative rates as

$$\frac{1}{\tau_n} = \frac{1}{\tau_r} + \frac{1}{\tau_{nr}} \tag{3.9.21}$$

The simple $\delta n_p / \tau_n$ form is valid, for example, for minority carrier recombination ($p \gg n$). These equations are therefore used widely to discuss minority carrier injection.

3.10 PROBLEMS

Problem 3.1 The electron mobility of Si at 300 K is 1400 cm^2/V·s. Calculate the mean free path and the energy gained in a mean free path at an electric field of 1 kV/cm. Assume that the mean free path = $v_{th} \cdot \tau_{sc}$, where v_{th} is the thermal velocity of the electron ($v_{th} \sim$ 2.0 × 10^7 cm/s).

Problem 3.2 The mobility of electrons in the material InAs is $\sim$ 35,000 cm^2/V·s at 300K compared to a mobility of 1400 cm^2/V·s for silicon. Calculate the scattering times in the two semiconductors. The electron masses are 0.02 m_0 and 0.26 m_0 for InAs and Si, respectively.

Problem 3.3 Calculate the ionized impurity limited mobility ($N_D = 10^{16}$ cm^{-3}; 10^{17} cm^{-3}) in GaAs from 77 K to 300 K.

Problem 3.4 If the measured room temperature mobility of electrons in GaAs doped n-type at 5 × 10^{17} cm^{-3} is 3500 cm^{2}V^{-1} s^{-1} calculate the relaxation time for phonon scattering.

Problem 3.5 Calculate the alloy scattering limited mobility in In$_{0.53}$Ga$_{0.47}$As as a function of temperature from 77 K to 400 K. Assume an alloy scattering potential of 1.0 eV.

Problem 3.6 The velocity of electrons in silicon remains $\sim$1 × 10^7 cm s^{-1} between 50 kVcm^{-1} and 200 kVcm^{-1}. Estimate the scattering times at these two electric fields.

Problem 3.7 The power output of a device depends upon the maximum voltage that the device can tolerate before impact-ionization-generated carriers become significant (say 10% excess carriers). Consider a device of length L, over which a potential V drops uniformly. What is the maximum voltage that can be tolerated by an Si and a diamond device for L = 2 μm and L = 0.5 μm? Use the values of the critical fields given in this chapter.

Problem 3.8 The electron concentration in a Si sample is given by

$$n(x) = n(0)\exp(-x/L_n); \;\; x > 0$$

with $n(0) = 10^{18}$ cm^{-3} and $L_n = 3.0$ μm. Calculate the diffusion current density as a function of position if D_n=35 cm^2/s.

Problem 3.9 Consider a GaAs sample doped n-type at 10^{16} cm^{-3} on which an experiment is done. At time $t = 0$ an external stimulus introduces excess electrons at a point $x = 0$. The excess charge is detected at $x = 10.0$ μm in the absence of any applied field after 2.5×10^{-9} s.
Use this information to answer the following:
• What is the diffusion coefficient of electrons?
• How much time will electrons travel (by drift) 1.0 μm under an applied field of 1.0 kV/cm? Assume that the velocity–field relation is linear.
• What is the conductivity of this sample? Assume that the electron effective mass is 0.067 m_0.

Problem 3.10 In a p-type GaAs doped at $N_a = 10^{18}$ cm^{-3}, electrons are injected to produce a minority carrier concentration of 10^{16} cm^{-3}. What is the rate of photon emission assuming that all e-h recombination is due to photon emission ? What is the optical output power? The photon energy is $\hbar\omega$ = 1.41 eV and the radiative lifetime is 1.0 ns.

Problem 3.11 Calculate the electron carrier density needed to push the electron Fermi level to the conduction bandedge in GaAs. Also calculate the hole density needed to push the hole Fermi level to the valence bandedge. Calculate the results for 300 K and 77 K.

Problem 3.12 A photodetector uses pure silicon as its active region. Calculate the dark conductivity of the detector (i.e., conductivity when no light is shining on the detector). Light with intensity 10^{-3} W/cm^2 shines on the device. Calculate the conductivity in presence of light.

$$\begin{aligned} \mu_n &= 1000 \text{ cm}^2/\text{V}\cdot\text{s} \\ \mu_p &= 400 \text{ cm}^2/\text{V}\cdot\text{s} \\ \alpha &= 10^3 \text{ cm}^{-1} \\ \tau_r &= 10^{-7} \text{ s} \end{aligned}$$

Problem 3.13 Electrons are injected into a $p-$typesilicon sample at 300 K. The electron-hole radiative lifetime is 1 ensuremathμs. The sample also has midgap traps with a cross-section of 10^{-15} $\mathrm{cm}^{-2}$and a density of 10^{16} cm^{-3}. Calculate the diffusion length for the electrons if the diffusion coefficient is 30 $\mathrm{cm}^2\mathrm{s}^{-1}$.

Problem 3.14 Assume that silicon has a midgap impurity level with a cross-section of 10^{-14} cm^2. The radiative lifetime is given to be 1 ensuremathμsat 300 K. Calculate the maximum impurity concentration that will ensure that $\tau_r < \tau_{nr}$.

Problem 3.15 When holes are injected into an $n-$typeohmic contact, they decay within a few hundred angstroms. Thus one can assume that the minority charge density goes to zero at an ohmic contact. Discuss the underlying physical reasons for this boundary condition.

Problem 3.16 Electrons are injected into $p-$typeGaAs at 300 K. The radiative lifetime for the electrons is 2 ns. The material has 10^{15} impurities with a cross-section of 10^{-14} cm^2. Calculate the distance the injected minority charge will travel before 50% of the electrons recombine with holes. The diffusion coefficient is 100 cm^2/s.

Problem 3.17 Electrons are injected into $p-$typesilicon at $x = 0$. Calculate the fraction of electrons that recombine within a distance L where L is given by (a) 0.5 μm, (b) 1.0 μm, and (c) 10.0 μm. The diffusion coefficient is 30 $\mathrm{cm}^2\mathrm{s}^{-1}$and the minority carrier lifetime is 10^{-7}s.

Problem 3.18 Consider a Si sample of length L. The diffusion coefficient for electrons is 25 $\mathrm{cm}^2\mathrm{s}^{-1}$ and the electron lifetime is 0.01 μs. An excess electron concentration is maintained at $x = 0$ and $x = L$. The excess concentrations are:

$$\delta n(x=0) = 2.0 \times 10^{18} \quad \mathrm{cm}^{-3}; \; \delta n(x=L) = -1.0 \times 10^{14} \quad \mathrm{cm}^{-3}$$

Calculate and plot the excess electron distribution from $x = 0$ to $x = L$. Do the calculations for the following values of L:

$$\begin{aligned} L &= 10.0\,\mu\mathrm{m} \\ L &= 5.0\,\mu\mathrm{m} \\ L &= 1.0\,\mu\mathrm{m} \\ L &= 0.5\,\mu\mathrm{m} \end{aligned}$$

Note that the excess electron distribution starts out being nonlinear in space for the long structure, but becomes linear between the two boundary values for the short structure.

Problem 3.19 An experiment is carried out at 300 K on n-type Si doped at $N_d = 10^{17}$ cm^{-3}. The conductivity is found to be 10.0 $(\Omega\ \mathrm{cm})^{-1}$.

When light with a certain intensity shines on the material the conductivity changes to 11.0 $(\Omega\ \mathrm{cm})^{-1}$. The light is turned off at time 0 and it is found that at time 1.0 μs the conductivity is 10.5 $(\Omega\ \mathrm{cm})^{-1}$. The light-induced excess conductivity is found to decay as

$$\delta\sigma = \delta\sigma(0)\exp\left(-\frac{t}{\tau}\right)$$

where τ is the carrier lifetime. Calculate the following:
• What fraction of the donors is ionized?
• What is the diffusion length of holes in this material?

Use the following data:

$$\mu_n = 1100 \ \text{cm}^2/\text{V} \cdot \text{s} : \ \ \mu_p = 400 \ \text{cm}^2/\text{V} \cdot \text{s}$$

Problem 3.20 Calculate the area density of surface states that would lead to *surface generation* rate of a fully depleted surface to equal *twice the generation rate in the surface depletion region.* Consider the states to be characterized by a capture cross-section of $10^{-15} cm^2$ and the thermal velocity to be $10^7 \frac{cm}{s}$. Assume the surface depletion region to be $1\mu m$ wide and the time constant $\tau_0 = 1\mu s$.

3.11 FURTHER READING

- **Transport in crystalline materials**
  - M. Lundstrom, Fundamentals of Carrier Transport (Modular Series on Solid State Devices), eds. by G.W, Neudeck and R.F. Pierret, Addison-Wesley, Reading, MA, vol. X (1990).
  - J. Singh, Modern Physics for Engineers, Wiley-Interscience, New York (1999).
  - K. Seeger, Semiconductor Physics, Springer Verlag, New York (1985).
  - J. Singh, Electronic and Optoelectronic Properties of Semiconductors, Cambridge University Press (2003).
- **Transport in disordered materials**
  - N.F. Mott and E.A. Davis, Electronic Processes in Non-Crystalline Materials, Clarenden Press, Oxford (1971).
  - A.J. Moulson and J.M. Herbert, Electroceramics: Materials, Properties, and Applications, Chapman & Hall (1992).

Chapter 4

JUNCTIONS IN SEMICONDUCTORS: *P-N* DIODES

4.1 Introduction

In the introduction to this textbook we examined how semiconductor devices are driving modern information technology. Cell phones, computer, internet, etc. all depend upon devices that will be discussed in the next several chapters. Semiconductor diodes (and Schottky diodes) discussed in this and the next chapter have rectifying properties. p-n diodes can be used as detectors and light emitters. Devices such as field effect and bipolar transistors are used for amplification and signal generation. They can also be used in digital technology as ON/OFF switches.

Semiconductor devices operate on the basis of the basic principle that the conducting and optical properties of semiconductors can be altered easily and rapidly. One way this can be done is through the use of junctions between dissimilar materials. Junctions can form between $n-$type and $p-$type materials, between materials with different bandgaps, and between metals and semiconductors. In this chapter we will discuss the p-n junction.

4.2 *P-N* JUNCTION IN EQUILIBRIUM

The p-n junction is one of the most important junctions in solid-state electronics. The junction is used as a device in applications such as rectifiers, waveform shapers, variable capacitors, lasers, detectors, etc. The key ingredient of the bipolar transistor, which is one of the most important electronic devices is a p-n junction,. To understand how a p-n junction operates we need to know: i) What are the carrier distributions for electrons and holes in the material? ii) What are the physical processes responsible for current flow in the structure?

Let us start with p- and n-type semiconductors without forming a junction as shown in figure 4.1(a). The electron affinity $e\chi$, defined as the energy difference between the conduction band and vacuum level, is shown along with the work function ($e\phi_{sp}$ or $e\phi_{sn}$). The work function represents the energy required to remove an electron from the semiconductor to the "free" vacuum level and is the difference between the vacuum level and the Fermi level.

Let us now examine what happens when the p- and n-type materials are made to form a junction. In the absence of any applied bias, there is no net current in the system and the Fermi level is uniform throughout the structure. In figure 4.1(b) we show a schematic of the band diagram of a p-n junction.

Majority carriers near the interface on both sides diffuse across the junction (holes from p side and electrons from n side), as a result of the difference in electron and hole densities across the junction. Most of the electrons which diffuse to the p-side recombine with holes, and most of the holes which diffuse to the n-side recombine with electrons. As a result, a region is formed near the junction that has been depleted of mobile carriers. An electric field exists in this depletion region that sweeps out any electrons and holes that enter the region.

Three regions can be identified as seen in figure 4.1(b):

i) The p-type region where the material is neutral and the bands are flat. The density of acceptors exactly balances the density of holes (assuming that all of the acceptors are ionized);

ii) The n-type region away from the junction where again the material is neutral and the density of immobile donors exactly balances the free electron density. Again we assume that all of the donors are ionized. In general the majority carrier (holes in the p-region and electrons in the n-region) densities are equal to the density of ionized dopants as long as minority carrier densities are negligible.

iii) Around the junction there is a depletion region where the bands are bent and a field exists that sweeps the mobile carriers, leaving behind negatively charged acceptors in the p-region and positively charged donors in the n-region. The depletion region extends a distance W_p in the p-region and a distance W_n in the n-region.

Due to the field in the depletion region electrons or holes which enter the depletion region are swept away. Thus, once equilibrium is established, a drift current exists that counterbalances the diffusion current. Let us calculate the width of the depletion region, and the electric field. To obtain analytical results we make some simplifying assumptions:

i) The junction is uniformly doped.

ii) The mobile charge density in the depletion region is not zero, but it is much smaller than the background dopant density. To solve the Poisson equation we will assume that the mobile carrier density is essentially zero, the depletion approximation.

The various current flow terms in the diode are as follows: the electron drift current and electron diffusion current as well as the hole drift and hole diffusion current, as shown in figure 4.2(b). When there is no applied bias, these currents cancel each other individually. Let us consider these current components. The hole current density is

$$J_p(x) = e\left[\underbrace{\mu_p p(x)\mathcal{E}(x)}_{\text{drift}} - \underbrace{D_p \frac{dp(x)}{dx}}_{\text{diffusion}}\right] = 0 \qquad (4.2.1)$$

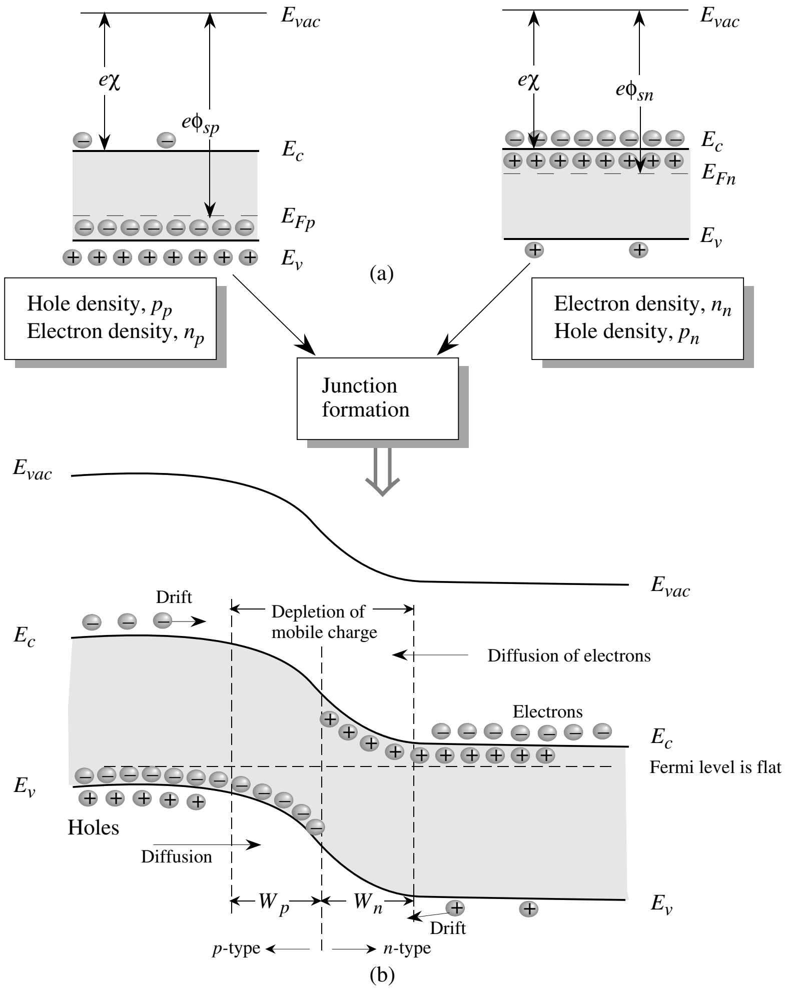


Figure 4.1: Forming a p-n junction (a) The p- and n-type regions before junction formation. The electron affinity $e\chi$ and work functions $e\phi_{sp}$ and $e\phi_{sn}$ are shown along with the Fermi levels. (b) A schematic of the junction and the band profile showing the vacuum level and the semiconductor bands.

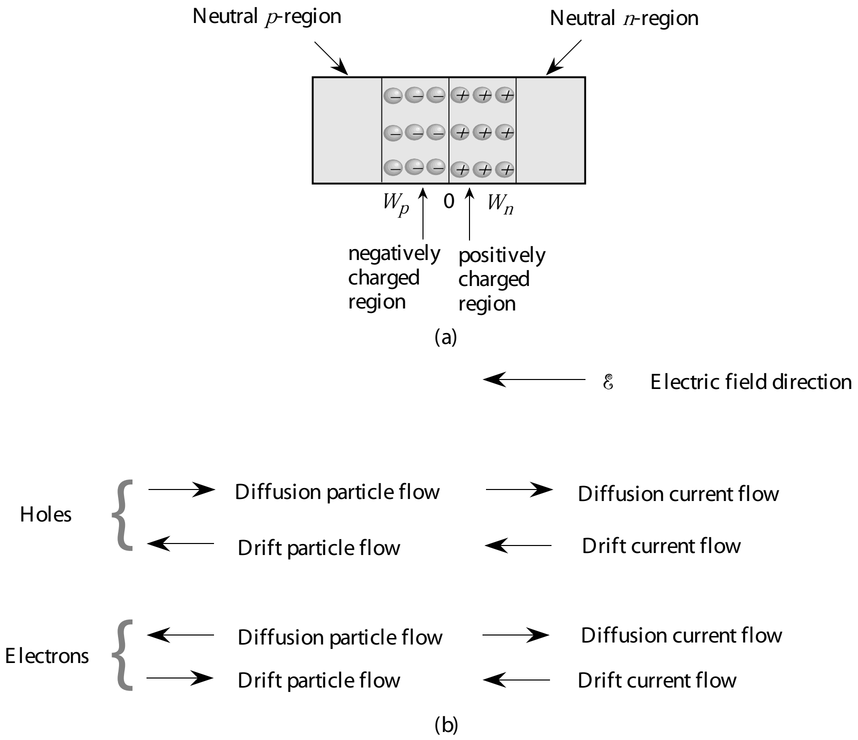


Figure 4.2: (a) Region of a p-n junction without bias, showing the neutral and depletion areas. (b) A schematic showing various current and particle flow components in the p-n diode at equilibrium.

The ratio of μ_p and D_p is given by the Einstein relation

$$\frac{\mu_p}{D_p} = \frac{e}{k_B T} \tag{4.2.2}$$

Since the Fermi level is uniform in the structure as we go from the p-side to the n-side, as shown in figure 4.3. As a result of bringing the p and n type semiconductors, a built-in voltage, V_{bi},is produced between the $n-$side and the $p-$side of the structure. As indicated in figure 4.3, the built-in voltage is given by

$$eV_{bi} = E_g - (E_c - E_F)_n - (E_F - E_v)_p$$

where the subscripts n and p refer to the n-side and p-side of the device. Using the Boltzmann approximation for the Fermi level (see equation 2.4.4)

$$(E_c - E_F)_n = -k_B T \ln(\frac{n_{no}}{N_c})$$

where n_{no} is the electron density on the n-side of the device. Assuming that all of the donors are ionized,

$$n_{no} = N_d$$

Similarly,

$$(E_F - E_v)_p = -k_B T \ln(\frac{p_{p0}}{N_v})$$

where p_{po} is the hole density on the p-side and is given by

$$p_p = N_a$$

This gives the built-in voltage

$$eV_{bi} = E_g + k_B T \ln(\frac{n_{n0} p_{p0}}{N_c N_v})$$

Using the relation for intrinsic carrier density

$$n_i^2 = N_c N_v \ \exp\left(-\frac{E_g}{k_B T}\right)$$

we get

$$V_{bi} = \frac{k_B T}{e} \ \ln(\frac{n_{n0} p_{p0}}{n_i^2}) \tag{4.2.3}$$

If n_{n0} and n_{p0} are the electron densities in the n-type and p-type regions, the law of mass action (i.e., the product np is constant) tells us that

$$n_{n0} p_{n0} = n_{p0} p_{p0} = n_i^2 \tag{4.2.4}$$

This gives for the built-in potential, V_{bi} = $V_n - V_p$ (figure 4.3)

$$V_{bi} = \frac{k_B T}{e} \ \ln \ \frac{p_{p0}}{p_{n0}} \tag{4.2.5}$$

or

$$V_{bi} = \frac{k_B T}{e} \ \ln \ \frac{n_{n0}}{n_{p0}} \tag{4.2.6}$$

We have the following equivalent expressions:

$$\frac{p_{p0}}{p_{n0}} = \ \exp\left(eV_{bi}/k_B T\right) = \frac{n_{n0}}{n_{p0}} \tag{4.2.7}$$

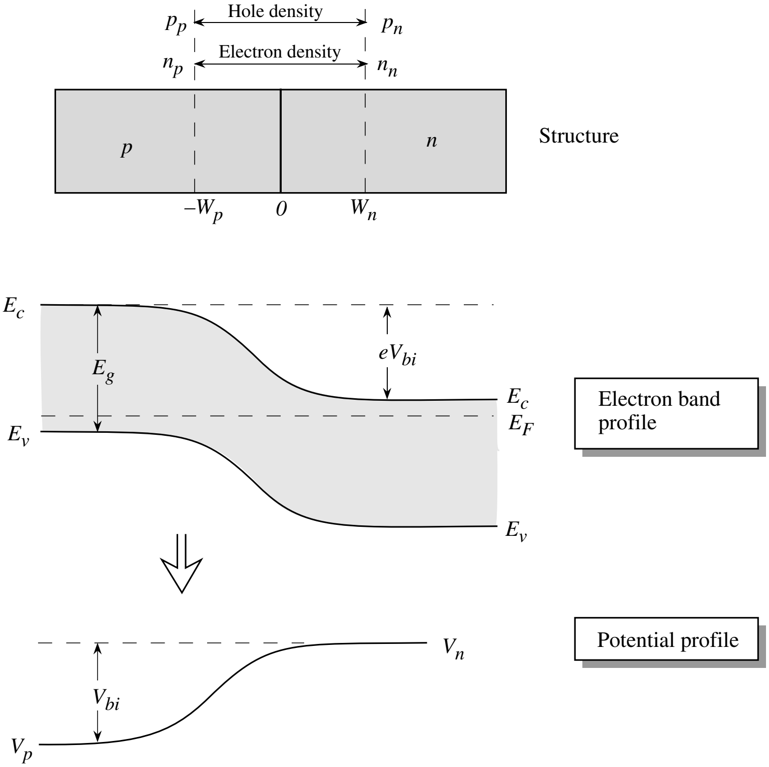


Figure 4.3: A schematic showing the p-n diode and the potential and band profiles. The voltage V_{bi} is the built-in potential at equilibrium.

In this relation V_{bi} is the built-in voltage in the absence of any external bias. Under the approximations discussed later, a similar relation holds when an external bias V is applied to alter V_{bi} to $V_{bi} - V$, and will be used when we calculate the effect of external potentials on the current flow.

We need to solve the Poisson equation to calculate the width of the depletion region for the diode under no applied bias. The calculation in the presence of a bias V will follow the same approach and V_{bi} will simply be replaced by $V_{bi} - V$, the total potential across the junction. Note that we have the equality

$$A\, W_p N_a = A\, W_n N_d \tag{4.2.8}$$

where A is the cross-section of the p-n structure and N_a and N_d are the uniform doping densities for the acceptors and donors.

The Poisson equation in the depletion approximation for various regions is

$$\begin{aligned}\frac{d^2V(x)}{dx^2} &= 0 \qquad -\infty < x < -W_p && (4.2.9)\\ \frac{d^2V(x)}{dx^2} &= \frac{eN_a}{\epsilon} \qquad -W_p < x < 0 && (4.2.10)\\ \frac{d^2V(x)}{dx^2} &= -\frac{eN_d}{\epsilon} \qquad 0 < x < W_n && (4.2.11)\\ \frac{d^2V(x)}{dx^2} &= 0 \qquad W_n < x < \infty && (4.2.12)\end{aligned}$$

Solving these equations gives the electric field in the p-side of the depletion region

$$\mathcal{E}(x) = -\frac{dV}{dx} = -\frac{eN_a x}{\epsilon} - \frac{eN_a W_p}{\epsilon} \qquad -W_p < x < 0 \qquad (4.2.13)$$

The electric field reaches a peak value at x = 0. The potential is given by integrating the field,

$$V(x) = \frac{eN_a x^2}{2\epsilon} + \frac{eN_a W_p x}{\epsilon} + \frac{eN_a W_p^2}{2\epsilon} + V_p \qquad -W_p < x < 0 \qquad (4.2.14)$$

For the n-side of the depletion region and n-side of the neutral region, we use the conditions

$$\begin{aligned}V(x) &= V_n \qquad W_n < x < \infty\\ \mathcal{E}(x) &= 0 && (4.2.15)\end{aligned}$$

where V_n is the potential at the neutral n-side. The electric field and potential on the n-side is found to be

$$\mathcal{E}(x) = \frac{eN_d x}{\epsilon} - \frac{eN_d W_n}{\epsilon} \qquad 0 < x < W_n \qquad (4.2.16)$$

$$V(x) = -\frac{eN_d x^2}{2\epsilon} + \frac{eN_d W_n x}{\epsilon} - \frac{eN_d W_n^2}{2\epsilon} + V_n \qquad 0 < x < W_n \qquad (4.2.17)$$

The potential difference between points $-W_p$ and 0 is

$$V(0) - V(-W_p) = \frac{eN_a W_p^2}{2\epsilon} \qquad (4.2.18)$$

Similarly,

$$V(W_n) - V(0) = \frac{eN_d W_n^2}{2\epsilon} \qquad (4.2.19)$$

Thus the built-in potential is

$$V(W_n) - V(-W_p) = V_{bi} = \frac{eN_d W_n^2}{2\epsilon} + \frac{eN_a W_p^2}{2\epsilon} \qquad (4.2.20)$$

Please note that in the above equation, and throughout this chapter, the $p-$type semiconductor (the semiconductor on the left hand side) is the reference electrode. In the case of the MOSFET as we will see in chapter 9 a different reference is used. As noted earlier charge neutrality gives us

$$N_d W_n = N_a W_p \tag{4.2.21}$$

and we get

$$W_p(V_{bi}) = \left\{ \frac{2\epsilon V_{bi}}{e} \left[\frac{N_d}{N_a(N_a + N_d)} \right] \right\}^{1/2} \tag{4.2.22}$$

$$W_n(V_{bi}) = \left\{ \frac{2\epsilon V_{bi}}{e} \left[\frac{N_a}{N_d(N_a + N_d)} \right] \right\}^{1/2} \tag{4.2.23}$$

$$W(V_{bi}) = \left[\frac{2\epsilon V_{bi}}{e} \left(\frac{N_a + N_d}{N_a N_d} \right) \right]^{1/2} \tag{4.2.24}$$

The expressions derived above can be extended to find the electric fields, potential, and depletion widths for arbitrary values of V_p and V_n under certain approximations. Thus we can use these equations directly when the diode is under external bias V, by simply replacing V_{bi} by V_{bi} - V.

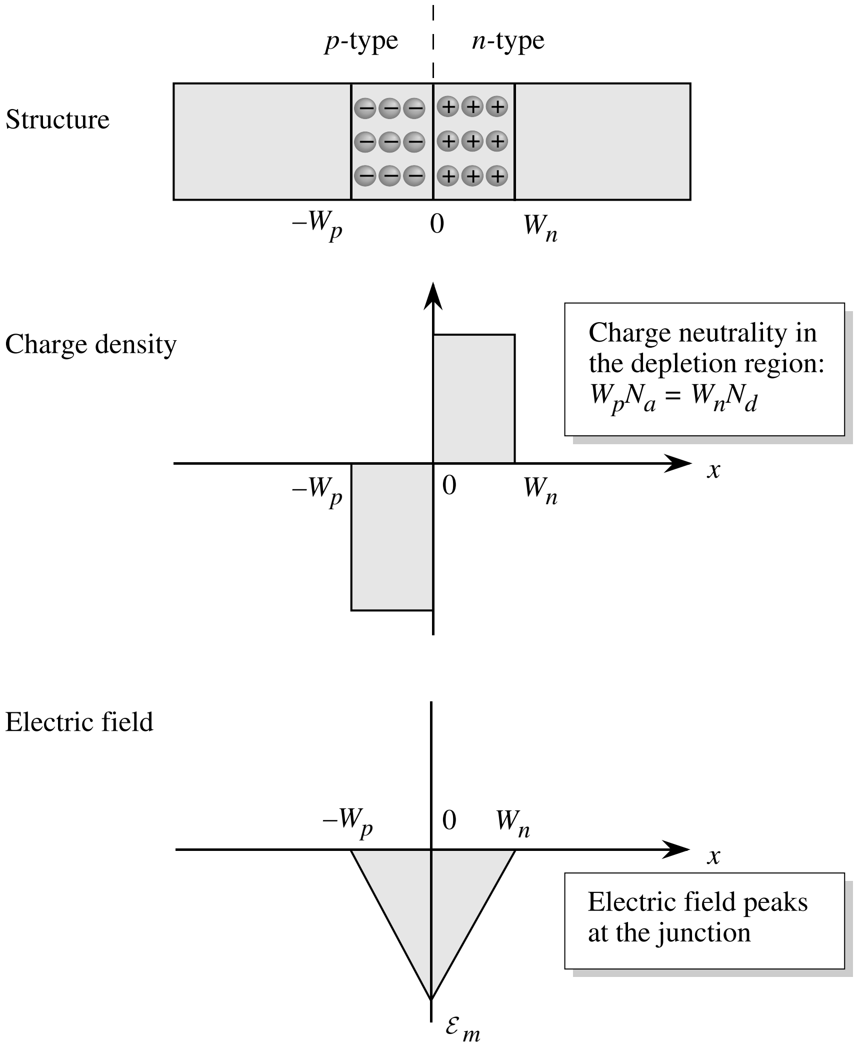
In figure 4.4 we show the charge and electric field profile. The electric field is nonuniform in the depletion region, peaking at the junction with a peak value.

$$\mathcal{E}_m = -\frac{eN_d W_n}{\epsilon} = -\frac{eN_a W_p}{\epsilon} \tag{4.2.25}$$

The sign of the field simply reflects the fact that in our study the field points toward the negative x-axis. It is important to note that if $N_a \gg N_d$, the depletion width W_p is much smaller than W_n. Thus a very strong field exists over a very narrow region in the heavily doped side of the junction. In such junctions (p^+n or n^+p) the depletion region exists primarily on the lightly doped side.

Example 4.1 A diode is fabricated on an n-type ($N_d = 10^{16}$ cm^{-3}) silicon substrate, on which a p-type region doped to 10^{18} cm^{-3} is created. Calculate the Fermi level positions in the p- and n-regions, determine the contact potential in the diode, and calculate the depletion widths on the p- and n-side. Using the effective density of states relations, we have ($N_c = 2.8 \times 10^{19}$ cm^{-3}; $N_v = 1 \times 10^{19}$ cm^{-3}at 300 K)

$$\begin{aligned} E_{Fn} &= E_c + k_B T \ \ln \frac{n_{n0}}{N_c} \\ &= E_c - (0.026 \text{ eV}) \times 7.937 \\ &= E_c - 0.206 \text{ eV} \\ E_{Fp} &= E_v - k_B T \ \ln \frac{p_{p0}}{N_v} \\ &= E_v + (0.026 \text{ eV}) \times 2.3 \\ &= E_v + 0.06 \text{ eV} \end{aligned}$$



Charge neutrality in the depletion region: $W_pN_a = W_nN_d$

Electric field peaks at the junction

Figure 4.4: The p-n structure, with the charge and the electric field profile in the depletion region. The electric field peaks at the junction as shown.

The built-in potential is given by

$$\begin{aligned} eV_{bi} &= E_g - 0.06 - 0.206 \text{ eV} \\ &= 1.1 - 0.06 - 0.206 \\ &= 0.834 \text{ eV} \end{aligned}$$

The depletion width on the p-side is given by

$$\begin{aligned} W_p(V_{bi}) &= \left\{ \frac{2\epsilon V_{bi}}{e} \left[\frac{N_d}{N_a(N_a + N_d)} \right] \right\}^{1/2} \\ &\cong \left\{ \frac{2 \times (11.9 \times 8.84 \times 10^{-12} \text{ F/m}) \times 0.834 \text{ (volts)}}{(1.6 \times 10^{-19} \text{ C})} \right. \\ & \qquad \left. \times \frac{10^{22} \text{ m}^{-3}}{10^{24} \text{ m}^{-3} \times (1.01 \times 10^{24} \text{ m}^{-3})} \right\}^{1/2} \\ &= 3.2 \times 10^{-9} \text{ m} = 32 \text{ Å} \end{aligned}$$

The depletion width on the n-side is 100 times longer:

$$W_n(V_o) = 0.32 \ \mu\text{m}$$

4.3 *P-N* DIODE UNDER BIAS

We have noted that in the absence of an applied bias, even though there is no current flowing in the diode, there are drift and diffusion currents that flow and exactly cancel each other. In the presence of the applied bias, the balance between the drift and diffusion currents is disturbed and a net current will flow. Under the following simplifying assumptions, one can use the formalism of the previous section to study the biased diode. These approximations are found to be valid under usual diode operating conditions.

- We assume that the change in carrier densities is small so that we can use the concept of quasi-equilibrium. The diode is made up of quasi-neutral regions and the depletion region. In the depletion region, the electron and hole distributions are essentially described by a Boltzmann distribution and that the concept of a quasi-Fermi level (see section 3.7 and section 4.5.2) is valid for electrons and holes. The quasi-Fermi levels for the electrons and holes extend from the quasi-neutral regions as shown in figure 4.5.
- The external potential drops mainly across the depletion region because the major barrier to current flow is the p-n junction dipole.

The key to the p-n diode operation is that a bias applied across the diode can shrink or increase the barrier seen by electrons and holes. This is shown schematically in figure 4.5. When a forward bias V_f is applied, the p-side is at a positive potential with respect to the n-side. In the reverse bias case, the p-side is at a negative potential $-V_r$ with respect to the n-side.

In the forward bias case, the potential difference between the n- and p-side is (applied bias $V = V_f$)

$$V_{Tot} = V_{bi} - V = V_{bi} - V_f \tag{4.3.1}$$

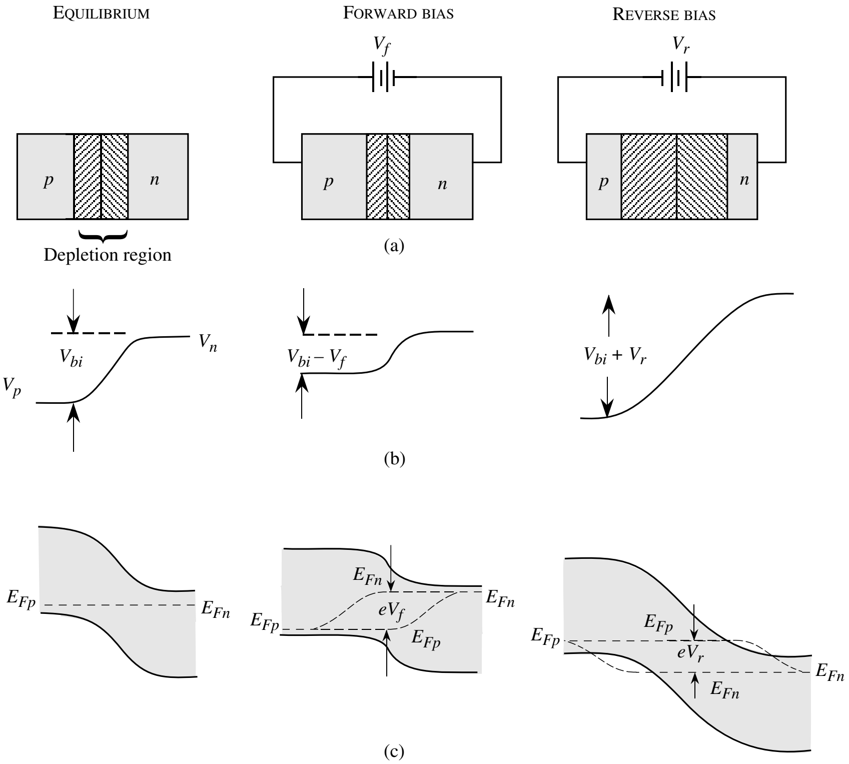


Figure 4.5: A schematic showing (a) the biasing of a p-n diode in the equilibrium, forward, and reverse bias cases; (b) the energy band profiles. In forward bias, the potential across the junction decreases, while in reverse bias it increases. The quasi-Fermi levels are shown in the depletion region.

while for the reverse bias case it is (the applied potential V is negative, $V = -V_r$, where V_r has a positive value)

$$V_{Tot} = V_{bi} - V = V_{bi} + V_r \tag{4.3.2}$$

Under the approximations given above, the equations for electric field profile, potential profile, and depletion widths we calculated in the previous section are directly applicable except that V_{bi} is replaced by V_{Tot}. Thus the depletion width and the peak electric field at the junction decrease under forward bias, while they increase under reverse bias, as can be seen from equation 4.2.24 and 4.2.25 if V_{bi} is replaced by V_{Tot}.

4.3.1 Drift and Diffusion Currents in the Biased Diode

The p-n diode characteristics are dominated by minority carrier flow i.e. electrons entering the p-side and holes entering from the p-side. When the diode is forward biased the barrier that electrons (holes) need to overcome to enter the p-side (n-side) from the n-side (p-side) decreases. This allows a higher minority charge injection. In reverse bias, the minority carrier injection is suppressed. This is shown schematically in figure 4.6. The presence of the bias increases or decreases the electric field in the depletion region. However, under moderate external bias, the electric field in the depletion region is always higher than the field for carrier velocity saturation ($E \gtrsim 10$kV cm^{-1}) for most semiconductors. Thus the change in electric field does not alter the drift part of the electron or hole current in the depletion region and the magnitude is determined by the rate of supply of minority carriers by diffusion from the bulk to the depletion region as will be described in section 4.5. When no bias is applied we have

$$\frac{p_{p0}}{p_{n0}} = \exp\left(eV_{bi}/k_BT\right) \tag{4.3.3}$$

In the presence of the applied bias, under the assumptions of quasi-equilibrium, we get

$$\frac{p(-W_p)}{p(W_n)} = \exp\left(e(V_{bi}-V)/k_BT\right) \cong \frac{p_{p0}}{p(W_n)} \tag{4.3.4}$$

We have assumed that the injection of mobile carriers is small (low-level injection) so that the majority carrier densities are essentially unchanged because of injection, i.e., $p(-W_p) = p_{p0}$. Taking the ratio of the two equations

$$\frac{p_n(W_n)}{p_{n0}} = \exp\left(eV/k_BT\right) \tag{4.3.5}$$

This equation suggests that the hole minority carrier density at the edge of the n-side depletion region can be increased or decreased dramatically by applying a bias.

A similar consideration gives, for the electrons injected as a function of applied bias,

$$\frac{n_p(-W_p)}{n_{p0}} = \exp\left(eV/k_BT\right) \tag{4.3.6}$$

From these equation we can see that the excess carriers created due to injection across the depletion regions are

$$\Delta p_n = p(W_n) - p_n = p_{n0}(\,\exp\left(eV/k_BT\right) - 1) \tag{4.3.7}$$

$$\Delta n_p = n_p(-W_p) - n_{po} = n_{po}(\,\exp\left(eV/k_BT\right) - 1) \tag{4.3.8}$$

The excess minority carriers that are introduced will decay into the majority region due to recombination with the majority carriers. For long diodes, the decay is simply given by the appropriate diffusion lengths (L_p for holes, L_n for electrons). Using results derived in section 3.9,

$$\begin{aligned} \delta p(x) &= \Delta p_n \,\exp\left((-(x-W_n)/L_p)\right) \\ &= p_{n0}\left[\,\exp\left(eV/k_BT\right) - 1\right]\cdot\exp\left[-(x-W_n)/L_p\right] \end{aligned} \tag{4.3.9}$$

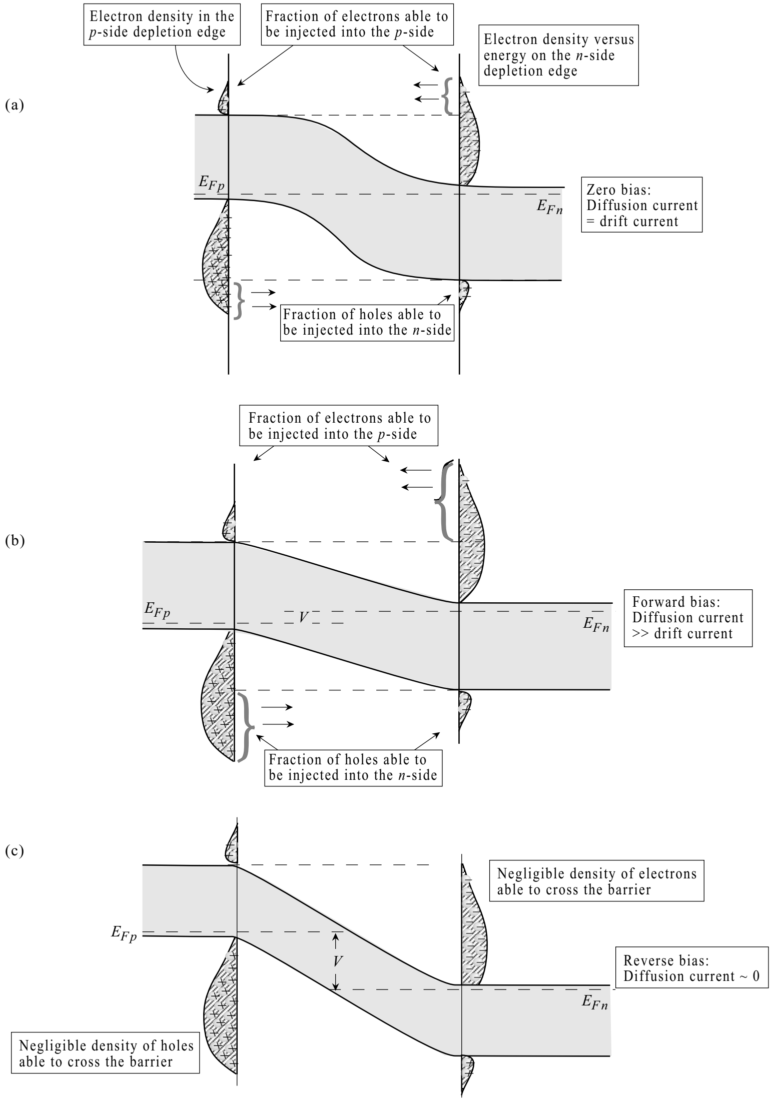

(a)
Electron density in the p-side depletion edge
Fraction of electrons able to be injected into the p-side
Electron density versus energy on the n-side depletion edge
E_{Fp}
E_{Fn}
Zero bias: Diffusion current = drift current
Fraction of holes able to be injected into the n-side
(b)
Fraction of electrons able to be injected into the p-side
E_{Fp}
V
E_{Fn}
Forward bias: Diffusion current >> drift current
Fraction of holes able to be injected into the n-side
(c)
Negligible density of electrons able to cross the barrier
E_{Fp}
V
Reverse bias: Diffusion current ~ 0
E_{Fn}
Negligible density of holes able to cross the barrier

Figure 4.6: A schematic of the minority and majority charge distribution in the n- and p-sides. The minority carrier injection (electrons from n-side to p-side or holes from p-side to n-side) is controlled by the applied bias as shown.

for $x > W_n$

$$\begin{aligned} \delta n_p(x) &= \Delta n_{po}\ \exp\left((x + W_p)/L_n\right) \\ &= n_{po}\left[\ \exp\left(eV/k_BT\right) - 1\right]\exp\left[(x + W_p)/L_n\right] \end{aligned} \tag{4.3.10}$$

$$\left\{ \begin{array}{l} x < -W_p \\ x \text{ is negative} \\ W_p \text{ is positive} \end{array} \right\}.$$

Holes are injected into the n-side and the value of this diffusion current is

$$I_p(x) = -eAD_p\frac{d(\delta p(x))}{dx} = eA\frac{D_p}{L_p}(\delta p(x)) \quad x > W_n \tag{4.3.11}$$

The hole current injected into the n-side is given by the hole current at $x = W_n$ (after using the value of $\delta p(x = W_n)$ from equation 4.3.9)

$$I_p(W_n) = e\,\frac{AD_p}{L_p}\,p_n\left(\ \exp\left(eV/k_BT\right) - 1\right) \tag{4.3.12}$$

Using similar arguments, the total electron current injected across the depletion region into the p-side region is given by

$$I_n(-W_p) = \frac{eAD_n}{L_n}\,n_{po}\left(\ \exp\left(eV/k_BT\right) - 1\right) \tag{4.3.13}$$

In this section we will assume that the diode is ideal which essentially means there is no e-h recombination within the depletion region. In the next section we will discuss the case where recombination occurs for a real diode. For the ideal diode the total current can be simply obtained by adding the hole current injected across W_n and electron current injected across $-W_p$, which is clear from figure 4.7c. The sum of the electron and hole currents in the depletion region, $I = I_p + I_n$ is given by $I_p(W_n) + I_n(-W_p)$ as the currents do not change in the depletion region due to the assumption of no generation - recombination. The diode current is therefore

$$\begin{aligned} I(V) &= I_p(W_n) + I_n(-W_p) \\ &= eA\left[\frac{D_p}{L_p}p_{n0} + \frac{D_n}{L_n}n_{po}\right]\left(\ \exp\left(eV/k_BT\right) - 1\right) \\ I(V) &= I_o\left(\ \exp\left(eV/k_BT\right) - 1\right) \end{aligned} \tag{4.3.14}$$

This is the diode equation. Under reverse bias, the current simply goes toward the value $-I_o$, where I_o is the prefactor of the diode equation.

$$I_o = eA\left(\frac{D_pp_{n0}}{L_p} + \frac{D_nn_{po}}{L_n}\right) \tag{4.3.15}$$

Notice that the diode current increases rapidly when a forward bias is applied and has a small value at negative bias. This gives the diode its rectification properties.

4.3.2 Minority and Majority Currents in the p-n Diode

The p-n diode is a bipolar device in which electrons and holes both carry current. To obtain the diode current we have simply added the electron current and hole current injection across the depletion region. This current was evaluated at its peak value at the edges of the depletion region. The diffusion current decreases rapidly in the majority region because of recombination. As the holes recombine with electrons in the n-region, an equal number of electrons are injected into the region. These electrons provide a drift current in the n-side to exactly balance the hole current that is lost through recombination. Let us consider the hole diffusion current in the n-type region. This current is, from equation 4.3.11, using the value of $\delta p(x)$ from equation 4.3.10,

$$I_p(x) = e\,A\,\frac{D_p}{L_p}\,p_{n0}\;\;\exp\left(-\frac{x-W_n}{L_p}\right)\left[\exp\left(\frac{eV}{k_BT}\right)-1\right]$$

$$x > W_n$$

We have also seen that the total current is

$$I = e\,A\left(\frac{D_p}{L_p}\,p_{n0} + \frac{D_n}{L_n}\,n_{po}\right)\left[\exp\left(\frac{eV}{k_BT}\right)-1\right] \tag{4.3.16}$$

Thus the electron current in the n-type region is

$$I_n(x) = I - I_p(x) \qquad\qquad (\text{for } x > W_n)$$

$$= eA\left\{\frac{D_p}{L_p}p_{n0}\left[1-\;\exp\left(-\frac{x-W_n}{L_p}\right)\right]\;+\;\frac{D_n}{L_n}n_{p0}\right\}\left[\exp\left(\frac{eV}{k_BT}\right)-1\right]$$

As the hole current decreases from W_n into the n-side, the electron current increases correspondingly to maintain a constant current. A similar situation exists on the p-side region. As the electron injection current decays, the hole current compensates. The electron and hole currents flowing in the diode have a behavior shown schematically in figure 4.7.

The rectifying properties of a diode are shown in figure 4.8. The reverse current saturates to a value I_o given by equation 4.3.15. Since this value is quite small, the diode is essentially nonconducting. On the other hand, when a positive bias is applied, the diode current increases exponentially and the diode becomes strongly conducting. The forward bias voltage at which the diode current density becomes significant ($\sim 10^3$ A·cm^{-2}) is called the cut-in voltage. This voltage is $\sim$ 0.8V for Si diodes and $\sim$ 1.2 V for GaAs diodes. The cut-in voltage is approximately 80 $\%$ of the material bandgap.

4.3.3 Narrow Diode Current

In the discussion for the diode current we have assumed the n and p-sides have lengths that are much greater than the minority carrier diffusion lengths. Often this is not the case. This is true for high speed diodes and for p-n junctions in bipolar transistors. In this such case we

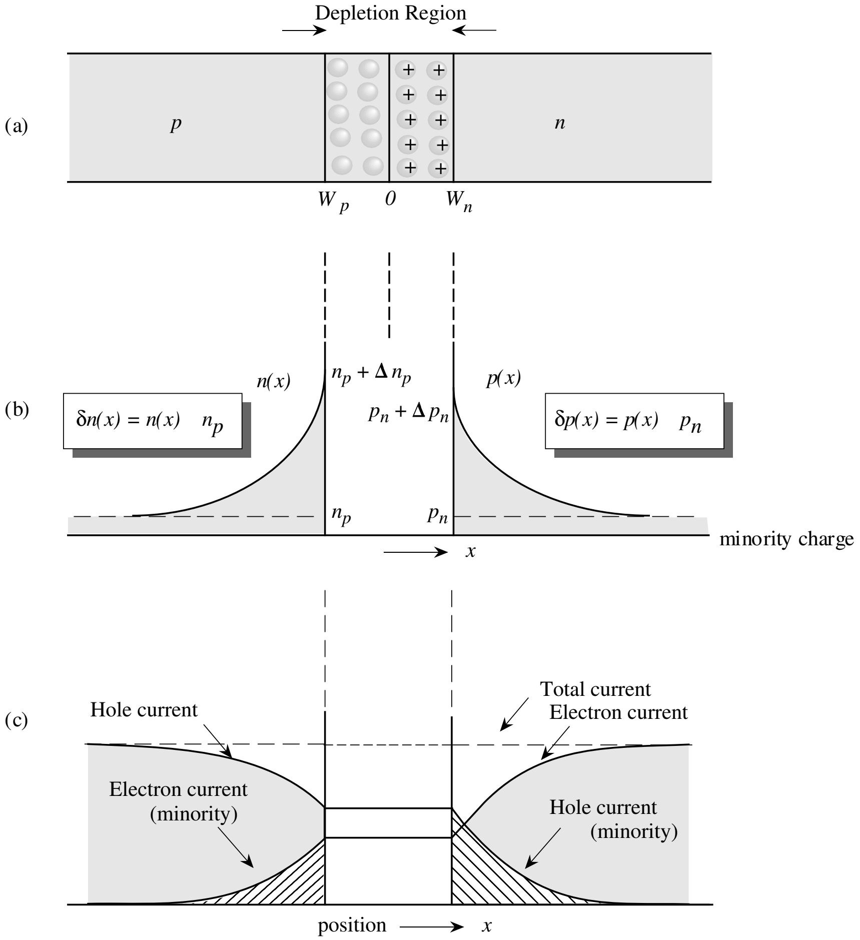


Figure 4.7: (a) A schematic showing the p-n structure under forward bias. (b) The minority carrier distribution (c) Minority and majority current.

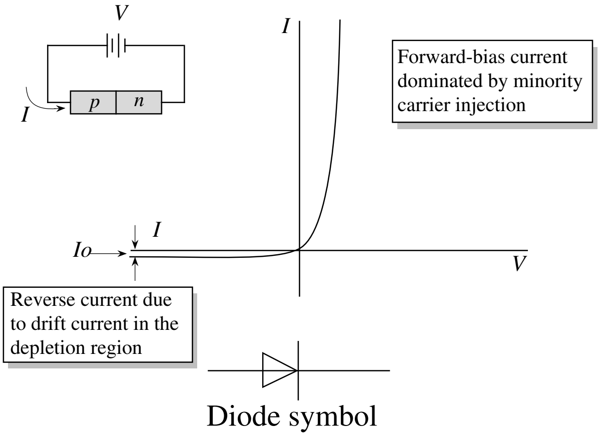


Figure 4.8: Rectifying I-V current of the p-n diode.

cannot assume that the injected excess minority carrier density will simply decay exponentially as $\exp\left\{-\left(x-W_n\right)/L_p\right\}$ (for holes). In fact, for the narrow diode one has to consider the ohmic boundary conditions where at the contacts the excess minority carrier density goes to zero.

In figure 4.9 we show a case where the diode extends a distance $W_{\rm ln}$ and $W_{\rm ln}$ as shown in the n- and p-sides. We know from section 3.9 if the diode is narrow the injected minority carrier concentration goes from its value at the depletion edge toward zero at the contact in a linear manner. The hole current injected across W_n becomes (note that $\delta p_n(W_{\ell n})=0$)

$$\begin{aligned} I_p(W_n) &= -eAD_p\frac{d(\delta p(x))}{dx} = -eAD_p\left[\frac{\delta p_n(W_n)-\delta p_n(W_{\ell n})}{W_{\ell n}-W_n}\right] \\ &= \frac{-eAD_p}{W_{\ell n}-W_n}p_n\left[\exp\left(\frac{eV}{k_BT}\right)-1\right] \end{aligned} \quad (4.3.17)$$

A similar expression results in this linear approximation for the electron distribution. The net effect is that the prefactor of the diode current changes (i.e., the term L_n or L_p in the denominator is replaced by a smaller term $(W_{\ell n}-W_n)$ or $(|W_{\ell p}-W_p|)$. The prefactor becomes

$$I_o = eA\left[\frac{D_p p_{n0}}{(W_{\ell n}-W_n)}+\frac{D_n n_{p0}}{(|W_{\ell p}-W_p|)}\right] \quad (4.3.18)$$

The narrow diode therefore has a higher saturation current than a long diode. The advantage of the narrow diode lies in its superior time-dependent response—a topic we will consider later.

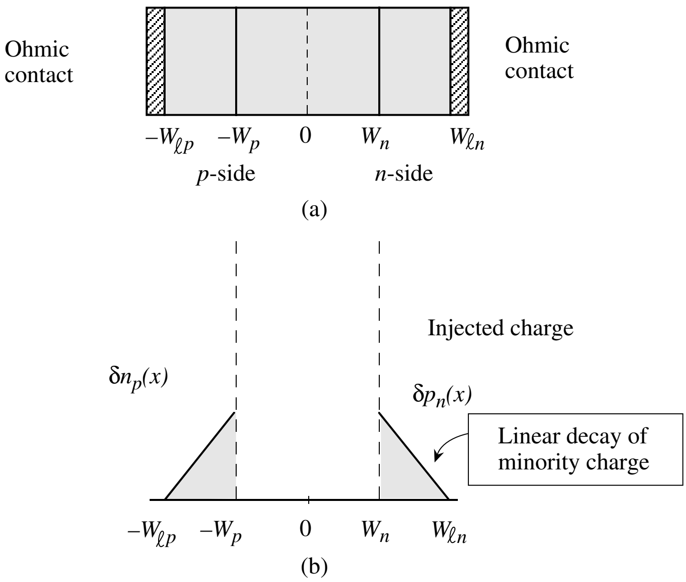


Figure 4.9: (a) A schematic of the narrow p-n diode with ohmic contacts at the boundaries. (b) The injected charge distribution.

4.4 REAL DIODES: CONSEQUENCES OF DEFECTS AND CARRIER GENERATION

In the discussion so far we have assumed the diode we are dealing with is ideal, i.e., there are no defects and associated bandgap states that may lead to trapping, recombination, or generation terms. In section 3.7 we discussed the effects of bandgap states produced by defects. In a real diode, a number of sources may lead to bandgap states. The states may arise if the material quality is not very pure so that there are chemical impurities present. Let us assume that the density of such deep level states is N_t. We will assume that the deep level is at the center of the bandgap.

We learned in chapter 3 in the SRH analysis that under the approximation of:

1. $\sigma = \sigma_n = \sigma_p$ and
2. $E_t = E_i$, and
3. $e_n, e_p, \sigma_n, \sigma_p$ are unperturbed in non-equilibrium

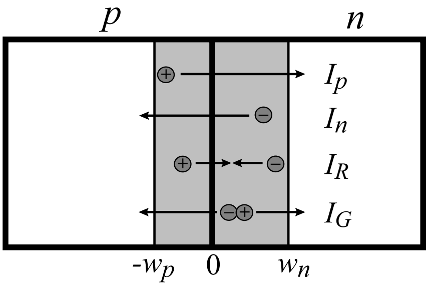


Figure 4.10: Qualitative diagram of all current components flowing in a p-n diode.

We get

$$U = \frac{1}{\tau} \cdot \frac{pn - n_i^2}{n + p + 2n_i} \tag{4.4.1}$$

where

$$\tau = \frac{1}{\sigma v_{th} N_t}$$

and $pn - n_i^2 =$ driving force for recombination and $n + p + 2n_i =$ resistance to recombination. This applies to any semiconductor with or without band bending. Note that the values of n and p are functions of band bending, photon flux, etc. Also, note that U is maximized for a certain level of perturbation when the denominator is minimized. As electrons and holes enter the depletion region, one possible way they can cross the region without overcoming the potential barrier is to recombine with each other. This leads to an additional flow of charged particles. This current, called the generation-recombination current, must be added to the current calculated so far. In figure 4.10 we show a qualitative diagram of all current components flowing in the diode.

4.4.1 Generation-Recombination Currents

To calculate the recombination currents in a diode, the Sah-Noyce-Shockley current, J_{SNS}, we consider a forward biased diode shown in figure 4.11b. Under a forward bias of V, the product of np is a constant across the depletion layer and is

$$np = n_i^2 \exp\left(\frac{eV}{k_B T}\right)$$

This is easily seen by recognizing that

$$n(x) = n_{n0} \exp\left(\frac{-e\psi'(x)}{k_B T}\right) \tag{4.4.2}$$

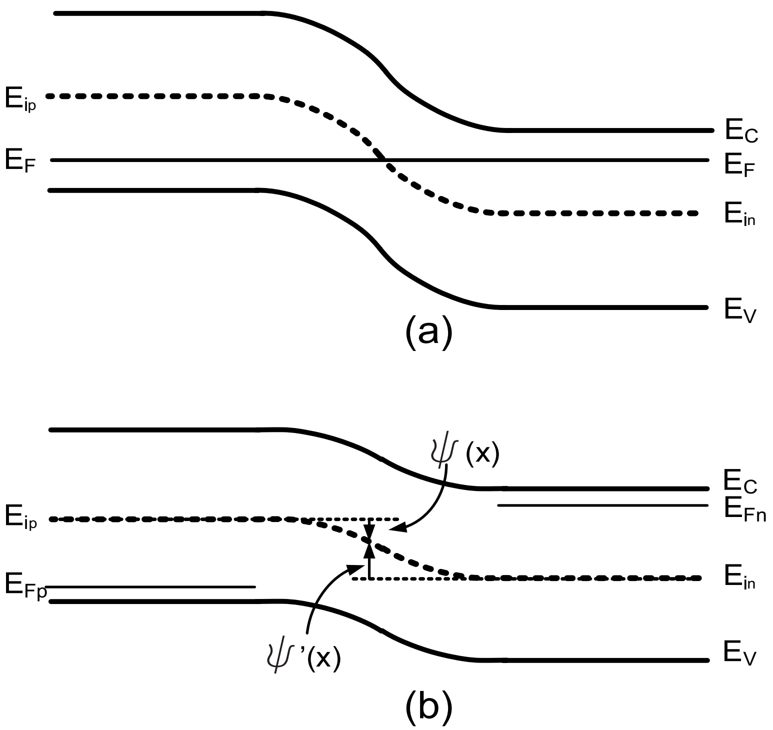


Figure 4.11: Band diagram of a p-n junction (a) in equilibrium and (b) under forward bias

$$p(x) = p_{p0}\ \exp\left(\frac{-e\psi(x)}{k_BT}\right) \tag{4.4.3}$$

Note that $\psi(x)$ is the voltage measured downwards from E_{ip} and $\psi'(x)$ is measured upwards from E_{in} and that

$$n(x)\cdot p(x) = n_{n0}\ p_{p0}\exp\left[\frac{-e(\psi'(x)+\psi(x))}{k_BT}\right]$$

$$\psi(x)+\psi'(x) = V_{bi}+V \tag{4.4.4}$$

Therefore,

$$n(x)\cdot p(x) = n_{n0}p_{n0}\ \exp\left(\frac{eV}{k_BT}\right)$$

or

$$n(x)\cdot p(x) = n_i^2\ \exp\left(\frac{eV}{k_BT}\right) \tag{4.4.5}$$

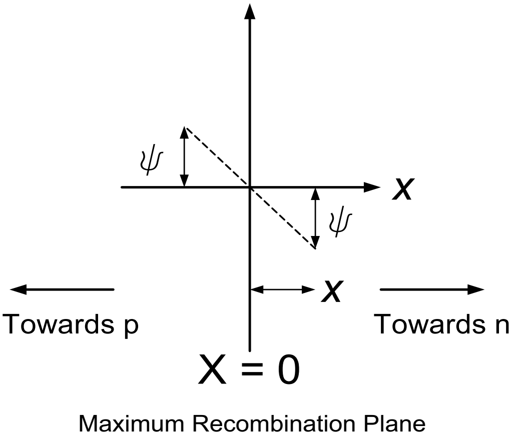


Figure 4.12: The maximum recombination plane is where the recombination is peaked within the depletion region

Under steady state bias of V the term $n + p + 2n_i$ is minimized when $n = p$. The value of x where this occurs is chosen to be zero. This is the maximum recombination plane (MRP).

$$n(0) = p(0) = n_i \exp\left(\frac{eV}{2k_BT}\right) \tag{4.4.6}$$

If we move away from the MRP, the electron and hole concentrations change proportionally to the term

$$\exp\left(\pm\frac{e\psi(x)}{k_BT}\right)$$

as shown in figure 4.12, Where

$$n(x) = n(0) \exp\left(\frac{e\psi(x)}{k_BT}\right) \quad \text{and} \quad p(x) = p(0) \exp\left(-\frac{e\psi(x)}{k_BT}\right)$$

Assuming the distance x is small so that we can assume the electric field is constant for purposes of the analysis to be $\mathcal{E} = \mathcal{E}(0)$. Then $\psi = \pm\mathcal{E}x$ and therefore,

$$U = \frac{1}{\tau}\cdot\frac{pn - n_i^2}{n + p + 2n_i} = \frac{1}{\tau}\cdot\frac{n_i^2 \exp\left(eV/k_BT\right) - n_i^2}{n(0) \exp\left(e\psi/k_BT\right) + p(0) \exp\left(-e\psi/k_BT\right) + 2n_i}$$

Neglecting n_i^2 in the numerator and $2n_i$ in the denominator we get

$$U = \frac{1}{\tau}\cdot\frac{n_i \exp\left(eV/2k_BT\right)}{2\cosh(e\mathcal{E}x/k_BT)} \tag{4.4.7}$$

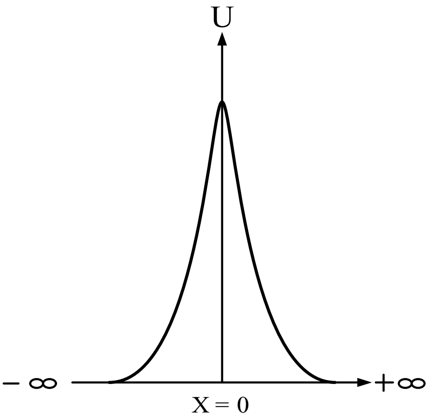


Figure 4.13: Schematic of the net recombination rate as a function of distance from the MRP

To calculate the total recombination current we need to integrate over the volume of the depletion region. Since the recombination rate curve is highly peaked about $x = 0$, the MRP, as shown in figure 4.13, the following approximations remain valid.

1. Linearizing the potential $\psi = \pm \mathcal{E} x$ since only small values of x contribute to the integral.
2. $\int_{-W_p}^{W_n} \to \int_{\infty}$

Therefore,

$$I_{SNS} = I_R = eA \int U(x) dx \tag{4.4.8}$$

Making these substitutions and solving equation 4.4.8, we find:

$$\begin{aligned} I_R =& \frac{en_i A}{2\tau} \exp\left(\frac{eV}{2k_B T}\right) \int_{-\infty}^{+\infty} \frac{dx}{\cosh\left[e\mathcal{E}(0)x/k_B T\right]} && (4.4.9)\\ =& \frac{en_i A}{2\tau} \cdot \frac{\pi k_B T}{e\mathcal{E}(0)} \exp\left(\frac{eV}{2k_B T}\right) && (4.4.10)\\ =& I_{GR}^{\circ} \exp\left(\frac{eV}{2k_B T}\right) && (4.4.11) \end{aligned}$$

At zero applied bias, a generation current of I_G balances out the recombination current.

The generation-recombination current therefore has an exponential dependence on the voltage as well, but the exponent is different. The generation-recombination current is

$$\begin{aligned} I_{GR} &= I_R - I_G = I_R - I_R(V=0) \\ &= I_{GR}^{\circ}\left[\exp\left(\frac{eV}{2k_BT}\right)-1\right] \end{aligned} \tag{4.4.12}$$

where V is small so that the MRP is assumed constant.

The total device current now becomes

$$I = I_o\left[\exp\left(\frac{eV}{k_BT}\right)-1\right] + I_{GR}^{\circ}\left[\exp\left(\frac{eV}{2k_BT}\right)-1\right]$$

or

$$\begin{aligned} I &\cong I_S\left[\exp\left(\frac{eV}{nk_BT}\right)-1\right] \\ &= I_S\left[\exp\left(\frac{V}{n}\cdot\frac{e}{k_BT}\right)-1\right] \end{aligned} \tag{4.4.13}$$

where n is called the diode ideality factor or the voltage partitioning factor because the factor of 2 in equation 4.4.12 is a consequence of recombination occurring at the maximum recombination plane. The prefactor I_{GR}^{o} can be much larger than I_o for real devices. Thus at low applied voltages the diode current is often dominated by the second term. However, as the applied bias increases, the diffusion current starts to dominate. We thus have two regions in the forward I-V characteristics of the diode, as shown in figure 4.14. One of the reasons it is experimentally difficult to measure an IV curve with an ideality factor of 2 is because the MRP is actually changing with applied bias.

In figure 4.15, we show the effects that material defects can have on the diode current characteristics. We can see that defects such as threading dislocations can cause large undesirable reverse leakage currents that are not predicted by the ideal diode characteristics calculated in this section.

Example 4.2 Consider the p-n diodes examined in problem 4.12. In that example, the diode prefactor was calculated assuming that there is no recombination in the depletion region. Calculate the effect of the generation-recombination current assuming a lifetime of 10^{-6} s.

The prefactor of the generation-recombination current is

$$I_{GR}^{o} = \frac{eAn_i}{2\tau}\frac{\pi k_BT}{e\mathcal{E}(0)}$$

At zero applied bias, we know that the MRP occurs where $p_{p0}e^{-\frac{\psi(x)}{k_BT}} = n_i$. This allows

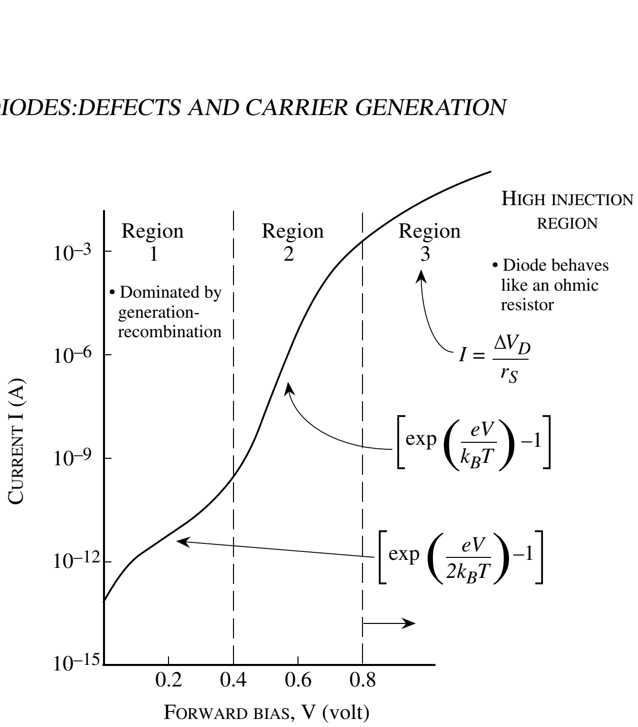


Figure 4.14: A log plot of the diode current in forward bias. At low biases, the recombination effects are quite pronounced, while at higher biases the slope becomes closer to unity. At still higher biases the behavior becomes more ohmic.

us to solve for $\psi(x)$ and thus $\mathcal{E}(0)$. This gives

$$\begin{aligned} I_{GR}^{\circ} &= \frac{\left(1.6 \times 10^{-19}\text{ C}\right)\left(10^{-3}\ \text{cm}^2\right)\left(1.5 \times 10^{10}\text{ cm}^{-3}\right)}{2\left(10^{-6}\text{ s}\right)} \cdot \frac{(3.14)\,(.026V)}{(3.2 \times 10^4\ V/cm)} \\ &= 3.0 \times 10^{-12}\text{ A} \end{aligned}$$

and

$$I_{GR} = I_{GR}^{o}\left[\exp\left(\frac{\text{eV}}{2k_BT}\right) - 1\right]$$

We can see that the generation-recombination prefactor is much larger than the prefactor due to the diffusion term. Thus the reverse current will be dominated by the generation-recombination effects.

In forward bias, the diffusion current is initially much smaller than the generation-recombination term. However, at higher forward bias the diffusion current will start to dominate. For example, we see that at a forward bias of 0.2 V, the diffusion current is 2.2×10^{-11} A, while the generation-recombination current is 1.65×10^{-10} A. At a

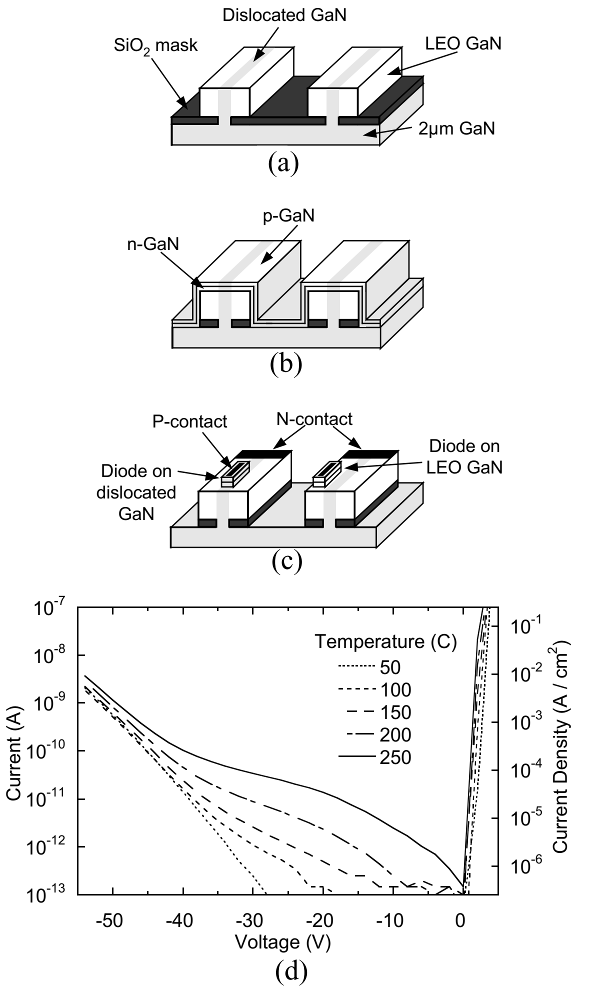


Figure 4.15: Effects of threading dislocations on reverse leakage current in p-n diodes. GaN p-n diodes were fabricated on the same wafer, some of them being placed on areas with high dislocation density, and some placed in areas with virtually no dislocations. The fabrication process is shown in (a) - (c). Current characteristics for a number of diodes are shown in (d). We see that the reverse leakage current in the devices on dislocated material is 3-4 orders of magnitude higher than that of devices on non-dislocated material, indicating that the dislocations provide a leakage path for current to travel. Figures taken from the PhD dissertation of Peter Kozodoy, UCSB.

forward bias of 0.6 V, the diffusion current is 1.07×10^{-4} A, while the generation-recombination current is 8.45×10^{-7} A.

Example 4.3 Consider a long p-n diode on silicon with the following parameters:

$$\begin{aligned}
n\text{-side doping} &= 10^{17}\ \text{cm}^{-3} \\
p\text{-side doping} &= 10^{17}\ \text{cm}^{-3} \\
\text{Minority carrier lifetime } \tau_n &= \tau_p = 10^{-7}\ \text{s} \\
\text{Electron diffusion constant} &= 30\ \text{cm}^2/\text{s} \\
\text{Hole diffusion constant} &= 10\ \text{cm}^2/\text{s} \\
\text{Diode area} &= 10^{-4}\ \text{cm}^2 \\
\text{Carrier lifetime in the depletion region} &= 10^{-8}\ \text{s}
\end{aligned}$$

Calculate the diode current at a forward bias of 0.5 V and 0.6 V at 300 K. What is the ideality factor of the diode in this range?

For this diode structure we have the following:

$$\begin{aligned}
n_p &= 2.25 \times 10^3\ \text{cm}^{-3} \\
p_n &= 2.25 \times 10^3\ \text{cm}^{-3} \\
L_n &= 17.32\ \mu\text{m} \\
L_p &= 10.0\ \mu\text{m} \\
V_{bi} &= 0.817\text{V}
\end{aligned}$$

The prefactor in the ideal diode equation is

$$\begin{aligned}
I_0 &= eA\left(\frac{D_p p_n}{L_p} + \frac{D_n n_p}{L_n}\right) \\
&= 9.83 \times 10^{-16}
\end{aligned}$$

The prefactor to the recombination-generation current is

$$I_{GR}^0 = \frac{eAn_i}{2\tau} \cdot \frac{\pi k_B T}{e\mathcal{E}(0)}$$

where τ is the lifetime in the depletion region.

The $\mathcal{E}(0)$ at a forward bias of 0.5 V is found to be

$$\mathcal{E}(0) = 6.94 \times 10^4\ \text{V/cm}$$

The $\mathcal{E}(0)$ at a forward bias of 0.6 V is found to be

$$\mathcal{E}(0) = 5.74 \times 10^4\ \text{V/cm}$$

The prefactors to the recombination-generation current is

$$I^0_{GR}(0.5\text{ V}) = 1.4 \times 10^{-11}\text{ A}$$
$$I^0_{GR}(0.6\text{ V}) = 1.7 \times 10^{-11}\text{ A}$$

The current is now

$$\begin{aligned} I(0.5\text{ V}) &= 9.83 \times 10^{-16}\ \exp\left(\frac{0.5}{0.026}\right) + 1.4 \times 10^{-11}\ \exp\left(\frac{0.5}{0.052}\right) \\ &= 4.33 \times 10^{-7}\text{ A} \end{aligned}$$

and

$$\begin{aligned} I(0.6\text{ V}) &= 9.83 \times 10^{-16}\ \exp\left(\frac{0.6}{0.026}\right) + 1.7 \times 10^{-11}\ \exp\left(\frac{0.6}{0.052}\right) \\ &= 1.21 \times 10^{-5}\text{ A} \end{aligned}$$

We can write the diode current as

$$I \cong I_S\ \exp\left(\frac{\text{eV}}{nk_BT}\right)$$

Thus

$$\frac{I(V_2)}{I(V_1)} \cong \exp\left(\frac{e(V_2 - V_1)}{nk_BT}\right)$$

Using this relation, we find that

$$n \cong 1.15$$

4.5 Reverse Bias Characteristics

The case for reverse bias is very different. Here the application of bias increases barriers. The only carriers that can flow are those that can diffuse to the depletion region and are swept across by the field; these are minority carriers, holes in the n-region and electrons in the p-region (figure 4.16).

4.5.1 First Observation

Since we are only dealing with minority carrier currents we know that minority carrier drift can be neglected, hence only minority carrier diffusion is relevant. To calculate diffusion currents we need to know the charge profile. Charge profiles are obtained by solving the continuity equation as shown in chapter 3. We assume that the large electric field in the reverse biased p-n junction sweeps minority carriers away from the edge of the junction. Using the Schockley Boundary Conditions:

$$n_p\left(-W_p\right) = 0 \quad \text{And} \quad p_n\left(+W_n\right) = 0$$

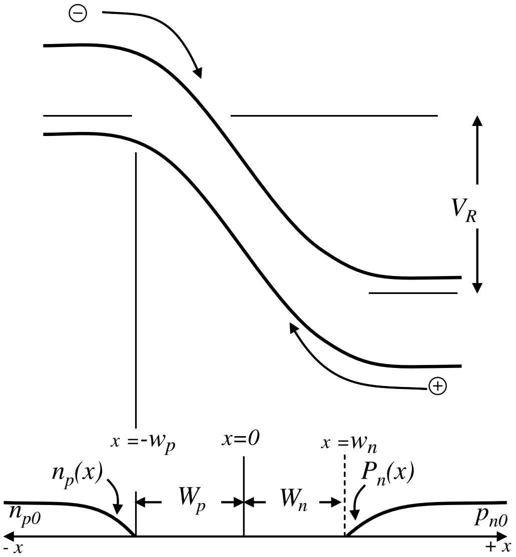


Figure 4.16: Here the minority carriers are electrons injected from the p-region to the n-region (opposite to the forward-biased case)

We also know that the minority carrier concentration in the bulk is n_{p0} ($p-$type) and p_{n0} ($n-$type). Therefore, the shape of the curve will be qualitatively as shown (figure 4.16) reducing from the bulk value to zero at the depletion region edge. We now consider the flow of minority holes. The charge distribution is obtained by solving:

$$D_p \frac{d^2 p_n}{dx^2} + G_{\text{th}} - R = 0 \tag{4.5.1}$$

Where G_{th} is the generation due to thermal emission of carriers, and R is the recombination rate for excess carriers. The process of carrier recombination is driven by excess carriers. This dependence may be written as: (where α_r is a material-dependent rate constant).

$$R = np\alpha_r = \frac{p_n}{\tau_p} \tag{4.5.2}$$

where in a p-type semiconductor $\tau_p = \frac{1}{\alpha_r n}$. Clearly, in an intrinsic semiconductor without excess minority carriers the expressions for R and G_{th} become equivalent - expressing the equilibrium of the system:

$$G_{th} - R = 0$$

or

$$G_{th} = \alpha_r n_i^2$$

In non equilibrium recognizing that G_{th} is a function of temperature and invariant under injection

$$G_{th} - R = \alpha_r n_i^2 - \alpha_r n_p = \frac{p_{n0} - p_n}{\tau_p} \tag{4.5.3}$$

Therefore, from equation 4.5.1

$$D_p \frac{d^2 p_n}{dx^2} + \frac{p_{n0} - p_n}{\tau_p} = 0 \tag{4.5.4}$$

Note that the second term in this sum is a generation term because $p_n < p_{n0}$ for all $x > W_n$. This is natural because both generation and recombination are mechanisms by which the system returns to its equilibrium value. When the minority carrier concentration is above the equilibrium minority carrier value then recombination dominates and when the minority carrier concentration is less than that at equilibrium, then generation dominates. Again, using $\Delta p_n (x - W_n) = p_{n0} - p_n (x - W_n)$, we have

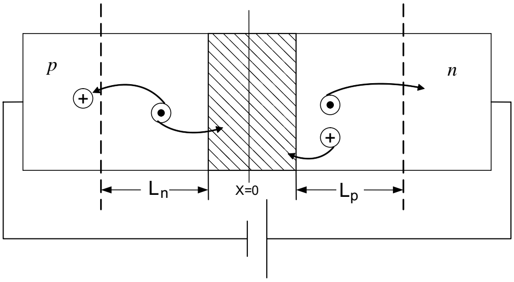


Figure 4.17: Minority carriers generated within a diffusion length of the depletion region enter and are swept away.

$$D_p \frac{d^2 \Delta p_n (x)}{dx^2} + \frac{\Delta p_n (x)}{\tau_p} = 0 \tag{4.5.5}$$

and,

$$\Delta p_n (x - W_n) = C_1 \exp\left(+\frac{x - W_n}{L_p}\right) + C_2 \exp\left(-\frac{x - W_n}{L_p}\right) \tag{4.5.6}$$

Where we know that:

$$\begin{cases} C_1 = 0 & \text{for physical reasons,} \\ \Delta p_n (x - W_n) \to 0 & x - W_n = \infty, \\ \Delta p_n (x - W_n) \to p_{n0} - p_n(0) = p_{n0} & x - W_n = 0. \end{cases}$$

$$\therefore \Delta p_n\left(x-W_n\right)=p_{n0}\ \exp\left(-\frac{x-W_n}{L_p}\right)$$

which can be rewritten for $x \geq W_n$

$$\boxed{p_n\left(x\right)=p_{n0}\left[1-\ \exp\left(-\frac{x-W_n}{L_p}\right)\right]} \tag{4.5.7}$$

which implies that the flux of holes entering the depletion region is

$$J_p(W_n)=eD_p\frac{dp_n(W_n)}{dx}=e\frac{D_p}{L_p}p_{n0}$$

and similarly, $J_n=e\frac{D_n}{L_n}\cdot n_{p0}$. Assuming no generation in the depletion region, the net current flowing is:

$$J_s=e\left(D_p\frac{p_{n0}}{L_p}+D_n\frac{n_{p0}}{L_n}\right) \tag{4.5.8}$$

This result is remarkable because we get the same answer if we took the forward bias equation (valid only in forward bias) and arbitrarily allowed V to be large and negative (for reverse bias)- i.e.

$$J=J_s\left[\,\exp\left(qV/kT\right)-1\right] \tag{4.5.9}$$

if V is large and negative, $J_R=-J_S$ which is the answer we derived in equation 4.5.8. This can be understood as follows. As shown in figure 4.17 any minority carrier electrons generated within a diffusion length of the n depletion edge can diffuse to the edge of the junction and be swept away. Minority electrons generated well beyond a length L_n will recombine with holes resulting in the equilibrium concentration, n_{p0}. Similarly holes generated within L_p, a diffusion length, of the depletion region edge could diffuse into the depletion region. It is important to note (from the first term of equation 4.5.8) that the slope of the minority carrier profile at the depletion region edge :

$$slope=\frac{p_{n0}}{L_p}=\frac{\text{difference from bulk value}}{L_p} \tag{4.5.10}$$

This is always true when recombination and generation dominate. Recall that even in forward bias (shown in figure 4.18) the slope of the carrier profile is again

$$\frac{\text{difference from bulk value}}{L_p}=\frac{\Delta p_n(W_n)}{L_p} \tag{4.5.11}$$

4.5.2 Quasi Fermi Levels

The Quasi Fermi Level is a very useful concept as it accurately represents the occupancy of states of the system that it refers to. It is important to recognize that semiconductor devices are composed of several interacting systems. For example, the conduction band containing free electrons, the valence band containing free holes and trap states in the gap have an occupancy,

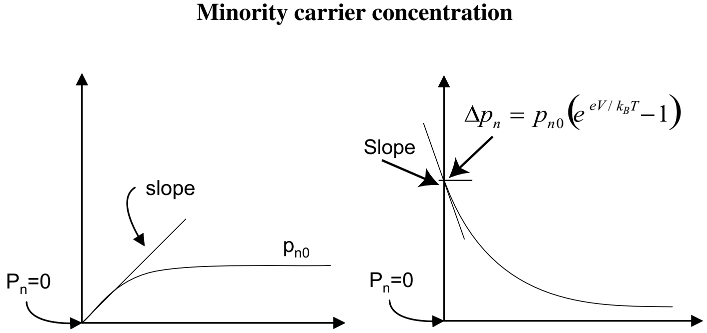


Figure 4.18: left: Reverse bias minority carrier concentration. right: Forward bias minority carrier concentration

with each set of traps constituting a separate system. Each of these systems will be represented by their own quasi Fermi level, $\mathrm{E_{Fn}}$ for the electrons, $\mathrm{E_{Fp}}$ for the holes and $\mathrm{E_{FTi}}$ for the traps of energy $\mathrm{E_{Ti}}$. If thermal energy is the only energy source determining the occupancy of the different states and the systems are all freely interacting then of course all the quasi Fermi levels merge into the Fermi level of the system at equilibrium. We have also seen in figure 4.5 that the quasi Fermi levels vary across a device in non-equilibrium. The variation of the Fermi level is determined by the current flow in the system and the interaction of the various systems. Let us look at the variation of $\mathrm{E_{Fn}}$ and $\mathrm{E_{Fp}}$ in the case of a forward biased diode in figure 4.5c. The electron quasi Fermi level is determined by the electron concentration point by point and is determined by the level set by the reservoir of electrons which is the $n-$type layer. The large electron concentration in the $n-$type layer ensures that only a small gradient in $\mathrm{E_{Fn}}$ can sustain current of relevant magnitudes and hence is pictured flat in the bulk. The same applies for $\mathrm{E_{Fp}}$ on the $p-$side. As electrons and holes flow across the depletion region the quasi Fermi levels remain almost flat because the length of the depletion region is very small allowing large gradients to be present to satisfy needed current flow with low absolute values of $\Delta\mathrm{E_{Fn}}$ and $\Delta\mathrm{E_{Fp}}$. In the depletion region, since there is no recombination assumed, the electron concentration is determined by the $n-$region and the hole concentration by the $p-$region, and in both instances substantially by the thermal supply as given by the Boltzmann distribution as shown in figure 4.6. However, in the bulk regions the electrons and holes recombine and hence the $\mathrm{E_{Fn}}$ in the $p-$region decreases in a manner determined by the recombination rate and not the Boltzmann tail of the majority electrons. In the case of reverse bias (figure 4.5c) the injected electrons and holes are sourced as thermally generated minority carriers in the bulk regions and swept across the junction to constitute the reverse saturation current. The quasi-femi levels reflect the change in the minority carrier concentration in the bulk regions on the application of the reverse bias. The minority carrier electron concentration at the edge of the depletion region of the p-side under zero bias is $n_{n0}\ \exp\left(-eV_{bi}/k_BT\right)$ or n_{p0}. On application of the reverse bias of V_r, the electron concentration at the junction edge supplied by thermionic emission from the n-side is given by $n_{no}\ \exp\left(-e(V_{bi} - V_r)/k_B T\right)$ or $n_{p0}\ \exp\left(eV_r/k_B T\right)$, but since V_r is negative this is very small. The linear decrease in the value of $\mathrm{E_{Fn}}$ from the $p-$bulk to the edge of the junction reflects the exponential decrease in the carrier concentration because of diffusion toward the junction as shown in equation 4.5.7. The reader should be aware that this picture reflects the case of no velocity saturation of either electrons or holes in the materials. If that happens, the electron concentration at the edge of the junction cannot decrease arbitrarily and the problem has to be solved such that the carriers allow current to be continuous through the structure and is left to the reader as an exercise.

4.6 HIGH-VOLTAGE EFFECTS IN DIODES

In deriving the current-voltage relation we have made two important assumptions: i) the excess carrier density injected across the depletion region is small compared to the majority charge density; ii) the reverse current saturates since it is due to the carriers drifting across the depletion region and is limited by the diffusive flux of minority carriers to the junction.

4.6.1 Forward Bias: High Injection Region

We have so far assumed that the injection density of minority carriers was low so that the voltage all dropped across the depletion region. However, as the forward bias is increased, the injection level increases and eventually the injected minority carrier density becomes comparable to the majority carrier density. When this happens, an increasingly larger fraction of the external bias drops across the undepleted region. The diode current will then saturate, as shown as Region 3 in figure 4.14. The minority carriers transport is not only due to diffusion, but also due to the electric field that is now present in the undepleted region. As the forward bias increases, the devices start to behave like a resistor, where the current-voltage relation is given by a simple linear expression. The current is now controlled by the resistance of the n- and p-type regions as well as the contact resistance.

4.6.2 Reverse Bias: Impact Ionization

We have noticed that under reverse bias conditions the electric field across the depletion region increases. As a result electrons and holes forming the reverse current can acquire very high energies. Once this excess energy reaches the value of the bandgap we can have impact ionization as discussed in chapter 3. The final result is that one initial electron can create two electrons in the conduction band and one hole in the valence band. This results in current multiplication and the initial current reverse bias I_o becomes

$$I_o^{'} = M(V) I_o \tag{4.6.1}$$

here M is a factor that depends upon the impact ionization rate which we now derive.

4.7 Avalanche Breakdown in a *p-n* junction

Consider a $p-i-n$ junction where the applied voltage is such that the electric field on the intrinsic region which is a constant is assumed to be large enough to saturate the electron and hole velocities. We assume in our analysis that $v_{se} = v_{sh} = v_s$. As is shown schematically in figure 4.19, a few lucky electrons (minority carriers) injected from the p-side into the high field region can gain enough kinetic energy ($> E_g$) to collide with the lattice creating electron-hole pairs. This process is called impact ionization. These electrons and holes accelerate again leading to more collisions and further generation. The same applies to holes injected from the $n-$side. To analyze the resultant current due to impact ionization one solves the continuity equation for electrons and holes:

$$\frac{\partial n}{\partial t} = \frac{1}{e}\frac{\partial J_n}{\partial x} + G_n(e) + G_h(e) \tag{4.7.1}$$

where $G_n(e) =$ the rate of generation of secondary electrons by accelerated electrons $= \alpha_n n(x) v_s$, where $\alpha(cm^{-1})$, the ionization coefficient of electrons, is the number of electron-hole pairs generated per electron per cm, and $n(x)(cm^{-3})$ is the local concentration of electrons, and v_s (cm/s) is the saturated electron velocity. The ionization coefficient is much less than 1 because only lucky electrons create electron-hole pairs and the above equation assumes that all electrons are participating in the process. $G_h(e)$ is the rate of generation of secondary electrons by accelerated holes:

$$G_h(e) = \alpha_p \cdot p(x) \cdot v_s \tag{4.7.2}$$

where α_p is the ionization coefficient for holes. $p(x)\left(cm^{-3}\right)$ = local concentration of holes. v_s is the saturated hole velocity. Assuming $\alpha_n = \alpha_p = \alpha$ and that $\alpha \neq f(\mathcal{E})$, the latter being a good approximation in a $p-i-n$ structure,

$$\frac{\partial n}{\partial t} = \frac{1}{e}\frac{\partial n}{\partial x} + \alpha\left(n(x)v_s + p(x)v_s\right) = 0 \tag{4.7.3}$$

in steady state which we now consider. The impact ionization process causes the electron current to increase from its reverse saturation value,

$$J_{n0} = -e\frac{D_n}{L_n}n_{p0} \tag{4.7.4}$$

to a larger value at the p-side. The same is true for holes as shown in figure 4.20. To solve the steady state continuity equation we also note that

$$J_n(x) = -en(x)v_s \tag{4.7.5}$$

and

$$J_p(x) = ep(x)\left(-v_s\right) = -ep(x)v_s \tag{4.7.6}$$

which gives

$$\frac{\partial J_n}{\partial x} - \alpha\left(env_s + epv_s\right) = 0 \tag{4.7.7}$$

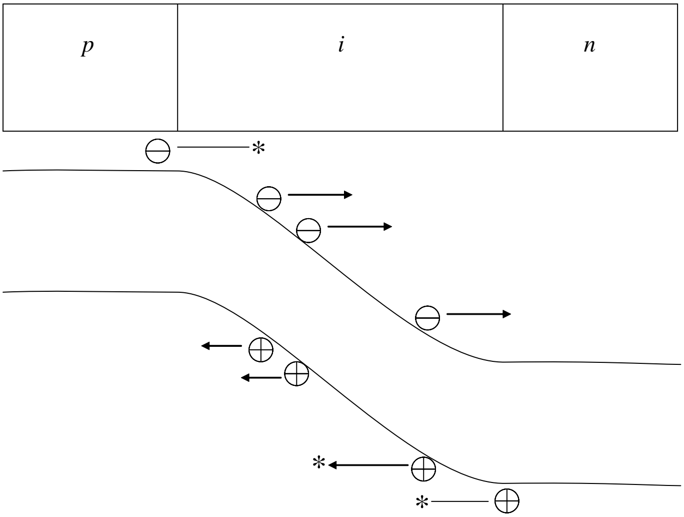


Figure 4.19: Band diagram for a $p - i - n$ diode showing the avalanche ionization and multiplication process where an injected minority carrier causes generation of electron-hole pairs through impact ionization. Multiple ionization events may be caused by a single carrier. The star represents a collision which generates an electron-hole pair.

or

$$\frac{\partial J_n}{\partial x} = \alpha \left(J_p + J_n\right) = \alpha J_{Total} \tag{4.7.8}$$

Integrating over the length of the multiplication region, w_a, we get

$$\int dJ_n = J_{Total} \int_0^{w_a} \alpha \, dx \tag{4.7.9}$$

$$J_n\left(w_a\right) - J_n(0) = J_{Total} \int_0^{w_a} \alpha \, dx \tag{4.7.10}$$

Recognizing

$$J_n\left(w_a\right) = J_{Total} - J_{p0} \tag{4.7.11}$$

and $J(0) = J_{n0}$, one gets

$$J_{Total} - J_{P0} - J_{n0} = J_{Total} \int_0^{w_a} \alpha \, dx \tag{4.7.12}$$

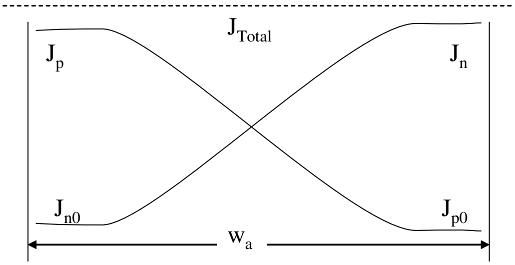


Figure 4.20: Majority and minority currents in a reverse-biased $p - i - n$ junction

Defining a multiplication factor, M,

$$M = \frac{J_{Total}}{J_{p0} + J_{n0}} = \frac{J_{Total}}{J_s} \tag{4.7.13}$$

the equation reduces to:

$$1 - \frac{1}{M} = \int_0^{w_a} \alpha \, dx \tag{4.7.14}$$

or

$$M = \left[1 - \int_0^{w_a} \alpha \, dx\right]^{-1} \tag{4.7.15}$$

Breakdown is defined as the case where $J_{Total} \to \infty$ or $M \to \infty$. This condition is achieved when

$$1 - \int_0^{w_a} \alpha \, dx \to 0 \tag{4.7.16}$$

or

$$\int_0^{w_a} \alpha \, dx \to 1 \tag{4.7.17}$$

We recognize that α is a $f(\mathcal{E})$ in general. In the case of a constant α, the breakdown condition reduces to

$$\alpha \cdot w_a \to 1$$

which represents the case of every electron (hole) injected into the high field region generating an electron-hole pair before exiting. This process is self-sustaining.

4.7.1 Reverse Bias: Zener Breakdown

Impact ionization or avalanche breakdown is one mechanism for breakdown in diodes. There is another one that can be important for narrow gap diodes or heavily doped diodes. This mechanism is due to the quantum-mechanical process of tunneling. The tunneling process, allows

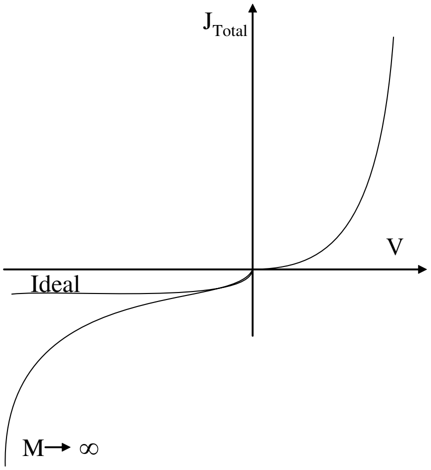


Figure 4.21: Current-Voltage characteristics for a $p - i - n$ diode in avalanche breakdown showing the ideal (non-avalanche) case as well as the limit where M becomes large.

electrons in the valence band to tunnel into the conduction band and vice versa. Electrons tunneling through the diode do not have to go over the barrier and as a result the diode reverse current can increase dramatically.

To examine how tunneling occurs let us examine the band profile in a reverse-biased p-n junction. Assume that the diode is heavily doped so that the Fermi level on the n-side and the Fermi level on the p-side are in the conduction and valence bands, respectively. The heavy doping ensures that electrons in the conduction band can tunnel into "available" empty states in the valence band. A typical electron sees a potential barrier between points x_2 and x_1, as shown in figure 4.22b. The tunneling probability is given under such conditions by

$$T \approx \exp\left(-\frac{4\sqrt{2m^*}E_g^{3/2}}{3e\hbar\mathcal{E}}\right) \tag{4.7.18}$$

where E_g is the bandgap of the semiconductor, m^* is the reduced mass of the electron-hole system, and $\mathcal{E}$ is the field.

There is a special class of diodes called Zener diodes where tunneling is exploited. The depletion width can be controlled by the doping density. If the junction is made from heavily doped materials, the Zener tunneling can start at a reverse bias of V_z, which could be as low as a few tenths of a volt. The voltage across the junction is then clamped at V_z, and the current is controlled by the external circuit as shown in figure 4.23. This clamping property provides a very useful application for the Zener diodes. If V_z is breakdown voltage (due to impact ionization or

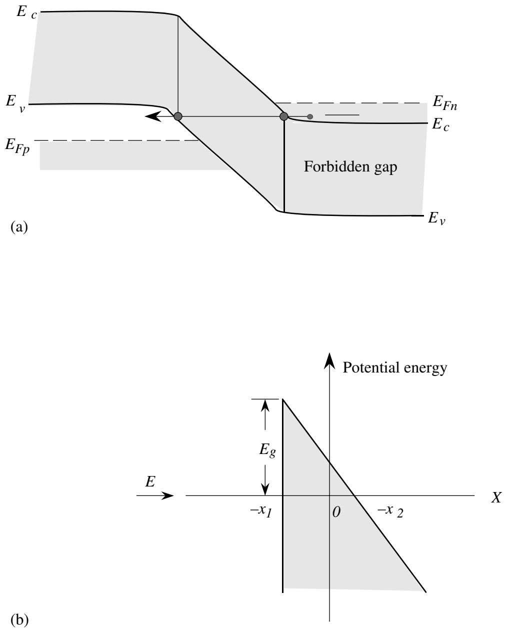


Figure 4.22: (a) A schematic showing the band diagram for a reverse-biased p-n junction along with how an electron in the valence band can tunnel into an unoccupied state in the conduction band.(b) The potential barrier seen by the electron during the tunneling process.

Zener breakdown), the current for reverse bias voltages greater than V_z is

$$I = \frac{|V - V_z|}{R_L} \tag{4.7.19}$$

Example 4.4 A silicon p^+n diode has a doping of $N_a = 10^{19}$ cm^{-3}, $N_d = 10^{16}$ cm^{-3}. Calculate the 300 K breakdown voltage of this diode. If a diode with the same ϵ/N_d value were to be made from diamond, calculate the breakdown voltage.

The critical fields of silicon and diamond are (at a doping of 10^{16} cm^{-3}) $\sim 4 \times 10^5$ V/cm and 10^7 V/cm. The breakdown voltage is

$$V_{BD}(Si) = \frac{\epsilon(\mathcal{E}_{crit})^2}{2eN_d} = \frac{(11.9)(8.85 \times 10^{-14}\mathrm{F/cm})(4 \times 10^5 \ \mathrm{V/cm})^2}{2(1.6 \times 10^{-19} \ \mathrm{C})(10^{16} \ \mathrm{cm}^{-3})} = 51.7 \ \mathrm{V}$$

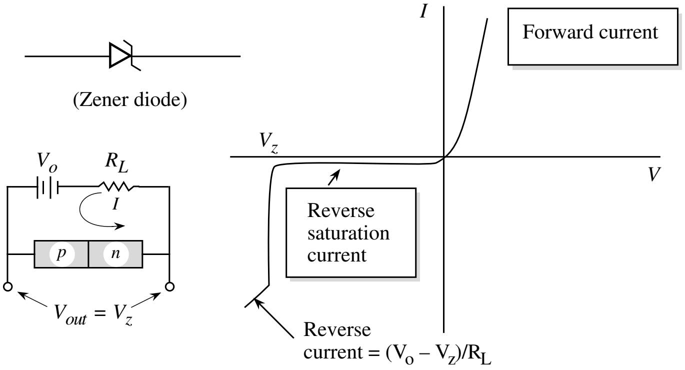


Figure 4.23: (a) Tunneling breakdown effect in the reverse-biased p-n diode for a voltage-clamping circuit. The circuit is thus very useful as a voltage regulator and zener diode circuit symbol.

The breakdown for diamond is

$$V_{BD}(\mathrm{C}) = 51.7 \times \left(\frac{10^7}{4 \times 10^5}\right)^2 = 32.3\,\mathrm{kV!}$$

One can see the tremendous potential of diamond for high-power applications where the device must operate under high applied potentials. At present, however, diamond-based diodes are not commercially available.

4.8 DIODE APPLICATIONS: AN OVERVIEW

4.8.1 Applications of p-n diodes

The p-n junction (or the Schottky diode) is the fundamental building block of semiconductor devices. The applications are based on certain properties of the junction

I. The injection of electron-hole pairs to generate light via recombination (eg. LEDs and LASERs)

II. The separation of electron-hole pairs at the junction to constitute a current source (eg. solar cell)

III. The temperature dependence of the I-V characteristic (eg. a temperature sensor)

IV. The non-linear nature of the I-V characteristic (eg. frequency multipliers and mixers)

V. The device as a switch (eg. rectifiers, inverters, power supplies etc)

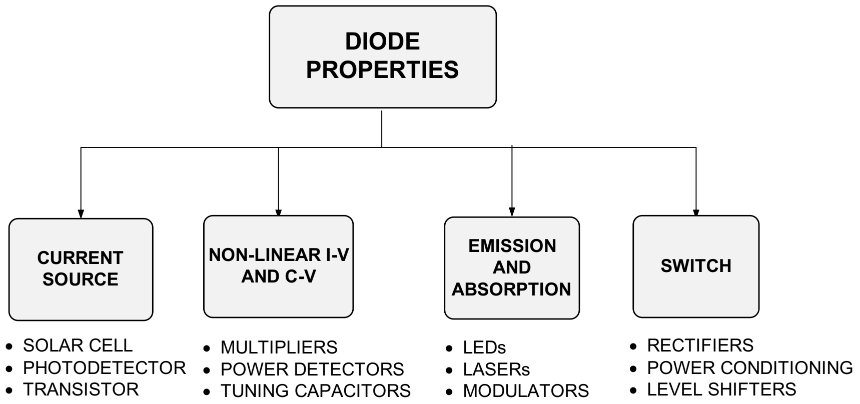


Table 4.1: Some important applications of semiconductor diodes in electronics and optoelectronics.

We now go through these applications briefly to help explain how these properties are harnessed. The goal is not to provide full details but to elucidate methodology. The diode has many uses when employed as a current source. The diode when operated under reverse bias has the properties of a current source (infinite output resistance or equivalently a constant current with voltage). Consider figure 4.24. If a current source is available which can be controlled then it can form the basis of several critical and valuable applications. If large changes in the current source can be effected by a small change in input voltage, $\Delta V_{\rm in}$) then the resultant change in output voltage, $\Delta V_{\rm out}$, could be large if the current is delivered to a large load resistance. The resultant voltage gain , $\Delta V_{\rm out}/\Delta V_{\rm in}$, forms the basis of transistor operation and explains why the output of a transistor is always represented by a current source. If the current source can be controlled by incident photons, then the resultant current is basis of the operation of a photodetector or a solar cell. The transistor is described in detail in later chapters and we describe the solar cell and photodetector below.

4.8.2 The Solar Cell and Photodetector

Consider a reverse biased diode which is subjected to illumination with photons with energy larger than the bandgap.

Generation Currents: p-n Junctions Illuminated With Light

For a reverse biased junction, equation 4.3.15 can be understood as follows. Any minority carrier electrons generated within a diffusion length of the depletion edge can diffuse to the edge

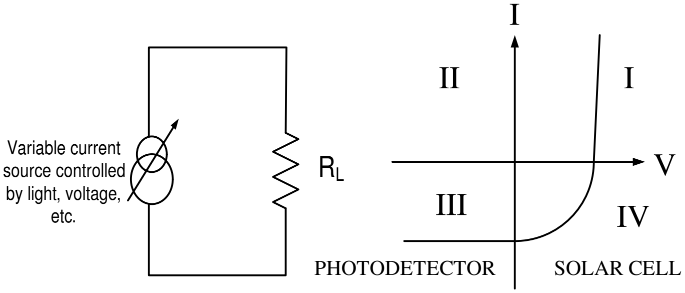


Figure 4.24: Simple circuit diagram (left) and current-voltage ($I - V$) plot showing regimes of operation for a $p - n$ diode under illumination. When operated in quadrant III, the device acts as a photodetector, whereas in quadrant IV it behaves as a solar cell

of the junction and be swept away. Minority electrons generated well beyond a length L_n will recombine with holes resulting in the equilibrium concentration, n_{p0}. Similarly holes generated within, L_p, a diffusion length, of the depletion region edge will be swept into the depletion region.

In the event that there is light shining on the p-n junction, as shown in figure 4.25a, then the charge profile is perturbed in the following manner. Far in the bulk region, an excess minority carrier concentration is generated, where $\Delta n_p = G_L \tau_n$ and $\Delta p_n = G_L \tau_p$. This is shown in figure 4.25b. The new equation to be solved for reverse saturation current differs from the one previously used in that a light generation term is added.

$$D_p \frac{d^2 p}{dx^2} + G_{th} - R + G_L = 0 \tag{4.8.1}$$

or

$$D_p \frac{d^2 p}{dx^2} + \frac{p_{n0} - p_n}{\tau_p} + G_L = 0 \tag{4.8.2}$$

with boundary conditions similar to before.

$$p_n(\infty) = p_{n0} + \tau_p G_L \tag{4.8.3}$$

$$p_n(W_n) = 0 \tag{4.8.4}$$

Solving these equations, we get

$$p_n(x) = (p_{n0} + \tau_p G_L) \left[1 - \exp \left(-\frac{x - W_n}{L_p} \right) \right] \tag{4.8.5}$$

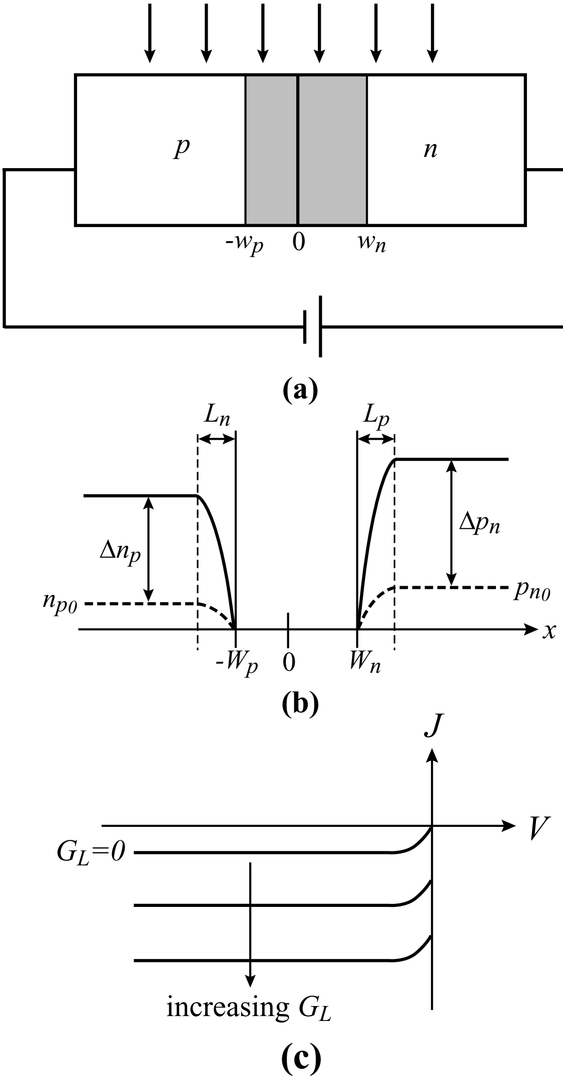


Figure 4.25: (a) Schematic diagram of a reverse-biased p-n junction illuminated with light. (b) Minority carrier profile in the structure. (c) Reverse bias current increases as the light intensity is increased.

A similar set of equations for electrons gives us the following expression for $n_p(x)$ in the neutral p-region:

$$n_p(x) = (n_{p0} + \tau_n G_L) \left[1 - \exp\left(\frac{x + W_p}{L_n}\right)\right] \tag{4.8.6}$$

The slope of the charge profile at the edge of the depletion region is

$$\frac{dp_n(W_n)}{dx} = \frac{p_{n0} + \tau_p G_L}{L_p} \tag{4.8.7}$$

Therefore

$$J_p(x = W_n) = eD_p \left(\frac{p_{n0} + \tau_p G_L}{L_p}\right) \tag{4.8.8}$$

Similarly,

$$J_n(x = W_p) = eD_n \left(\frac{n_{p0} + \tau_n G_L}{L_n}\right) \tag{4.8.9}$$

The reverse saturation current J_R is then given by

$$\boxed{J_R = e\left[D_n \frac{n_{p,bulk}}{L_n} + D_p \frac{p_{n,bulk}}{L_p}\right]} \tag{4.8.10}$$

where $n_{p,bulk}(p_{n,bulk})$ is the minority carrier concentration in the bulk in non-equilibrium (steady state). Here $n_{p,bulk} = n_{p0} + \tau_n G_L$. By changing the slope of the minority profile at the edge of the junction, such as by shining light on the diode, it is possible to control the reverse current across the diode. This is shown schematically in figure 4.25c. Controlling and monitoring the current flowing across a reverse bias diode forms the basis of a large number of devices, including photodetectors and bipolar transistors. As the incident light intensity is enhanced, or equivalently the electron-hole pair generation rate is increased, the reverse current increases as is shown schematically in figure 4.24, where the $I - V$ plane is demarcated into four quadrants. The photodetector operation is in the third quadrant. Notice here that the current and voltage have the same sign (negative) and hence the device dissipates power (a positive product of current and voltage). However, if a positive voltage is applied to the diode while light is incident on the junction then the sign of the current is negative and the sign of the voltage across the diode is positive. This results in a negative product of current and voltage or the diode is a source of power and not a dissipater of power. This is the regime of operation of the solar cell and is in the fourth quadrant of the $I - V$ plane. The current characteristic is best analyzed by employing the rule that the current through the diode is always the sum of forward and reverse currents. In the absence of any energy source (other than thermal) carriers contributing to both forward and reverse currents are generated thermally (either by the thermal ionization of dopants, or band-to-band generation). At zero bias these currents balance each other. In a solar cell under optical excitation the forward current is unchanged and continues to be provided by the thermal injection of carriers across the junction (as has been described before) whereas the reverse current changes dramatically and is carried dominantly by photo-generated carriers. This is the reason why the net current is not zero at zero applied bias in an illuminated solar cell. This current is called

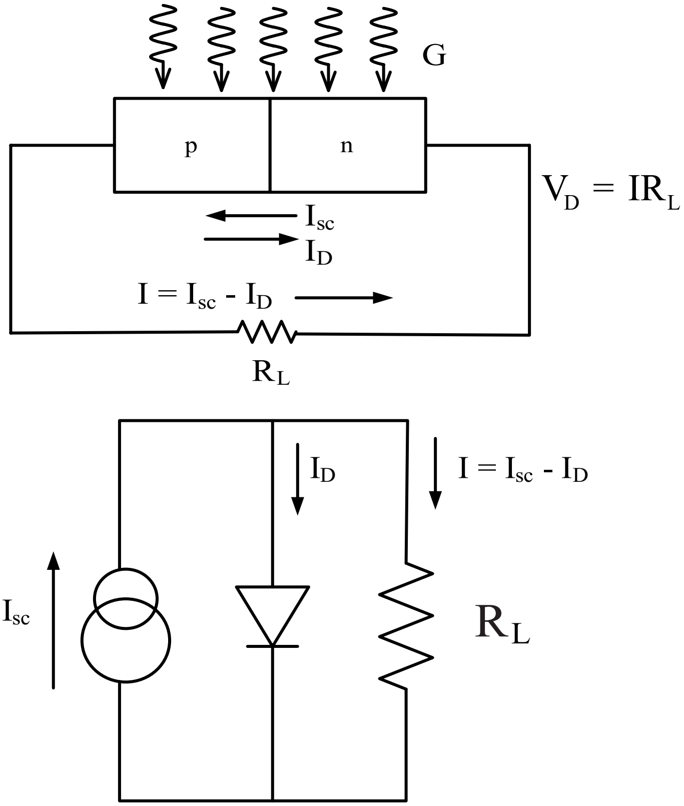


Figure 4.26: Equivalent circuit of a solar cell

the short circuit current, I_{sc}. The forward voltage increases the forward thermionic/diffusion currents exponentially as given by the diode law whereas the reverse current remains a constant with the net current being given by

$$J \cdot A = I = I_f - I_r = I_s \left[\exp\left(\frac{eV}{k_B T} \right) - 1 \right] - I_{sc} \tag{4.8.11}$$

In a solar cell configuration the forward bias is not explicitly applied across the cell. It is generated by the flow of the current across the load (which may be the resistance of a light bulb for instance). This is shown schematically in figure 4.26 along with the equivalent circuit of the solar cell. The total cell current goes to zero at a voltage, $\mathrm{V_{oc}}$, termed the open circuit voltage, when the forward diode current is equal and opposite to the generated current. From equation 4.8.11

$$V_{oc} = \frac{k_B T}{e} \ln\left(\frac{I_{sc}}{I_s} + 1\right)$$

To obtain the maximum power from a cell it is desirable to have the largest product of voltage and current possible in the fourth quadrant of the $I - V$ plane. The maximum power point is that bias at which the maximum power is available from the cell, or, is the bias at which the largest rectangle can be accommodated within the I-V curve. The power at any bias point is given by the IV product

$$P = I \cdot V = (I_{sc} - I_D) \cdot V = \left[I_{sc} - I_s \ \exp\left(\frac{eV}{k_B T}\right)\right] \cdot V$$

and the maximum power point is obtained by maximizing the product. This is left as an exercise. The maximum power is also alternately represented by

$$P = V_{oc} I_{sc} \cdot F$$

where F is the defined as the Fill Factor of the cell. Hence to get the maximum power from a cell it is desirable to obtain the largest $\mathrm{V_{oc}}$ and $\mathrm{I_{sc}}$ possible which is best achieved by using a tandem cell which comprise of a series connection of cells with different bandgaps that maximize solar absorption (while maintaining a large open circuit voltage) coupled with concentrator lenses that maximize input photon intensity.

4.8.3 The uses of diode non-linearity (Mixers, Multipliers, Power Detectors)

A mixer is a frequency translation device that translates an input signal band of frequencies to a different band of output frequencies. There are two main uses of the mixer: down conversion and up conversion. Down conversion, used in receivers, takes a higher input RF frequency and shifts it down to a lower frequency where the channel selection can be performed and interfering signals can be filtered out. Up conversion takes a lower frequency band limited signal and shifts it to a higher frequency. This is typically the transmitter application.

A mixer does not really "mix" or sum signals; it multiplies them. For example, the analog multiplier performs the frequency translation function:

$$A = A \sin \omega_1 t \qquad B = B \sin \omega_2 t$$

$$(A \sin \omega_1 t)(B \sin \omega_2 t) = (AB/2)\left[\cos(\omega_1 - \omega_2)t - \cos(\omega_1 + \omega_2)t\right] \tag{4.8.12}$$

Note that both sum and difference frequencies are obtained by the multiplication of the two

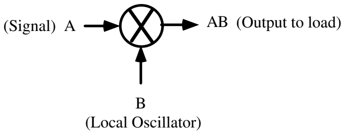


Figure 4.27: Mixer symbol. A represents the signal input; B the reference input.

input sinusoidal signals as shown in equation 4.8.12. One of these is the input signal (A) whose amplitude and phase generally vary with time. The other input (B) is a reference signal, locally generated, called the local oscillator, normally with fixed amplitude and phase. With the ideal analog multiplication process shown in figure 4.27, no harmonics or spurious signals are produced. Also, there is no feed through of A or B to the output. But, in reality, mixers always produce many spurious outputs that consist of harmonics of A and B and additional mixing products $m\omega_1 \pm\ n\omega_2$, where m and n are integers. A "good" mixer is designed such that it suppresses these spurious outputs and provides a highly linear amplitude and phase relationship between signal input (A) and the output.

The forward I-V characteristic of the diode can be represented by a series expansion. For example, in the case of a simple exponential diode characteristic, equation 4.8.13 can represent the current voltage characteristic. Coefficients a_i will vary with DC bias, series resistance, and the shape of the $I - V$ characteristic.

$$\begin{aligned} I_D &= I_S \left[\exp\left(\frac{eV_D}{k_BT)} \right) - 1 \right] \\ &\simeq I_S \left[a_1 V_D + \frac{1}{2} a_2 V_D^2 + \frac{1}{6} a_3 V_D^3 + \ldots \right] \end{aligned} \tag{4.8.13}$$

$$\begin{aligned} V_{RF} + V_{LO} - V_D - I_S \left(a_1 V_D + \frac{1}{2} a_2 V_o^2 \right) (R_S + R_L) = 0 \\ V_o(t) = I_S (a_1 V_D + \frac{1}{2} a_2 V_D^2) R_L \\ a_2 V_D^2 \sin^2 \omega t = a_2 V_D^2 \left[1 - \cos 2\omega t \right] \end{aligned} \tag{4.8.14}$$

Now, suppose that two inputs are summed as shown in figure 4.27 and the diode current produces an output $V_o(t)$ across resistor R_L. One input V_{RF} is the signal; the other VLO is the reference local oscillator. The diode voltage, V_D, can be found using the series approximation equation 4.8.13, and the output voltage, $V_o(t)$, is calculated from the diode current I_D. If only first and second order terms are used, a quadratic equation is easily solved.

While only the outputs shown in equation 4.8.12 are desired, the mixer output will also contain a DC term, RF and LO feed through, and terms at all harmonics of the RF and LO frequencies. Only the second-order product term produces the desired outputs. It can be seen in equation

4.8.14 that the second-order nonlinearity also produces a second harmonic and a DC term. The second harmonic generation is the property used in frequency multipliers. Also, the DC term amplitude is proportional to the square of the input voltage, hence input power. This is the principle of operation of diode power detectors.

4.8.4 Power Devices

A DC-to-DC converter is a module that accepts a DC input voltage and produces a DC output voltage typically at a different voltage level or of different polarity. These modules have become ubiquitous in modern electronic systems. For example, laptops use them to convert the mains power supply voltage to the battery voltage (18 V), which in turn is converted to the supply voltage for the computing electronics (1.5-3.5 V) and the voltage for the display (voltage variable depending on type of display). All are different! In addition, DC-to-DC converters are used to provide bus isolation, power bus regulation, etc. There are several topologies to achieve the desired conversion and we will briefly discuss a Buck or Step- Down Converter to appreciate the functional requirements of the transistor switch and diode that this employed. As in most power conversion circuits it is imperative to not have current flow with a large voltage across dissipative elements such as transistor switches. This will cause power dissipation and excessive heating in the circuit. To reduce the voltage across a switch while it is conducting, an inductor is typically employed in circuits. Furthermore a capacitor is used at the output to stabilize the output voltage through the switching cycle. In the Buck/Step-Down circuit (figure 4.28), an input transistor is turned on causing the input voltage $\mathrm{V_{in}}$ (which has to be stepped-down) to appear at one end of the inductor while the other remains at the output. This voltage will cause the inductor current to rise, storing energy as magnetic flux. During this process the diode is reverse biased and turned off and the current flows through the transistor and the inductor to the output capacitor and load. When the transistor is turned off, the current through the inductor will continue flowing but now be forced through the diode causing the diode to turn on. This process is called free-wheeling. The voltages V_x and V_o will follow standard L and C charging/discharging relationships as shown in figure 4.28.

Figure 4.28 shows schematically the change in the current and voltage across the inductor over a switching cycle of the transistor. From the relation

$$V_x - V_o = L\frac{di}{dt} \tag{4.8.15}$$

the change of current satisfies

$$i_f - i_i = \int_{\text{ON}} (V_x - V_o)\,dt + \int_{\text{OFF}} (V_x - V_o)\,dt \tag{4.8.16}$$

where $\mathrm{i_i}$ and $\mathrm{i_f}$ are the currents through the inductor at the beginning and the end of a cycle. For steady state operation it is required that the current at the start and end of a period T be the same. To get a 'simple relation' between voltages we assume 'no voltage drop across transistor or diode' while ON and a perfect switch change. Thus during the ON time $V_x = V_{in}$ and in the

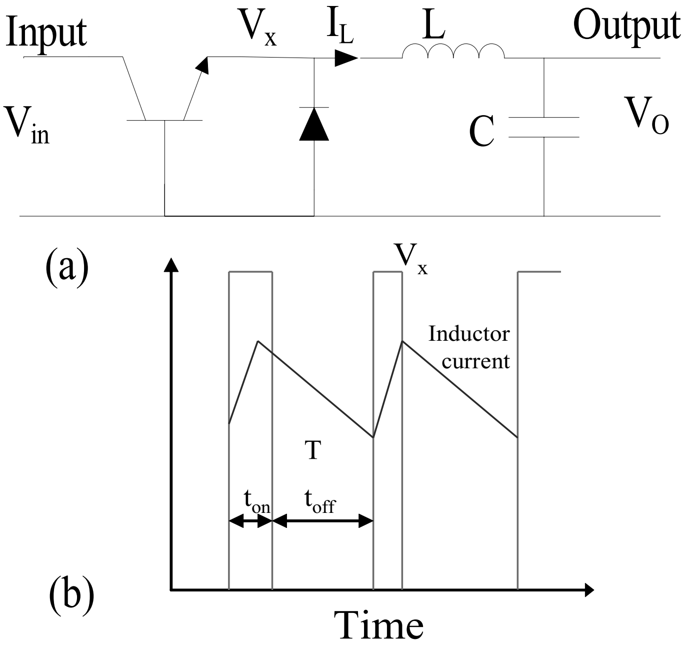


Figure 4.28: (a)Buck converter schematic (b) Voltage and current changes

OFF time $V_x = 0$. Thus in steady state

$$i_f - i_i = 0 = \int_0^{t_{\text{ON}}} (V_{in} - V_o)\,dt + \int_{t_{\text{ON}}}^{t_{\text{ON}}+t_{\text{OFF}}} (-V_o)\,dt \tag{4.8.17}$$

which gives

$$(V_{in} - V_o)\,t_{\text{ON}} - V_o t_{\text{OFF}} = 0 \tag{4.8.18}$$

or

$$\frac{V_o}{V_{in}} = \frac{t_{\text{ON}}}{T} \tag{4.8.19}$$

Using D as the “duty cycle” gives

$$D = \frac{t_{\text{ON}}}{T} \tag{4.8.20}$$

the voltage relationship becomes $V_o = DV_{in}$. Since the circuit is ’lossless’ and the input and output powers must match on the average $V_oI_o = V_{in}I_{in}$. This requires the diode to have no voltage drop in the forward direction and no leakage current in the reverse direction. In the presence of a reverse leakage current I_{rev}, additional loss of $I_{rev}V_{in}$ occurs in the free-wheeling diode when the transistor is ON; and if the forward diode drop is non- negligible, $I_oV_{\rm diode}$ will be lost in the diode when the transistor is OFF. Both these losses are important since in the first case V_{in} is large and in the second I_o is large. It is imperative that in these applications the device behave as a nearly- perfect diode with $\mathrm{V_{diode}}$ and $\mathrm{I_{rev}}$ both being as small as possible. Furthermore, the diode should switch off faster than the transistor, to reduce transient dissipation in it. Schottky diodes which are unipolar and have short switching times are emerging as preferred diodes in free-wheeling applications.

One of the most important applications is in the area of optoelectronic devices. Essentially all the semiconductor devices catering to optoelectronics use the diode concept. These include detectors, avalanche photodetectors, optical modulators, as well as light-emitting diodes and semiconductor lasers. In this section we discuss the operation of the light emitting diode

4.9 Light emitting diode (LED)

The simplicity of the light-emitting diode (LED) makes it a very attractive device for display and communication applications. The basic LED is a p-n junction that is forward biased to inject electrons and holes into the p- and n-sides respectively. The injected minority charge recombines with the majority charge in the depletion region or the neutral region. In direct band semiconductors, this recombination leads to light emission since radiative recombination dominates in high-quality materials. In indirect gap materials, the light emission efficiency is quite poor and most of the recombination paths are nonradiative, which generates heat rather than light. In the following section we will examine the important issues that govern the LED operation.

We will briefly outline some of the important considerations in choosing a semiconductor for LEDs or laser diodes.

4.9.1 Emission Energy

The light emitted from the device is very close to the semiconductor bandgap, since the injected electrons and holes are described by quasi-Fermi distribution functions. The desire for a particular emission energy may arise from a number of motivations. In figure 4.29 we show the response of the human eye to radiation of different wavelengths. Also shown are the bandgaps of some semiconductors. If a color display is to be produced that is to be seen by people, one has to choose an appropriate semiconductor. Very often one has to choose an alloy, since there is a greater flexibility in the bandgap range available. In figure 4.30 we show the loss characteristics of an optical fiber. As can be seen, the loss is least at 1.55 μm and 1.3 μm. If optical communication sources are desired, one must choose materials that can emit at these wavelengths. This is especially true if the communication is long haul, i.e., over hundreds or even thousands of kilometers. InP-based materials are used for these applications. Materials like GaAs that emit at 0.8 μm can still be used for local area networks (LANs), which involve communicating within a building or local areas. The area of displays and lighting is filled dominantly by GaN-based

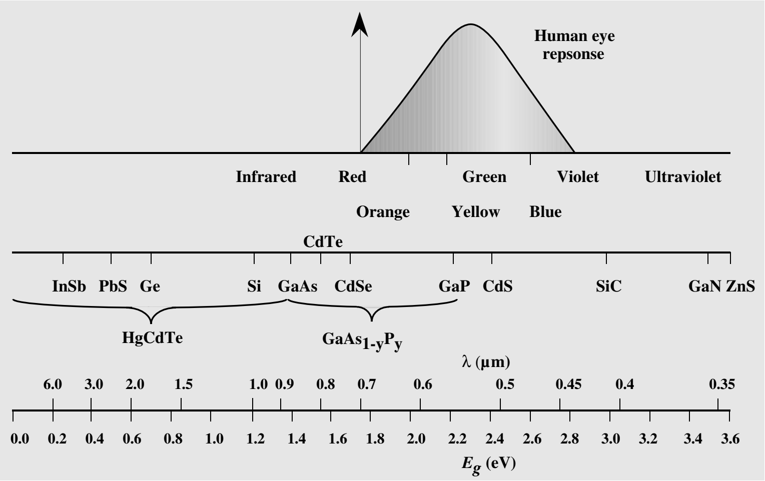


Figure 4.29: The bandgap and cutoff wavelengths for several semiconductors. The semiconductor bandgaps range from 0 (for $\mathrm{Hg_{0.84}Cd_{0.15}Te}$) to well above 3 eV, providing versatile detection systems.

materials using InGaN as the emission region for blue and green and GaAs-based AlGaInP for the red region.

Substrate Availability:

Almost all optoelectronic light sources depend upon epitaxial crystal growth techniques where a thin active layer (a few microns) is grown on a substrate (which is $\sim$ 200 μm). The availability of a high-quality substrate is extremely important in epitaxial technology. If a substrate that lattice-matches to the active device layer is not available, the device layer may have dislocations and other defects in it. These can seriously hurt device performance. One of the most important opto-electronic materials for LEDs that has emerged lately is GaN. In spite of the lack of a native substrate, GaN-based LEDs grown on either sapphire or SiC have become multi-billion dollar industry. The reason is that the InGaN quantum well which is used as the emission region has fluctuations which cause local energy minima for electron and holes. Thus radiative recombination is encouraged within this region and diffusion to and non-radiative recombination at a dislocation minimized. This is shown schematically in figure 4.31. Furthermore, the dislocation propagation and generation of dislocation sin GaN is very high because of the high bond energies in the material. This eliminates one of the failure mechanisms in conventional LEDs and lasers, that of generation and propagation of dislocations caused by absorption of emitted the photon energy. The important substrates that are available for conventional light-emitting technology

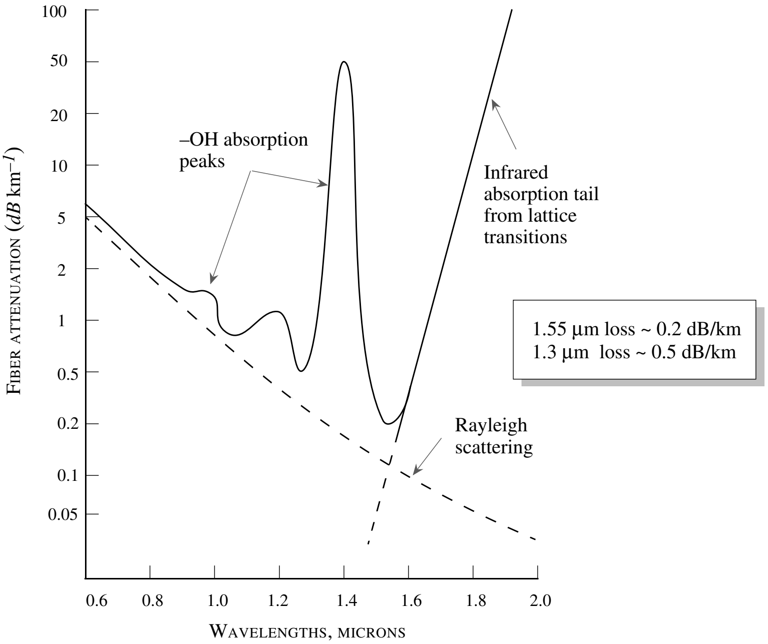


Figure 4.30: Optical attenuation vs. wavelength for an optical fiber. Primary loss mechanisms are identified as absorption and scattering.

(which do not benefit from the above mentioned advantages of GaN and InGaN) are GaAs and InP. A few semiconductors and their alloys can match these substrates. The lattice constant of an alloy is the weighted mean of the lattice constants of the individual components, i.e., the lattice constant of the alloy A_xB_{1-x} is

$$a_{all} = xa_A + (1-x)a_B \tag{4.9.1}$$

where a_A and a_B are the lattice constants of A and B. Semiconductors that cannot lattice-match with GaAs or InP have an uphill battle for technological success. The crystal grower must learn the difficult task of growing the semiconductor on a mismatched substrate without allowing dislocations to propagate into the active region.

Important semiconductor materials exploited in optoelectronics are the alloy $Ga_xAl_{1-x}As$, and AlGaInP which is a quaternary material which is lattice-matched very well to GaAs substrates; $In_{0.53}Ga_{0.47}As$ and $In_{0.52}Al_{0.48}As$, which are lattice-matched to InP; InGaAsP, whose composition can be tailored to match with InP and can emit at 1.55 μm; and GaAsP, which has a wide range of bandgaps available.

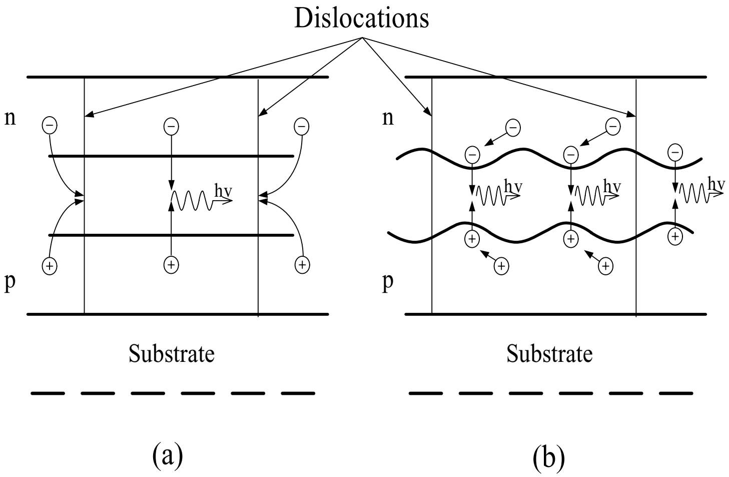


Figure 4.31: (a) e-h diffuse to dislocations and recombine. (b) In the presence of energy fluctuations such as in the InGaN the electrons recombine efficiently.

In general, the electron-hole recombination process can occur by radiative and nonradiative channels. Under the condition of minority carrier recombination or high injection recombination, as shown in section 3.8.1and section 3.8.2, one can define a lifetime for carrier recombination. If τ_r and τ_{nr} are the radiative and nonradiative lifetimes, the total recombination time is (for, say, an electron)

$$\frac{1}{\tau_n} = \frac{1}{\tau_r} + \frac{1}{\tau_{nr}} \tag{4.9.2}$$

The internal quantum efficiency for the radiative processes is then defined as

$$\eta_{Qr} = \frac{\frac{1}{\tau_r}}{\frac{1}{\tau_r} + \frac{1}{\tau_{nr}}} = \frac{1}{1 + \frac{\tau_r}{\tau_{nr}}} \tag{4.9.3}$$

In high-quality direct gap semiconductors, the internal efficiency is usually close to unity. In indirect materials the efficiency is of the order of 10^{-2} to 10^{-3}.

Before starting the discussion of light emission, let us remind ourselves of some important definitions and symbols used in this chapter:

I_{ph} : photon current = number of photons passing a cross-section/second.
J_{ph} : photon current density = number of photons passing a unit area/second.
P_{op} : optical power intensity = energy carried by photons per second per area.

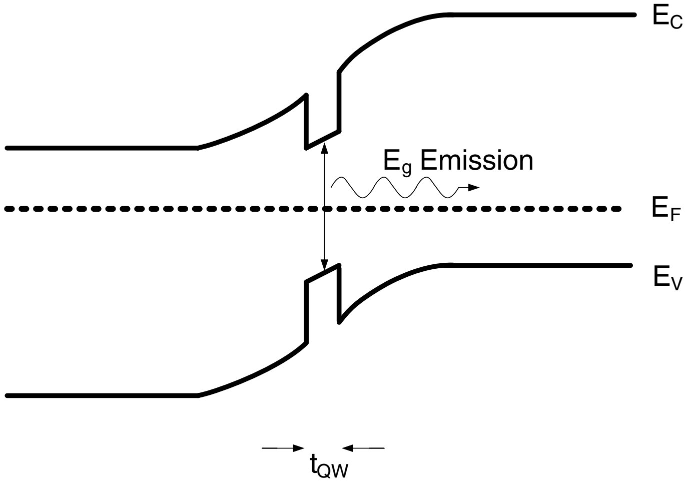


Figure 4.32: Band diagram of a single quantum-well LED with the advantages of increased carrier density, enhanced confinement and reduced probability of re-absorption of emitted photons in bulk layers.

4.9.2 Carrier Injection and Spontaneous Emission

The LED is essentially a forward-biased p-n diode, with a quantum well emission region as shown in figure 4.32. The reason for using a quantum well is to (i) increase the electrons and hole density in the recombination region increasing the direct recombination rate and leading to higher light output, (ii) having an emission region that is lower in energy that the injection (cladding) regions which allows the generated photons to escape without being re-absorbed in the injection regions, (iii) minimizing the overflow of electrons into the cladding regions where the injected carriers either recombine non-radiatively or generate light of an undesired wavelength. The current flow in a p-n junction was discussed in detail earlier in this chapter.The basis of that derivation was that electrons and holes are injected across the junction and recombine either in the bulk(long base case) or at contacts (short base case). Neither of those conditions apply to an LED. Here the current flow occurs via recombination in the quantum well. The turn-on voltage of the LED is therefore given by the bandgap of the emission region and is not explicitly related to the built-in voltage of the p-n junction. An example of this is an InGaN LED grown within GaN p-and n regions. The built-in voltage of this device is close to the bandgap of GaN (3.4V) though the turn-on voltage is 2.8V close to the emission energy of the photons. The current flow mechanism is shown in figure 4.33. The current is given by $J = e \cdot R_{spon}$ where R_{spon} is the spontaneous recombination rate in the well. The efficiency of the process is the

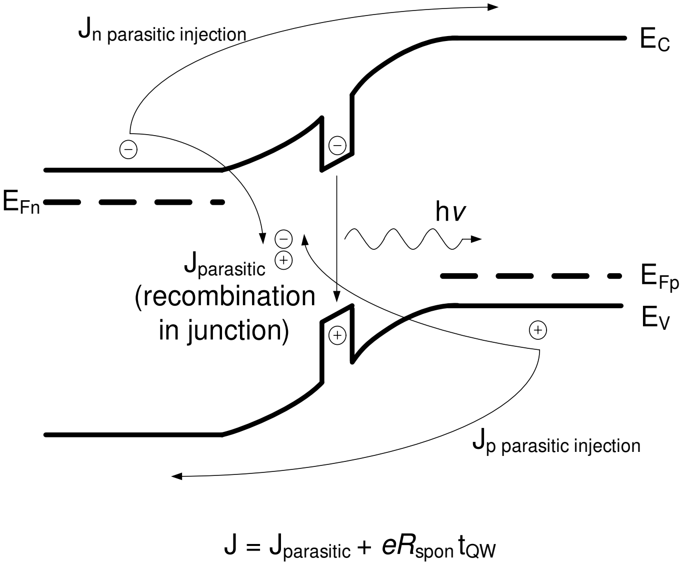

J_n parasitic injection
E_C
E_{Fn}
$h\nu$
$J_{parasitic}$
(recombination in junction)
E_{Fp}
E_V
J_p parasitic injection
$J = J_{parasitic} + eR_{spon}t_{QW}$

Figure 4.33: Current flow mechanisms in a LED.

ratio of the current generating photons of the desired wavelength to the total current. The current calculated for the p-n junction in the earlier sections are the wasted currents in the LED as they calculate currents in the bulk and due to non-radiative centers. What remains to be calculated is the spontaneous recombination rate $R_{spon}.a$

As discussed in section 3.8.1, the radiative process is "vertical," i.e., the k-value of the electron and that of the hole are the same in the conduction and valence bands, respectively. From figure 4.34 we see that the photon energy and the electron and hole energies are related by

$$\hbar\omega - E_g = \frac{\hbar^2 k^2}{2}\left[\frac{1}{m_e^*} + \frac{1}{m_h^*}\right] = \frac{\hbar^2 k^2}{2m_r^*} \tag{4.9.4}$$

where m_r^* is the reduced mass for the e-h system.

If an electron is available in a state k and a hole is also available in the state k (i.e., if the Fermi functions for the electrons and holes satisfy f^e(k) = f^h(k) = 1), the radiative recombination rate is found to be

$$\boxed{W_{em} \sim 1.5 \times 10^9 \hbar\omega \text{ [eV s}^{-1}]} \tag{4.9.5}$$

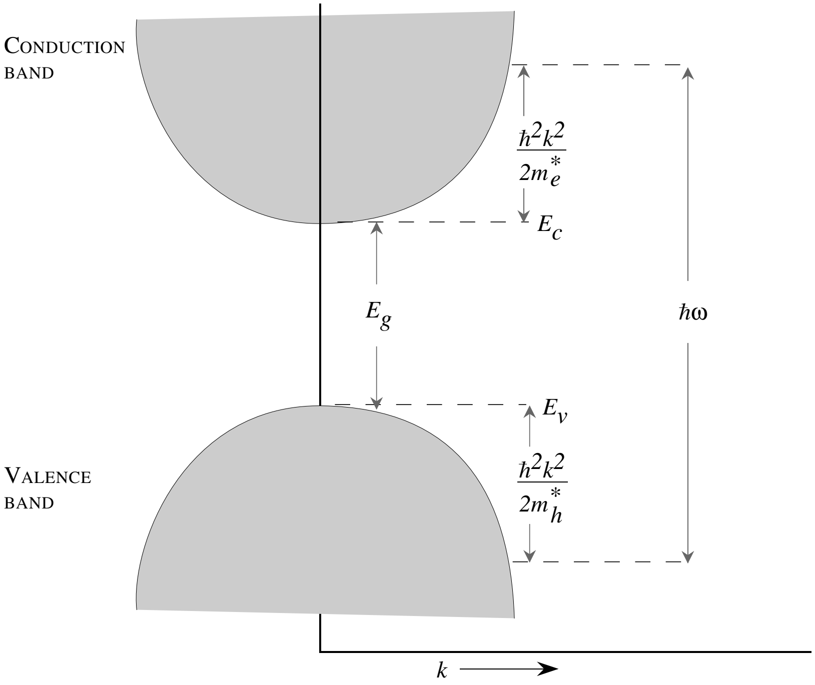


Figure 4.34: A schematic of the E-k diagram for the conduction and valence bands. Optical transitions are vertical; i.e., the k-vector of the electron in the valence band and in the conduction band is the same.

and the recombination time becomes ($\hbar\omega$ is expressed in electron volts)

$$\boxed{\tau_o = \frac{0.67}{\hbar\omega[\text{eV}]}\ \text{ns}} \tag{4.9.6}$$

The recombination time discussed above is the shortest possible spontaneous emission time since we have assumed that the electron has a unit probability of finding a hole with the same k-value.

When carriers are injected into the semiconductors, the occupation probabilities for the electron and hole states are given by the appropriate quasi-Fermi levels. The emitted photons leave the device volume so that the photon density never becomes high in the e-h recombination region. In a laser diode the situation is different, as we shall see later. The photon emission rate is given by integrating the emission rate W_{em} over all the electron-hole pairs after introducing the appropriate Fermi functions.

There are several important limits of the spontaneous rate:

i. In the case where the electron and hole densities n and p are small (non degenerate case), the Fermi functions have a Boltzmann form $(\exp(-E/k_BT))$. The recombination rate is found to be

$$R_{spon} = \frac{1}{2\tau_o}\left(\frac{2\pi\hbar^2 m_r^*}{k_BTm_e^*m_h^*}\right)^{3/2} np \tag{4.9.7}$$

The rate of photon emission depends upon the product of the electron and hole densities. If we define the lifetime of a single electron injected into a lightly doped $(p = N_a \leq 10^{17}\text{cm}^{-3})$ p-type region with hole density p, it would be given from equation 4.9.7 by

$$\boxed{\frac{R_{spon}}{n} = \frac{1}{\tau_r} = \frac{1}{2\tau_o}\left(\frac{2\pi\hbar^2 m_r^*}{k_BTm_e^*m_h^*}\right)^{3/2} p} \tag{4.9.8}$$

The time τ_r in this regime is very long (hundreds of nanoseconds), as shown in figure 4.35, and becomes smaller as p increases.

ii. In the case where electrons are injected into a heavily doped p-region (or holes are injected into a heavily doped n-region), the function $f^h(f^e)$ can be assumed to be unity. The spontaneous emission rate is

$$R_{spon} \sim \frac{1}{\tau_o}\left(\frac{m_r^*}{m_h^*}\right)^{3/2} n \tag{4.9.9}$$

for electron concentration n injected into a heavily doped p-type region and

$$R_{spon} \sim \frac{1}{\tau_o}\left(\frac{m_r^*}{m_e^*}\right)^{3/2} p \tag{4.9.10}$$

for hole injection into a heavily doped n-type region.

The minority carrier lifetimes (i.e., n/R_{spon}) play a very important role not only in LEDs but also in diodes and bipolar devices. In this regime the lifetime of a single electron (hole) is independent of the holes (electrons) present since there is always a unity probability that the electron (hole) will find a hole (electron). The lifetime is now essentially τ_o, as shown in figure 4.35.

iii. Another important regime is that of high injection, where $n = p$ is so high that one can assume $f^e = f^h$ = 1 in the integral for the spontaneous emission rate. The spontaneous emission rate is

$$\boxed{R_{spon} \sim \frac{n}{\tau_o} \sim \frac{p}{\tau_o}} \tag{4.9.11}$$

and the radiative lifetime $(n/R_{spon} = p/R_{spon})$ is τ_o.

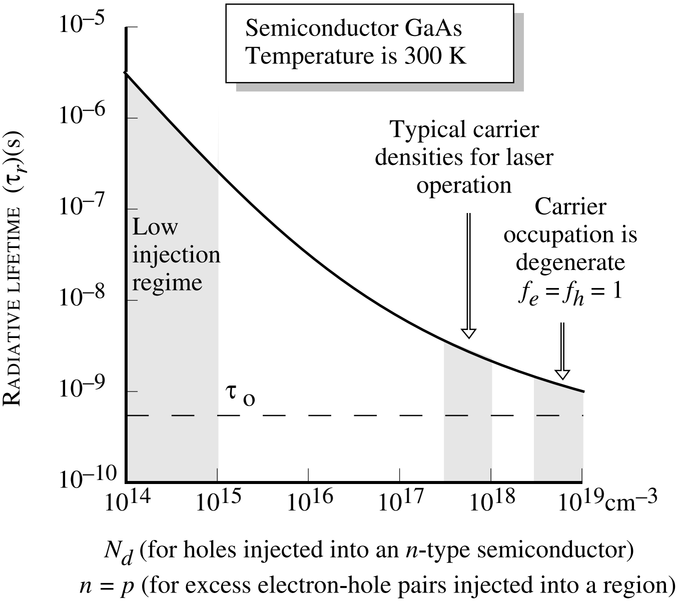


Figure 4.35: Radiative lifetimes of electrons or holes in a direct gap semiconductor as a function of doping or excess charge. The figure gives the lifetimes of a minority charge (a hole) injected into an n-type material. The figure also gives the lifetime behavior of electron-hole recombination when excess electrons and holes are injected into a material as a function of excess carrier concentration.

iv. A regime that is quite important for laser operation is one where sufficient electrons and holes are injected into the semiconductor to cause "inversion." As will be discussed later, this occurs if $f^e + f^h \geq 1$. If we make the approximation $f^e \sim f^h = 1/2$ for all the electrons and holes at inversion, we get the relation

$$R_{spon} \sim \frac{n}{4\tau_o} \tag{4.9.12}$$

or the radiative lifetime at inversion is

$$\boxed{\tau \sim \frac{\tau_o}{4}} \tag{4.9.13}$$

This value is a reasonable estimate for the spontaneous emission rate in lasers near threshold.

The radiative recombination depends upon the radiative lifetime τ_r and the non-radiative lifetime τ_{nr}. To improve the efficiency of photon emission one needs a value of τ_r as small as possible and τ_{nr} as large as possible. To increase τ_{nr} one must reduce the material defect density. This includes improving surface and interface qualities.

The LED current is then given by

$$J = eR_{spon}t_{QW} + J_0 \exp\left[\frac{e\left(V_{bi} - V_{turnon)}\right)}{k_BT}\right] + J_{SNS}$$

The parasitic currents are the second and third terms in the expression. The second term represents current injected over the barrier and the third term the current recombining at the maximum recombination plane.

Example 4.5 Calculate the e-h recombination time when an excess electron and hole density of 10^{15}cm^{-3} is injected into a GaAs sample at room temperature.

Since 10^{15} cm^{-3} or 10^{21} m^{-3} is a very low level of injection, the recombination time is given by equation 4.9.8 as

$$\begin{aligned} \frac{1}{\tau_r} &= \frac{1}{2\tau_o}\left(\frac{2\pi\hbar^2 m_r^*}{k_BTm_e^*m_h^*}\right)^{3/2} p \\ &= \frac{1}{2\tau_o}\left(\frac{2\pi\hbar^2}{k_BTm_e^* + m_h^*}\right)^{3/2} p \end{aligned}$$

Using τ_o = 0.6 ns and k_BT = 0.026 eV, we get for m_e^* = 0.067 $m_o, m_h^* = 0.45\, m_o$,

$$\begin{aligned} \frac{1}{\tau_r} &= \frac{10^{21}\text{ m}^{-3}}{2 \times (0.6 \times 10^{-9}\text{ s})}\left[\frac{2 \times 3.1416 \times (1.05 \times 10^{-34}\text{ Js})^2}{(0.026 \times 1.6 \times 10^{-19}\text{ J}) \times (0.517 \times 9.1 \times 10^{-31}\text{ kg})}\right]^{3/2} \\ \tau_r &= 5.7 \times 10^{-6}s \cong 9.5 \times 10^3\tau_o \end{aligned}$$

We see from this example that at low injection levels, the carrier lifetime can be very long. Physically, this occurs because at such a low injection level, the electron has a very small probability of finding a hole to recombine with.

Example 4.6 In two n^+p GaAs LEDs, $n^+ \gg p$ so that the electron injection efficiency is 100% for both diodes. If the nonradiative recombination time is 10^{-7}s, calculate the 300 K internal radiative efficiency for the diodes when the doping in the p-region for the two diodes is 10^{16} cm^{-3} and 5×10^{17} cm^{-3}.

When the p-type doping is 10^{16} cm^{-3}, the hole density is low and the e-h recombination time for the injected electrons is given by equation 4.9.8 as

$$\frac{1}{\tau_r} = \frac{1}{2\tau_o}\left(\frac{2\pi\hbar^2 m_r^*}{k_BTm_e^*m_h^*}\right)^{3/2} p$$

From the previous example, we can see that for p equal to 10^{16} cm^{-3}, we have (in the previous example the value of p was ten times smaller)

$$\tau_r = 5.7 \times 10^{-7} \text{ s}$$

In the case where the p-doping is high, the recombination time is given by the high-density limit (see equation 4.9.10) as

$$\begin{aligned} \frac{1}{\tau_r} &= \frac{R_{spon}}{n} = \frac{1}{\tau_o}\left(\frac{m_r^*}{m_h^*}\right)^{3/2} \\ \tau_r &= \frac{\tau_o}{0.05} \sim 20\tau_o \sim 12 \text{ ns} \end{aligned}$$

For the low-doping case, the internal quantum efficiency for the diode is

$$\eta_{Qr} = \frac{1}{1+\frac{\tau_r}{t_{nr}}} = \frac{1}{1+(5.7)} = 0.15$$

For the more heavily doped p-region diode, we have

$$\eta_{Qr} = \frac{1}{1+\frac{10^{-7}}{20\times 10^{-9}}} = 0.83$$

Thus there is an increase in the internal efficiency as the p doping is increased.

Example 4.7 Consider a GaAs p-n diode with the following parameters at 300 K:

Electron diffusion coefficient,	D_n	$=$	$30 \text{ cm}^2/\text{V}\cdot\text{s}$
Hole diffusion coefficient,	D_p	$=$	$15 \text{ cm}^2/\text{V}\cdot\text{s}$
p-side doping,	N_a	$=$	$5\times 10^{16} \text{ cm}^{-3}$
n-side doping,	N_d	$=$	$5\times 10^{17} \text{ cm}^{-3}$
Electron minority carrier lifetime,	τ_n	$=$	10^{-8} s
Hole minority carrier lifetime,	τ_p	$=$	10^{-7} s

Calculate the injection efficiency of the LED assuming no recombination due to traps.

The intrinsic carrier concentration in GaAs at 300 K is 1.84×10^6 cm^{-3}. This gives

$$\begin{aligned} n_p &= \frac{n_i^2}{N_a} = \frac{(1.84\times 10^6)^2}{5\times 10^{16}} = 6.8\times 10^{-5} \text{ cm}^{-3} \\ p_n &= \frac{n_i^2}{N_d} = \frac{(1.84\times 10^6)^2}{5\times 10^{17}} = 6.8\times 10^{-6} \text{ cm}^{-3} \end{aligned}$$

The diffusion lengths are

$$\begin{aligned} L_n &= \sqrt{D_n\tau_n} = \left[(30)(10^{-8})\right]^{1/2} = 5.47\ \mu\text{m} \\ L_p &= \sqrt{D_p\tau_p} = \left[(15)(10^{-7})\right]^{1/2} = 12.25\ \mu\text{m} \end{aligned}$$

The injection efficiency is now (assuming no recombination via traps)

$$\gamma_{inj} = \frac{\frac{eD_n n_{po}}{L_n}}{\frac{eD_n n_{po}}{L_n} + \frac{eD_p p_{no}}{L_p}} = 0.98$$

Example 4.8 Consider the p-n^+ diode of the previous example. The diode is forward biased with a forward-bias potential of 1 V. If the radiative recombination efficiency η_{Qr} = 0.5, calculate the photon flux and optical power generated by the LED. The diode area is 1 mm^2.

The electron current injected into the p-region will be responsible for the photon generation. This current is

$$\begin{aligned} I_n &= \frac{AeD_n n_{po}}{L_n}\left[\exp\left(\frac{eV}{k_BT}\right) - 1\right] \\ &= \frac{(10^{-2}\ \mathrm{cm}^2)(1.6\times10^{-19}\ \mathrm{C})(30\ \mathrm{cm}^2/\mathrm{s})(6.8\times10^{-5}\ \mathrm{cm}^{-3})}{5.47\times10^{-4}\ \mathrm{cm}}\left[\exp\left(\frac{1}{0.026}\right) - 1\right] \\ &= 0.30\ \mathrm{mA} \end{aligned}$$

The photons generated per second are

$$\begin{aligned} I_{ph} = \frac{I_n}{e}\cdot\eta_{Qr} &= \frac{(0.30\times10^{-3}\ \mathrm{A})(0.5)}{1.6\times10^{-19}\ \mathrm{C}} \\ &= 9.38\times10^{14}\ \mathrm{s}^{-1} \end{aligned}$$

Each photon has an energy of 1.41 eV (= bandgap of GaAs). The optical power is thus

$$\begin{aligned} \mathrm{Power} &= (9.38\times10^{14}\ \mathrm{s}^{-1})(1.41)(1.6\times10^{-19}\ \mathrm{J}) \\ &= 0.21\ \mathrm{mW} \end{aligned}$$

4.10 PROBLEMS

- **Section 4.2**

Problem 4.1 Why does the potential in a p-n diode fall mainly across the depletion region and not across the neutral region?

Problem 4.2 An abrupt GaAs p-n diode has $N_a = 10^{17}$ cm^{-3} and $N_d = 10^{15}$ cm^{-3}.
(a) Calculate the Fermi level positions at 300 K in the p and n regions.
(b) Draw the equilibrium band diagram and determine the contact potential V_{bi}.

Problem 4.3 Consider an Si p-n diode doped at $N_a = 10^{17}$ cm^{-3}; $N_d = 5\times10^{17}$ cm^{-3} at 300 K. Plot the band profile in the neutral and depletion region. Also, plot the electron and hole concentration from the p- to the n-sides of equilibrium. How good is the depletion approximation?

Problem 4.4 Consider the sample discussed in problem 4.2. The diode has a diameter of 50 μm. Also calculate the charge in the depletion regions and plot the electric field profile in the diode.

Problem 4.5 An abrupt silicon p-n diode at 300 K has a doping of $N_a = 10^{18}$ cm^{-3}, $N_d = 10^{15}$ cm^{-3}. Calculate the built-in potential and the depletion widths in the n and p regions.

Problem 4.6 A Ge p-n diode has $N_a = 5 \times 10^{17}$ cm^{-3} and $N_d = 10^{17}$ cm^{-3}. Calculate the built-in voltage at 300 K. At what temperature does the built-in voltage decrease by 1%?

Problem 4.7 Consider a p-n junction with $N_A = N_D = 10^{17} cm^{-3}$. When the capacitance is measured to be twice the value expected.
The reason is an unintentional interfacial dipole between the p and n layers.

1. What is the magnitude of the dipole moment?
2. Draw the band diagrams of the ideal p-n junction and the non-ideal one. Include the electric field profiles and depletion region widths.

Assume that the dipole is supported by negative and positive charges separated by a very small distance, δ.

- **Section 4.3**

Problem 4.8 Explain, using physical arguments, why the reverse current in a p-n diode does not change with bias (before breakdown). Would this be the case if the electrons and holes had a constant mobility independent of the electric field?

Problem 4.9 The diode of problem 4.3 is subjected to bias values of: (a) V_f = 0.1 V; (b) V_f = 0.5 V; (c) V_r = 1.0V; (d) V_r = 5.0 V. Calculate the depletion widths and the maximum field F_m under these biases.

Problem 4.10 Consider a p^+n Si diode with $N_a = 10^{18}$ cm^{-3} and $N_d = 10^{16}$ cm^{-3}. The hole diffusion coefficient in the n-side is 10 cm^2/s and $\tau_p = 10^{-7}$ s. The device area is 10^{-4} cm^2. Calculate the reverse saturation current and the forward current at a forward bias of 0.8 V at 300 K.

Problem 4.11 Consider a p^+n silicon diode with area 10^{-4} cm^2. The doping is given by $N_a = 10^{18}$ cm^{-3} and $N_d = 10^{17}$ cm^{-3}. Plot the 300 K values of the electron and hole currents I_n and I_p at a forward bias of 0.8 V. Assume $\tau_n = \tau_p = 1$ μs and neglect recombination effects. $D_n = 20$ cm^2/s and $D_p = 10$ cm^2/s.

Problem 4.12 A GaAs LED has a doping profile of $N_a = 10^{17}$ cm^{-3}, $N_d = 10^{18}$ cm^{-3} at 300 K. The minority carrier time is $\tau_n = 10^{-8}$ s; $\tau_p = 5 \times 10^{-9}$ s. The electron diffusion coefficient is 100 cm^2 s^{-1} while that of the holes is 20 cm^2 s^{-1}. Calculate the ratio of the electron-injected current (across the junction) to the total current.

Problem 4.13 The diode of problem 4.12 has an area of 1 mm^2 and is operated at a forward bias of 1.2 V. Assume that 50% of the minority carriers injected recombine with the majority charge to produce photons. Calculate the rate of the photon generation in the n- and p-side of the diode.

Problem 4.14 Consider a GaAs p-n diode with a doping profile of $N_a = 10^{16}$ cm^{-3}, $N_d = 10^{17}$ cm^{-3} at 300 K. The minority carrier lifetimes are $\tau_n = 10^{-7}$ s; $\tau_p = 10^{-8}$ s. The electron and hole diffusion coefficients are 150 cm^2/s and 24 cm^2/s, respectively. Calculate and plot the minority carrier density in the quasi-neutral n and p regions at a forward bias of 1.0 V.

Problem 4.15 Consider a p-n diode made from InAs at 300 K. The doping is $N_a = 10^{16}$ cm$^{-3} = N_d$. Calculate the saturation current density if the electron and hole density of states masses are $0.02m_o$ and $0.4m_o$, respectively. Compare this value with that of a silicon p-n diode doped at the same levels. The diffusion coefficients are $D_n = 800$ cm^2/s; $D_p = 30$ cm^2/s. The carrier lifetimes are $\tau_n = \tau_p = 10^{-8}$s for InAs. For the silicon diode use the values $D_n = 30$ cm^2/s; $D_p = 10$ cm^2/s; $\tau_n = \tau_p = 10^{-7}$s.

Problem 4.16 Consider a p-n diode in which the doping is linearly graded. The doping is given by

$$N_d - N_a = Gx$$

so that the doping is p-type at $x < 0$ and n-type at $x > 0$. Show that the electric field profile is given by

$$\mathcal{E}(x) = \frac{e}{2\epsilon} G \left[x^2 - \left(\frac{W}{2} \right)^2 \right]$$

where W is the depletion width, given by

$$W = \left[\frac{12\epsilon \left(V_{bi} - V \right)}{eG} \right]^{1/3}$$

Problem 4.17 A silicon diode is being used as a thermometer by operating it at a fixed forward-bias current. The voltage is then a measure of the temperature. At 300 K, the diode voltage is found to be 0.6 V. How much will the voltage change if the temperature changes by 1 K?

Problem 4.18 Compare the dark currents (i.e., reverse saturation current) in p-n diodes fabricated from GaAs, Si, Ge, and In$_{0.53}$Ga$_{0.47}$As. Assume that all the diodes are doped at $N_d = N_a = 10^{18}$ cm^{-3}. The material parameters are (300 K):

$$\begin{array}{rcl}
\text{GaAs} & : & \tau_n = \tau_p = 10^{-8}\ \text{s}; D_n = 100\ \text{cm}^2/\text{s}; D_p = 20\ \text{cm}^2/\text{s} \\
\text{Si} & : & \tau_n = \tau_p = 10^{-7}\ \text{s}; D_n = 30\ \text{cm}^2/\text{s}; D_p\ = 15\ \text{cm}^2/\text{s} \\
\text{Ge} & : & \tau_n = \tau_p = 10^{-7}\ \text{s}; D_n = 50\ \text{cm}^2/\text{s}; D_p\ = 30\ \text{cm}^2/\text{s}
\end{array}$$

When p-n diodes are used as light detectors, the dark current is a noise source.

Problem 4.19 When we derived the law of the junction, we assumed that the electron and hole quasi-fermi levels were constant across the depletion region. Inherent in this assumption is another assumption, that the electron and hole mobilities are high enough that most reasonable current densities can be provided by a minimum change in the quasi-fermi level across the depletion edge, or ΔE_{Fn} is small. $J_n = q\mu_n n \frac{\Delta E_{Fn}}{\Delta x}$ What if this were not true and I had a p-n junction made of a semiconductor where the hole mobility was very low? Assuming no recombination in the junction calculate and plot the hole concentration at the edge of the depletion region as a fuction of bias for $\mu_p = 10\frac{cm^2}{Vs}$ and compare to the the value obtained from the law of the junction. State your assumptions.

Problem 4.20 Consider the GaAs diode shown in figure 4.36, where the n-type region has a small width $W_N << L_P$ while the p-region is thick.

1. Plot the minority and majority carrier and currents distributions in the n and p regions of this diode.
2. Now the diode is illuminated leading to an optical generation of $10^{20} cm^{-3}s^{-1}$. Plot the carrier distributions and currents in the n and p regions. Calculate the current in the diode under forward bias and reverse bias voltages of 0.5 V and -1 V respectively. Both sides are doped at $10^{17} cm^{-3}$. Assume that there are ohmic contacts on both sides, and that they have infinite recombination velocities.

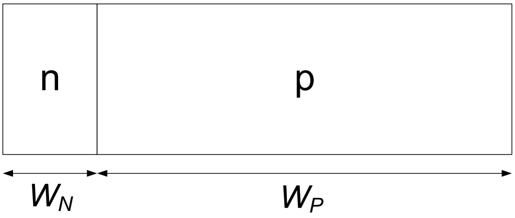


Figure 4.36: Figure for problem 4.20.

• **Section 4.4**

Problem 4.21 Consider a Si p-n diode at 300 K. Plot the I-V characteristics of the diode between a forward bias of 1.0 V and a reverse bias of 5.0 V. Consider the following cases for the impurity-assisted electron-hole recombination time in the depletion region: (a)

1.0 μs; (b) 10.0 ns; and (c) 1.0 ns. Use the following parameters:

$$\begin{aligned}
A &= 10^{-3}\,\text{cm}^2 \\
N_a &= N_d = 10^{18}\,\text{cm}^{-3} \\
\tau_n &= \tau_p = 10^{-7}\,\text{s} \\
D_n &= 25\,\text{cm}^2/\text{s} \\
D_p &= 6\,\text{cm}^2/\text{s}
\end{aligned}$$

Problem 4.22 Consider a GaAs p-n diode with $N_a = 10^{17}$ cm^{-3}, $N_d = 10^{17}$ cm^{-3}. The diode area is 10^{-3} cm^2 and the minority carrier mobilities are (at 300 K) $\mu_n = 3000$ cm^2/V·s; $\mu_p = 200$ cm^2/V·s. The electron-hole recombination times are 10^{-8}s ($\tau_p = \tau_n = \tau$). Calculate the diode current at a reverse bias of 5 V. Plot the diode forward-bias current including generation-recombination current between 0.1 V and 1.0 V.

Problem 4.23 A long base GaAs abrupt p-n junction diode has an area of 10^{-3} cm^2, $N_a = 10^{18}$ cm^{-3}, $N_d = 10^{17}$ cm^{-3}, $\tau_p = \tau_n = 10^{-8}$ s, $D_p = 6$ cm^2 s^{-1} and $D_n = 100$ cm^2 s^{-1}. Calculate the 300 K diode current at a forward bias of 0.3 V and a reverse bias of 5 V. The electron-hole recombination time in the depletion regions is 10^{-7}s.

Problem 4.24 Two different processes are used to fabricate a Si p-n diode. The first process results in a electron-hole recombination time via impurities in the depletion region of 10^{-7} s while the second one gives a time of 10^{-9} s. Calculate the diode ideality factors for the two cases near a forward bias of 0.9 V. Use the following parameters:

$$\begin{aligned}
N_a &= N_d = 10^{18}\ \text{cm}^{-3} \\
\tau_n &= \tau_p = 10^{-7}\ \text{s} \\
D_n &= 25\ \text{cm}^2/\text{s} \\
D_p &= 8\ \text{cm}^2/\text{s}
\end{aligned}$$

Problem 4.25 Consider a Si diode with the following parameters:

$$\begin{aligned}
A &= 10^{-3}\ \text{cm}^2 \\
N_a &= N_d = 10^{18}\ \text{cm}^{-3} \\
\tau_n &= \tau_p = 10^{-7}\ \text{s} \\
D_n &= 25\ \text{cm}^2/\text{s} \\
D_p &= 8\ \text{cm}^2/\text{s}
\end{aligned}$$

The length of the n- and p-sides are 1.0 μm each and the electron-hole impurity-assisted recombination time in the depletion region is 10^{-8} s. Plot the I-V relation of the diode from -5.0 V to 1.0 V. Compare the results for the case where a long diode is made from the same material technology.

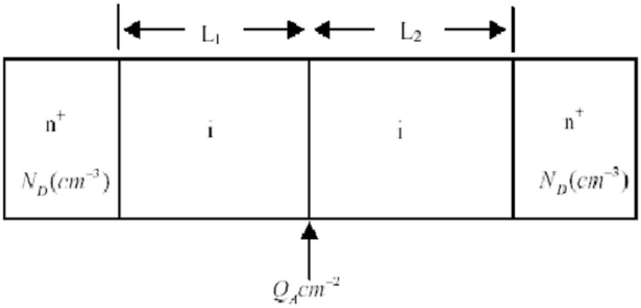


Figure 4.37: Diode for problem 4.28.

Problem 4.26 Consider a GaAS p-n junction with $N_A = N_D = 10^{17} cm^{-3}$. Assume a mid-gap trap in the material that causes the minority carrier lifetime to be $0.1ns$. Calculate and plot the electron and hole currents (including the recombination current) in the depletion region. Explain the features on the graph.

Problem 4.27 Consider a Si p-n junction biased as a solar cell. Light falls on this solar cell leading to optical generation $G_{OP} = 10^{20}/s$. What is the optically generated current(I_{OP}) for the diode? What is the open-circuit voltage(V_{OC})? Plot the minority carrier profiles when the voltage across the junction is V_{OC}, $V_{OC}/2$ and 0. Consider generation in the depletion region. Use $\tau_p = \tau_n = 10^{-6}s, \mu_n = \mu_p = 1000 cm^2/V.s$ and $N_D = N_A = 10^{17}$.

Problem 4.28 Consider the diode in figure 4.37. A sheet of acceptors of areal density Q_A is placed in an intrinsic region of GaAs such that it is at a distance L_1 from one n^+ region and L_2 from another.
(a) Calculate an expression for the potential across the structure in terms of Q_A, L_1, L_2, N_D and other material parameters of GaAs.
(b) Sketch the band diagram for the case where L_1 = 0.1 μm, L_2 = 0.2 μm, Q_A = $5 \times 10^{11} cm^{-2}$ and N_D = N_C. What will the turn-on voltage of the diode be in each direction?
(c) Calculate the maximum value of Q_A, $Q_{A,MAX}$ that gives the highest turn-on voltage in each direction/polarity.
(d) If I now set Q_A=2$Q_{A,MAX}$, what will the turn-on voltage of the diode be? Explain what happens.

Problem 4.29 Consider a p-i-n junction in AlInAs (E_g = 1.4 eV, n_i = 10^7 cm^{-2}) that is grown by MOCVD. To prevent the acceptor atoms from diffusing, the temperature of

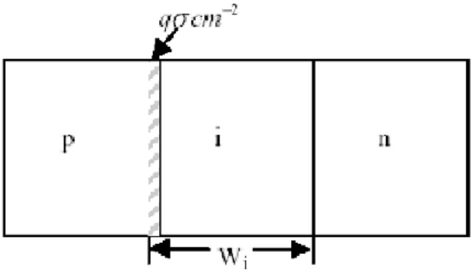


Figure 4.38: Diode for problem 4.29.

growth is dropped after growth of the p-layer is completed. The growth is then completed with an i-layer of thickness W_i and the subsequent n-layer, as shown in figure 4.19. When the capacitance of the diode is measured, it is determined that the diode is actually a p-n junction and not the p-i-n that was designed. The reason is that while waiting for the temperature to drop after growth of the p-layer, oxygen (a donor) incorporated with density $q\sigma$ cm $^{-2}$ at the p-i interface, (see figure 4.19).
(a) Derive the relation between the doping densities, W_i, and $q\sigma$ so that the measurement is explained. Assume $N_A = N_D$ for simplicity.
(b) Next, calculate a numerical value for W_i. Assume $N_A = N_D = 10^{17}$ cm^{-3} and $q\sigma$ = 5 × 10^{11} cm^{-2}.

• **Section 4.5**

Problem 4.30 The critical field for breakdown of silicon is 4 ×10^5 V/cm. Calculate the n-side doping of an abrupt p^+n diode that allows one to have a breakdown voltage of 30 V.

Problem 4.31 Consider an abrupt p^+n GaAs diode at 300 K with a doping of $N_d = 10^{16}$ cm^{-3}. Calculate the breakdown voltage. Repeat the calculation for a similarly doped p^+n diode made from diamond. Use Appendix B for the data you may need.

Problem 4.32 What is the width of the potential barrier seen by electrons during band-to-band tunneling in an applied field of 5 × 10^5 V/cm in GaAs, Si and $\mathrm{In_{0.53}Ga_{0.47}As}$ (E_g = 0.8 V)?

Problem 4.33 Consider an Si p-n diode with $N_a = 10^{18}$ cm^{-3}; $N_d = 10^{18}$ cm^{-3}. Assume that the diode will break down by Zener tunneling if the peak field reaches 10^6 V/cm. Calculate the reverse bias at which the diode will break down.

Problem 4.34 Punch through diode: For junction diodes that have to operate at high reverse biases, one needs a very thick depletion region. However, in forward-bias conditions this region is undepleted and leads to a series resistance. One uses a p^+-n-n^+

structure in such cases. The width of the n-region is smaller than the depletion region width at breakdown.

Consider two Si p^+-n-n^+ diodes with the n region having a doping of 10^{14} cm^{-3}. In one case the n-region is 150 μm long while in the other case it is 80 μm. What are the reverse-bias voltages that the diodes can tolerate before punch through occurs?

• Section 4.8

Problem 4.35 Consider a Si p-n junction where the p and n regions are much shorter than the diffusion length. Assume that the doping on both sides is $10^{17} cm^{-3}$. Use $m_p = m_n = 1000 \frac{cm^2}{Vs}, N_A = N_D = 10^{17} cm^{-3}$ and that the neutral region width on each side is $W_n = W_p = 0.1 \mu m$ Use $E_g = 1.1 eV, N_C = N_V = 10^{19} cm^{-3}$ and $t_{GEN} = 1 \mu s.$

1. What is the reverse current in the this diode when no light shines on it? Assume a large reverse bias, $qV >> kT$, and room temperature.
2. Now, light incident on the devide leads to an optical EHP generation rate $G_{OP} = 10^{22} cm^{-3} s^{-1}$. What is the reverse current in the diode? Assume room temperature, as before.
3. Now, the temperature of the diode is reduced with the light left on. At what temperature will the reverse current be equal to that calculated in the first part of the problem?

4.11 DESIGN PROBLEMS

Problem 4.1 Consider a Si long diode that must be able to operate up to a reverse bias of 10 V. The maximum electric field that the diode can tolerate anywhere within the structure is 5×10^5 V/cm. Design the diode so that the reverse current is as small as possible within the given specifications. Assume that $N_a = N_d$. What is the doping density you will use?

Problem 4.2 Consider a Si short p-n diode with the following parameters:

n-side length	=	2.0×10^{-4} cm
p-side length	=	2.0×10^{-4} cm
n-side doping	=	10^{17} cm^{-3}
p-side doping	=	10^{17} cm^{-3}
minority carrier lifetime τ_n	=	$\tau_p = 10^{-7}$ s
electron diffusion constant	=	25 cm^2/s
hole diffusion constant	=	10 cm^2/s
diode area	=	10^{-3} cm^2

Calculate the diode current (assuming that the diode is non-ideal) at a forward bias of 0.1 V and at 0.7 V at 300 K. What are the diode ideality factors near the two biasing values?

Problem 4.3 Give a short discussion on why the reverse-bias current in an ideal p-n diode has no voltage dependence. Discuss also the voltage dependence of the reverse-bias current in a non-ideal diode (i.e., a diode with defects).

Problem 4.4 Consider a Si short (or narrow) p-n diode with the following parameters:

$$\begin{aligned}
n\text{-side thickness} &= 3.0\ \mu/\text{m} \\
p\text{-side thickness} &= 4.0\ \mu/\text{m} \\
n\text{-side doping} &= 10^{18}\ \text{cm}^{-3} \\
p\text{-side doping} &= 10^{18}\ \text{cm}^{-3} \\
\text{minority carrier lifetime } \tau_n &= \tau_p = 10^{-7}\ \text{s} \\
\text{electron diffusion constant} &= 30\ \text{cm}^2/\text{s} \\
\text{hole diffusion constant} &= 10\ \text{cm}^2/\text{s} \\
\text{diode area} &= 10^{-4}\ \text{cm}^2/\text{s}
\end{aligned}$$

Calculate the diode current at a forward bias of 0.5 V at 300 K. Also calculate the total excess hole charge (in coulombs) injected into the n-side (from W_n to the diode n-side contact) at this biasing.

Problem 4.5 Consider a Si long p-n diode with the following parameters:

$$\begin{aligned}
n\text{-side doping} &= 10^{18}\ \text{cm}^{-3} \\
p\text{-side doping} &= 10^{18}\ \text{cm}^{-3} \\
\text{minority carrier lifetime } \tau_n &= \tau_p = 10^{-7}\ \text{s} \\
\text{electron diffusion constant} &= 30\ \text{cm}^2/\text{s} \\
\text{hole diffusion constant} &= 10\ \text{cm}^2/\text{s} \\
\text{diode area} &= 10^{-4}\ \text{cm}^2/\text{s}
\end{aligned}$$

Calculate the diode current at a forward bias of 1.0 V at 300 K.

An electron comes from the p-side into the depletion region and is swept away by the field to the n-side. $Estimate$ the time it takes the electron to cross the depletion region at zero applied bias and a reverse bias of 1.0 volt.

Problem 4.6 Consider a Si long p-n diode with the following parameters:

$$\begin{aligned}
n\text{-side doping} &= 10^{17}\ \text{cm}^{-3} \\
p\text{-side doping} &= 10^{17}\ \text{cm}^{-3} \\
\text{minority carrier lifetime } \tau_n &= \tau_p = 10^{-7}\ \text{s} \\
\text{electron diffusion constant} &= 30\ \text{cm}^2/\text{s} \\
\text{hole diffusion constant} &= 10\ \text{cm}^2/\text{s} \\
\text{diode area} &= 10^{-4}\ \text{cm}^2 \\
\text{carrier lifetime in the depletion region} &= 10^{-8}\ \text{s}
\end{aligned}$$

Calculate the diode current at a forward bias of 0.5 V and 0.6 V at 300 K. What is the ideality factor of the diode in this range?

Problem 4.7 Consider a narrow diode with the same parameters as given above. Calculate the total electron- and hole-injected charge in the n- and p- sides at a forward bias of 0.4 V. The widths of the n- and p-sides are both 1.0 μm.

Problem 4.8 Discuss how a p-n diode can be used as a temperature sensor. Assuming an ideal Si p-n diode, calculate the value of x and y where

$$x = \frac{1}{I_o}\frac{dI_o}{dT}, \; y = \frac{1}{I}\frac{dI}{dT}$$

In real diodes the value of x and y is smaller than what is expected for an ideal diode. Discuss the reason for this.

Problem 4.9 Assume that a Si diode suffers Zener breakdown at a field of 2×10^5 V/cm if both n- and p-sides are doped above 10^{18} cm^{-3}. Design a diode that suffers Zener breakdown at a reverse bias of 5 V. Draw the I-V characteristics for this diode assuming reasonable material parameters.

Problem 4.10 Consider a 20 μm diameter p-n diode fabricated in silicon. The donor density is 10^{16} cm^{-3} and the acceptor density is 10^{18} cm^{-3}. Calculate the following in this diode at 300 K: i) The depletion widths and the electric field profile under reverse biases of 0, 2, 5, and 10 V, and under a forward bias of 0.5 V. ii) What are the charges in the depletion region for these biases?

Problem 4.11 Consider the diode discussed in design problem 4.10. Calculate the average field in the depletion region at the four reverse-bias values considered. Calculate the velocity of the electrons at these average fields using the velocity-field results given in chapter 3 What can be said about the change in the drift components of the diode current with the change in bias?

Problem 4.12 Consider an ideal diode model for a silicon p-n diode with $N_d = 10^{16}$ cm^{-3} and $N_a = 10^{18}$ cm^{-3}. The diode area is 10^{-3} cm^2.

The transport properties of the diode are given by the following values at 300 K:

$$n - \text{side}\left\{\begin{array}{ll} \mu_p = 300\ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1}; & \mu_n = 1300\ \ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1} \\ D_p = 7.8\ \text{cm}^2\ \text{s}^{-1}; & D_n = 33\ \text{cm}^2\ \text{s}^{-1} \end{array}\right.$$

$$p - \text{side}\left\{\begin{array}{ll} \mu_p = 100\ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1}; & \mu_n = 280\ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1} \\ D_p = 2.6\ \text{cm}^2\ \text{s}^{-1}; & D_n = 7.3\ \text{cm}^2\ \text{s}^{-1} \end{array}\right.$$

(Note that the mobility is a lot lower in the heavily doped p-side because of the increased ionized impurity scattering.) Assume that $\tau_n = \tau_p = 10^{-6}$ s. Calculate the diode current.

Problem 4.13 We will define a p-n diode to be "turned on" when the current density reaches 10^3 A/cm^2 (this is an approximate criterion). Calculate the turn-on or cut-in voltage for a GaAs and a Si p-n diode with following parameters (same for both diodes):

$$\begin{aligned} N_d &= N_a = 10^{17}\ \mathrm{cm}^{-3} \\ \tau_n &= \tau_p = 10^{-8}\ \mathrm{s} \end{aligned}$$

Use table 3.1 to determine diffusion coefficients. Assume that the diodes are long and T = 300 K.

Problem 4.14 An important use of a forward-biased p-n diode is as an emitter in a bipolar transistor. In the emitter it is desirable that the current be injected via only one kind of charge. The diode efficiency is thus defined as (J_n is the current density carried by electron injection into the p-side)

$$\gamma_{inj} = \frac{J_n}{J_{Tot}} = \frac{1}{1 + J_p/J_n}$$

Consider a GaAs p-n diode with the following parameters:

Electron diffusion coefficient,	D_n	$=$	$30\ \mathrm{cm}^2/\mathrm{s}$
Hole diffusion coefficient,	D_p	$=$	$15\ \mathrm{cm}^2/\mathrm{s}$
p-side doping,	N_a	$=$	$5 \times 10^{16}\ \mathrm{cm}^{-3}$
n-side doping,	N_d	$=$	$5 \times 10^{17}\ \mathrm{cm}^{-3}$
Electron minority carrier time,	τ_n	$=$	$10^{-8}\ \mathrm{s}$
Hole minority carrier time,	τ_p	$=$	$10^{-7}\ \mathrm{s}$

Calculate the diode injection efficiency (this is called the emitter efficiency in a bipolar transistor).

Problem 4.15 Consider the p-n diode in problem 4.12. In that problem we examined the prefactor of the diode current using the long diode conditions. Calculate the prefactor for the case of a short diode in which both the n- and p-side widths are 5.0 μm.

4.12 FURTHER READING

- **General**
  - M. S. Tyagi, Introduction to Semiconductor Materials and Devices (John Wiley and Sons, New York, 1991).
  - B. G. Streetman and S. Banerjee, Solid State Electronic Devices (Prentice-Hall, Englewood Cliffs, NJ, 1999).
  - G. W. Neudeck, "Modular Series on Solid State Devices," Vol. 11, The P-N Junction Diode, (Addison-Wesley, Reading, MA, 1983).

- R. S. Muller and T. I. Kamins, Device Electronics for Integrated Circuits (John Wiley and Sons, New York, 1986).

- **Diode Breakdown**
  - M. H. Lee and S. M. Sze, "Orientation Dependence of Breakdown Voltage in GaAs," Solid State Electronics **23**, 1007 (1980).
  - S. M. Sze, Physics of Semiconductor Devices (John Wiley and Sons, New York, 1981).
  - S. M. Sze and G. Gibbons, Applied Physics Letters **8**, 112 (1986).

- **Temporal Response of Diodes**
  - R. H. Kingston, "Switching Time in Junction Diodes and Junction Transistors," Proc. IRE **42**, 829 (1954).
  - M. S. Tyagi, Introduction to Semiconductor Materials and Devices (John Wiley and Sons, New York, 1991).
  - D. A. Neamen, Semiconductor Physics and Devices: Basic Principles (Irwin, Homewood, IL, 1992).

Chapter 5

SEMICONDUCTOR JUNCTIONS

5.1 INTRODUCTION

The discussions in chapter 4 suggest that when two different materials form a junction (e.g. $n-$type and $p-$type semiconductors) interacting electrical effects arise. We have seen how the p-n diode has nonlinear I-V characteristics and tunable C-V characteristics. We can form junctions between metals and semiconductors, between semiconductors with different gaps etc. These junctions also have special properties useful for devices. Metals by themselves are necessary to connect the semiconductors to the “outside world” of voltage sources and circuits. They are also able to produce rectifying junctions. Insulators are also an integral part of electronics. These materials provide an isolation between two regions of a device, can be used for bandstructure tailoring, can be used as capacitors, etc. In this chapter we will examine some important properties of a variety of junctions.

5.2 METAL INTERCONNECTS

Metals form an important part of semiconductor technology. As shown in figure 5.1, they are used as interconnects (i.e. low resistance conductors), they form Schottky barriers and Ohmic contacts, and they form gates in field effect transistors. We have discussed in section 2.7 that due to the high density of mobile electrons, the resistivity of metals is very low. In table 5.1 we show the resistivities of some important metals used in electronics. In semiconductor circuits, interconnects provide pathways through which charge travels from one point to another. While these interconnects are obviously passive elements of the circuit they are extremely important and play a role in circuit performance. The metal strips making up the interconnect must be able to carry adequate current and make good contact with the devices. Interconnects are deposited on insulators and touch the active devices only through windows that are opened at select points. Aluminum is a commonly used interconnect material. In bulk, Al is a good conductor, with resistivity of 2.7×10^{-6} $\Omega - \mathrm{cm}$. In thin-film form, the resistivity can be up to a factor of 20

MATERIAL	RESISTIVITY (μΩ-cm)
Aluminum (Al)	
Bulk	2.7
Thin Film	0.2-0.3
Alloys, Δρ	
per %Si	+0.7%Si
per %Cu	+0.3%Cu
Titanium (Ti)	40.0
Tungsten (W)	5.6
Ti-W	15-50
Gold (Au)	2.44
Silver (Ag)	1.59
Copper (Cu)	1.77
Platinum (Pt)	10.0
Silicides	
PtSi	28-35
NiS_2	50

WORK FUNCTIONS OF SOME METALS

Element	Work function, ϕ_m (volt)
Ag, silver	4.26
Al, aluminum	4.28
Au, gold	5.1
Cr, chromium	4.5
Mo, molybdenum	4.6
Ni, nickel	5.15
Pd, palladium	5.12
Pt, platinum	5.65
Ti, titanium	4.33
W, tungsten	4.55

ELECTRON AFFINITY OF SOME SEMICONDUCTORS

Element	Electron affinity, χ (volt)
Ge, germanium	4.13
Si, silicon	4.01
GaAs, gallium arsenide	4.07
AlAs, aluminum arsenide	3.5

Table 5.1: Resistivities of some metals used in solid state electronics

lower, allowing the thin interconnect film to carry very high current densities, of the order of ($\sim 10^6$ Acm^{-2}).

Example 5.1 In this example we will study some important concepts in thin-film resistors, which form an important part of semiconductor device technology. The resistors are often made from polysilicon that is appropriately doped. In thin-film technology it is usual to define sheet resistance instead of the resistance of the material. Consider, as shown in figure 5.1b, a material of length L, width W, and depth D. The resistance of the material is

$$R = \frac{\rho L}{WD} = \frac{\rho L}{A} \tag{5.2.1}$$

As we have discussed in chapter 3, the resistivity ρ is given by

$$\rho = \frac{1}{ne\mu} \tag{5.2.2}$$

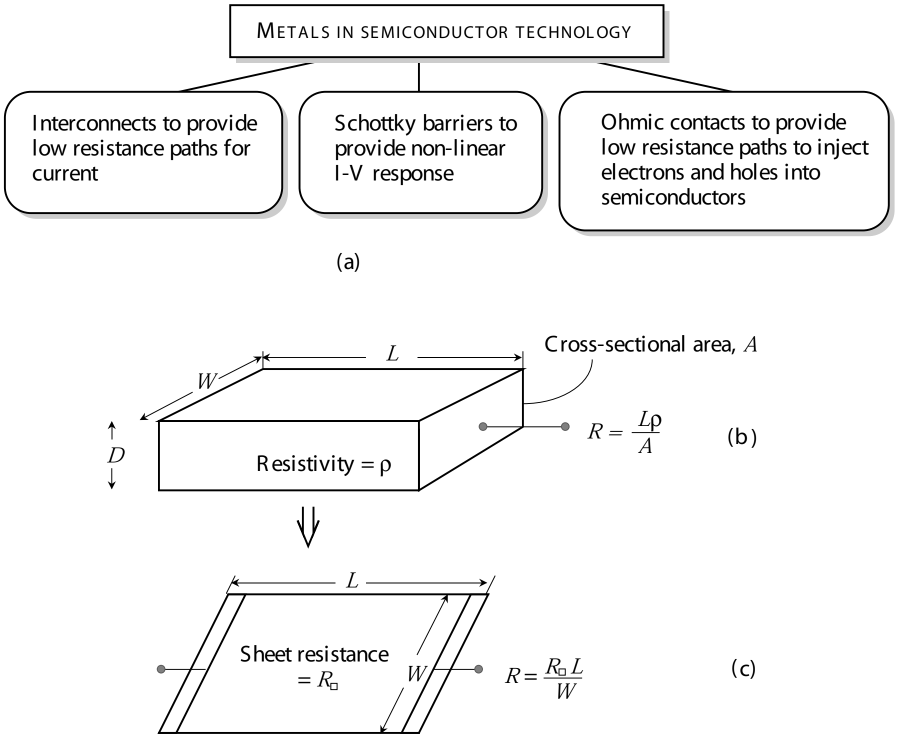

METALS IN SEMICONDUCTOR TECHNOLOGY
Interconnects to provide low resistance paths for current
Schottky barriers to provide non-linear I-V response
Ohmic contacts to provide low resistance paths to inject electrons and holes into semiconductors
(a)
L
W
D
Cross-sectional area, A
Resistivity = ρ
R = Lρ/A
(b)
L
Sheet resistance = R□
W
R = R□ L/W
(c)

Figure 5.1: (a) Metals serve three important functions in semiconductor technology. (b) A resistor of dimensions $L \times W \times D$. (c) Representation of the resistors in terms of sheet resistance.

where n is the free carrier density and μ is the mobility of the carriers (the equation can be modified for a p-type material).

The sheet resistance is a measure of the characteristics of a uniform sheet of film. It is defined as ohms per square, as shown in figure 5.1c, and is related to the film resistance by

$$R_{\square} = R\frac{W}{L} \tag{5.2.3}$$

5.3 METAL SEMICONDUCTOR JUNCTION: SCHOTTKY BARRIER

The metal-semiconductor junction can result in a junction that has non-linear diode characteristics similar to those of the p-n diode except that for many applications it has a much faster response since carrier transport is unipolar. Such a junction is called a Schottky barrier diode.

5.3.1 Schottky Barrier Height

The working of the Schottky diode depends upon how the metal-semiconductor junction behaves in response to external bias. Let us pursue the approximation we used for the p-n junction and examine the band profile of a metal and a semiconductor. A metal semiconductor structure is shown in figure 5.2a. In figure 5.2b and figure 5.2c the band profiles of a metal and a semiconductor are shown. Figure 5.2b shows that the band profile and Fermi level positions when the metal is away from the semiconductor. In figure 5.2c the metal and the semiconductor are in contact. The Fermi level E_{Fm} in the metal lies in the band, as shown. Also shown is the work function $e\phi_m$. In the semiconductor, we show the vacuum level along with the position of the Fermi level E_{Fs} in the semiconductor, the electron affinity, and the work function.

We will assume an ideal surface for the semiconductor in the first calculation. Later we will examine the effect of surface defects. We will assume that $\phi_m > \phi_s$ so that the Fermi level in the metal is at a lower position than in the semiconductor. This condition leads to an $n-$type Schottky barrier. When the junction between the two systems is formed, the Fermi levels should line up at the junction and remain flat in the absence of any current, as shown in figure 5.2c. At the junction, the vacuum energy levels of the metal side and semiconductor side must be the same. To ensure the continuity of the vacuum level and align the Fermi levels. Electrons move out from the semiconductor side to the metal side. Note that since the metal side has an enormous electron density, the metal Fermi level or the band profile does not change when a small fraction of electrons are added or taken out. As electrons move to the metal side, they leave behind positively charged fixed dopants, and a dipole region is produced in the same way as for the p-n diode.

In the ideal Schottky barrier with no bandgap defect levels, the height of the barrier at the semiconductor-metal junction (figure 5.2c), is defined as the difference between the semiconductor conduction band at the junction and the metal Fermi level. This barrier is given by (see figure 5.2c)

$$e\phi_b = e\phi_m - e\chi_s \tag{5.3.1}$$

The electrons coming from the semiconductor into the metal face a barrier denoted by eV_{bi} as shown in figure 5.2c. The potential eV_{bi} is called the built-in potential of the junction and is given by

$$eV_{bi} = -(e\phi_m - e\phi_s) \tag{5.3.2}$$

It is possible to have a barrier for hole transport if $\phi_m < \phi_s$. In figure 5.3 we show the case of a metal-p-type semiconductor junction where we choose a metal so that $\phi_m < \phi_s$. In this case, at equilibrium the electrons are injected from the metal to the semiconductor, causing a negative

Schottky Metal	*n* Si	*p* Si	*n* GaAs
Aluminum, Al	0.7	0.8	
Titanium, Ti	0.5	0.61	
Tungsten, W	0.67		
Gold, Au	0.79	0.25	0.9
Silver, Ag			0.88
Platinum, Pt			0.86
PtSi	0.85	0.2	
NiSi$_2$	0.7	0.45	

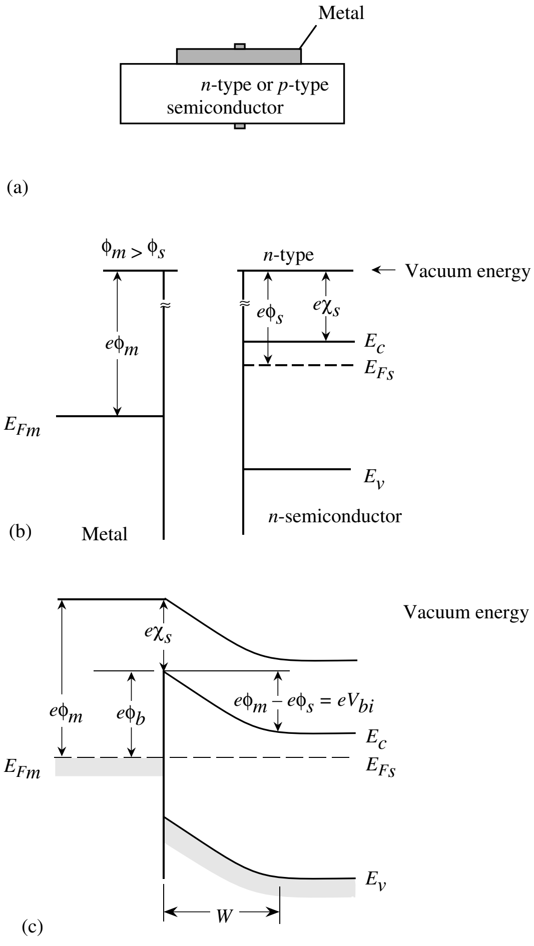
Table 5.2: Schottky barrier heights (in volts) for several metals on n- and p-type semiconductors.

charge on the semiconductor side. The bands are bent once again and a barrier is created for hole transport. The height of the barrier seen by the holes in the semiconductor is

$$eV_{bi} = e\phi_s - e\phi_m \tag{5.3.3}$$

The Schottky barrier height for n- or p-type semiconductors depends upon the metal and the semiconductor properties. This is true for an ideal case. It is found experimentally that the Schottky barrier height is often independent of the metal employed, as can be seen from table 5.2 This can be understood qualitatively in terms of a model based upon non ideal surfaces. In this model the metal-semiconductor interface has a distribution of interface states that may arise from the presence of chemical defects from exposure to air or broken bonds, etc. We have seen in chapter 3 that defects can create bandgap states in a semiconductor. Surface defects can create $\sim 10^{13}$ cm^{-2}defects if there is 1 in 10 defects at the surface. Surface defects lead to a distribution of electronic levels in the bandgap at the interface, as shown in figure 5.4. The distribution may be characterized by a neutral level ϕ_o having the property that states below it are neutral if filled and above it are neutral if empty. If the density of bandgap states near ϕ_o is very large, then addition or depletion of electrons to the semiconductor can not alter the Fermi level position at the surface without large changes in surface charges (beyond the numbers demanded by charge neutrality considerations). Thus, the Fermi level is said to be pinned. In this case, as shown in figure 5.4, the Schottky barrier height is

$$e\phi_b = E_g - e\phi_o \tag{5.3.4}$$



METAL-SEMICONDUCTOR JUNCTION AT EQUILIBRIUM

Figure 5.2: (a) A schematic of a metal-semiconductor junction. (b) The various important energy levels in the metal and the semiconductor with respect to the vacuum level. (c) The junction potential produced when the metal and semiconductor are brought together. Due to the built-in potential at the junction, a depletion region of width W is created.

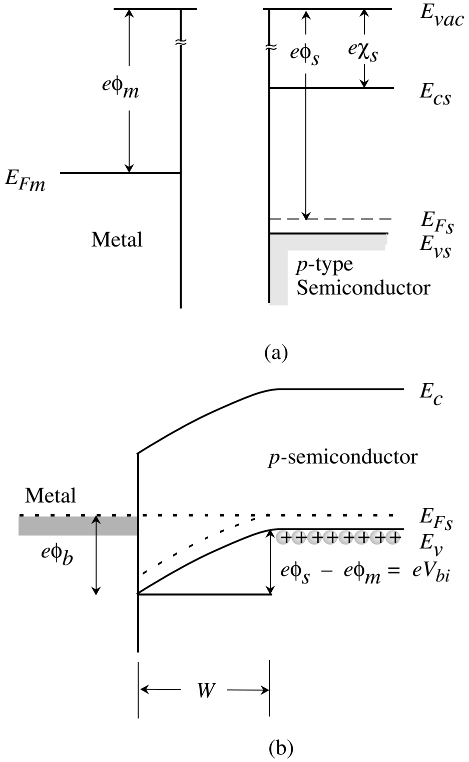


Figure 5.3: A schematic of the ideal p-type Schottky barrier formation. (a) The positions of the energy levels in the metal and the semiconductor; (b) the junction potential and the depletion width.

and is almost independent of the metal used. The model discussed above provides a qualitative understanding of the Schottky barrier heights. However, the detailed mechanism of the interface state formation and Fermi level pinning is quite complex. In table 5.2 we show Schottky barrier heights for some common metal-semiconductor combinations. In some materials such as GaN and AlGaN,the surface retains its ideal behavior and the Schottky barrier is indeed controlled by the metal work function.

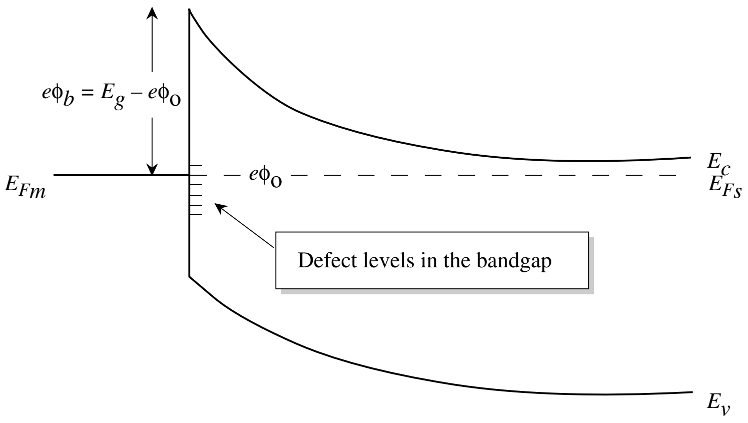


Figure 5.4: Interface states at a real metal-semiconductor interface. A neutral level ϕ_o is defined so that the interface states above ϕ_o are neutral if they are empty and those below ϕ_o.

5.3.2 Capacitance Voltage Characteristics

Once the Schottky barrier height is known, the electric field profile, depletion width, depletion capacitance, etc., can be evaluated the same way we obtained the values for the p-n junction. The problem for a Schottky barrier on an n-type material is identical to that for the abrupt p^+n diode, since there is no depletion on the metal side. One again makes the depletion approximation; i.e., there is no mobile charge in the depletion region and the semiconductor is neutral outside the depletion region. Then the solution of the Poisson equation gives the depletion width W for an external voltage applied to the metal V

$$W = \left[\frac{2\epsilon(V_{bi} - V)}{eN_d}\right]^{1/2} \tag{5.3.5}$$

Here N_d is the doping of the n-type semiconductor. Note that there is no depletion on the metal side because of the high electron density there. The potential V is the applied potential, which is positive for forward bias and negative for reverse bias.

5.3.3 Current Flow across a Schottky Barrier: Thermionic Emission

Consider the Schottky barrier band diagram shown on figure 5.5 at zero bias.

The Schottky barrier between a metal and semiconductor is shown in equilibrium (at zero bias) with the electron distribution shown on the right

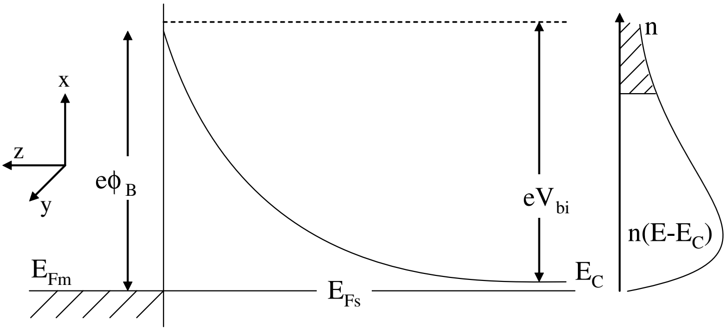


Figure 5.5: Schottky Barrier in equilibrium

Also shown is the electron distribution:

$$n\left(E-E_C\right)=2f\left(E-E_C\right)\cdot N\left(E-E_C\right) \tag{5.3.6}$$

similar to the case of a $p-n$ junction, the factor of 2 in accounting for electron spin. Thermionic emission assumes that all electrons in the semiconductor with kinetic energy in the $+z$ direction greater than eV_{bi} ($E_z > eV_{bi}$) and $k_z > 0$, are capable of surmounting the barrier and contributing to current flow from the semiconductor to the metal, $J_{s\rightarrow m}$. Note that the total kinetic energy $E-E_C = E_x + E_y + E_z$. At thermal equilibrium the current from the metal to the semiconductor, $J_{m\rightarrow s}$, will be equal in magnitude and opposite in sign to $J_{s\rightarrow m}$, making the net current zero. To calculate $J_{s\rightarrow m}$ one needs to sum the current carried by every allowed electron:

$$J_{s\rightarrow m}=e\sum n\left(E-E_c\right)\cdot v_z \tag{5.3.7}$$

for $E_z > eV_{bi}$ and $v_z > 0$. The methodology employed is to calculate the number of electrons at energy E in a volume of k-space $(dk)^3$, multiply the number with the electron velocity in the direction along the barrier, and sum or integrate over energy. Assuming a crystal of length L, periodic boundary conditions yield allowed k values given by

$$k=2\pi N \tag{5.3.8}$$

where N is an integer and the separation between allowed k's is $\Delta k = 2\pi/L$. The number of electrons in a volume element dk_x, dk_y, dk_z is therefore

$$dN=2f\left(E-E_C\right)\frac{dk_x dk_y dk_z}{\Delta k^3} \tag{5.3.9}$$

Assuming $(E-E_C) \gg E_F$ and writing $E-E_F = E-E_C+E_C-E_F$ gives

$$dN=2\ \exp\left(\frac{-((E-E_C)+(E_C-E_F))}{k_BT}\right)\frac{dk_x dk_y dk_z}{\Delta k^3} \tag{5.3.10}$$

The current density contributed by these electrons is

$$J_z = -ev_z \frac{dN}{L^3} \tag{5.3.11}$$

if $k_z > 0$ and $E_z > eV_{bi}$. Note that all values of E_x and E_y are allowed as they represent motion in the $x - y$ plane which is not constrained by the barrier in the $+z$ direction. Note that

$$(E_x - E_C) = \frac{\hbar^2 k_x^2}{2m^*} \tag{5.3.12}$$

with similar relationships for $(E_y - E_C)$ and $(E_z - E_C)$. Also employing the condition $(E_z - E_c) > eV_{bi}$ yields a minimum value of

$$k_{min} = \sqrt{eV_{bi}\left(\frac{2m^*}{\hbar^2}\right)} \tag{5.3.13}$$

Also,

$$v_z = \frac{\hbar k_z}{m^*} \tag{5.3.14}$$

Therefore,

$$J_z = \frac{-e}{(2\pi)^3} \int_{-\infty}^{+\infty} dk_x \int_{-\infty}^{+\infty} dk_y \int_{k_{min}}^{+\infty} \frac{\hbar k_z}{m^*} dk_z \cdot$$

$$2\exp\left[-\left(E_x + E_y + E_z\right)/k_B T\right] \cdot \exp\left[-\left(E_C - E_F\right)/k_B T\right] \exp\left(\frac{E_C}{k_B T}\right)$$

$$= -\frac{2e}{(2\pi)^3} \int_x \cdot \int_y \cdot \int_z \exp\left(-\frac{E_C - E_F}{k_B T}\right) \tag{5.3.15}$$

where

$$\int_x = \int_y = \int_{-\infty}^{\infty} \exp\left(\frac{\hbar^2 k_x^2}{2m^* k_B T}\right) dk_x = \frac{\sqrt{2\pi m^* k_B T}}{\hbar} \tag{5.3.16}$$

and

$$\int_z = \int_{k_{min}}^{\infty} \exp\left(-\frac{\hbar^2 k_z^2}{k_B T}\right) \cdot \frac{\hbar k_z}{m^*} \cdot dk_z \tag{5.3.17}$$

$$= \frac{k_B T}{\hbar} \exp\left(-\hbar^2 k_{min}^2 / k_B T\right) = \frac{k_B T}{\hbar} \exp\left(\frac{-eV_{bi}}{k_B T}\right) \tag{5.3.18}$$

Therefore,

$$J_z = \frac{4\pi}{(2\pi\hbar)^3} \cdot em^* k_B^2 T^2 \exp\left(-\frac{(eV_{bi} + (E_C - E_F))}{k_B T}\right) \tag{5.3.19}$$

or

$$J_z = A^* \cdot T^2 \exp\left(\frac{-e\phi_B}{k_B T}\right) = J_{s\to m}\left(V = 0\right) \tag{5.3.20}$$

where

$$A^* = \frac{4\pi e m^* k_B^2}{2\pi\hbar^3} = 120\ A\ cm^{-2}\ K^{-2} \times \frac{m^*}{m_0} \tag{5.3.21}$$

is the Richardson constant and $\phi_B = V_{bi} + (E_C - E_F)$, the barrier seen by electrons in the metal of the Schottky barrier height. We have calculated $J_{s\rightarrow m}$ at $V = 0$. The analysis can be easily extended to a forward bias of V_F, the only change being replacing the barrier, V_{bi} by the new barrier $V_{bi} - V_F$. This changes I_z to

$$I_z = \frac{k_B T}{\hbar} \exp\left(-\frac{eV_{bi}}{k_B T}\right) \cdot \exp\left(\frac{eV_F}{k_B T}\right) \tag{5.3.22}$$

or

$$J_{s\rightarrow m}\left(V = V_F\right) = J_{s\rightarrow m}\left(V = 0\right) \cdot \exp\left(\frac{eV_F}{k_B T}\right) \tag{5.3.23}$$

Since the current flow from the metal to the semiconductor is unchanged:

$$J\left(V = V_F\right) = J_{s\rightarrow m}\left(V = V_F\right) - J_{m\rightarrow s}\left(V = V_F\right) \tag{5.3.24}$$

$$= A^* T^2 \exp\left(\frac{-q\phi_B}{k_B T}\right)\left[\exp\left(\frac{eV_F}{k_B T}\right) - 1\right] \tag{5.3.25}$$

Example 5.2 In a W-n-type Si Schottky barrier the semiconductor has a doping of 10^{16} cm^{-3} and an area of 10^{-3} cm^2.
(a) Calculate the 300 K diode current at a forward bias of 0.3 V.
(b) Consider an Si $p^+ - n$ junction diode with the same area with doping of $N_a = 10^{19}$ cm^{-3} and $N_d = 10^{16}$ cm^{-3}, and $\tau_p = \tau_n = 10^{-6}$ s. At what forward bias will the p-n diode have the same current as the Schottky diode? $D_p = 10.5$ cm^2/s.

From table 5.2 the Schottky barrier of W on Si is 0.67 V. Using an effective Richardson constant of 110 A cm$^{-2}K^{-1}$, we get for the reverse saturation current

$$\begin{aligned} I_s &= (10^{-3}\ \text{cm}^2) \times (110\ \text{A cm}^{-2}K^{-2}) \times (300K)^2 \exp\left(\frac{-0.67(\text{eV})}{0.026(\text{eV})}\right) \\ &= 6.37 \times 10^{-8}\ \text{A} \end{aligned}$$

For a forward bias of 0.3 V, the current becomes (neglecting 1 in comparison to exp (0.3/0.026))

$$\begin{aligned} I &= 6.37 \times 10^{-8} A\ \exp(0.3/0.026) \\ &= 6.53 \times 10^{-3}\ \text{A} \end{aligned}$$

In the case of the p-n diode, we need to know the appropriate diffusion coefficients and lengths. The diffusion coefficient is 10.5 cm^2/s, and using a value of $\tau_p = 10^{-6}$s we get $L_p = 3.24 \times 10^{-3}$ cm. Using the results for the abrupt $p^+ - n$ junction, we get for the saturation current ($p_n = 2.2 \times 10^4$ cm^{-3}) (note that the saturation current is essentially due to hole injection into the n-side for a p^+-n diode)

$$\begin{aligned} I_o &= (10^{-3}\ \mathrm{cm}^2) \times (1.6 \times 10^{-19}\ \mathrm{C}) \times \frac{(10.5\ \mathrm{cm}^2/\mathrm{s}^{-1})}{(3.24 \times 10^{-3}\ \mathrm{cm})} \times (2.25 \times 10^4\ \mathrm{cm}^{-3}) \\ &= 1.17 \times 10^{-14}\ \mathrm{A} \end{aligned}$$

This is an extremely small value of the current. At 0.3 V, the diode current becomes

$$I = I_s \exp\left(\frac{eV}{k_BT}\right) = 1.2 \times 10^{-9}\ \mathrm{A}$$

a value which is almost six orders of magnitude smaller than the value in the Schottky diode. For the p-n diode to have the same current that the Schottky diode has at 0.3 V, the voltage required is 0.71 V.

This example highlights the important differences between Schottky and junction diodes. The Schottky diode turns on (i.e., the current is ∼1 mA) at 0.3 V while the p-n diode turns on at closer to 0.7 V.

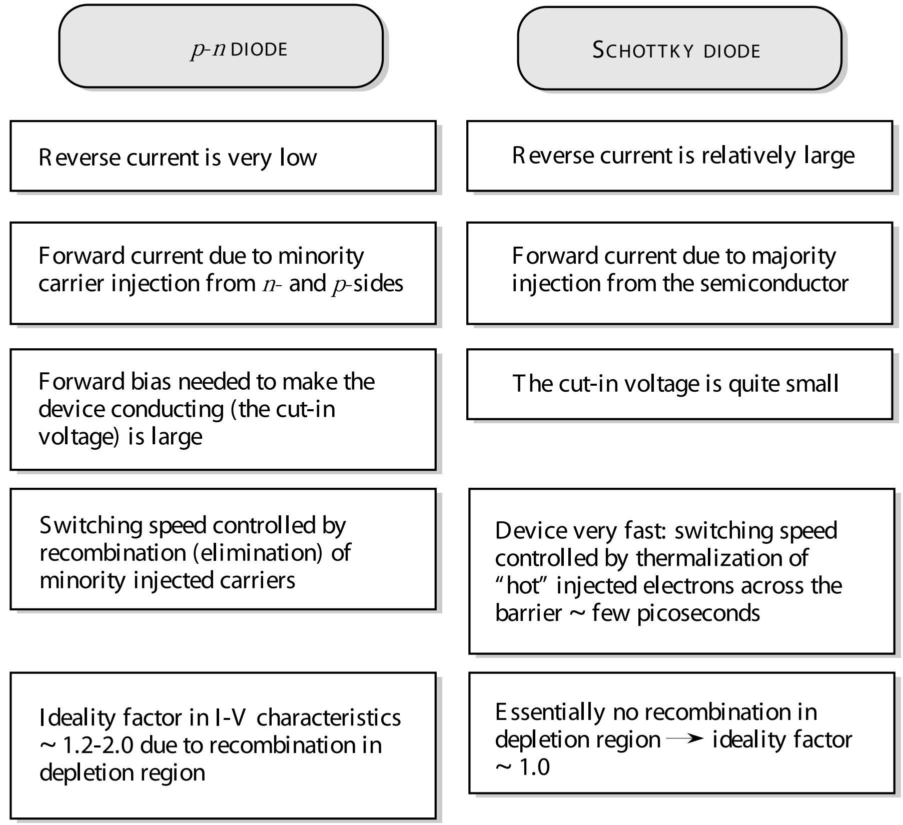
5.3.4 Comparison of Schottky and p-n diodes

Both the $p - n$ diode and the Schottky diode can be used for rectification and non-linear $I - V$ response. One may ask which provides superior performance. The answer depends upon specific applications. The questions of turn on voltage, speed needed, reverse leakage, etc. are important in deciding whether a $p - n$ diode or Schottky diode should be used. The Schottky diodes have a number of important advantages over $p - n$ diodes. Some of these are listed in figure 5.6. The temperature dependence of the Schottky barrier current is quite weak compared to that of a p-n diode. This is because in a p-n diode, the currents are controlled by the diffusion current of minority carriers, which in turn depends on minority carrier concentration that has a rather strong temperature dependence.

The fact that the Schottky barrier is a majority carrier device gives it a tremendous advantage over p-n diodes in terms of the device speed. Device speed is no longer dependent upon extracting minority charge via diffusion or recombination. By making small devices, the RC time constant of a Schottky barrier can approach a few picoseconds, which is orders of magnitude faster than that of p-n diodes.

Another important advantage of the Schottky diode is the fact that there is essentially no recombination in the depletion region and the ideality factor is very close to unity. In p-n diodes, there is significant recombination in the depletion region and ideality factors range from 1.2 to 2.0.

The main disadvantage of Schottky diodes is a higher reverse current density. The thermionic-emission-controlled prefactor gives a current density in the range of $\sim 10^{-7}$ Acm^{-2}, which is three to four orders of magnitude higher than that of the $p - n$ diode. Thus for a given applied bias, the Schottky barrier has much higher current than the $p - n$ diode. As a result the Schottky diode is preferred as a low-voltage high-current rectifier. Since, the reverse current in a Schottky



p-n DIODE	SCHOTTKY DIODE
Reverse current is very low	Reverse current is relatively large
Forward current due to minority carrier injection from *n*- and *p*-sides	Forward current due to majority injection from the semiconductor
Forward bias needed to make the device conducting (the cut-in voltage) is large	The cut-in voltage is quite small
Switching speed controlled by recombination (elimination) of minority injected carriers	Device very fast: switching speed controlled by thermalization of "hot" injected electrons across the barrier ~ few picoseconds
Ideality factor in I-V characteristics ~ 1.2-2.0 due to recombination in depletion region	Essentially no recombination in depletion region → ideality factor ~ 1.0

Figure 5.6: A comparison of some pros and cons of the p-n diode and the Schottky diode.

barrier is also quite large, which is a disadvantage for many applications. Another issue is technology related. The Schottky barrier quality depends critically on the surface quality, the processing steps are quite critical. For many semiconductors, it is not possible to have a good Schottky contact since the contact is very "leaky" due to defects. For such materials, the only way to have rectification is by using a p-n junction.

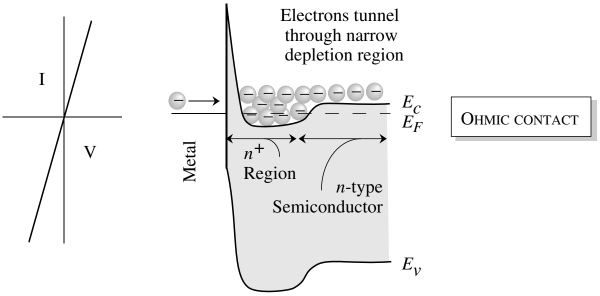


Figure 5.7: Current-voltage characteristics of an ohmic contact along with the band diagrams of metal -n^+-n contact. The heavy doping reduces the depletion width to such an extent that the electrons can tunnel through the spiked barrier easily in either direction.

5.4 METAL SEMICONDUCTOR JUNCTIONS FOR OHMIC CONTACTS

In our discussion on $p-n$ diodes and Schottky diodes we have discussed how a bias is applied across the device to cause current flow. It is important to ask how a connection is made from a power supply to the semiconductor. How do electrons or holes flow into and out of a semiconductor? There is a large barrier (the work function) that restricts the flow of electrons. We have also seen from the previous section that at least in some cases a metal-semiconductor junction also provides a barrier to flow of electrons. However, it is possible to create metal-semiconductor junctions that have a linear non-rectifying I-V characteristic, as shown in figure 5.7. Such junctions or contacts are called ohmic contacts.

There are two possibilities for creating ohmic contacts. In the previous section, to produce a Schottky barrier on an $n-$type semiconductor, we needed (for the ideal surface) a metal with a work function larger than that of the semiconductor. Thus, in principle, if we use a metal with a work function smaller than the semiconductor, one should have no built-in barrier. However, this approach is not often useful in practice because the Fermi level at the surface of real semiconductors is pinned because of the high interface density in the gap.

The Schottky barrier discussed earlier can be altered to create an ohmic contact. This is done through heavy doping and use of tunneling to get large current across the interface. Let us say we have a built-in potential barrier, V_{bi}. The depletion width on the semiconductor side is

$$W = \left[\frac{2\epsilon V_{bi}}{eN_d}\right]^{1/2} \tag{5.4.1}$$

Now if near the interface region the semiconductor is heavily doped, the depletion width could be made extremely narrow. In fact, it can be made so narrow that even though there is a potential barrier, the electrons can tunnel through the barrier with ease, as shown in figure 5.7. The quality of an ohmic contact is usually defined through the resistance R of the contact over a certain area A. The normalized resistance is called the specific contact resistance r_c and is given by

$$r_c = R \cdot A \tag{5.4.2}$$

Under conditions of heavy doping where the transport is by tunneling, the specific contact resistance has the following dependence for tunneling, probability T, through a triangular barrier):

$$\ln\ (r_c) \propto \frac{1}{\ell n(T)} \propto \frac{(V_{bi})^{3/2}}{F} \tag{5.4.3}$$

where the field is

$$\mathcal{E} = \frac{V_{bi}}{W} \ \propto\ (V_{bi})^{1/2}\,(N_d)^{1/2} \tag{5.4.4}$$

Thus,

$$\begin{aligned} \ell n\ (r_c) &\propto V_{bi} \\ &\propto \frac{1}{\sqrt{N_d}} \end{aligned} \tag{5.4.5}$$

The resistance can be reduced by using a low Schottky barrier height and doping as heavily as possible. The predicted dependence of the contact resistance on the doping density is, indeed, observed experimentally. It is observed from experiments that it is usually more difficult to obtain contacts with $p-$type semiconductors with low resistance. This is due to the difficulty in p-doping. It is also due to the fact that in many materials the relatively high effective mass of holes, leads to reduced tunneling currents. Also, in the case of many wide bandgap semiconductors such as GaN, the barrier heights between available metals and the valence band is much greater than that of the conduction band.

5.5 INSULATOR-SEMICONDUCTOR JUNCTIONS

In chapter 2 we have called materials with large bandgaps insulators. Usually these materials don't have high crystalline quality and are difficult to dope. These materials have very high resistivity and are used to isolate regions to prevent current flow. Most insulator-semiconductor combinations involve structures that are not lattice-matched. In most cases the insulator and the semiconductor do not even share the same basic lattice type. In this section we will briefly review a few such combinations. Important issues in these junctions are listed in figure 5.8. The key issues here revolve around producing an interface with very low density of trapping states and low interface leakage.

5.5.1 Insulator-Silicon

The most important junction in solid state electronics is the SiO_2-Si system. In spite of the severe mismatch between SiO_2 structure and Si structure, the interface quality is quite good.

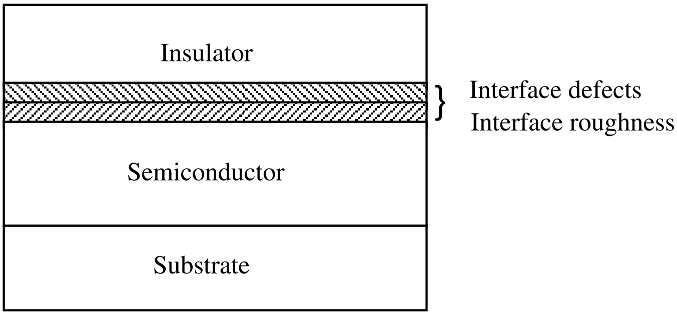


Figure 5.8: Insulator-semiconductor junctions are dominated by interface quality and defect levels in insulators.

Midgap interface density as low as 10^{10} eV^{-1} cm^{-2}can be readily obtained. The ability to produce such high-quality interfaces is responsible for the remarkable success of the metal-oxide-silicon (MOS) devices. Due to the low interface densities, there is very little trapping of electrons (holes) at the interface so that high-speed switching can be predictably used. In has to be recognized though that the interface is still rough with islands with a height of 5 Å over lateral extents of ~50 Å. Typical electron mobility in Si MOSFETs is ~600 $\mathrm{cm}^2/(\mathrm{V} \cdot \mathrm{s})$ compared to a mobility of ~1100 $\mathrm{cm}^2/(\mathrm{V} \cdot \mathrm{s})$ (300 K) for bulk pure Si. We will discuss the MOS structure in detail in chapter 9.

Silicon nitride (Si_3N_4) is another important film that forms modest-quality junctions with Si. Silicon nitride can be used in a metal-insulator-semiconductor device in Si technology, but its applications are limited. The film is used more as a mask for oxidation of the Si film. It also makes a good material for passivation of finished devices. Silicon oxy-nitride on the other hand forms high-quality interfaces with silicon and can be used in FETs.

Although not an insulator or a metal, we include polycrystalline silicon ("poly") in this chapter because of its importance in Si technology. Polysilicon can be deposited by the pyrolysis (heat-induced decomposition) of silane:

$$\mathrm{SiH_4} \longrightarrow \mathrm{Si} + 2\mathrm{H_2} \tag{5.5.1}$$

Depending upon the deposition temperature, micro crystallites of different grain sizes are produced. Typical grain size is ~ 0.1 μm.

Poly films can be doped to low resistivity to produce useful conductors for a number of applications. Poly is often used as a gate of an MOS transistor, as a resistor, or as a link between a metal and the Si substrate to ensure an ohmic contact.

5.6 SEMICONDUCTOR HETEROJUNCTIONS

A growing number of modern devices are based on semiconductor *heterojunctions*, or junctions formed between two different materials. Modern bipolar transistors employ a p-n heterojunction in order to improve the emitter injection efficiency (see chapter 7), while in HFET technology a heterojunction is used to form a high mobility channel (see chapter 8). In this section we discuss the properties of p-n heterojunctions. Specifically, we will focus on the junction formed between an n-type wide bandgap material (such as AlGaAs) and a p-type narrower bandgap material (such as GaAs).

5.6.1 Abrupt p-n heterojunction

Electrostatics

To construct a band diagram for an abrupt p-n heterojunction, we proceed in the same manner as for the p-n homojunction. We begin with two separate materials (figure 5.9a) and consider what the equilibrium conditions must be when a junction is formed between them (figure 5.9b). In figure 5.9a, the material on the left (material 1) is n-type and has a wide bandgap, while the material on the right (material 2) is p-type and has a narrower bandgap. The doping in the p-type material is much higher than that of the n-type material (this is the typical emitter-base structure in a III-V npn heterojunction bipolar transistor). The two materials have different electron affinities (χ_1 and χ_2), bandgaps (E_{g1} and E_{g2}), and dielectric constants (ϵ_1 and ϵ_2).

Figure 5.9b shows a band diagram of the system once a junction is formed between the two materials. Since the materials have different bandgaps, there must exist a discontinuity in the conduction band (ΔE_c) and/or the valence band (ΔE_v) at the interface . The difference in the bandgap between the two materials is equal to the sum of the conduction band and valence band discontinuities, or

$$\Delta E_g = E_{g1} - E_{g2} = \Delta E_c + \Delta E_v \tag{5.6.1}$$

By examination of figure 5.9, it is tempting to assume that ΔE_c is simply the difference in the electron affinities of the two materials. However, there also exist dipole charges at the heterointerface which cause a shift in the relative band discontinuities. These dipole charges result from the locally different atomic and electronic structures of the two materials at the heterointerface as compared to their bulk atomic structure. While electron affinity rules accurately predict discontinuities in a limited number of material systems in which these dipole effects are small, in most heterostructures these dipole charges are significant and must be accounted for. Band line-ups for a number of materials were shown in figure 2.31.

Similar to the case of a p-n homojunction, when the p- and n-type semiconductors are brought together, a built-in voltage, V_{bi}, is produced between the two sides of the structure. The built-in voltage is equal to the sum of the band bending on the n-side (V_{d1}) and the bend bending on the p-side (V_{d2}). By examination of figure 5.9, the built-in voltage can be shown to be

$$eV_{bi} = eV_{d1} + eV_{d2} = E_{g2} - (E_F - E_v)_p - (E_c - E_F)_n + \Delta E_c$$

where the subscripts n and p refer to the n-side and p-side of the device. Comparing this expression to that of the p-n homojunction , we see that the only difference is the additional ΔE_c

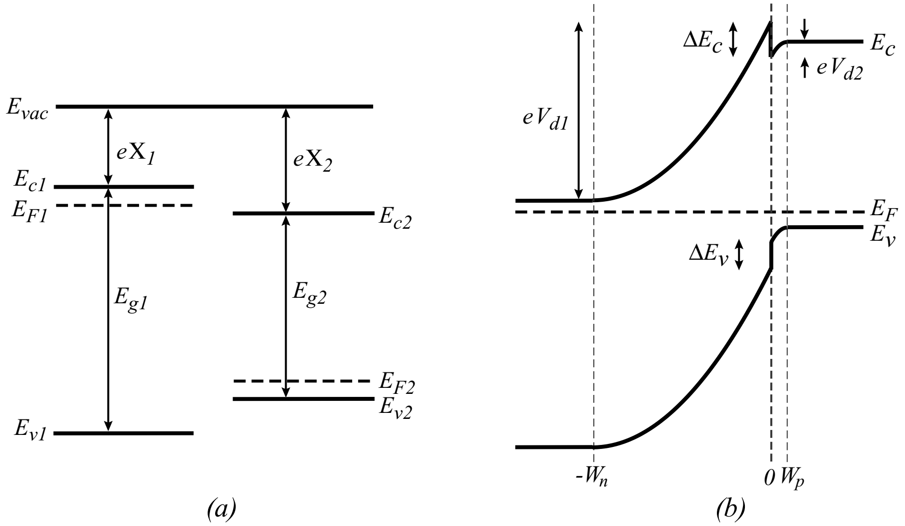


Figure 5.9: (a) Band line-ups of two distinct materials prior to the formation of a junction. (b) Band diagram of a heterojunction formed between the two materials.

term. Making the same substitutions for $(E_F - E_v)_p$ and $(E_c - E_F)_n$ as we made in the p-n homojunction case gives us the built-in potential as

$$\boxed{V_{bi} = \frac{1}{e}\left(\Delta E_c + E_{g2}\right) - \frac{k_B T}{e}\,\ell n\left[\frac{N_{c1}N_{v2}}{n_1 p_2}\right]} \tag{5.6.2}$$

where N_{c1} is the conduction band density of states in material 1, N_{v2} is the valence band density of states in material 2, n_1 = N_{d1} is the electron concentration in material 1 (assuming full ionization), and p_2 = N_{a2} is the hole concentration in material 2 (also assuming full ionization).

The depletion region width $W\,(V_{bi})$ and the electric field profile in the depletion region can be found in the same way as for a p-n homojunction, except that $\epsilon_1 \neq \epsilon_2$, so the electric field is not continuous at the material interface. The charge density and electric field profiles in the structure are shown in figure 5.10. Gauss' Law states that the displacement field $D = \epsilon\mathcal{E}$ ($\mathcal{E}$ is the electric field) must be continuous at the interface. This gives the relationship

$$\epsilon_1\mathcal{E}_{1,m} = \epsilon_2\mathcal{E}_{2,m} \tag{5.6.3}$$

where $\mathcal{E}_{1,m}$ is the maximum electric field in material 1 and $\mathcal{E}_{2,m}$ is the maximum electric field in material 2 (see figure 5.10).

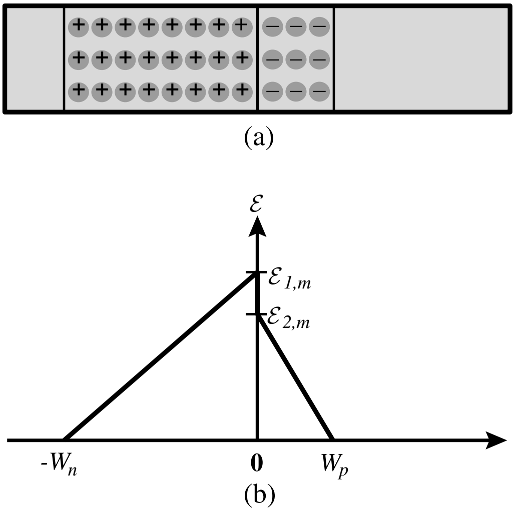


Figure 5.10: (a) Space-charge density and (b) electric field profiles in an n-p heterostructure.

To find $\mathcal{E}_{1,m}$, $\mathcal{E}_{2,m}$, W_n, and W_p, the following equations, along with equation 5.6.3, can be used:

$$N_d W_n = N_a W_p \tag{5.6.4}$$

$$\frac{\mathcal{E}_{1,m}}{W_n} = \frac{qN_d}{\epsilon_1} \text{ and } \frac{\mathcal{E}_{2,m}}{W_p} = \frac{qN_a}{\epsilon_2} \tag{5.6.5}$$

$$V_{bi} = \text{area under } \mathcal{E} = \frac{1}{2}\left[W_n \mathcal{E}_{1,m} + W_p \mathcal{E}_{2,m}\right] = V_{d1} + V_{d2} \tag{5.6.6}$$

These are essentially the same set of equations used to derive W_n, W_p, and $\mathcal{E}$ for a p-n homojunction . Equation 5.6.4 is our charge neutrality condition, and equation 5.6.5 and equation 5.6.6 result from solving Poisson's equation.

From these equations, we obtain the following set of relationships:

$$\boxed{W_p(V_{bi}) = \left\{\frac{2\epsilon_1\epsilon_2 V_{bi}}{e}\left[\frac{N_d}{N_a(N_a\epsilon_2 + N_d\epsilon_1)}\right]\right\}^{1/2}} \tag{5.6.7}$$

$$\boxed{W_n(V_{bi}) = \left\{\frac{2\epsilon_1\epsilon_2 V_{bi}}{e}\left[\frac{N_a}{N_d(N_a\epsilon_2 + N_d\epsilon_1)}\right]\right\}^{1/2}} \tag{5.6.8}$$

$$W(V_{bi}) = W_p(V_{bi}) + W_n(V_{bi}) = \left[W_n^2(V_{bi}) + W_p^2(V_{bi}) + 2W_n(V_{bi})W_p(V_{bi})\right]^{1/2}$$

$$\boxed{W(V_{bi}) = \left[\frac{2\epsilon_1\epsilon_2 V_{bi}}{e}\left(\frac{[N_a + N_d]^2}{N_a N_d\,[N_a\epsilon_2 + N_d\epsilon_1]}\right)\right]^{1/2}} \tag{5.6.9}$$

$$\boxed{\mathcal{E}_{1,m} = \frac{eN_d W_n}{\epsilon_1}} \tag{5.6.10}$$

$$\boxed{\mathcal{E}_{2,m} = \frac{eN_a W_p}{\epsilon_2}} \tag{5.6.11}$$

We can also solve for the band bending on either side of the junction V_{d1} and V_{d2}:

$$\boxed{V_{d1} = \frac{1}{2}W_n\mathcal{E}_{1,m} = \frac{eN_d W_n^2}{2\epsilon_1}} \tag{5.6.12}$$

$$\boxed{V_{d2} = \frac{1}{2}W_p\mathcal{E}_{2,m} = \frac{eN_a W_p^2}{2\epsilon_2}} \tag{5.6.13}$$

Current flow in abrupt p-n heterostructure

In a p-n homojunction, the ratio of current carried by electrons to current carried by holes I_n/I_p can be made large by making N_d larger than N_a. However, in bipolar transistor technology, it is desirable to have N_a be large while simultaneously maintaining a large value of I_n/I_p. This can only be achieved by employing a p-n heterostructure. We will now calculate the current characteristics of a p-n heterojunction and show how the ratio I_n/I_p can be controlled.

In figure 5.11a, we show the band diagram of the p-n heterostructure from figure 5.9b under forward bias. In determining the current characteristics of this structure, we make the following assumptions:

1. The electron and hole components of the current can each be described by thermionic emission, similar to the treatment given in section 5.3.3 for the electron current in an n-type Schottky barrier. The barrier to hole injection from the p side to the n side is labeled $e\phi_{Bh}$ in figure 5.11a, and the barrier to electron injection from the n side to the p side is labeled $e\phi_{Be}$. The electron current $I_n \propto \exp[-e\phi_{Be}/k_BT]$, and the hole current $I_p \propto \exp[-e\phi_{Bh}/k_BT]$.

2. The downwards notch in the conduction band immediately to the right of the junction does not capture electrons or in any way affect the electron current.

The general idea behind the use of heterostructures is that we would like to increase the barrier that holes must overcome ϕ_{Bh} relative to the barrier that electrons must overcome ϕ_{Be}. In p-n homojunctions, $\phi_{Be} = \phi_{Bh} = V_{bi} - V_F$. In p-n heterojunctions, these barriers are no longer equal.

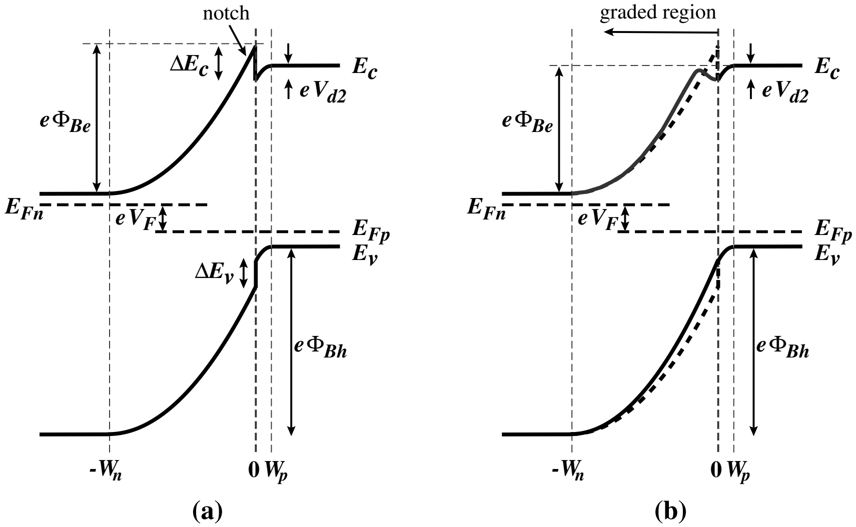


Figure 5.11: Band diagrams of (a) a forward biased abrupt p-n junction and (b) a forward biased graded p-n junction

Referring to figure 5.11a, for an abrupt p-n heterojunction, the barriers to hole injection (ϕ_{Bp}) and to electron injection (ϕ_{Bn}) are given by

$$e\phi_{Bh} = E_{g1} - (E_F - E_v)_p - (E_c - E_F)_n - eV_F \tag{5.6.14}$$

$$e\phi_{Be} = e(V_{d1} - V_F) \tag{5.6.15}$$

$$= E_{g2} - (E_F - E_v)_p - (E_c - E_F)_n - eV_F - eV_{d2} + \Delta E_c \tag{5.6.16}$$

Since we are assuming $N_a >> N_d$, the term eV_{d2} in the expression for ϕ_{Be} is small and can be omitted. The difference in the two barriers is given by

$$e\phi_{Bh} - e\phi_{Be} = E_{g1} - E_{g2} - \Delta E_c = \Delta E_v \tag{5.6.17}$$

The ratio of electron current to hole current in a p-n homojunction is given by

$$\left(\frac{I_n}{I_p}\right)_{hom} = \frac{D_n N_d L_p}{D_p N_a L_n} \tag{5.6.18}$$

In the p-n homojunction, the barrier seen by electrons is the same as that seen by holes, and the only means of controlling this ratio is through doping. In an abrupt p-n heterojunction with negligible barrier discontinuity at the conduction band edge, the barrier seen by electrons is smaller than that seen by holes by an amount ΔE_v, and so the corresponding ratio of electron current to hole current becomes

$$\left(\frac{I_n}{I_p}\right)_{het} = \frac{D_n N_d L_p}{D_p N_a L_n} \exp\left(\frac{\Delta E_v}{k_B T}\right) \tag{5.6.19}$$

We can see that even if N_a is kept high, the ratio of electron current to hole current in a heterojunction can still be kept large, even for a relatively modest bandgap discontinuity.

5.6.2 Graded p-n heterojunction

Although the abrupt p-n heterostructure discussed in the previous section did result in an increase in the barrier to hole injection, the notch in the conduction band at the interface also caused an undesirable increase in the barrier to electron injection. While the net effect was still an increase in the ratio I_n/I_p, eliminating this notch further increases I_n/I_p to a value

$$\left(\frac{I_n}{I_p}\right) = \frac{D_n N_d L_p}{D_p N_a L_n} \exp\left(\frac{\Delta E_g}{k_B T}\right) \tag{5.6.20}$$

In order to reduce this notch, the bandgap of the p material can be graded upwards from the junction, as shown in figure 5.11b. For example, in an n-AlGaAs/p-GaAs graded heterostructure, the n material is GaAs at the junction and is graded to the final AlGaAs composition over a short distance. The final shape of the notch depends on the length and profile of the grade; longer grading typically gives a smaller notch. However, it is important that the grade is contained well within the depletion region. If the grade ends outside the depletion region, then the barrier seen by holes decreases, thus reducing the benefits of the heterojunction. Note that the barrier to holes in both abrupt and graded heterojunctions is the same. It is just the barrier for electron flow that is reduced in the graded structures, allowing for the increased ratio of I_n/I_p.

Example 5.3 Designing a p-n heterojunction grade
Consider four different n-p$^+$ Al$_{0.3}$Ga$_{0.7}$As/GaAs heterojunctions with N$_D = 10^{17}$ and $N_A = 5 \times 10^{18}$. The AlGaAs in these junctions is graded from $x = 0$ to $x = 0.3$ over $X_{Grade} = 0$(abrupt), $X_{Grade} = 100\AA$, $X_{Grade} = 300\AA$, and $X_{Grade} = 1\mu$. Calculate and plot the energy band diagrams for the above four cases.

Assume the dielectric constant of AlGaAs to be the same as that of GaAs.
E$_g$ = 1.8 eV for Al$_{0.3}$Ga$_{0.7}$As, and E$_g$ = 1.4 eV for GaAs. ΔE$_g$ = 0.374 eV, ΔE$_C$ = 0.237 eV, and ΔE$_V$ = 0.137 eV. On the AlGaAs emitter side, far away from the junction,

$$\phi_n = E_C - E_F = k_B T \ln\left(\frac{N_C}{n}\right) = 0.0323 \text{ eV} \tag{5.6.21}$$

Since the the p-GaAs is degenerately doped, Joyce-Dixon statistics must be applied:

$$\phi_p = E_V - E_F =$$

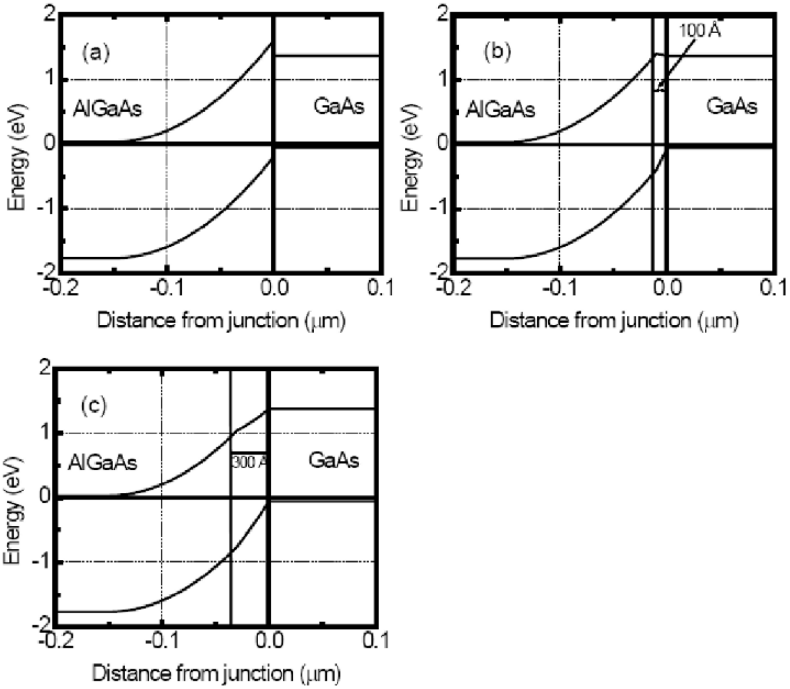


Figure 5.12: Solutions to example 5.3

$$k_BT\left[\ln\left(\frac{p}{N_V}\right)+A_1\frac{p}{N_V}+A_2\left(\frac{p}{N_V}\right)^2\right]=0.011\text{ eV}$$

The solutions are plotted in figure 5.12.

5.6.3 Quasi-electric fields

In a homogeneous semiconductor, the separation between the conduction and valence bands is everywhere equal to the semiconductor bandgap. Any electric field applied to the material therefore results in an equal slope in the conduction and valence bands, as indicated in figure 5.13a. When a hole or electron is placed in this structure, a force of magnitude $e\mathcal{E}$ will act on the particle. The magnitude of the force is equal to the slope of the bands and is the same for both electrons and holes. However, the direction of the force is opposite for the two particles.

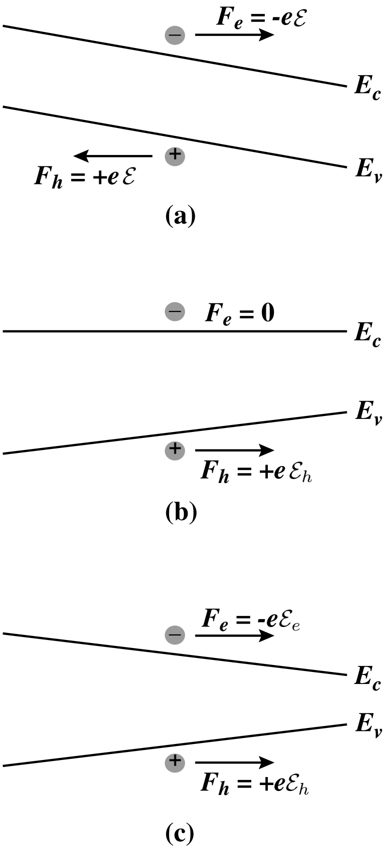


Figure 5.13: (a) Band diagram when an electric field is applied to a homogeneous semiconductor. (b) Quasi-field causes a force on holes but not on electrons. (c) Quasi-field in which electrons and holes feel a force in the same direction.

An interesting phenomenon arises in semiconductors with graded bandgaps, such as the bipolar transistor emitter-base structure shown in figure 5.11b. In the graded region, the bandgap is not constant, so the slopes in the conduction and valence bands are no longer equal. Hence the forces acting on electrons and holes in this region are no longer equal in magnitude. It is in general possible for a force to act on only one type of carrier, as shown in figure 5.13b, or for forces to act in the same direction for both electrons and holes, as in figure 5.13c. Such behavior cannot be achieved by pure electric fields in homogeneous materials. These fields, which were first described by Herbert Kroemer in 1957, are therefore referred to as quasi-electric fields.

In a given material, the total field acting on a hole or an electron is always the sum of the applied field and the quasi field, or

$$\mathcal{E}_{e,tot} = \mathcal{E}_{app} + \mathcal{E}_{e,quasi} \tag{5.6.22}$$

$$\mathcal{E}_{h,tot} = \mathcal{E}_{app} + \mathcal{E}_{h,quasi} \tag{5.6.23}$$

The applied field, which results from applying a voltage difference between the ends of the material, will always be the same for both electrons and holes, but the quasi field could be different for both. The band profiles in figure 5.13b and figure 5.13c can therefore be achieved in a number of different ways. For example, the profile in figure 5.13b could be achieved in the following two ways:

1. An undoped (intrinsic) material with a graded composition and zero applied electric field typically results in the profile in figure 5.13c. If an electric field $\mathcal{E}_{app} = -\mathcal{E}_{e,quasi}$ is then applied to this material, the resulting profile will be the one shown in figure 5.13b.

2. A uniformly doped n-type material with a graded composition and zero applied electric field will also result in the profile in figure 5.13b. In this case, the doping ensures that the separation between the conduction band and the Fermi level remains approximately constant. Notice that the resulting quasi-electric field in this structure acts only on minority carriers.

Quasi-electric fields provide engineers with additional tools that can be exploited in device design. They have proven to be very useful in decreasing transit times in devices that rely on minority carrier transport. For example, in bipolar technology, a highly doped graded base layer is often used to speed up the transport of minority carriers from the emitter to the collector. For a base with uniform bandgap, minority carriers injected from the emitter must diffuse across the base, a process that is generally slow. By using a highly doped graded base to generate a quasi-electric field, such as was described in the second example above, minority carriers can be swept across much more quickly, thus reducing the base transit time and improving the device RF performance.

5.7 PROBLEMS

Temperature is 300 K unless stated otherwise.

- **Section 5.2**

Problem 5.1 A 2 μm thick ×10 μm Al interconnect is used in a semiconductor chip. If the length of a particular interconnect is 1 cm, calculate the resistance of the line. Use table 5.1 for the resistivity data.

Problem 5.2 If a current density of 10^5 A/cm^2 flows in the interconnect of problem 5.1 to, calculate the potential drop.

Problem 5.3 Use the resistivity of Al and Cu given in table 5.1 and use the Drude model (chapter 1) to calculate the mobility of electrons in these two materials.

Problem 5.4 Discuss why in the analysis of the conductivity of metals we do not consider hole conductivity.

- **Section 5.3**

Problem 5.5 Assume the ideal Schottky barrier model with no interface states for an n-type Si with $N_d = 10^{16}$ cm^{-3}. The metal work function is 4.5 eV and the Si electron affinity is 4 eV. Calculate the Schottky barrier height, built-in voltage, and depletion width at no external bias.

Problem 5.6 A Schottky barrier is formed between Al and n-type silicon with a doping of 10^{16} cm^{-3}. Calculate the theoretical barrier if there are no surface states. Compare this with the actual barrier height. Use the data in the text.

Problem 5.7 Assume that at the surface of GaAs, 50% of all bonds are "defective" and lead to a uniform distribution of states in the bandgap. Each defective bond contributes one bandgap state. What is the two-dimensional density of bandgap states (units of eV^{-1}cm^{-2})? Assume that the neutral level ϕ_o is at midgap. Approximately how much will the Fermi level shift if a total charge density of 10^{12} cm^{-2} is injected into the surface states? This example gives an idea of "Fermi level pinning".

Problem 5.8 The capacitance of a Pt-n-type GaAs Schottky diode is given by

$$\frac{1}{(C(\mu F))^2} = 1.0 \times 10^5 - 2.0 \times 10^5 \text{ V}$$

The diode area is 0.1 cm^2. Calculate the built-in voltage V_{bi}, the barrier height, and the doping concentration.

Problem 5.9 Calculate the mean thermal speed of electrons in Si and GaAs at 77 K and 300 K. $m^*_{Si} = 0.3m_o$; $m^*_{GaAs} = 0.067m_o$.

Problem 5.10 Calculate the saturation current density in an Au Schottky diode made from n-type GaAs at 300 K. Use the Schottky barrier height values given in table 5.2.

Problem 5.11 Consider an Au n-type GaAs Schottky diode with 50 μm diameter. Plot the current voltage characteristics for the diode between a reverse voltage of 2 V and a forward voltage of 0.5 V.

Problem 5.12 Calculate and plot the I-V characteristics of a Schottky barrier diode between W and n-type Si doped at 5 × 10^{16} cm^{-3} at 300 K. The junction area is 1 mm^2. Plot the results from a forward current of 0 to 100 mA.

Problem 5.13 In some narrow-bandgap semiconductors, it is difficult to obtain a good Schottky barrier (with low reverse current) due to the very small barrier height. Consider an n-type InGaAs sample. Describe, on physical bases, how the "effective" Schottky barrier height can be increased by incorporating a thin p-type doped region near the surface region.

Problem 5.14 In the text, when we discussed the current flow in a Schottky barrier, we assumed that the current was due to thermionic emission only. This is based on classical physics where it is assumed that only particles with energy greater than a barrier can pass through. Consider a W-n-type GaAs Schottky barrier in which the Schottky barrier triangular potential is described by a field of 10^5 V/cm. The Schottky barrier height is 0.8 V. Calculate the tunneling probability through the triangular barrier as a function of electron energy from $E = 0.4$ eV to $E = 0.8$ eV. The tunneling current increases the Schottky reverse current above the value obtained by thermionic current considerations.

Problem 5.15 Consider an Al-n-type Si Schottky diode. The semiconductor is doped at 10^{16} cm^3. Also consider a p-n diode made from Si with the following parameters (the diode is ideal):

$$\begin{aligned}
N_d &= N_a = 10^{18}\ \text{cm}^{-3} \\
D_n &= 25\ \text{cm}^2/\text{s} \\
D_p &= 8\ \text{cm}^2/\text{s} \\
\tau_n &= \tau_p = 10\, ns
\end{aligned}$$

Calculate the turn-on voltages for the Schottky and p-n diode. Assume that the current density has to reach 10^5 A/cm^2 for the diode to be turned on.

Problem 5.16 An important problem in very high-speed transistors (to be discussed in chapter 8) based on the InAlAs/InGaAs system is the reliability of the Schottky barrier. Consider a Schottky barrier formed on an InAlAs doped n-type at 10^{16} cm^{-3}. Calculate the saturation current density if the Schottky barrier height is (i) 0.7 V; (ii) 0.6 V at 300 K.

The mass of the electrons in InAlAs is 0.08 m$_o$. The Richardson constant has a value

$$R^* = 120 \left(\frac{m^*}{m_o}\right) = 9.6A\,\text{cm}^{-2}K^{-2}$$

The saturation current density then becomes

$$\begin{aligned}
J_s\,(\phi_b = 0.7\text{ V}) &= R^*T^2 \exp\left[-\left(\frac{e\phi_b}{k_BT}\right)\right] \\
&= 1.8 \times 10^{-6}\ \text{A/cm}^2 \\
J_s(\phi_b = 0.6\text{ V}) &= 8.2 \times 10^{-5}\ \text{A/cm}^2
\end{aligned}$$

Thus the current density varies by a very large value depending upon the Schottky barrier value. The Schottky barrier height depends upon the metal-semiconductor interface quality and can be easily affected by fabrication steps. In a p-n diode, on the other hand, the built-in voltage is fixed by doping and is more controllable.

Problem 5.17 A metal-i-GaN-n-GaN Schottky junction is shown in the figure 5.14 (a) below. Fixed positive and negative polarization charges across the GaN create a 2DEG in this structure. The metal semiconductor barrier height ϕ_B is 0.5 eV. Assume that the i-GaN layer is 50 nm thick, and that the relative dielectric constant of AlGaN is 10.

1. Draw the equilibrium band diagram. Also draw the band diagram for a reverse bias of 10V.
2. Now, a phantom material (see figure 5.14(b)) with a high dielectric constant of 100 and thickness of 50 nm is inserted between the gate and the GaN layer. Draw the band diagram at equilibrium and at a reverse bias of 10 V across the junction. Assume that the barrier height on this material is the same as in the case of GaN, and that the bands of this material align perfectly with GaN.
3. The tunneling probability across a barrier of the form in figure 5.14 (c) is given by
$$P = e^{-\frac{4\pi\sqrt{2m}d(E_0^{3/2}-E_1^{3/2})}{2(E_0-E_1)h}}$$
Use the above expression to estimate the ratio between the tunneling probabilities for the cases in part (a) and (b) at equilibrium, and when a reverse bias of 10 V is applied.

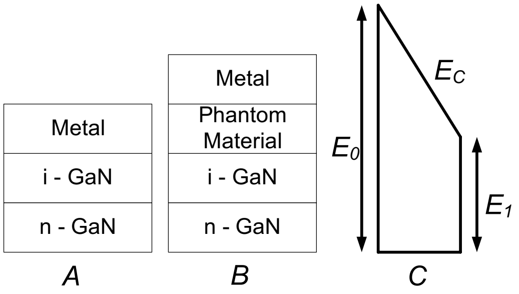


Figure 5.14: Figure for problem 5.17.

• **Section 5.4**

Problem 5.18 A gold contact is deposited on GaAs doped at $N_d = 5 \times 20$ cm^{-3}. Calculate the tunneling probability of the electrons to go into the semiconductor.

Problem 5.19 A metal with a work function of 4.2 V is deposited on an n-type silicon semiconductor with an electron affinity of 4.0 V. Assume that there are no interface states. Calculate the doping density for which there is no space charge region at zero applied bias.

Problem 5.20 To fabricate very high-quality ohmic contacts on a large-bandgap material one often deposits a heavily doped low-bandgap material. For example, to make an n-type ohmic contact on GaAs, one may deposit n^+InAs. Discuss why this would help to improve the resistance of the contact.

Problem 5.21 Consider a W-n Si Schottky barrier on silicon doped at 10^{18} and 10^{20} cm^{-3}. Calculate the tunneling probability of electrons for electrons with energies near the conduction band in the two doping cases.

- **Section 5.5**

Problem 5.22 A (001) Si-SiO$_2$ interface has $\sim 10^{11}$ cm^{-2} interface states. Assume that each state corresponds to one defective bond at the interface. Calculate the fraction of defective bonds for the interface.

Problem 5.23 Calculate the sheet resistance of a 0.5 μm thick poly film doped n-type at 10^{19} cm^{-3}. This film is used to form a resistor of width 20 μm and length 0.1 cm. Calculate the resistance of the contact if the electron mobility is 150 cm^2/V·s.

- **Section 5.6**

Problem 5.24 The measured value of ΔE_C for the AlGaAs/GaAs system is approximately $0.6\Delta E_G$. However, electron affinity rules predict a different value for ΔE_C. Show, using band diagrams and potential profiles, that an interfacial dipole present at the AlGaAs/GaAs interface can explain this. Discuss both n-AlGaAs/n-GaAs and n-AlGaAs/p-GaAs junctions.

Problem 5.25 Consider four different n-p^+ Al$_{0.3}$Ga$_{0.7}$As/GaAs heterojunctions with $N_D = 10^{17}$ and $N_A = 5 \times 10^{18}$. The AlGaAs in these junctions is graded from $x = 0$ to $x = 0.3$ over $X_{Grade} = 0$ (abrupt), $X_{Grade} = 100$ Å, $X_{Grade} = 300$ Å, and $X_{Grade} = 1\ \mu m$.
(a) Calculate and plot the energy band diagrams for the above four cases.
(b) When a forward bias is applied, how will the minority carrier current ratio $I_{n-GaAs}/I_{p-AlGaAs}$ vary in these four heterojunctions? Which one would you use as the base-emitter junction in an n-p-n HBT?
Assume the dielectric constant of AlGaAs to be the same as that of GaAs.

Problem 5.26 In an attempt to increase the collector breakdown voltage of an n-AlInAs/p-GaInAs/i-GaInAs/n+-GaInAs HBT, I replace the i-GaInAs collector with i-InP. Unfortunately, this introduced a potential barrier in the conduction band of $\Delta E_C = 0.2eV$. I decide to linearly grade the barrier over a distance of 50nm for GaInAs ($E_g = 0.7eV$) to InP ($E_g = 1.4eV$). Design the electrostatics so that there is no barrier to electron from over the graded region. Assume that $E_F = E_V$ in the p base and $E_F = E_C$ in the n+ subcollector. How would the desing change if I decided to grade the region parabolically as shown in figure 5.15.

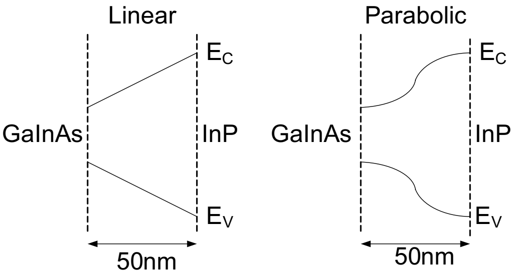


Figure 5.15: Figure for problem 5.26.

5.8 FURTHER READING

- D. A. Neamen, Semiconductor Physics and Devices: Basic Principles (Irwin, Boston, 1997).
- M. S. Tyagi, Introduction to Semiconductor Materials and Devices (John Wiley and Sons, New York, 1991).
- R. S. Muller and T. I. Kamins, Device Electronics for Integrated Circuits (Wiley, New York, 1986).

Chapter 6

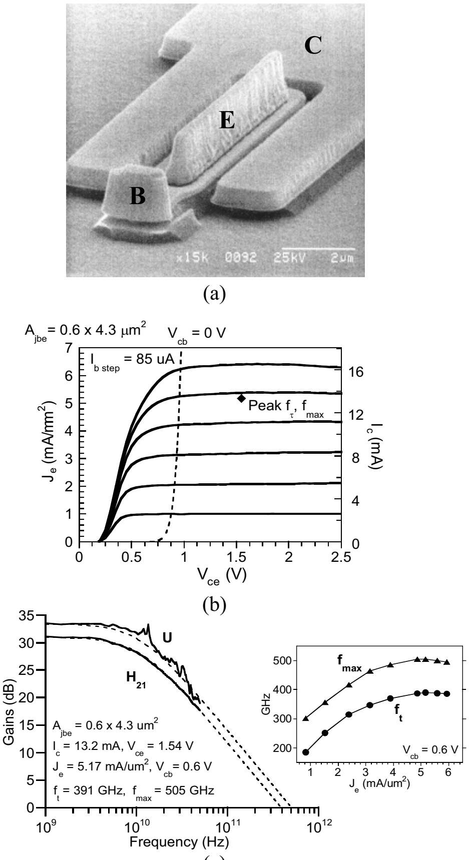
BIPOLAR JUNCTION TRANSISTORS

6.1 INTRODUCTION

The bipolar junction transistor was the first three-terminal device in solid state electronics and continues to be a device of choice for many digital and microwave applications. For a decade after its invention, the bipolar device remained the only three-terminal device in commercial applications. However, as the Si-SiO$_2$ interface improved, the MOSFET has become dominant. Heterojunction bipolar devices now have very high performance in terms of frequency and gain. In figure 6.1 we show the structure and device performance parameters of a state of the art InGaAs/InP heterojunction bipolar transistor. In a three terminal device the goal is to use a small input to control a large output. The input could be an incoming weak signal to be amplified, or a digital signal. The workings of a three terminal device can be understood by examining how the flow of water can be controlled. In one case, let's say the water was to flow in a pipe of fixed diameter while in another, it could flow over an open channel. In figure 6.2 we show two different ways one could design a system to control the water flow. On the left-hand side sequence of figure 6.2 we show how a change in the ground potential can be used to modify the water flow. Only the fraction of water that is above the bump will flow across the potential profile. The value of the potential bump could be controlled by an independent control input.

Water flow can also be controlled by a faucet in which the faucet controls the constriction of the pipe and thus the water flow. In a bipolar device one controls the potential profile in the current flow channel by using the base current as a controlling agent. In a FET on the other hand one controls the channel constriction by applying a gate bias.

As noted earlier, an important requirement for an electronic device is that a small change in the input should cause a large change in the output, i.e., the device should have a high gain. This requirement is essential for amplification of signals, tolerance of high noise margins in digital devices, and the ability to have a large fan-out (i.e., the output can drive several additional devices). Another important requirement is that the input should be isolated from the output. For



(a)

(b)

(c)

Figure 6.1: State of the art n-InP/p^+-InGaAs/n^--InP double heterojunction bipolar transistor (DHBT). (a)SEM image of device, (b) dc I-V characteristics, and (c) high frequency current gain and power gain. For this device, the average $\beta \approx 36$ and $V_{BR,CEO} = 5.1$ V (measured at $I_C = 50\ \mu A$). Figures courtesy of M. Rodwell and Z. Griffith, UCSB.

the faucet example, these two requirements mean we should be able to turn the faucet on and off with little effort and the water should not leak out of the faucet head!

In this chapter we will discuss static characteristics of the bipolar transistor.

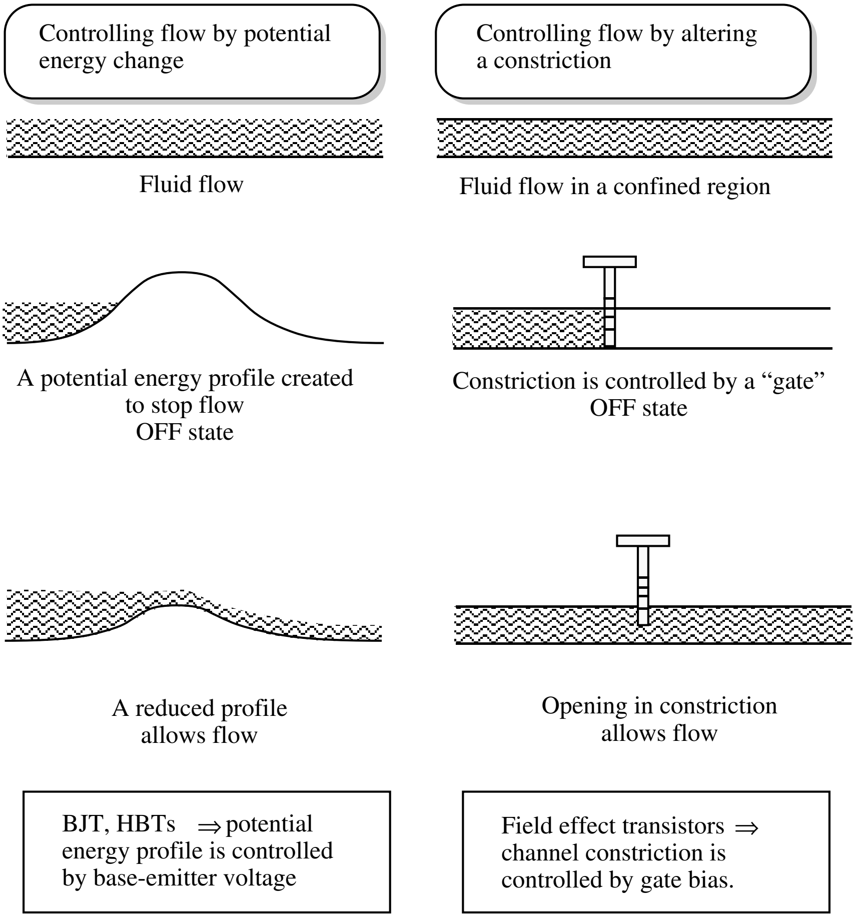

Controlling flow by potential energy change
Controlling flow by altering a constriction
Fluid flow
Fluid flow in a confined region
A potential energy profile created to stop flow OFF state
Constriction is controlled by a "gate" OFF state
A reduced profile allows flow
Opening in constriction allows flow
BJT, HBTs ⇒ potential energy profile is controlled by base-emitter voltage
Field effect transistors ⇒ channel constriction is controlled by gate bias.

Figure 6.2: Two different ways to control flow of a fluid. The bipolar and field effect transistors use these two approaches to control current flow.

6.2 BIPOLAR TRANSISTOR: A CONCEPTUAL PICTURE

The bipolar junction transistor employs two back to back $p - n$ diodes which with clever design rules can have a high amplification and can operate at high frequency. It can also act as a digital device and as a microwave device.

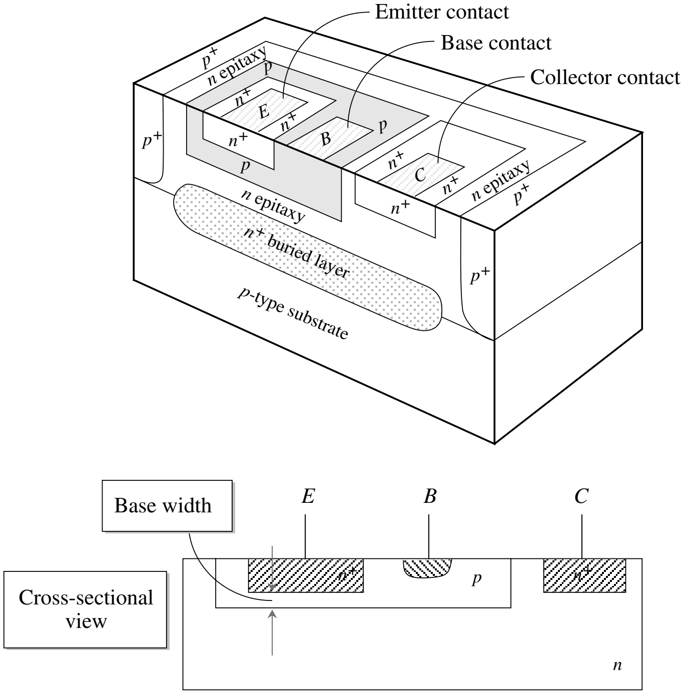


Figure 6.3: A schematic of the structure and doping profiles of a bipolar junction transistor along with a simplified view of the cross-section.

We have shown a state of the art bipolar device. A schematic of the device is shown in figure 6.1. The device could have a doping of the form $n^+ - p$-n or $p^+ - n$-p. We will focus on the $n^+ - p$-n device. The emitter is heavily doped $n-$type, the p-region forms the base, and the lower n region is the collector. The emitter doping N_{de} is much larger than the base doping N_{ab} to ensure that the device has a high current gain, i.e., that a small base current change produces a large collector current change.

To understand how the device can have gain, let us consider a BJT where the emitter base junction (EBJ) is forward biased and the base collector junction (BCJ) is reverse biased. This biasing creates the forward active mode. The band profile of the device is shown in figure 6.4. Note that the base width W_b is much smaller than the diffusion length of electrons in the p-type base region. So that when electrons are injected from the emitter, most cross the base without recombining with holes. The strong electric field these electrons see once they reach the collector, cause them to be swept away and form the collector current.

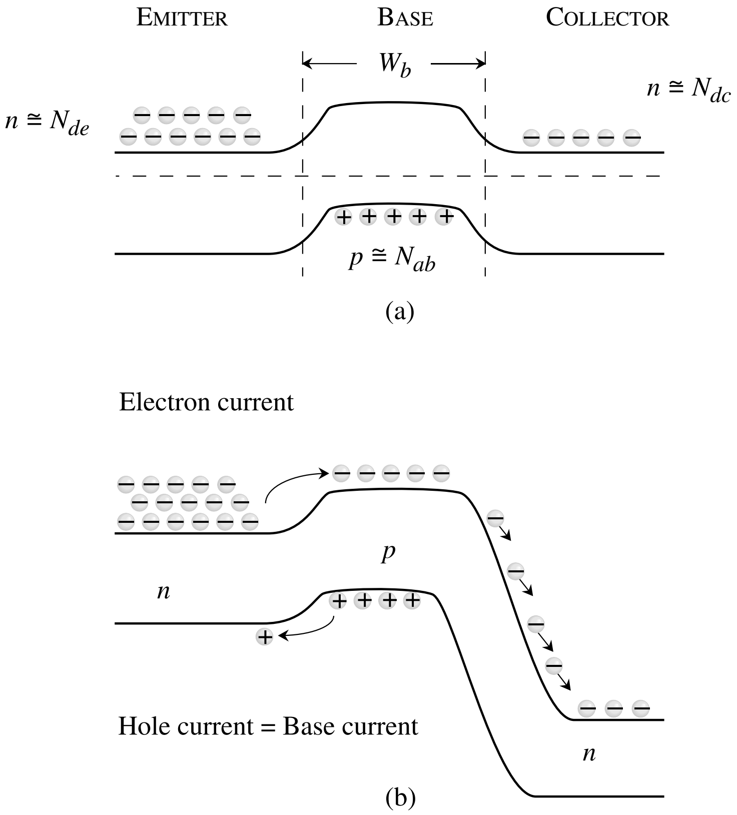


Figure 6.4: (a) Band profile of an unbiased n^+p-n BJT. (b) Band profile of a BJT biased in the forward active mode, where the EBJ is forward biased and the BJT is reverse biased.

We remind ourselves that if the EB diode is asymmetrically doped, the forward bias current is essentially made up of injection of electrons into the p-side. This forward-biased current can also be altered by a very small change in the forward bias voltage since the current depends exponentially on the forward bias value. The forward-biased n^+ emitter injects electrons into the p-base. Some of the electrons recombine in the base with the holes, but if the base region is less than the diffusion length of the minority carriers, most of them reach the depletion region of the p-n base-collector diode and are swept out to form the collector current. The collector current is proportional to the minority carriers (electrons) that reach the edge of the $p-n$ depletion region, as shown in figure 6.4b. Since the injected minority carriers are due to the emitter current, we have

$$I_C = BI_{En} \tag{6.2.1}$$

where I_{En} is the electron part of the emitter current and the factor B is called the base transport factor. In the absence of e-h recombination, the emitter current is made up of electrons injected from the n- to p-sides (I_{En}) and holes injected from p- to n-sides (I_{Ep}). Since the BCJ is reverse biased, the collector current is related only to the electrons injected and we define the emitter efficiency γ_e as

$$\boxed{\gamma_e = \frac{I_{En}}{I_{En} + I_{Ep}}} \tag{6.2.2}$$

For optimum devices, γ_e and B should be close to unity. The ratio between the collector and emitter currents is the current transfer ratio, α

$$\boxed{\frac{I_C}{I_E} = \frac{BI_{En}}{I_{En} + I_{Ep}} = B\gamma_e = \alpha} \tag{6.2.3}$$

This ratio is close to unity in good bipolar devices. In figure 6.5 we show a typical circuit for a BJT in the forward-bias active mode. A change in the base current alters the minority carrier density n_p in the base and causes a large change in the collector current. The ratio between the collector current and the controlling base current is of great importance since this represents the current amplification . The base current is made up of the hole current injected into the emitter (I_{Ep}) and the hole current due to the recombination in the base with injected electrons from the emitter (= $(1-B)I_{En}$). Thus

$$I_B = I_{Ep} + (1 - B)I_{En} \tag{6.2.4}$$

The base-to-collector current amplification factor, denoted by β is then

$$\begin{aligned} \beta = \frac{I_C}{I_B} &= \frac{BI_{En}}{I_{Ep} + (1 - B)I_{En}} = \frac{B(I_{En}/I_E)}{1 - B(I_{En}/I_E)} \\ &= \frac{B\gamma_e}{1 - B\gamma_e} \end{aligned} \tag{6.2.5}$$

This gives for the current gain

$$\boxed{\beta = \frac{\alpha}{1 - \alpha}} \tag{6.2.6}$$

The factor β can be quite large for the bipolar transistor. In the next section we will discuss the mathematical derivation of the device characteristics.

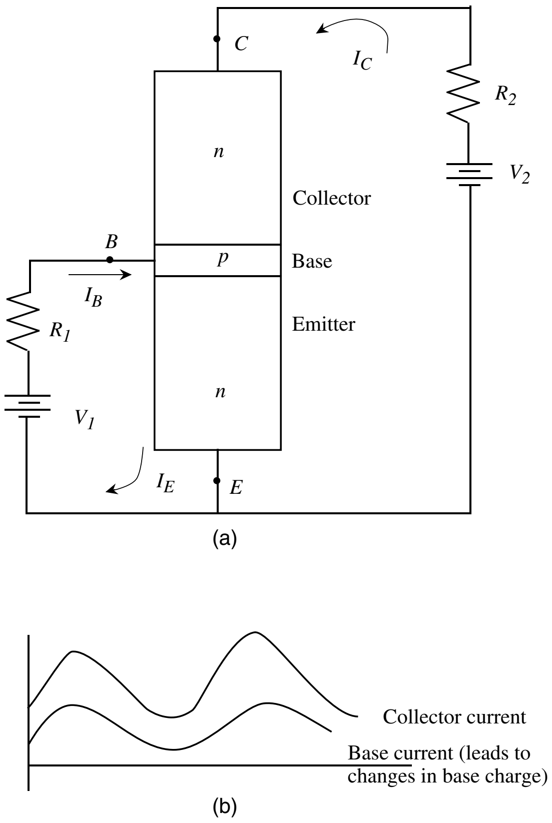


Figure 6.5: A schematic showing how the change in base current affects the majority carrier injection density and the collector current in a bipolar device. (a) A circuit using a bipolar transistor. (b) The effect of base current variation on the injected minority charge and the collector current. The collector current is much larger than the base current.

6.3 STATIC CHARACTERISTICS: CURRENT-VOLTAGE RELATION

We will now develop a model for the current flow in a BJT. Initially we will use a simple model which captures the essence of the device performance. Later we will discuss secondary issues. In the bipolar device carriers from the emitter are injected “vertically” across the base while the base charge is injected from the "side" of the device, as can be seen in figure 6.6. If we assume that the emitter width is wide, the device can be understood using a one-dimensional analysis. We will use the following simplifying assumptions.

1. The electrons injected from the emitter diffuse across the base region and the field across the base is small enough that there is no drift.

2. The electric fields are nonzero only in the depletion regions and are zero in the bulk materials.

3. The collector injection current is negligible when the BJT is reverse biased.

4. In describing voltages, we use the following notation. The first subscript of the voltage symbol represents the contact with respect to which the potential is measured. For example, $V_{BE} > 0$ means the base is positive with respect to the emitter.

In general, a number of currents can be identified in the bipolar device, (figure 6.6) as follows:

- Base current:Is made of holes that recombine with electrons injected from the emitter (Component I) and holes that are injected across the emitter-base junction into the emitter (Component II). Once again we ignore the BCJ for the forward active region.

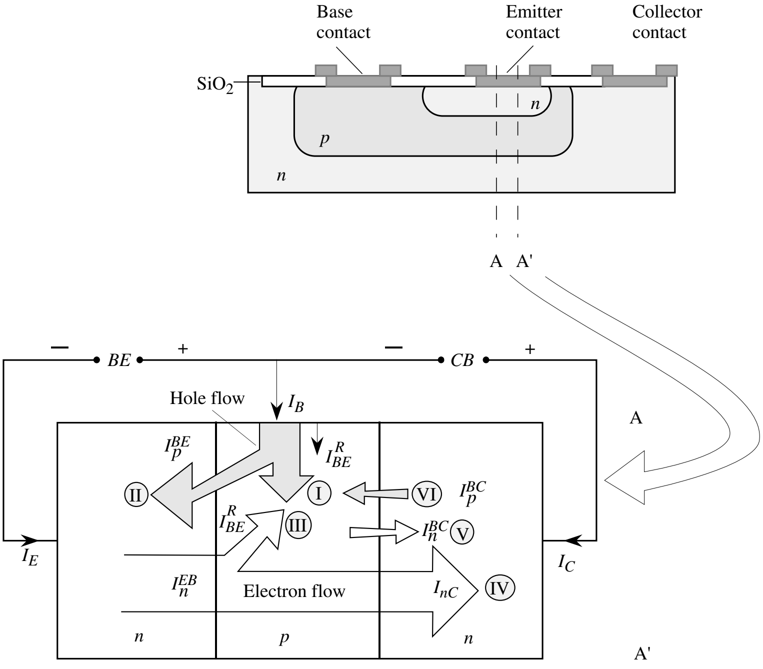
- Emitter current: Consists of the electron current that recombines with the holes in the base region (III), the electron current which is injected into the collector (IV), and the hole current injected from the base into the emitter (II).

Minority electron (V) and hole (VI) currents flow in the base-collector junction and are important when the emitter current goes toward zero. In our analysis, we will assume that all the dopants are ionized and the majority carrier density is simply equal to the doping density. The symbols for the doping density are (for the npn device): N_{de}—donor density in the emitter; N_{ab}—acceptor density in the base; N_{dc}—donor density in the collector. If the ionization of the dopants is not complete we need to adjust for the ionization efficiency.

The back to back $p - n$ diodes in the bipolar device can operate in four possible biasing modes as shown in table 6.1. Depending upon the applications, the bipolar device operation may span one or all of these modes. For example, for small-signal applications where one needs amplification one only operates in the forward active mode, while for switching applications the device may have to operate under cutoff and saturation modes and pass through the active mode during the switching.

6.3.1 Current Flow in a BJT

Since the bipolar device is based on $p - n$ diodes, we will use our understanding of current flow of $p - n$ diodes. Note that we will assume the emitter width is long compared to hole



I_n^{EB} = Emitter current injected into the base ≡ I_{En}

I_p^{BE} = Base current injected into the emitter ≡ I_{Ep}

I_{BE}^R = Recombination current in the base region

I_p^{BC} = Hole current injected across reverse-biased base collector junction

I_n^{BC} = Electron current injected across reverse-biased base collector junction

I_{nC} = Electron current coming from the emitter ($\cong I_C$)

Figure 6.6: A schematic of an Si BJT showing the three-dimensional nature of the structure and the current flow. Along the section AA', the current flow can be assumed one-dimensional. The various current components in a BJT are discussed in the text.

diffusion length while the base width is small compared to the electron diffusion length. We will use the different axes and origins shown in figure 6.7. The distances are labeled x_e, x_b, and x_c as shown and are measured from the edges of the depletion region. The base width is W_b, but the width of the "neutral" base region is W_{bn} as shown. We assume that W_b and W_{bn} are equal.

Mode of operation	**EBJ bias**	**CBJ bias**
Forward active	Forward ($V_{BE} > 0$)	Reverse($V_{CB} > 0$)
Cutoff	Reverse ($V_{BE} < 0$)	Reverse ($V_{CB} > 0$)
Saturation	Forward ($V_{BE} > 0$)	Forward ($V_{CB} < 0$)
Reverse active	Reverse ($V_{BE} < 0$)	Forward ($V_{CB} < 0$)

Table 6.1: Operation modes of the npn bipolar transistor. Depending upon the particular application, the transistor may operate in one or several modes.

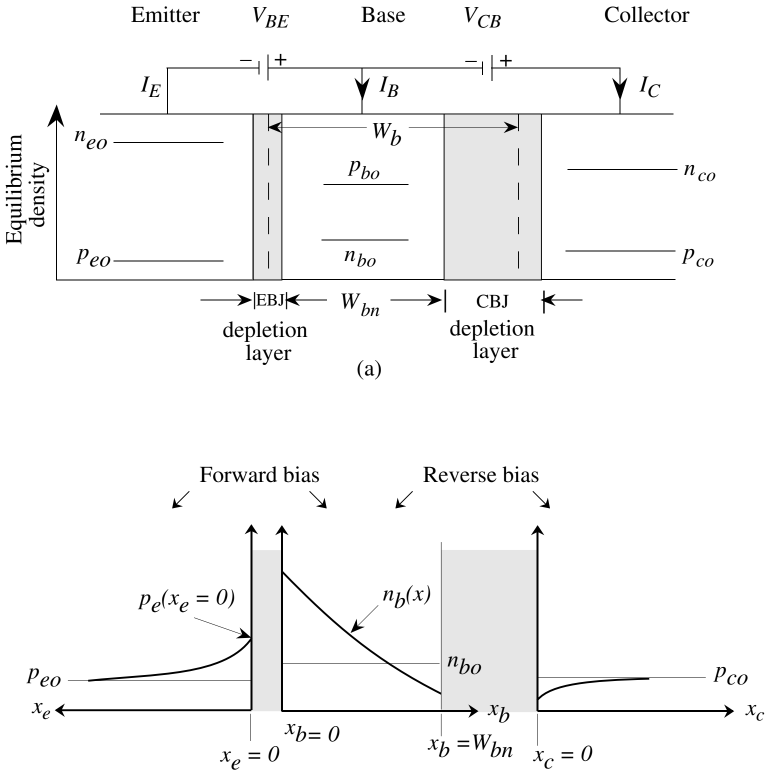
Later we will study the effect of the two widths being different. Using the p-n diode theory, we have the following relations for the excess carrier densities in the various regions are(see our discussion on carrier decay in chapter 3 and chapter 4):

$$\begin{aligned}
\delta p_e(x_e = 0) &= \text{excess hole density at the emitter side of the EBJ} \\
&= p_{eo} \left[\exp\left(eV_{BE}/k_BT\right) - 1\right] \qquad (6.3.1) \\
\delta n_b(x_b = 0) &= \text{excess electron density on the base side of the EBJ} \\
&= n_{bo} \left[\exp\left(eV_{BE}/k_BT\right) - 1\right] \qquad (6.3.2) \\
\delta n_b(x_b = W_{bn}) &= \text{excess electron density at the base side of the CBJ} \\
&\quad \text{(collector-base junction)} \\
&= n_{bo} \left[\exp\left(-eV_{CB}/k_BT\right) - 1\right] \qquad (6.3.3) \\
\delta p_c(x_c = 0) &= \text{excess hole density at the collector side of the CBJ} \\
&= p_{co} \left[\exp\left(-eV_{CB}/k_BT\right) - 1\right] \qquad (6.3.4)
\end{aligned}$$

As shown in figure 6.7 in these expressions the subscripts p_{eo}, n_{bo}, and p_{co} represent the minority carrier equilibrium densities in the emitter, base, and collector, respectively. The total minority carrier concentrations p_e in the emitter, n_b in the base, and p_c in the collector are shown schematically in figure 6.7b. Assuming 100% ionization of the dopants, the majority carrier densities are $n_{eo} = N_{de}$, $p_{bo} = N_{ab}$, and $n_{co} = N_{dc}$ for the emitter, base, and collector. We will assume that the emitter and collector regions are longer than the hole diffusion lengths L_p, so that the hole densities decrease exponentially away from base regions.

To find the current flow we have to calculate the spatial variation of carrier densities. In the base region, the excess electron density is given at the edges of the neutral base region by equation 6.3.2 and equation 6.3.3 To obtain the electron density in the base we must solve the continuity equation using these two boundary conditions, as discussed in section 3.9. The excess minority carrier density in the base region is given by

$$\begin{aligned}
\delta n_b(x_b) &= \frac{n_{bo}}{\sinh\left(\frac{W_{bn}}{L_b}\right)} \left\{ \sinh\left(\frac{W_{bn} - x_b}{L_b}\right) \left[\exp\left(\frac{eV_{BE}}{k_BT}\right) - 1\right] \right. \\
&\quad \left. + \sinh\left(\frac{x_b}{L_b}\right) \left[\exp\left(-\frac{eV_{CB}}{k_BT}\right) - 1\right] \right\} \qquad (6.3.5)
\end{aligned}$$



(a)

(b)

Figure 6.7: A forward active mode BJT. (a) The equilibrium carrier concentrations of electrons and holes and positions of the junction depletion regions in the npn transistor. (b) Minority carrier distributions in the emitter, base, and collector regions.

The profile of the total minority carrier densities (i.e., background and excess) is shown in figure 6.7b. The electron distribution in the base is almost linear, as can be seen, and is assumed to be so for some simple applications. Once the excess carrier spatial distributions are known we can calculate the currents as we did for the p-n diode. We assume that the emitter-base currents are due to carrier diffusion once the device is biased. We have, for a device of area A and diffusion coefficients D_b and D_e in the base and emitter, respectively,

$$\begin{aligned} I_{En} = I_n^{EB} &= eAD_b \left. \frac{d\delta n_b(x)}{dx_b} \right|_{x_b=0} && (6.3.6) \\ I_{Ep} = I_p^{BE} &= -eAD_e \left. \frac{d\delta p(x)}{dx_e} \right|_{x_e=0} && (6.3.7) \end{aligned}$$

These are the current components shown in figure 6.6 and represent the emitter current components II, III, and IV. Assuming an exponentially decaying hole density into the emitter, we have, as in the case of a p-n diode ,

$$I_{Ep} = -A\left(\frac{eD_e p_{eo}}{L_e}\right)\left[\exp\left(\frac{eV_{BE}}{k_BT}\right) - 1\right] \tag{6.3.8}$$

Using the electron distribution derived in the base, we have for the electron part of the emitter current

$$\begin{aligned} I_{En} &= -\frac{eAD_b n_{bo}}{L_b \sinh\left(\frac{W_{bn}}{L_b}\right)} \left\{ \cosh\left(\frac{W_{bn} - x_b}{L_b}\right)\left[\exp\left(\frac{eV_{BE}}{k_BT}\right) - 1\right] \right. \\ &\quad \left. \left. - \cosh\left(\frac{x_b}{L_b}\right)\left[\exp\left(-\frac{eV_{CB}}{k_BT}\right) - 1\right] \right\} \right|_{at\ x_b=0} \\ &= -\frac{eAD_b n_{bo}}{L_b \sinh\left(\frac{W_{bn}}{L_b}\right)} \left\{ \cosh\left(\frac{W_{bn}}{L_b}\right)\left[\exp\left(\frac{eV_{BE}}{k_BT}\right) - 1\right] \right. \\ &\quad \left. - \left[\exp\left(-\frac{eV_{CB}}{k_BT}\right) - 1\right] \right\} \end{aligned} \tag{6.3.9}$$

For high emitter efficiency we want I_{En} to be much larger than I_{Ep}. This occurs if the emitter doping is much larger than the base doping. The total emitter current becomes

$$\begin{aligned} I_E &= I_{En} + I_{Ep} = -\left\{ \frac{eAD_b n_{bo}}{L_b} \coth\left(\frac{W_{bn}}{L_b}\right) + \frac{eAD_e p_{eo}}{L_e} \right\} \\ &\quad \left[\exp\left(\frac{eV_{BE}}{k_BT}\right) - 1\right] + \frac{eAD_b n_{bo}}{L_b \sinh\left(\frac{W_{bn}}{L_b}\right)} \left[\exp\left(-\frac{eV_{CB}}{k_BT}\right) - 1\right] \end{aligned} \tag{6.3.10}$$

The collector current components can be obtained by using the same approach. Thus we have

$$\begin{aligned} I_n^{BC} &= eAD_b \left. \frac{d\delta n_b(x_b)}{dx_b} \right|_{x_b=W_{bn}} && (6.3.11) \\ I_p^{BC} &= eAD_p \left. \frac{d\delta p(x_c)}{dx_c} \right|_{x_c=0} && (6.3.12) \end{aligned}$$

Using the results shown in the first part of equation 6.3.9 at $x_b = W_{bn}$, we have

$$\begin{aligned} I_n^{BC} = & - \frac{eAD_b n_{bo}}{L_b \sinh(W_{bn}/L_b)} \left[\exp\left(\frac{eV_{BE}}{k_B T} \right) - 1 \right] \\ & + \frac{eAD_b n_{bo}}{L_b} \coth\left(\frac{W_{bn}}{L_b} \right) \left[\exp\left(-\frac{eV_{CB}}{k_B T} \right) - 1 \right] \end{aligned} \qquad (6.3.13)$$

The hole current on the collector side is the same as for a reverse-biased p-n junction:

$$I_p^{BC} = -\frac{eAD_c p_{co}}{L_c} \left[\exp\left(-\frac{eV_{CB}}{k_B T} \right) - 1 \right] \qquad (6.3.14)$$

From the way we have defined the currents, the two current components flow along +x direction. If we define I_C as the total current flowing from the collector into the base, we have

$$\begin{aligned} -I_C = & \left[\frac{eAD_c p_{co}}{L_c} + \frac{eAD_b n_{bo}}{L_b} \coth\left(\frac{W_{bn}}{L_b} \right) \right] \left[\exp\left(-\frac{eV_{CB}}{k_B T} \right) - 1 \right] \\ & - \frac{eAD_b n_{bo}}{L_b \sinh\left(\frac{W_{bn}}{L_b} \right)} \left[\exp\left(\frac{eV_{BE}}{k_B T} \right) - 1 \right] \end{aligned} \qquad (6.3.15)$$

The base current is the difference between the emitter and collector currents: $(I_B = I_E - |I_C|)$. It is interesting to point out that if the base region W_{bn} is much smaller than the diffusion length, the electron gradient in the base region can be simplified by using the approximations

$$\begin{aligned} \sinh(\alpha) &= \frac{e^\alpha - e^{-\alpha}}{2} = \alpha + \frac{\alpha^3}{3!} + \frac{\alpha^5}{5!} + \cdots \\ \cosh(\alpha) &= \frac{e^\alpha + e^{-\alpha}}{2} = 1 + \frac{\alpha^2}{2!} + \frac{\alpha^4}{4!} \cdots \end{aligned}$$

For the forward active mode if we ignore the current flow in the reverse-biased BCJ we get

$$\begin{aligned} I_E = & \frac{-eAD_b n_{b0}}{L_b} \coth\left(\frac{W_{bn}}{L_b} \right) \left[\exp\left(\frac{eV_{BE}}{k_B T} \right) - 1 \right] \\ & - \frac{eAD_e p_{e0}}{L_e} \left[\exp\left(\frac{eV_{BE}}{k_B T} \right) - 1 \right] \end{aligned} \qquad (6.3.16)$$

Here the first part is due to electron injection from the emitter into the base (III and IV) and the second part is due to the hole injection from the base into the emitter (II). The collector current is

$$I_C = \frac{eAD_b n_{b0}}{L_b \sinh\left(\frac{W_{bn}}{L_b} \right)} \left[\exp\left(\frac{eV_{BE}}{k_B T} \right) - 1 \right] \qquad (6.3.17)$$

Assuming that $W_{bn} \ll L_b$, we can expand the hyperbolic functions as noted above. The base current is the difference between the emitter and collector current. We find that

$$I_B = \frac{eAD_e p_{e0}}{L_e} \left[\exp\left(\frac{eV_{BE}}{k_B T} \right) - 1 \right] + \frac{eAD_b n_{b0} W_{bn}}{2L_b^2} \left[\exp\left(\frac{eV_{BE}}{k_B T} \right) - 1 \right] \qquad (6.3.18)$$

The first part represents the hole current injected from the base into the emitter and the second part represents the hole current recombining with electrons injected from the emitter .

Having derived the current components, in the next section we will examine how material properties and doping levels can be manipulated to improve device performance. It is useful to recast the prefactor of the first term in the emitter current (equation 6.3.16) in a different form. The prefactor, which we will denote by I_S (we assume that $W_{bn} \ll L_b$ so that $\coth\alpha = 1/\alpha$), is

$$I_S = \frac{eAD_b n_{bo}}{W_{bn}} = \frac{e^2A^2D_b n_i^2}{eAN_{ab}W_{bn}} = \frac{e^2AD_b n_i^2}{eQ_G}$$

where Q_G is called the Gummel number for the transistor. It has a value

$$Q_G = N_{ab}W_{bn} \tag{6.3.19}$$

and denotes the charge in the base region of the device (assuming full ionization). As we will see later, the Gummel number has an important effect on device performance.

To understand the operation of a BJT as an amplifier or a switching device it is useful to examine the device under conditions of saturation, forward active (or reverse active), and cutoff. In figure 6.8 we show the band profile and the minority carrier distribution for each of these modes. Note that in saturation where both EBJ and BCJ are forward biased, a large minority carrier density (electrons for the npn device) is injected into the base region. This plays an important role in device switching, as will be discussed later. In the cutoff mode there is essentially no minority charge in the base, since the EBJ and BCJ are both reverse biased. In the forward active mode, the mode used for amplifiers, the EBJ is forward biased while the BCJ is reverse biased. Under this mode $I_C \gg I_B$, providing current gain.

6.3.2 BJT Biasing in circuits

The three terminal bipolar transistor can be biased in one of three different configurations shown in figure 6.9a. The configuration chosen depends upon the applications. As shown, one of the terminals can be chosen as a common terminal between the input and output terminals. The full I-V characteristics of a BJT in the common-base and the common-emitter configuration are shown in figure 6.9b. In the common-base configuration the cutoff mode occurs when the emitter current is zero. Note that the emitter current is finite, the collector current does not go to zero at V_{CB} = 0. The BCJ has to be forward biased at the turn on voltage ($\sim$ 0.7 V for Si devices) to balance the injected emitter current.

In the common-emitter mode, the cutoff mode occurs when the base current is zero and indicates the region where the EBJ is no longer forward biased. The saturation region is represented by the region where $V_{CE} = V_{BE}$ and both EBJ and BCJ are forward biased.

6.3.3 Current-Voltage: The Ebers-Moll Model

It is useful in circuit applications to represent the $I - V$ characteristics derived by us in terms of a simple physical model. Several models have been developed to do so. Here we will discuss

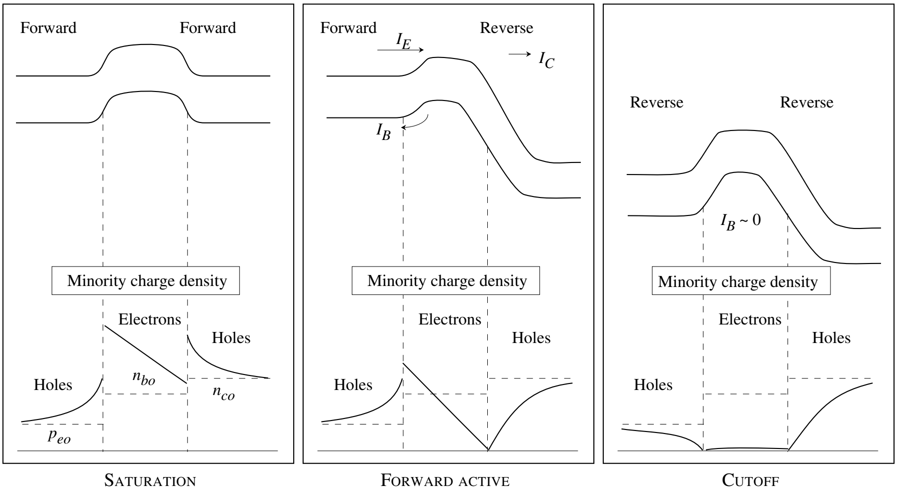


Figure 6.8: The band profile and minority charge distribution in a BJT under saturation, forward active, and cutoff modes.

the Ebers-Moll model. We can write the currents given in equation 6.3.10 and equation 6.3.17 as

$$I_E = -I_{ES}\left[\exp\left(\frac{eV_{BE}}{k_BT}\right)-1\right] + \alpha_R I_{CS}\left[\exp\left(-\frac{eV_{CB}}{k_BT}\right)-1\right] \tag{6.3.20}$$

$$I_C = \alpha_F I_{ES}\left[\exp\left(\frac{eV_{BE}}{k_BT}\right)-1\right] - I_{CS}\left[\exp\left(-\frac{eV_{CB}}{k_BT}\right)-1\right] \tag{6.3.21}$$

where

$$I_{ES} = \frac{eAD_bn_{bo}}{L_b}\coth\left(\frac{W_{bn}}{L_b}\right) + \frac{eAD_ep_{eo}}{L_e} \tag{6.3.22}$$

$$I_{CS} = \frac{eAD_cp_{co}}{L_c} + \frac{eAD_bn_{bo}}{L_b}\coth\left(\frac{W_{bn}}{L_b}\right) \tag{6.3.23}$$

$$\alpha_F I_{ES} = \alpha_R I_{CS} = \frac{eAD_bn_{bo}}{L_b\sinh\left(\frac{W_{bn}}{L_b}\right)} \tag{6.3.24}$$

Notice the symmetry underlying these equations. This symmetry allows us to develop a simple model described in figure 6.10. The parameter α_F represents the common-base current gain in the forward active mode, I_{CS} gives the reverse-bias BCJ current, α_R is the common-base current gain for the inverse active mode (i.e., EBJ is reverse biased and CBJ is forward biased) and I_{ES}

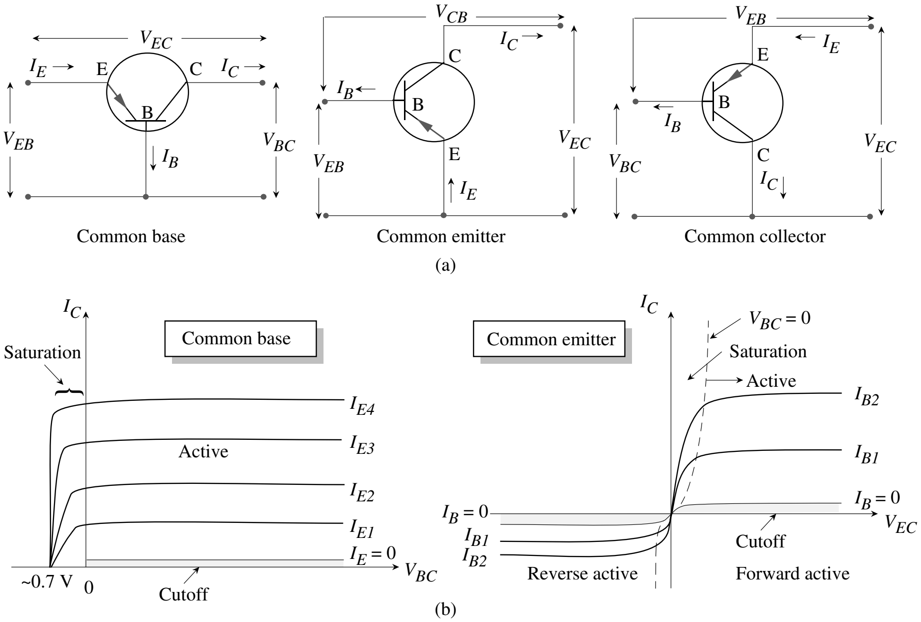


Figure 6.9: (a) Three possible configurations under which a BJT can be used in circuits. (b) A schematic of the current-voltage characteristics of a BJT in the common-base and common-emitter configuration.

gives the reverse-bias EBJ current. These equations represent two diodes that are coupled to each other. The Ebers-Moll model is primarily useful to develop a physical description of the bipolar device.

An important application of the Ebers-Moll model is to find the conditions for the saturation mode. In the common-emitter mode, the saturation condition is given by

$$V_{CE}(sub) = V_{BE} + V_{CB} = V_{BE} - V_{BC} \tag{6.3.25}$$

Note that both V_{BE} and V_{BC} (= $-V_{CB}$) are positive.

We also have the current conservation expression:

$$I_E + I_B + I_C = 0 \tag{6.3.26}$$

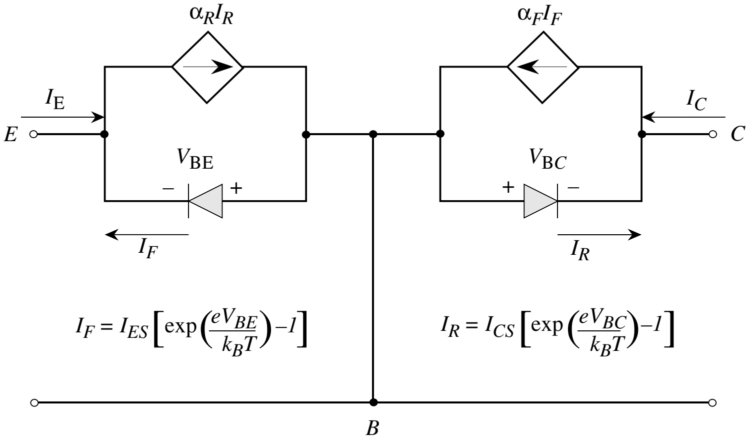

$\alpha_R I_R$
$\alpha_F I_F$
I_E
I_C
E
C
V_{BE}
V_{BC}
$-$
$+$
$+$
$-$
I_F
I_R
$I_F = I_{ES}\left[\exp\left(\frac{eV_{BE}}{k_BT}\right)-1\right]$
$I_R = I_{CS}\left[\exp\left(\frac{eV_{BC}}{k_BT}\right)-1\right]$
B

Figure 6.10: The Ebers-Moll equivalent circuit of a bipolar transistor looks at the device as made up of two coupled diodes.

Using this equation to eliminate I_E from equation 6.3.20, we can obtain the values of V_{BE} and V_{BC} in terms of I_C, I_B, and the parameters I_{ES}, I_{CS}, α_R, and α_F. This gives for $V_{CE(sat)}$

$$V_{CE(sat)} = V_{BE} - V_{BC} = \frac{k_BT}{e}\ell n\left[\frac{I_C(1-\alpha_R)+I_B}{\alpha_F I_B - (1-\alpha_F)I_C}\cdot\frac{I_{CS}}{I_{ES}}\right] \tag{6.3.27}$$

Substituting for I_{CS}/I_{ES} from equation 6.3.24, we get

$$V_{CE(sat)} = \frac{k_BT}{e}\ell n\left[\frac{I_C(1-\alpha_R)+I_B}{\alpha_F I_B - (1-\alpha_F)I_C}\cdot\frac{\alpha_F}{\alpha_R}\right] \tag{6.3.28}$$

Typical values of $V_{CE(sat)}$ are 0.1 to 0.2 V, as can be seen in example 6.2.

6.4 DEVICE DESIGN AND DEVICE PERFORMANCE PARAMETERS

In this section we will examine how device design influences performance of a BJT. Through material and geometric parameters we can control are doping densities, base width, device area, and in some cases material choice (e.g. Si or GaAs etc.). Usually it would be difficult to change the material system since it is difficult to alter the processing technology. The main performance parameters one wants to improve are the current gain, and device operation frequency. Additionally there are issues related to high voltage biasing that we will discuss later. We will focus on the forward active mode of the device so that we have the conditions

$$eV_{BE} \gg k_BT \tag{6.4.1}$$

$$eV_{CB} \gg k_BT \tag{6.4.2}$$

In a well-designed bipolar transistor we always have $W_b \ll L_b$.

Emitter Injection Efficiency

Bipolar transistor gain is intimately tied to emitter efficiency. The emitter injection efficiency is the ratio of the electron current (in the npn BJT) due to the electron injection from the emitter to the total emitter current. Thus,

$$\gamma_e = \frac{I_{En}}{I_{En} + I_{Ep}} \tag{6.4.3}$$

For high emitter efficiency, I_{Ep} should be minimal. Under the voltage approximations made we have from Eqns. 6.3.16 and 6.3.18,

$$\begin{aligned} I_{Ep} &= -\frac{eAD_ep_{eo}}{L_e} \exp\left(\frac{eV_{BE}}{k_BT}\right) && (6.4.4) \\ I_{En} &\cong -\frac{eAD_bn_{bo}}{L_b \tanh\left(\frac{W_{bn}}{L_b}\right)} \exp\left(\frac{eV_{BE}}{k_BT}\right) && (6.4.5) \end{aligned}$$

Thus the emitter efficiency becomes

$$\gamma_e = \frac{1}{1 + (p_{eo}D_eL_b/n_{bo}D_bL_e)\tanh\left(W_{bn}/L_b\right)} \tag{6.4.6}$$

If the base width is small compared to the electron diffusion length, the tanh (W_{bn}/L_b) can be replaced by (W_{bn}/L_b) and we have

$$\gamma_e \cong \frac{1}{1 + (p_{eo}D_eW_{bn}/n_{bo}D_bL_e)} \sim 1 - \frac{p_{eo}D_eW_{bn}}{n_{bo}D_bL_e} \tag{6.4.7}$$

Thus for γ_e to be close to unity, we should design the device so that $W_{bn} \ll L_e$ and $p_{eo} \ll n_{bo}$. Thus a small base width and a heavy emitter doping compared to the base doping are essential. Of course, the base width cannot be arbitrarily reduced.

Base Transport Factor

The second part of the device gain is related to how electrons injected from the emitter move over the base. The base transport factor is the ratio of the electron current reaching the base-collector junction to the current injected at the emitter-base junction. As the electrons travel through the base, they recombine with the holes so that the base transport factor is less than unity . We have from equation 6.3.16 and equation 6.3.17 (in the forward active mode, the collector current is essentially due to electron injection from the emitter)

$$B = \frac{I_C}{I_{En}} \cong \frac{1}{\cosh\left(\frac{W_{bn}}{L_b}\right)} \tag{6.4.8}$$

For small base width we have

$$B \cong 1 - \frac{W_{bn}^2}{2L_b^2} \tag{6.4.9}$$

Note that the base transport factor depends upon the neutral base width, not the chemical base width. Thus it depends upon the bias conditions. This causes the Early effect discussed later.

Collector Efficiency γ_c

The collector efficiency is the ratio of the electron current that reaches the collector to the base-collector current. Due to the high reverse bias at the base-collector junction, essentially all the electrons are swept into the collector so that the collector efficiency can be taken to be unity.

Current Gain

Since we know how the expression for the emitter efficiency and base transport factor we can now examine the current gain. We are primarily interested in the ratio of the collector current and the base current. The parameter α defined as the ratio of the collector current to the emitter current is given by

$$\begin{aligned} \alpha &= \frac{I_C}{I_E} = \frac{BI_{En}}{I_{En}+I_{Ep}} = \gamma_e B \\ &= \left[1 - \frac{p_{eo}D_e W_{bn}}{n_{bo}D_b L_e}\right]\left[1 - \frac{W_{bn}^2}{2L_b^2}\right] \end{aligned} \tag{6.4.10}$$

The ratio of the collector current to the base current is extremely important since it is the base current that is used to control the device state. This is given by

$$\beta = \frac{\alpha}{1-\alpha} \tag{6.4.11}$$

We can see that heavy emitter doping and narrow base width are critical for high β. An important parameter characterizing the device performance is the transconductance, which describes the control of the output current (I_C) with the input bias (V_{BE}). The transconductance is ($I_C \;\propto\; \exp(\mathrm{e}V_{BE}/k_BT)$)

$$g_m = \frac{\partial I_C}{\partial V_{BE}} = \frac{eI_C}{k_BT} = \frac{e\beta I_B}{k_BT} \tag{6.4.12}$$

The transconductance of bipolar devices is extremely high compared to that of field-effect transistors of similar dimensions. This is because of the exponential dependence of I_C on V_{BE} in contrast to a weaker dependence of current on “gate bias” for field effect transistors.

6.5 BJT DESIGN LIMITATIONS: NEED FOR BAND TAILORING AND HBTs

So far in this chapter we have assumed that the emitter, base, and collector are made from the same material, Of course this need not be the case since as we have noted in previous chapters heterostructures can be fabricated with ease. In this section we will see the tremendous advantages of using heterostructure concepts in bipolar transistors. In the BJT, once a material system is chosen the only flexibility one has in the device design is the doping levels and the device dimensions. This is not optimum for high-performance devices. Let us examine the material parameters controlling the device performance parameters. We have seen that for the narrow base width case

$$\alpha = \left[1 - \frac{p_{eo} D_e W_{bn}}{n_{bo} D_b L_e}\right]\left[1 - \frac{W_{bn}^2}{2L_b^2}\right] \tag{6.5.1}$$

and the current gain β is

$$\beta = \frac{\alpha}{1-\alpha} \tag{6.5.2}$$

We have already noted that for β to be high, it is essential that: (i) the emitter doping be much higher than the base doping, i.e., for an npn device ($n_{eo} \gg p_{bo}$); and (ii) the base width be as small as possible. In fact, the product $p_{bo}W_b$, called the Gummel number, should be as small as possible. However, a small base with relatively low doping (usually in BJTs $n_{eo} \sim 10^2$-$10^3 p_{bo}$) introduces a large base resistance, which adversely affects the device performance. From this point of view, the Gummel number should be as high as possible.

One may argue that the emitter should be doped as much as possible maintaining $n_{eo} \gg p_{bo}$ and yet having a high enough base doping to ensure low base resistance. However, a serious problem arises from the bandgap shrinking of the emitter region that is very heavily doped.

If we assume that hole injection across the EBJ is a dominant factor, the current gain of the device becomes

$$\beta = \frac{\alpha}{1-\alpha} \simeq \frac{n_{bo} D_b L_e}{p_{eo} D_e W_{bn}} \tag{6.5.3}$$

If the emitter bandgap shrinks by $|\Delta E_g|$ due to doping, the hole density for the same doping changes by an amount that can be evaluated using the change in the intrinsic carrier concentration,

$$n_{ie}\left(E_g - |\Delta E_g|\right) = n_{ie}\left(E_g\right) \exp\left(\frac{|\Delta E_g|}{2k_BT}\right) \tag{6.5.4}$$

where ΔE_g is positive in our case. Thus the value of p_{eo} changes as

$$\begin{aligned} p_{eo}\left(E_g - |\Delta E_g|\right) &\propto n_{ie}^2\left(E_g - |\Delta E_g|\right) &(6.5.5)\\ &= p_{eo}\left(E_g\right)\exp\left(\frac{|\Delta E_g|}{k_BT}\right) &(6.5.6)\end{aligned}$$

The bandgap decrease with doping is given for Si by (N_d is in units of cm^{-3})

$$|\Delta E_g| = 22.5\left(\frac{N_d}{10^{18}} \cdot \frac{300}{T(K)}\right)^{1/2} \text{ meV} \tag{6.5.7}$$

The expression is reasonable up to a doping of 10^{19}cm^{-3}. At higher doping levels, the bandgap shrinkage is not so large. For example, at a doping of 10^{20} $\text{cm}^{-3}$the shrinkage is $\sim$ 160 meV and not 225 meV as given by the equation above. As a result of the bandgap decrease, the gain of the device decreases according to the equation

$$\beta = \frac{D_b N_{de} L_e}{D_e N_{ab} W_{bn}} \exp\left(-\frac{|\Delta E_g|}{k_B T}\right) \tag{6.5.8}$$

were we have assumed full ionization, i.e.

$$\frac{p_{eo}}{n_{bo}} = \frac{N_{ab}}{N_{de}} \tag{6.5.9}$$

As a result of this for a fixed base doping, as the emitter doping is increased, initially the current gain increases, but then as bandgap shrinkage increases, the current gain starts to decrease. From the discussion above, it is clear that the conflicting requirements of heavy emitter doping, low base doping, small base width, etc., as shown in figure 6.11, cannot be properly met by a BJT in which the same bandgap semiconductor is used for the emitter and the base. This led Shockley and Kroemer in the 1950s to conceive of the heterojunction bipolar transistor (HBJT or HBT), where the emitter is made from a wide-gap material. In a typical HBT the emitter is made from a material that has a bandgap that is, say, > 0.2 eV larger than the bandgap of the material used in the base. Near the base side, the emitter material composition is graded so that there is a smooth transition in the bandgap from the emitter side to the base side. A typical example of an HBT structural layout is shown in figure 6.12a. In the case shown, the emitter material is AlGaAs, which has a larger bandgap than GaAs, used for the base and the collector. We have discussed the heterojunction in detail in section 5.6. There we realized that the maximum benefit is obtained by grading the E-B junction such that the full bandgap differential can be used.

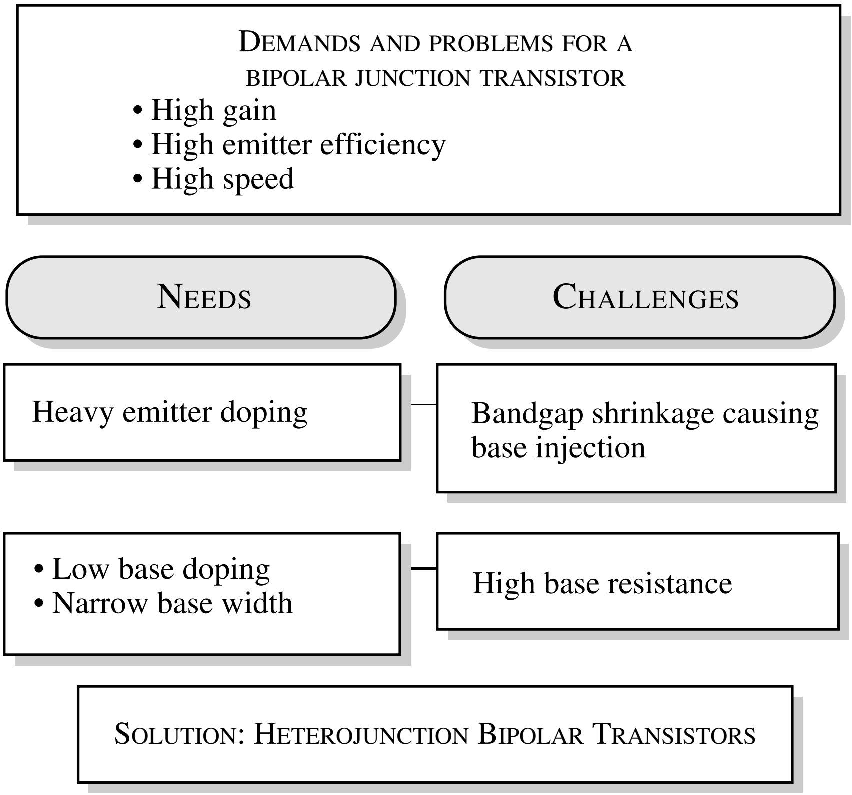
In figure 6.12b we show the band profile for the emitter and the base region. We can see that if ΔE_g is the bandgap difference between the emitter material bandgap and the base material bandgap, this difference appears across the valence band potential barrier, seen by holes. Thus, holes in the base see an increased barrier for injection into the emitter. As a result, the emitter efficiency dramatically increases. The suppression of hole injection current is given by

$$\frac{I_{Ep}(HBT)}{I_{Ep}(BJT)} = \exp\left(\frac{-\Delta E_g}{k_B T}\right)$$

The gain β in the device increases by the exponential factor. We have for β in an HBT

$$\beta = \frac{D_b N_{de} L_e}{D_e N_{ab} W_{bn}} \exp\left(\frac{\Delta E_g}{k_B T}\right)\left[1 - \frac{w_{bn}^2}{L_b^2}\right] \tag{6.5.10}$$

Typically the values of $\Delta E_g/k_B T$ are $\sim$ 10, so that β improves by$\sim$ 10^4. Due to the heavy doping now allowed in wide emitter HBTs, the base can be made narrow without too large a base resistance or the danger of punch through. This also avoids secondary effects such as Kirk effect and Early effect discussed later.



DEMANDS AND PROBLEMS FOR A BIPOLAR JUNCTION TRANSISTOR

- High gain
- High emitter efficiency
- High speed

NEEDS

CHALLENGES

Heavy emitter doping

Bandgap shrinkage causing base injection

- Low base doping
- Narrow base width

High base resistance

SOLUTION: HETEROJUNCTION BIPOLAR TRANSISTORS

Figure 6.11: Figure showing the conflicting requirements for high-performance BJTs. Heterostructure devices offer reconciliation of all these requirements.

Example 6.1 Consider an npn GaAs BJT that has a doping of $N_{de} = 5 \times 10^{17}$ cm^{-3}, $N_{ab} = 10^{17}$ cm^{-3}. Compare the emitter efficiency of this device with that of a similarly doped HBT where the emitter is Al$_{0.3}$Ga$_{0.7}$As and the base is GaAs. The following parameters characterize the devices at 300 K:

Electron diffusion constant in the base,	D_b	$=$	100 cm^{2}s^{-1}
Hole diffusion constant in the emitter,	D_e	$=$	15 cm^{2}s^{-1}
Base width,	W_b	$=$	0.5 μm
Bandgap discontinuity,	ΔE_g	$=$	0.36 eV
Minority carrier length for holes,	L_e	$=$	1.5 μm

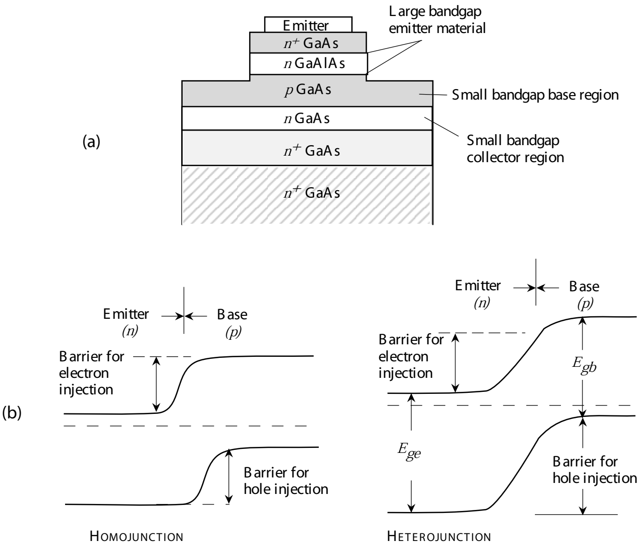

(a)
Emitter
n^+ GaAs
n GaAlAs
p GaAs
n GaAs
n^+ GaAs
n^+ GaAs
Large bandgap emitter material
Small bandgap base region
Small bandgap collector region
(b)
Emitter (n)
Base (p)
Barrier for electron injection
Barrier for hole injection
E_{gb}
E_{ge}
HOMOJUNCTION
HETEROJUNCTION

Figure 6.12: (a) A structural schematic of a heterojunction bipolar transistor made from GaAs and AlGaAs. (b) The band profile for a homojunction and heterojunction transistor. In the HBT, grading is used to avoid a potential "notch".

For GaAs we have the emitter and base minority carrier concentrations

$$\begin{aligned} p_{eo} &= \frac{n_i^2}{N_{de}} = \frac{(2.2\times10^6)^2}{5\times10^{17}} = 9.7\times10^{-6}\ \mathrm{cm}^{-3} \\ n_{bo} &= \frac{n_i^2}{N_{ab}} = \frac{(2.2\times10^6)^2}{10^{17}} = 4.84\times10^{-5}\ \mathrm{cm}^{-3} \end{aligned}$$

The emitter efficiency is

$$\begin{aligned} \gamma_e = 1 - \frac{p_{eo}D_eW_b}{n_{bo}D_bL_e} &= 1 - \frac{(9.7\times10^{-6})(15)(0.5\times10^{-4})}{(4.84\times10^{-5})(100)(1.5\times10^{-4})} \\ &= 0.99 \end{aligned}$$

In the HBT, the value of p_{eo} is greatly suppressed. The new value is approximately

$$\begin{aligned} p_{eo}(Al_{0.3}Ga_{0.7}As) &= \frac{n_i^2(GaAs)}{N_{de}}\exp\left(-\frac{\Delta E_g}{k_BT}\right) \\ &= p_{eo}(GaAs)\exp\left(-\frac{\Delta E_g}{k_BT}\right) = 9.4\times10^{-12}\ \mathrm{cm}^{-3} \end{aligned}$$

In this case the emitter efficiency is essentially unity.

6.5.1 The Generalized Moll-Ross Relationship

This very important relationship first developed by Moll and Ross, and subsequently generalized by Kroemer, is derived in this section. The Moll-Ross Relationship links the collector current density to the applied base-emitter voltage V_{BE} and to the Gummel number Q_G. It is a very powerful relationship, since it shows that the nature of the doping in the base is inconsequential as far as the output current is concerned. Rather, the total number of dopants in the base is the controlling factor. Kroemer's generalization expands this to heterostructure devices.

Let us assume an n-p-n transistor with a high current gain such that the hole current $J_p \approx 0$. In the base of the transistor, we can write

$$J_n = \mu_n n \frac{dE_{Fn}}{dx} \tag{6.5.11}$$

and

$$J_p = \mu_p p \frac{dE_{Fp}}{dx} \tag{6.5.12}$$

Since J_p is assumed to be approximately zero, $\frac{dE_{Fp}}{dx} \approx 0$, and so equation 6.5.12 can be rewritten as

$$J_n = \mu_n n \frac{d}{dx}\left(E_{Fn} - E_{Fp}\right) \tag{6.5.13}$$

Inserting Einstein's relationship

$$\mu_n = D_n \frac{e}{k_B T} \tag{6.5.14}$$

and using the relations

$$E_{Fn} - E_i = k_B T \ln\left(\frac{n}{n_i}\right) \tag{6.5.15}$$

and

$$E_i - E_{Fp} = k_B T \ln\left(\frac{p}{n_i}\right) \tag{6.5.16}$$

we get

$$\frac{d}{dx}\left[\ln\left(\frac{np}{n_i^2}\right)\right] = \frac{J_n}{eD_n n} \tag{6.5.17}$$

or

$$d\left[\ln\left(\frac{np}{n_i^2}\right)\right] = \frac{J_n}{e} \cdot \frac{p}{D_n n_i^2} \cdot dx \tag{6.5.18}$$

Let us integrate equation 6.5.18 from $x = 0$ to the edge of the neutral base $x = W_{bn}$.

$$\left.\frac{np}{n_i^2}\right|_{W_{bn}} - \left.\frac{np}{n_i^2}\right|_{x=0} = -\left.\frac{np}{n_i^2}\right|_{x=0} = \frac{J_n}{e}\int_0^{W_{bn}} \frac{p(x)}{D_n n_i^2} dx \tag{6.5.19}$$

Here, because of Shockley boundary conditions, we have assumed $np/n_i^2|_{W_{bn}}$ can be neglected. The quantity $np/n_i^2|_{x=0}$ is given by the law of the junction

$$\left.\frac{np}{n_i^2}\right|_{x=0} = \exp\left(\frac{eV_{BE}}{k_B T}\right) \tag{6.5.20}$$

This leads to the generalized Moll-Ross relation

$$\boxed{J_n = -\frac{e \cdot \exp\left(\frac{eV_{BE}}{k_B T}\right)}{\int_0^{W_{bn}} \left(\frac{p(x)}{D_n n_i^2}\right) dx}} \tag{6.5.21}$$

where the integration is over the extent of the neutral base. In the case where n_i and D_n are constant throughout the base, or equivalently the material is homogeneous,

$$J_n = -\frac{eD_n n_i^2}{\int_0^{W_{bn}} p(x)dx} \exp\left(\frac{eV_{BE}}{k_B T}\right) = -\frac{eD_n n_i^2}{Q_G} \exp\left(\frac{eV_{BE}}{k_B T}\right) \tag{6.5.22}$$

where

$$Q_G = \int_0^{W_{bn}} p(x)dx \simeq \int_0^{W_{bn}} N_a dx \tag{6.5.23}$$

is the Gummel number, defined as the total number of acceptor atoms in the neutral base.

6.5.2 How much β do we need?

This question is very important, but it really has no universal answer. Different applications have different minimum tolerances for β . This will be illustrated in the four examples shown below. Because an understanding of these applications requires some knowledge of bipolar frequency response, it is recommended that the reader examine chapter 7 before reading this section. We thank Prof. Mark Rodwell for discussions on this topic.

Microwave power amplifiers

In figure 6.13, we show a basic BJT small-signal model. As derived in chapter 7,

$$C_{in} = C_\pi + C_{BE} = \frac{\tau_B + \tau_C}{r_e} + C_{BE} \tag{6.5.24}$$

We will assume that $C_\pi >> C_{BE}$, so that C_{in} can be written as

$$C_{in} \approx \frac{\tau_B + \tau_C}{r_e} = (\tau_B + \tau_C)\, g_m \tag{6.5.25}$$

At a small signal frequency ω, the input current I_{in} is given by

$$I_{in} = \left\{ j\omega \left[(\tau_B + \tau_C)\, g_m\right] + \frac{g_m}{\beta} \right\} V_{in} \tag{6.5.26}$$

For an efficient transistor, one wants the first term in this expression to dominate, or

$$\omega\,(\tau_B + \tau_C) > \frac{1}{\beta} \tag{6.5.27}$$

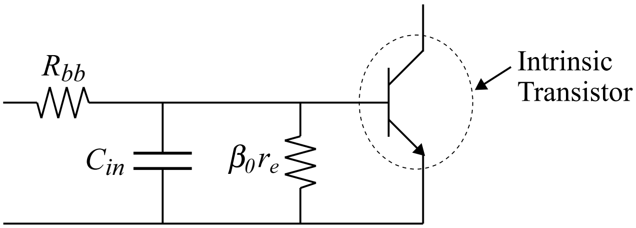


Figure 6.13: Basic BJT small-signal model for microwave power amplifiers.

If $\tau_B + \tau_C$ is the dominant delay, then we may assume $\tau_B + \tau_C \simeq (\omega_T)^{-1}$. Equation 6.5.27 can then be written as

$$\frac{\omega}{\omega_T} > \frac{1}{\beta} \tag{6.5.28}$$

Since power amplifiers are rarely operated at frequencies a factor of 20 below the transit frequency, $\beta \sim 20$ is in most instances adequate for these applications.

Microwave low noise amplifiers

The noise figure of low noise amplifiers is determined by the shot noise at the input. For BJTs this is the shot noise of the base current $2eI_B$, while for FETs it is that of the gate current $2eI_G$, were I_B is the base current of the BJT and I_G is the reverse leakage current of the gate diode of the FET. Since the forward bias base current $I_B = I_C/\beta$ is typically much larger than the reverse bias current I_G, the minimum noise figure attainable at low frequencies $(f << f_T)$ is limited by β, as shown in figure 6.14. Hence a high β (typically $\beta > 100$) is desirable for low noise applications.

Logic applications

To explain the relevant issues in logic circuits, we will use a representative logic family, CML. Let us assume one gate driving n gates, connected as shown in figure 6.15. Also assume that node **A** is at a high. In the absence of base current, the difference between the voltage high and the voltage low $\Delta V_L = I_o R_L$. However, in the presence of base current, since node **C** is high, a current nI_B is sourced by the node. Hence the voltage at node **C** is $V_{CC} - nI_B R_L$. This reduces the logic difference to

$$\Delta V_L = -\left(V_{CC} - I_o R_L\right) + \left(V_{CC} - nI_B R_L\right) = \left(I_o - nI_B\right) R_L = I_o \left(1 - \frac{n}{\beta}\right) R_L \tag{6.5.29}$$

To provide adequate noise margin, it is necessary that

$$\frac{n}{\beta} << 1$$

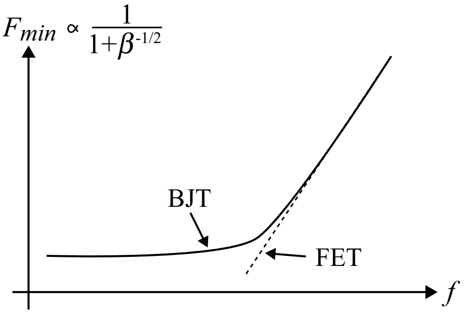


Figure 6.14: Minimum noise figure versus frequency for a BJT.

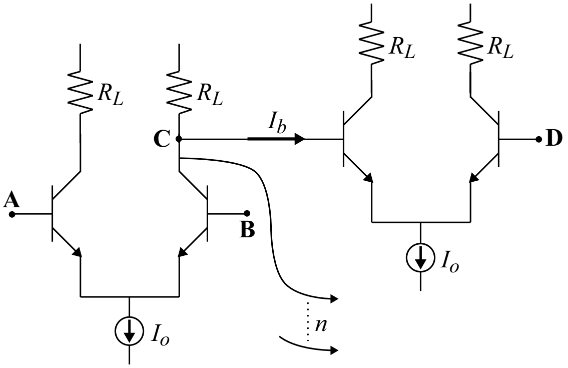


Figure 6.15: One gate driving n gates in a CML circuit.

Typically, $\beta \sim 50$ is desirable in such applications.

Flash analog-to-digital converters

An m-bit flash ADC is shown in figure 6.16 to illustrate the need for high β's in comparators using BJT-based differential amplifiers. These architectures are based on using $N = 2^m$ resistors in a reference ladder connected to a reference voltage V_{ref} and comparing each node voltage to the input voltage V_{in}. If the input voltage is between V_j and V_{j+1}, the comparators A_1 through A_j will produce a 1 at their output, and the rest will produce a zero. This output is connected via a decoder to a digital output. It is imperative that the voltage at any node, say V_j, be a predictable

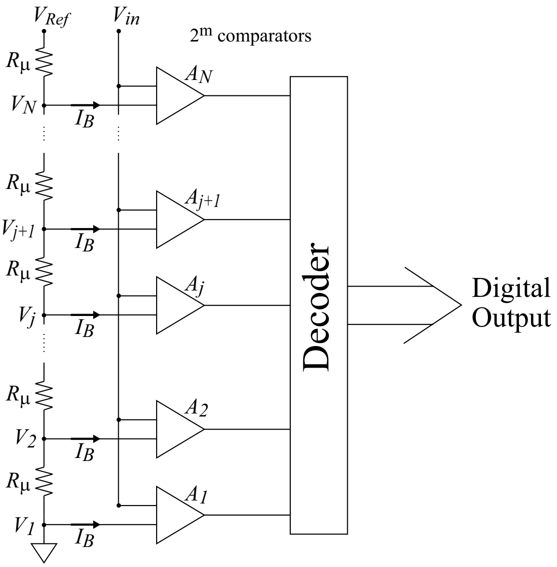


Figure 6.16: Circuit diagram of an m-bit flash analog-to-digital converter, $N = 2^m$.

function of the number of resistors, i.e.

$$V_j = \frac{(j-1)}{N} V_{ref}$$

However, the base current flowing into the comparators causes a deviation from this linear behavior. It is clear that the nodes 1 and N have minimum deviation, since they are proximal to the voltage supplies. However, as the nodes progress away from 1 and N toward the center of the array, the deviation increases because of a continuously increasing fraction of base current I_B drawn by the comparators. The maximum deviation is thus instinctively understood to be at the center node $j = N/2$ and is

$$\langle \Delta V \rangle = \frac{1}{8} N^2 R_u I_B \tag{6.5.30}$$

Therefore I_B should be reduced as much as possible, and hence β should be maximized. $\beta > 100$ is desirable for such applications.

6.6 SECONDARY EFFECTS IN REAL DEVICES

In the derivations of the bipolar device characteristics, we have made a number of simplifying assumptions. There are important secondary effects that make the device characteristics deviate from those derived so far. These deviations have important effects on circuit design as well as on the limits of device performance.

6.6.1 High Injection: The Kirk Effect

As will be shown in our high frequency analysis of the bipolar transistor in chapter 7, in order to achieve high frequency device operation, it is essential to operate the device at high current density. The reason for this in essence is that many important delays in the transistor have their origin in charging capacitances of the form

$$C = \frac{\epsilon A_j}{w_d} \tag{6.6.1}$$

where A_j is the area of the capacitor (typically the area of the p-n junction) and w_d is the junction depletion depth. Delays in the device are of the form

$$\tau = r_j \cdot C \tag{6.6.2}$$

where

$$r_j = \frac{\partial V}{\partial I} = \frac{k_B T}{eI} \tag{6.6.3}$$

is the ac resistance of the junction. The delay time τ can therefore be written as

$$\tau = \frac{k_B T}{eI} \cdot \frac{\epsilon A_j}{w_d} = \frac{k_B T}{e} \frac{\epsilon}{w_d} \cdot \frac{1}{J} \tag{6.6.4}$$

where $J = I/A_j$ is the current density. Thus it is imperative to increase J if one needs to reduce τ and hence increase the maximum device operating frequency.

There is, however, a maximum current density that the device can be operated at, above which the β of the transistor and the device frequency response drop catastrophically. Essentially, once the current density reaches this maximum value, the effective base length (i.e. the length between the emitter and the collector which electrons must diffuse across) becomes wider as a result of space-charge injection into the collector. This phenomenon is known as the Kirk Effect, and the associated current density at which it occurs is called the Kirk Threshold , J_{Kirk}. We will now explain why this occurs.

The basic analysis of bipolar transistors carried out in this chapter involved applying Shockley boundary conditions at the reverse biased base-collector junction. Under this assumption, the minority carrier density drops to zero at the collector edge of the base region and is zero everywhere within the base-collector depletion region. This of course is physically not possible, because having zero minority carriers within the junction requires carriers to travel at extremely high velocities as dictated by current continuity.

$$J_C = en_{p,C}v_e \tag{6.6.5}$$

$$\begin{aligned} \text{as} \quad n_{p,C} &\rightarrow 0 \\ v_e &\rightarrow \infty \end{aligned}$$

where $n_{p,C}$ is the electron concentration at the base-collector junction. Since in reality the bulk velocity in the base will saturate at some value v_s, $n_{p,C}$ cannot drop to zero, but instead drops to a value

$$n_{p,C} = \frac{J_C}{ev_s} \tag{6.6.6}$$

Because electrons in the depleted collector all travel at the saturated velocity v_s, the injected carrier concentration in the collector will everywhere be equal to $n_{p,C}$, and the net charge density in the collector

$$N_{C,net} = N_{d,C} - n_{p,C} \tag{6.6.7}$$

The resulting charge profile, as well as electric field profile in the collector region, for a device under dc bias with collector doping $N_{d,C}$ is shown in figure 6.17. Here we have assumed that the device is biased such that the n^- collector region is fully depleted (this is typically the case for bipolars in modern circuits). As indicated in figure 6.17c, the slope of the electric field in the collector region is given by

$$\frac{d\mathcal{E}}{dx} = \frac{eN_{C,net}}{\epsilon} \tag{6.6.8}$$

We will now consider the effect that increasing the collector current density J_C has on the charge distribution and electric field in the structure. We assume that the voltage across the base-collector junction maintains a constant value V_{CB}, implying that the total potential drop across the junction is $V_{CB} + V_{bi}$. Under this assumption, the area underneath the electric field curve in figure 6.17c always maintains a constant value $V_{CB} + V_{bi}$. The voltage drops in the p^+ and n^+ regions at the edge of the base-collector depletion region are very small relative to the voltage in the n^- layer, and hence the area of the shaded region in figure 6.17c is assumed to be $V_{CB} + V_{bi}$. Equivalently, the total base-collector depletion depth $w_{d,BC}$ can be assumed to be approximately equal to the collector width w_C.

As J_C increases, the injected charge density in the collector $n_{p,C}$ must increase to maintain current continuity, as indicated by equation 6.6.6. This causes the net charge in the collector $N_{C,net}$ to decrease (equation 6.6.7). Hence the slope of the electric field profile in the collector, which is proportional to $N_{C,net}$, decreases. At the critical current density

$$J_{crit} = eN_{d,C}v_e \tag{6.6.9}$$

the injected mobile charge $n_{p,C}$ exactly balances the ionized donor charge $N_{d,C}$, resulting in zero net charge as well as a constant electric field in the collector (see figure 6.18). Concurrently, the depletion region depth at the base edge x_{pC} decreases, since the electric field at the base edge of the collector must decrease in order for the shaded area in figure 6.17c to remain constant. Similarly, the depletion region in the n^+ subcollector x_{nC} increases, as the region has to terminate a higher electric field.

As $J_C > J_{crit}$, the slope of the electric field reverses sign, as $N_{C,net} = N_{d,C} - n_{p,C}$ is now negative. This process continues until another critical current threshold is reached, when the

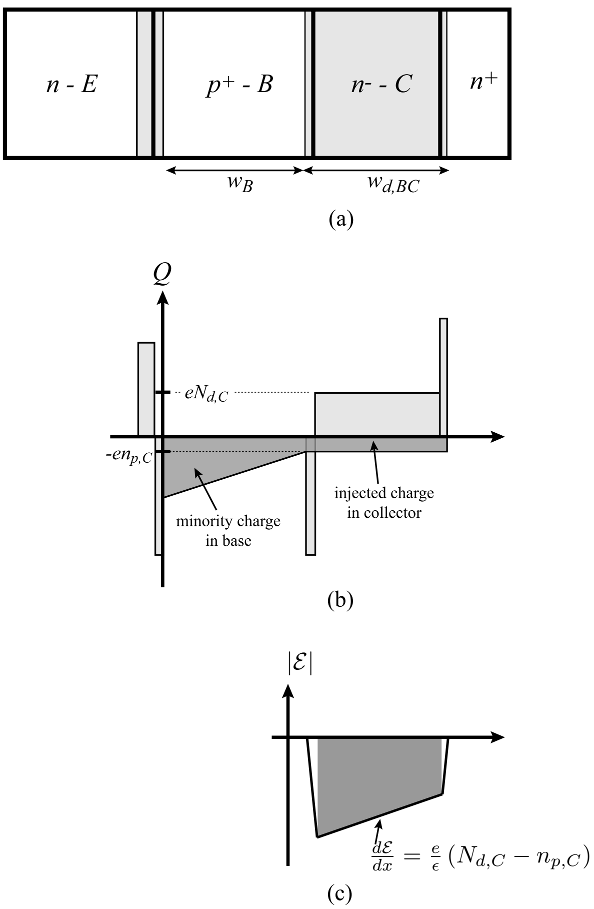


Figure 6.17: (a) Schematic diagram of a typical bipolar transistor structure. The EBJ is forward biased, and the BCJ is reverse biased such that the collector is fully depleted. (b) Corresponding charge profile in the device. The space charge in the device results from depletion charge (light gray) and injected mobile charge (dark gray). (c) Electric field in the collector region.

electric field at the base-edge reaches zero (and equivalently the depletion region width $x_{pC} = 0$). The current density when this condition occurs is called the Kirk threshold current density and is given the symbol J_{Kirk}. Its value can be readily calculated by solving equation 6.6.6-equation 6.6.8 subject to the conditions that the electric field at the base edge of the collector is

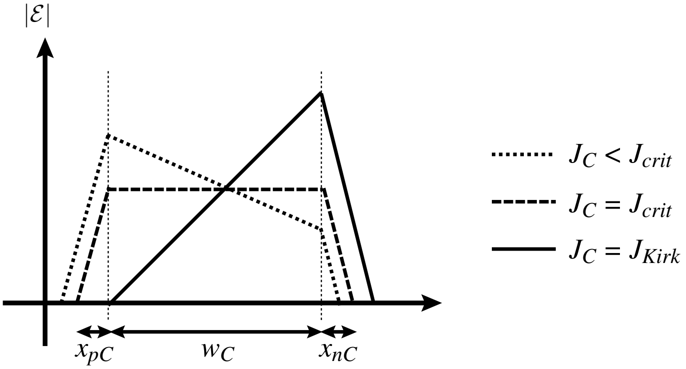


Figure 6.18: Electric field profile in the collector when $J_C < J_{crit}$, $J_C = J_{crit}$, and $J_C = J_{Kirk}$. The depletion extension into the base and subcollector are labeled x_{pC} and x_{nC}, respectively. x_{pC} and x_{nC} shown in this figure correspond to the profile with $J_C = J_{crit}$.

zero and the area underneath the electric field is $V_{CB} + V_{bi}$. From that we get the following set of equations.

$$\frac{1}{2}\mathcal{E}_{max} w_C = V_{CB} + V_{bi} \tag{6.6.10}$$

$$\frac{\mathcal{E}_{max}}{w_C} = \frac{d\mathcal{E}}{dx} = \frac{e}{\epsilon}\left(n_{Kirk} - N_{d,C}\right) \tag{6.6.11}$$

$$n_{Kirk} = \frac{J_{Kirk}}{ev_s} \tag{6.6.12}$$

Combining these equations and solving for J_{Kirk} gives us

$$\boxed{J_{Kirk} = \left[\frac{2\epsilon}{ew_C^2}\left(V_{CB} + V_{bi}\right) + N_{d,C}\right] ev_s} \tag{6.6.13}$$

Let us examine the consequence of reaching the Kirk threshold. If one assumed that the process of electric field modification with increasing current density described in this section were to continue, then a situation occurs where the direction of the electric field reverses in a region w_{rev} near the base-collector edge, as shown in figure 6.19a. This corresponds to the band diagram shown in figure 6.19b. However, this is an unphysical situation for homojunctions, as there is no blocking barrier for holes. Hence the holes from the base would flood the collector to achieve the physical situation of the holes being contained within a region Δw_B in the collector. Here, they neutralize the injected negative charge, resulting in zero electric field within this region. This is shown in figure 6.20.

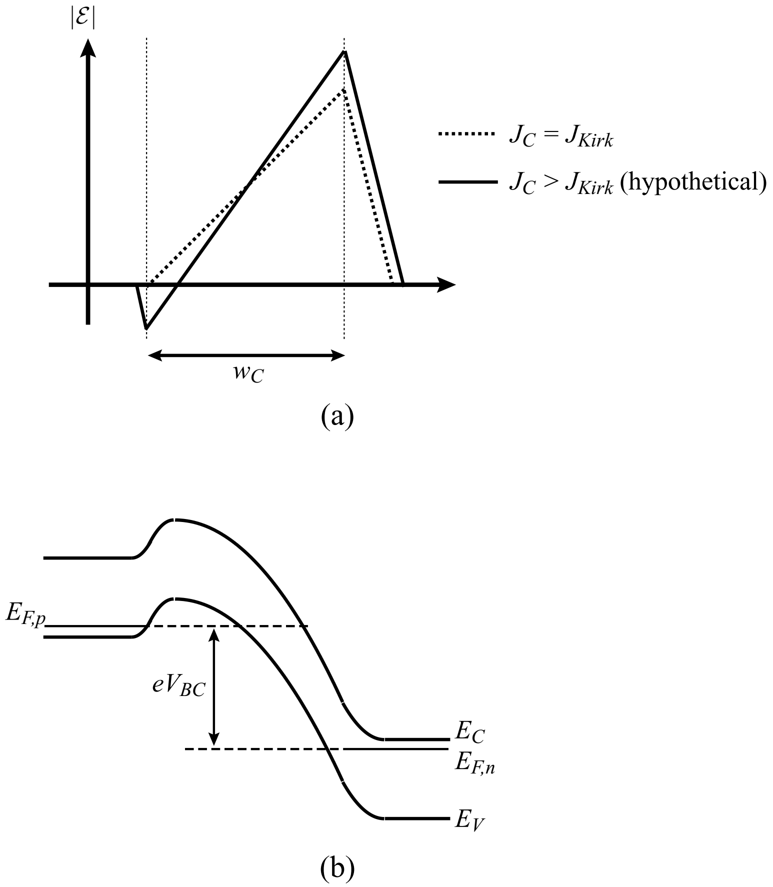


Figure 6.19: (a) Hypothetical electric field profile in the collector if the electric field were to continue evolving as described earlier after $J_C > J_{Kirk}$. (b) Band diagram corresponding to $J_C > J_{Kirk}$ above.

Now, electrons diffusing across the base must diffuse over a distance $w_B + \Delta w_B$ before they are swept into the collector by the electric field in the depletion region. This effective extension of the base region is called base widening. The consequence is a rapid increase in base recombination or a decrease in β for $J_C > J_{Kirk}$. It it therefore critical that J_{Kirk} be maximized. From equation 6.6.13, we can see that this can be achieved by

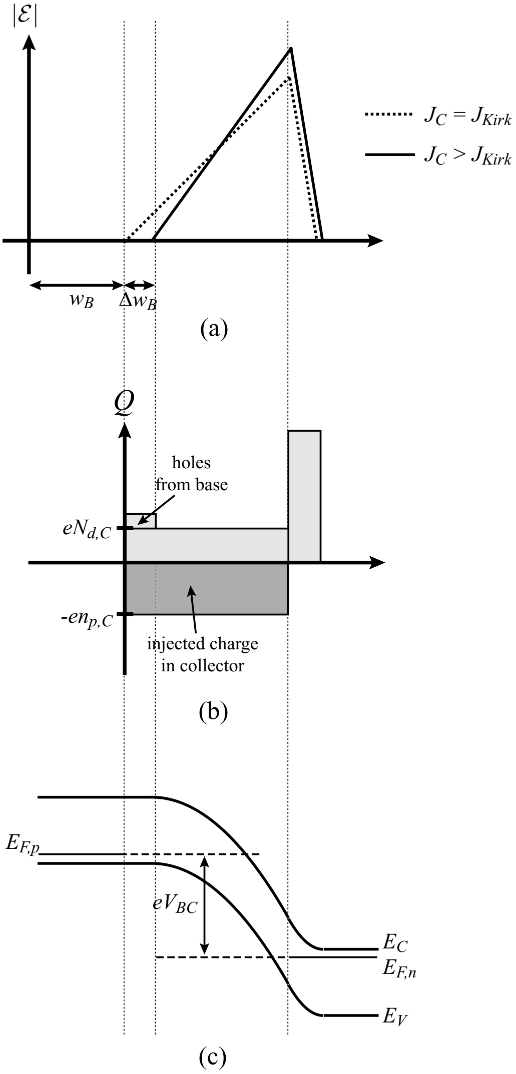


Figure 6.20: (a) Electric field profile, (b) charge profile, and (c) band diagram in the collector region when $J_C > J_{Kirk}$.

1. Minimizing w_C
2. Maximizing $N_{d,C}$
3. Maximizing v_s

Increasing $N_{d,C}$ and decreasing w_C both lead to higher electric fields in the collector, thus decreasing the breakdown voltage. Therefore, materials with larger bandgaps and higher electron velocity characteristics (such as InP) are preferred for collector regions.

In figure 6.1c, we showed the current-gain cutoff frequency f_T versus drain current J_C for a state-of-the-art InP-based bipolar transistor. Indeed, we see that initially, f_T increases with J_C. However, as J_C becomes larger, f_T saturates and eventually begins to drop once $J_C > J_{Kirk}$.

6.6.2 High Injection: Thermal Effects

At high injection levels there is thermal heating of the bipolar device since a power level of $I_C V_{BC}$ is dissipated. As the temperature of the device changes the current usually increases further. This is due to the exponential dependence of the injection current and the increase in the recombination time (which increases faster than the increase in the base transit time), which increases the base transport factor. As the current increases, a further increase in heat dissipation occurs until the device can be burnt out if proper design considerations are not met.

6.6.3 Base Width Modulation: The Early Effect

As seen clearly in our discussion of the Kirk Effect, the quantity W_{bn} that appeared in our derivation for the current-voltage characteristics is not the metallurgical base width W_b, but the distance which electrons must diffuse before they are swept into the collector by large electric fields. For $W_{bn} \ll L_b$ we can see from equation 6.3.17 that the collector current

$$I_C \propto \frac{1}{W_{bn}}$$

We saw that W_{bn} can be affected by the collector current density I_C for high injection conditions. Additionally, W_{bn} (and therefore I_C) also vary with the applied base-collector bias V_{BC}. This non-ideal behavior is known as the Early Effect. The physical reason for it is illustrated in figure 6.21. An increase in V_{BC} at a fixed emitter-base voltage results in an increase in the base-collector depletion width and a subsequent decrease in W_{bn}. This results in an increase in the slope of the minority carrier profile in the base, resulting in an increased collector current I_C. As can be seen in figure 6.21b, the Early Effect results in a finite output resistance of the transistor, which we will now show is

$$R_o = \frac{V_A}{I_C} \tag{6.6.14}$$

where V_A is called the Early voltage.

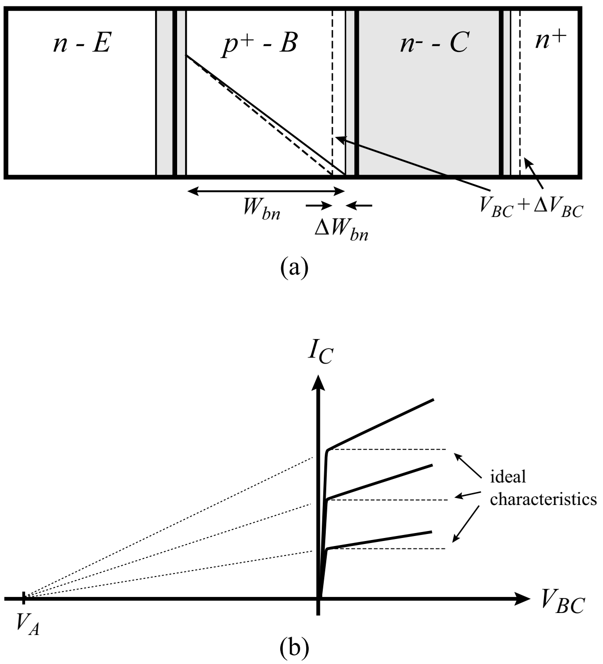


Figure 6.21: (a) Slope of the electron profile in the base increases when V_{BC} is increased, resulting in an increase in I_C. (b) Device dc I-V characteristics when Early Effect is accounted for.

Using the Moll-Ross relationship in equation 6.5.22, the variation of I_C with respect to V_{BC} can be written as

$$\begin{aligned} \frac{\partial I_C}{\partial V_{BC}} &= -eD_n n_i^2 A_E \exp\left(\frac{eV_{BE}}{k_B T}\right) \cdot \frac{\partial}{\partial x}\left(\frac{1}{\int_0^{W_{bn}} p(x)dx}\right) && (6.6.15) \\ &= +eD_n n_i^2 A_E \exp\left(\frac{eV_{BE}}{k_B T}\right) \cdot \frac{p(W_{bn})}{\left[\int_0^{W_{bn}} p(x)dx\right]^2} \frac{\partial(W_{bn})}{\partial V_{BC}} && (6.6.16) \end{aligned}$$

Collecting terms, we can simplify this expression to

$$\frac{\partial I_C}{\partial V_{BC}} = -I_C \left[p(W_{bn}) \cdot \frac{1}{\int_0^{W_{bn}} p(x)dx} \cdot \frac{\partial W_{bn}}{\partial V_{BC}}\right] = -\frac{I_C}{V_A} = \frac{I_C}{|V_A|} \qquad (6.6.17)$$

where

$$V_A = \frac{\int_0^{W_{bn}} p(x)dx}{p(W_{bn})\frac{\partial W_{bn}}{\partial V_{BC}}} \tag{6.6.18}$$

is defined as the Early voltage. It is clear that V_A is a bias-dependent quantity and hence is best defined for a particular base-collector voltage, which is chosen to be $V_{BC} = 0$. It turns out that for heavily doped base regions (used in most high performance devices), the variation of V_A with V_{BC} is small.

Let us now study the expression for V_A. The quantity

$$\int_0^{W_{bn}} p(x)dx$$

was defined as the Gummel number $Q_G(cm^{-2})$. If we take the derivative of Q_G with respect to V_{BC}, we get

$$\frac{\partial Q_G}{\partial V_{BC}} = \frac{\partial}{\partial V_{BC}} \left(\int_0^{W_{bn}} p(x)dx \right] = p(W_{bn}) \left(\frac{\partial W_{bn}}{\partial V_{BC}} \right) \tag{6.6.19}$$

So V_A can be rewritten as

$$V_A = \frac{Q_G}{\partial Q_G / \partial V_{BC}} \tag{6.6.20}$$

The change in base charge with respect to V_{BC} is by definition the base-collector depletion capacitance C_{BC}, or

$$C_{BC} = \frac{e \partial Q_G}{\partial V_{BC}} \ (Fcm^{-2}) \tag{6.6.21}$$

Thus V_A can be written as

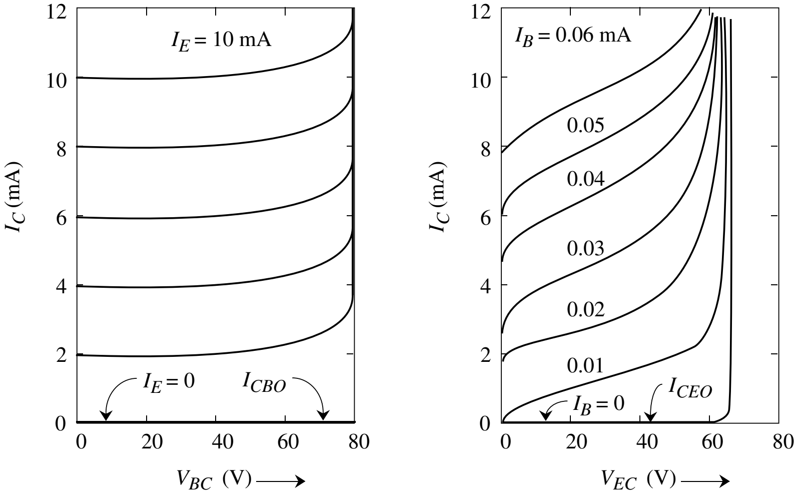
$$V_A = \frac{eQ_G}{C_{BC}} \tag{6.6.22}$$

Both Q_G and C_{BC} are measured at $V_{BC} = 0$. Variations in Q_G due to changes in V_{BC} are considered negligible, giving a constant V_A independent of bias. In actuality, the output conductance always increases with V_{BC} because the decrease in C_{BC} with bias tends to be smaller than the decrease in Q_G, since C_{BC} is determined dominantly by the depletion layer thickness in the collector.

To minimize the output conductance, or equivalently increase the Early voltage, one must increase the Gummel number Q_G and decrease C_{BC}. The path with least penalty is to increase Q_G, because a decrease in C_{BC} is equivalent to an increase in the collector depletion region thickness, which in high frequency transistors may result in an unacceptable collector transit delay.

6.6.4 Drift Effects in the Base: Nonuniform Doping

We have assumed so far that the base doping is uniform and consequently there is no built-in electric field in the base region. In real devices the doping can be quite nonuniform, especially if the doping is done by ion implantation. The nonuniform doping causes a built-in field that can



COMMON BASE:
During breakdown, emitter current is not affected

COMMON EMITTER:
At very low multiplication, base current starts to increase, causing the emitter current to rise, and the collector current starts to run away

Figure 6.22: Avalanche breakdown related characteristics of a bipolar transistor in the common-base and common-emitter configurations.

help or hinder the carriers injected into the base from the emitter. Of course, if the doping can be made non-uniform in a controlled manner, it can be exploited to shorten the base transit time.

6.6.5 Avalanche Breakdown

Just as in the case of the p-n diode, the avalanche process limits the collector-base voltage that the transistor can sustain. This then sets the limit on the power that can be obtained by the transistor. The breakdown due to the impact ionization (avalanching) is reflected in the I-V characteristics of the transistor in a manner shown in figure 6.22. In the common-base configuration, the breakdown occurs at a well defined collector-base voltage BV_{CBO}. On the other hand, for the common-emitter configuration, the breakdown is not as sharply reflected in the device output characteristics. The breakdown in the common-emitter configuration also occurs at a lower value of V_{CE} than it does in the common-base configuration.

In the common-base configuration, as V_{CB} is increased, the breakdown is essentially similar to that of a single p-n junction discussed in chapter 4. The current coming from the emitter has little effect on the breakdown. However, in the common-emitter configuration, as soon as the impact ionization process starts, say, in an npn BJT, the secondary holes are injected into the base and act as a base current, leading to increased emitter current and eventual current runaway as the process snowballs.

6.6.6 Low Injection Effects and Current Gain

In our calculations for the BJT junction currents we have assumed that in the space charge region, the junctions are "ideal," i.e., there is no current flow due to recombination-generation effects. In chapter 5, we discussed how non-ideal effects arising from recombination generation in the depletion region alter the current flowing in the junction. This effect is particularly important under low injection (i.e., low values of V_{EB}) conditions.

If we examine the forward-biased EBJ for a device operating in the forward active mode, the base current will have an "ideal" current component and a "non-ideal" current component arising from generation-recombination. We can write

$$I_B = \frac{eAD_e n_{oe}}{L} \exp\left(\frac{eV_{EB}}{k_BT}\right) + \frac{eAn_iW_{EBJ}}{2\tau} \exp\left(\frac{eV_{EB}}{2k_BT}\right)$$

where the second term is due to recombination in the emitter-base junction depletion region (W_{EBJ}). The recombination time is τ. The base current may be written as

$$I_B = I_S \exp\left(\frac{eV}{mk_BT}\right)$$

where m is the junction ideality factor.

The collector current is not greatly influenced by the recombination-generation process. At low injection the recombination-generation part of the base current dominates and as a result, the current gain β is reduced. As the injection (V_{EB} value) is increased, the recombination part becomes negligible and the value of β reaches its ideal value calculated earlier.

In section 6.3.3 we discussed the Ebers-Moll model for bipolar transistors. This model does not account for some of the issues discussed in this section. A more advanced model that includes more realistic effects is the Gummel-Poon model. Three important effects are incorporated in the Gummel-Poon model:

- Recombination current in the emitter depletion region under low injection levels.
- Reduction of current gain at high injection levels.
- Finite output conductance in terms of an Early voltage, V_A.

6.6.7 Current Crowding Effect

The picture we have developed for the BJT is a one-dimensional picture. In reality, the base current flows along the directions perpendicular to the emitter, as can be seen from figure 6.6.

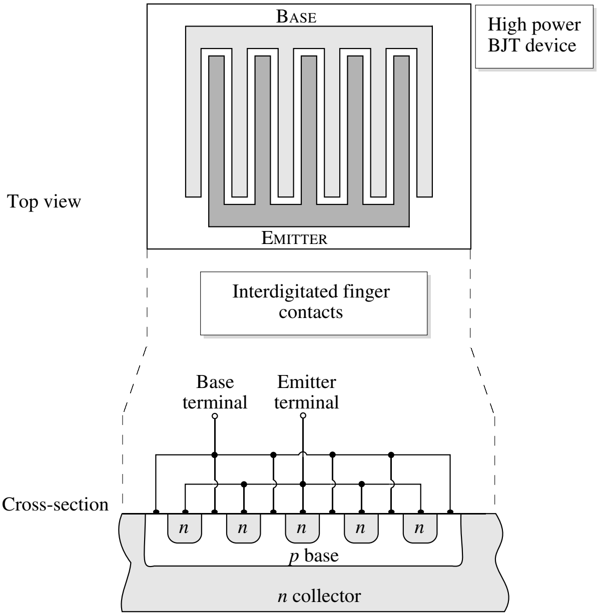


Figure 6.23: Top view and the cross-section of a typical device using an interdigitated emitter.

There is a voltage drop IR across the base cross-section that becomes increasingly important at high injections and high frequencies. As a result of this potential drop, the edge of the emitter may be forward biased but the "core" of the emitter region may not be forward biased. Higher current densities would thus flow along the edges of the emitter. This effect is called emitter crowding and, because of it, the high injection effects discussed above can be important even at low total current values.

Emitter crowding has an adverse effect on power transistors where high current values are required. It is essential for these transistors that the emitter be properly designed. Computer simulation techniques are used to study the current flow so that an optimum emitter can be used. The emitter crowding effects can be suppressed by increasing the perimeter-to-area ratio of the emitter. This is often done by using long fingers for the emitter and base contacts in the "interdigitated" approach shown in figure 6.23.

Example 6.2 Consider an npn silicon transistor at 300 K with a base doping of 5×10^{16} cm^{-3} and a collector doping of 5×10^{15} cm^{-3}. The width of the base region is 1.0 μm. Calculate the change in the base width as V_{CB} changes from 1.0 to 5.0 V. Also calculate how the collector current changes and determine the Early voltage. Assume that $D_b = 20$ cm^2/s, V_{BE} = 0.7 V, and $W_b \ll L_b$.

The depletion width at the base-collector junction is shared between the base and the collector region. The extent of depletion into the base side is given by

$$\Delta W_b = \left\{ \frac{2\epsilon_s (V_{bi} + V_{CB})}{e} \left(\frac{N_{dc}}{N_{ab}(N_{ab} + N_{dc})} \right) \right\}^{1/2}$$

The built-in voltage V_{bi} is given by

$$\begin{aligned} V_{bi} &= \frac{k_B T}{e} \ln \frac{N_{ab} N_{dc}}{n_i^2} = (0.026(V)) \times \ell n(1.1 \times 10^{12}) \\ &= 0.721 \text{ V} \end{aligned}$$

For an applied bias of 1 V, we get

$$\begin{aligned} \Delta W_b &= \left\{ \frac{2 \times (8.84 \times 10^{-14} \times 11.9 \text{ F/cm})(1.72 \text{ V})}{(1.6 \times 10^{-19} \text{ C})} \times \frac{1}{5.5 \times 10^{17} \text{ cm}^{-3}} \right\}^{1/2} \\ &= 6.413 \times 10^{-6} \text{ cm} = 0.064 \, \mu\text{m} \end{aligned}$$

The neutral base width is thus

$$W_{bn} = 0.936 \, \mu\text{m}$$

When the collector-base voltage increases to 5 V, we get

$$\Delta W_b = 0.117 \, \mu\text{m}$$

The neutral base width is

$$W_{bn} = 0.883 \, \mu\text{m}$$

In the limit of $W_{bn} \gg L_b$ we have for the collector current density (using equation 6.3.17 with sinh $(W_{bn}/L_b) \sim W_{bn}/L_b$)

$$J_C = \frac{e D_b n_{bo}}{W_{bn}} \text{ exp } \left(\frac{eV_{BE}}{k_B T} \right)$$

where

$$n_{bo} = \frac{n_i^2}{N_{ab}} = \frac{2.25 \times 10^{20}}{5 \times 10^{16}} = 4.5 \times 10^3 \text{ cm}^{-3}$$

For the base-collector bias of 1 V, we have $W_{bn} = 0.936 \, \mu$m. The collector current density is then

$$\begin{aligned} J_C &= \frac{(1.6 \times 10^{-19} \text{ C}) \times (20 \text{ cm}^2 \text{ s}^{-1}) \times (4.5 \times 10^3 \text{ cm}^{-3})(4.93 \times 10^{11})}{(0.936 \times 10^{-4} \text{ cm})} \\ &= 75.8 \text{ A/cm}^2 \end{aligned}$$

When the collector-base bias changes to 5 V, the current density becomes

$$J_C = 80.35 \text{ A/cm}^2$$

The slope of the J_C vs. V_{CE} curve is then

$$\frac{dJ_C}{dV_{CE}} \cong \frac{\Delta J_C}{\Delta V_{CE}} \cong \frac{80.35 - 75.8}{4} \text{ Acm}^{-2}\text{ V}^{-1}$$

We define the Early voltage V_A through the relation

$$\frac{dJ_C}{dV_{CE}} = \frac{J_C}{V_{CE} + V_A}$$

Equating the two relations, we get

$$V_A \cong 64.9 \text{ V}$$

Example 6.3 Consider a npn Si-BJT at 300 K with the following parameters:

$$\begin{aligned}
N_{de} &= 10^{18}\text{ cm}^{-3}\\
N_{ab} &= 10^{17}\text{ cm}^{-3}\\
N_{dc} &= 5\times 10^{16}\text{ cm}^{-3}\\
D_b &= 30.0\text{ cm}^2\text{s}^{-1}\\
L_b &= 15.0\ \mu\text{m}\\
D_e &= 10.0\text{ cm}^2\text{s}^{-1}\\
L_e &= 5.0\ \mu\text{m}
\end{aligned}$$

(i) Calculate the maximum base width, W_b, that will allow a current gain β of 100 when the EBJ is forward biased at 1.0 V and the BCJ is reverse biased at 5.0 V.
(ii) Describe two advantages and two disadvantages of making the base smaller.

Since the base width is much smaller than the base diffusion length, we have for the emitter and collector current

$$I_E = \left[\frac{eAD_bn_{b0}}{W_{bn}} + \frac{eAD_ep_{e0}}{L_e}\right]\left[\exp(\frac{eV_{BE}}{k_BT}) - 1\right]$$

$$I_C = \left[\frac{eAD_bn_{b0}}{W_{bn}}\right]\left[\exp(\frac{eV_{BE}}{k_BT}) - 1\right]$$

The base current is the difference of these and the current gain β is

$$\beta = \frac{I_C}{I_B} = \frac{D_bn_{b0}L_e}{D_ep_{e0}W_{bn}} = 100$$

This gives for W_{bn}

$$W_{bn} = 1.5\times 10^{-4}\text{ cm}$$

This is the neutral base width. The actual base width will be larger and we need to calculate the depletion on the base side at the BCJ due to the biasing of the device. Since the EBJ is strongly forward biased, there is essentially no depletion of the base at this junction.

The built-in voltage on the BCJ is

$$V_{bi} = \frac{k_B T}{e} \ln\left(\frac{N_{ab}N_{dc}}{n_i^2}\right) = 0.8 \text{ V}$$

Using the V_{bi} value we find that the depletion width on the base side of the EBJ for a 5 volt bias at the base collector junction is

$$\Delta W(V = 5 \text{ V}) = 1.59 \times 10^{-5} \text{ cm}$$

and the base width becomes

$$W_b = W_{bn} + 1.59 \times 10^{-5} = 1.659 \times 10^{-4} \text{ cm}$$

(ii) Two disadvantages of a shorter base:
• The output conductance will suffer and the collector current will have a stronger dependence on V_{CB}.
• The device may suffer punch through at a lower bias.

Two advantages:
• The current gain will be higher.
• The device speed will be faster.

Example 6.4 Consider a npn Si-BJT at 300 K with the following parameters:

$$\begin{aligned}
N_{de} &= 10^{18} \text{ cm}^{-3}\\
N_{ab} &= 10^{17} \text{ cm}^{-3}\\
N_{dc} &= 10^{16} \text{ cm}^{-3}\\
D_b &= 30.0 \text{ cm}^2\text{s}^{-1}\\
L_b &= 10.0\ \mu\text{m}\\
W_b &= 1.0\ \mu\text{m}\\
D_e &= 10 \text{ cm}^2\text{s}^{-1}\\
L_e &= 10.0\ \mu\text{m}\\
\text{Emitter thickness} &= 1.0\ \mu\text{m}\\
\text{Device area} &= 4.0 \times 10^{-6} \text{ cm}^2
\end{aligned}$$

Calculate the emitter efficiency and gain β when the EBJ is forward biased at 1.0 V and the BCJ is reverse biased at 5.0 V.

Calculate the output conductance of the device defined by

$$g_o = \frac{\Delta I_C}{\Delta V_{CB}}$$

To solve this problem we need to calculate the neutral base width in the device. Also note that since the emitter thickness is small compared to the carrier diffusion length in the emitter, we will use the narrow diode theory to calculate the emitter efficiency. .

Using the parameters given, the built-in voltage in the BCJ is

$$V_{bi} = \frac{k_BT}{e} \ln\left(\frac{10^{17}.10^{16}}{2.25 \times 10^{20}}\right) = 0.757 \text{ V}$$

The depletion width on the base side of the BCJ is found to be

$$\delta W(5.0 \text{ V}) = 8.296 \times 10^{-6} \text{ cm}$$

and

$$\delta W(6.0 \text{ V}) = 8.981 \times 10^{-6} \text{ cm}$$

Thus the neutral base width is

$$W_{bn}(5.0 \text{ V}) = 9.17 \times 10^{-5} \text{ cm}$$

The emitter efficiency is (for a narrow emitter of width W_e)

$$\gamma_e = 1 - \frac{p_{e0}D_eW_{bn}}{n_{b0}D_bW_e} = 0.969$$

We find that the base transport factor is

$$B = 1 - \frac{W_{bn}^2}{2L_b^2} = 0.996$$

This gives

$$\alpha = \gamma_e B = 0.9656$$

and the current gain is

$$\beta = \frac{\alpha}{1-\alpha} = 28$$

The collector current is

$$I_C = \frac{eAD_bn_{b0}}{W_{bn}}\left[\exp\left(\frac{eV_{BE}}{k_BT}\right) - 1\right] - \frac{eAD_bn_{b0}W_{bn}}{2L_b^2}\left[\exp\left(\frac{eV_{BE}}{k_BT}\right) - 1\right]$$

with the second part being negligible.

We find that

$$I_C(5.0 \text{ V}) = 23.79 \text{ A}$$

We now calculate the neutral base width when the BCJ is reverse biased at 6.0 V. This is

$$W_{bn}(6.0\ V) = 9.1 \times 10^{-5} \text{ cm}$$

This gives

$$I_C(6.0\ V) = 23.973 \text{ A}$$

The output conductance is now

$$g_o = 0.183\ \Omega^{-1}$$

6.7 PROBLEMS

Temperature is 300 K unless otherwise specified.

- **Section 6.3**

Problem 6.1 An npn HBT, shown in figure 6.24, is illuminated leading to an optical generation of $10^{20}\frac{1}{cm^3 s}$. $\tau_N = \tau_P = 1ns$. Assume Shockley boundary conditions.

1. Assume that all the light is absorbed in the collector depletion region. How will I_B, I_C, and I_E be different from the case where there is no illumination?
2. Now assume that all the light is absorbed in the (short) neutral base region. Again explain how I_B, I_C, and I_E be different from the case where there is no illumination? Assume that the emitter injection efficiency is 1.

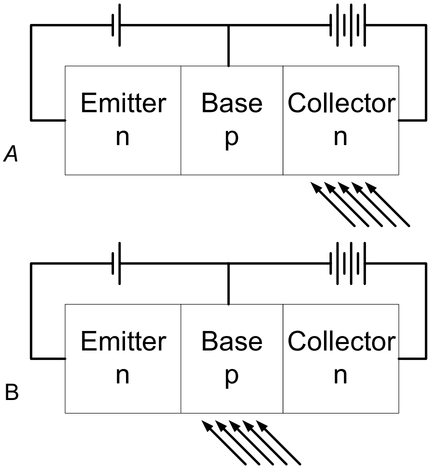


Figure 6.24: Figure for problem 6.1.

Problem 6.2 Consider an npn transistor with the following parameters:

$$\begin{aligned} D_b &= 20\ \mathrm{cm}^2\ \mathrm{s}^{-1} & D_e &= 10\ \mathrm{cm}^2\ \mathrm{s}^{-1} \\ N_{de} &= 5 \times 10^{18}\ \mathrm{cm}^{-3} & N_{ab} &= 5 \times 10^{16}\ \mathrm{cm}^{-3} \\ N_{dc} &= 5 \times 10^{5}\ \mathrm{cm}^{-3} & W_b &= 1.0\ \mu\mathrm{m} \\ \tau_B = \tau_E &= 10^{-7}\ \mathrm{s} & n_i^2 &= 2.25 \times 10^{20}\ \mathrm{cm}^{-6} \\ A &= 10^{-2}\ \mathrm{cm}^2 \end{aligned}$$

Calculate the collector current in the active mode with an applied emitter base bias of 0.5 V. What is the collector current when the base current is now increased by 20%?

Problem 6.3 An Si npn transistor at 300 K has an area of 1 mm^2, base width of 1.0 μm, and doping of $N_{de} = 10^{18}$ cm^{-3}, $N_{ab} = 10^{17}$ cm^{-3}, $N_{dc} = 10^{16}$ cm^{-3}. The minority carrier lifetimes are $\tau_E = 10^{-7} = \tau_B; \tau_C = 10^{-6}$ s. Calculate the collector current in the active mode for (a) V_{BE} = 0.5 V, (b) $I_E = 2.5$ mA, and (c) $I_B = 5\ \mu A$. The base diffusion coefficient is D_b = 20 cm^{2}s^{-1}.

Problem 6.4 An npn silicon transistor is operated in the inverse active mode (i.e., collector-base is forward biased and emitter base is reverse biased). The doping concentrations are $N_{de} = 10^{18}$ cm^{-3}; $N_{ab} = 10^{17}$ cm^{-3}, and $N_{dc} = 10^{16}$ cm^{-3}. The voltages are V_{BE} = -2 V, V_{BC} = 0.6 V. Calculate and plot the minority carrier distribution in the device. Also calculate the current in the collector and the emitter. The device parameters are: $W_b = 1.0\ \mu$m, $\tau_E = \tau_B = \tau_C = 10^{-7}$s, $D_b = 20$ cm^{2}s^{-1}, $D_e =$ 10 cm^{2}s^{-1}, $D_c = 25$ cm^{2}s^{-1}, A = 1 mm^2.

Problem 6.5 Calculate the error made in the emitter efficiency expression (i.e., equation 6.4.7 versus equation 6.4.6) when one makes the approximation given in the text for tanh. Obtain the error as a function of the ratio of L_b to W_{bn}.

Problem 6.6 Plot the dependence of the base transport factor in a bipolar transistor as a function of W_b/L_b over the range $10^{-2} \leq W_b/L_b \leq 10$. Assume that the emitter efficiency is unity. How does the common-emitter current gain vary over the same range of W_b/L_b?

Problem 6.7 In an npn bipolar transistor, calculate and plot the dependence of the emitter efficiency on the ratio of N_{ab}/N_{de} in the range $10^{-2} \leq N_{ab}/N_{de} \leq 1$. Calculate the results for the cases: (a) $D_e = D_b, L_e = L_b, W_b = L_b$, and (b) $D_e = 0.2D_b, L_e = 0.2L_b, W_b = 0.1L_b$.

Problem 6.8 In a uniformly doped npn bipolar transistor, the following current values are measured (see figure 6.6 for the current definitions):

$$\begin{array}{rcll} I_{En} & = & 1.2\,\text{mA} & \quad \mathrm{I_{Ep}} = 0.1\ \text{mA} \\ I_C = I_{nC} & = & 1.19\,\text{mA} & \quad \mathrm{I_{BE}^{R}} = 0.1\ \text{mA} \end{array}$$

Determine the parameters α, β, γ_e for the transistor.

Problem 6.9 Consider an npn bipolar transistor at 300 K with the following parameters:

$$\begin{array}{lll} N_{de} = 5 \times 10^{18}\ \text{s}; & N_{ab} = 5 \times 10^{16}\ \text{cm}^{-3}; & N_{dc} = 10^{15}\ \text{cm}^{-3} \\ D_e = 10\ \text{cm}^2\ \text{s}; & D_b = 15\ \text{cm}^2\ \text{s}; & D_c = 20\ \text{cm}^2\ \text{s} \\ \tau_E\ = 10^{-8}\ \text{s}; & \tau_B\ = 10^{-7} s; & \tau_C\ = 10^{-6}\ \text{s} \\ W_b = 1.0\ \mu\text{m}; & A\ = 0.1\ \text{mm}^2 & \end{array}$$

Calculate the emitter current and the collector current as well as the values of α, γ_e, β for V_{BE} = 0.6 V; V_{CE} = 5 V.

Problem 6.10 The mobility of holes in silicon is 100 cm^2/V·s. It is required that a BJT be made with a base width of 0.5 μm and base resistivity of no more than 1.0 Ω-cm. It is also desired that the emitter injection efficiency be at least 0.999. Calculate the emitter doping required. The various device parameters are

$$\begin{aligned} L_b &= 10\,\mu\text{m} \\ L_e &= 10\,\mu\text{m} \\ D_e &= 10\,\text{cm}^2\text{s}^{-1} \\ D_b &= 20\,\text{cm}^2\text{s}^{-1} \end{aligned}$$

What is the current gain β of the device? Assume $W_{bn} = W_b$.

Problem 6.11 Consider a npn Si-BJT at 300 K with the following parameters:

$$\begin{aligned} N_{de} &= 10^{18}\ \text{cm}^{-3} \\ N_{ab} &= 10^{17}\ \text{cm}^{-3} \\ N_{dc} &= 5 \times 10^{16}\ \text{cm}^{-3} \\ D_b &= 30.0\ \text{cm}^2\text{s}^{-1} \\ L_b &= 15.0\,\mu\text{m} \\ D_e &= 10.0\,\text{cm}^2\text{s}^{-1} \\ L_e &= 5.0\,\mu\text{m} \end{aligned}$$

(i) Calculate the maximum base width, W_b, that will allow a current gain β of 100 when the EBJ is forward biased at 1.0 V and the BCJ is reverse biased at 5.0 V.
(ii) Describe two advantages and two disadvantages of making the base smaller.

Problem 6.12 The $V_{CE}(sat)$ of an npn transistor decreases as the base current increases for a fixed collector current. In the Ebers-Moll model, assume α_F = 0.995, α_R = 0.1, and I_C = 1.0 mA. At 300 K, at what base current is the $V_{CE}(sat)$ value equal to (a) 0.2 V, (b) 0.1 V?

Problem 6.13 Consider an npn bipolar device in the active mode. Express the base current in terms of $\alpha_F, \alpha_R, I_{ES}, I_{CS}$, and V_{BE}, using the Ebers-Moll model.

Problem 6.14 Derive the expressions for the emitter and collector current for a pnp transistor in analogy with the equations derived in the text for the npn transistor.

Problem 6.15 An npn silicon bipolar device has the following parameters in the Ebers-Moll model at 300 K:

$$\alpha_F = 0.99;\ \alpha_R = 0.2$$

Calculate the saturation voltage V_{CE} for I_C = 5 mA and I_B = 0.2 mA. Why is I_C/I_B not equal to $\alpha_F/(1-\alpha_F)$?

• Section 6.5

Problem 6.16 Consider an npn silicon bipolar transistor in which $W_b = 2.0\ \mu$m, $L_e = L_b = 10.0\ \mu$m, and $D_e = D_b = 10$ cm^{2}s^{-1}. Assume that $N_{ab} = 10^{16}$ cm^{-3}. What is the emitter injection efficiency for $N_{de} = 10^{18}, 10^{19}$ and 10^{20} cm^{-3} when (a) bandgap narrowing is neglected, (b) bandgap narrowing is included?

Problem 6.17 A silicon npn bipolar transistor is to be designed so that the emitter injection efficiency at 300 K is γ_e = 0.995. The base width is 0.5 μm and $L_e = 10W_b, D_e = D_b$, and $N_{de} = 10^{19}$ cm^{-3}. Calculate the base doping required with and without bandgap narrowing effects.

Problem 6.18 Consider a GaAs/AlGaAs HBT in which an injector efficiency of 0.999 is required at 300 K. The emitter and base doping are both 10^{18} cm^{-3}. The base width is 0.1 μm. The carrier diffusion coefficients are $D_b = 60$ cm^{2}s, and $D_e = 20$ cm^{2}s. The carrier lifetimes are $\tau_B = \tau_E = 10^{-8}$s. Calculate the Al fraction needed in the emitter of the HBT.

Problem 6.19 Due to the high base doping possible, the base of an HBT can be very narrow. Consider a GaAs/AlGaAs HBT where the GaAs base is 500 Å. The minority charge diffusion coefficient is 100 cm^2/V·s in the base. Calculate the base transit time limited cutoff frequency of this device.

Problem 6.20 Consider an n-p-n bipolar transistor where the base is graded from Al$_{0.04}$Ga$_{0.96}$As at the emitter end to GaAs at the collector end. The emitter material is Al$_{0.22}$Ga$_{0.78}$As and it is graded to Al$_{0.04}$Ga$_{0.96}$As at the base.
(a) Sketch the detailed band diagram of the HBT in equilibrium, and under forward active bias. What is the quasi-electric field, $\boldsymbol{\mathcal{E}}_{\text{quasi,B}}$, in the base?
(b) Solve the drift-diffusion equation to obtain an expression for the base minority carrier concentration, $n(x)$, in terms of the total current J_C, $\boldsymbol{\mathcal{E}}_{\text{quasi,B}}$, and W_B. Assume μ = 1000 cm^2/Vs.

Problem 6.21 Due to an error during growth, the emitter-base grade in the transistor shown in the figure 6.25 below was started after the emitter n-type dopant cell shutter was opened. As a result, there was a thin n-type GaAs region between the p-type base and the grade.

1. Construct the equilibrium band diagram of this structure taking into account quasi-electric and electrostatic fields. What is the β of this transistor at zero emitter-base bias?
2. Now, the transistor is operated under high-injection conditions. If an emitter-base bias of 1V is applied, how will the β be affected? Draw field and potential profiles to explain your result.

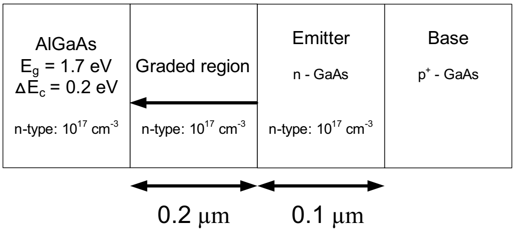


Figure 6.25: Figure for problem 6.21.

- **Section 6.6**

Problem 6.22 A silicon pnp transistor at 300 K has a doping of $N_{ae} = 10^{18}$ cm^{-3}, $N_{db} = 5 \times 10^{16}$ cm^{-3}, $N_{ac} = 10^{15}$ cm^{-3}. The base width is 1.0 μm. The value of D_b is 10 cm^2/s and $\tau_B = 10^{-7}$ s. The emitter base junction is forward biased at 0.7 V. Using the approximation that the minority carrier distribution in the base can be represented by a linear decay, calculate the hole diffusion current density in the base at (a) V_{CB} = 5 V (reverse bias), (b) V_{CB} = 15 V.

Problem 6.23 A uniformly doped npn bipolar transistor is fabricated to within $N_{de} = 10^{19}$ cm^{-3} and $N_{dc} = 10^{16}$ cm^{-3}. The base width is 0.5 μm. Design the base doping so that the punch through voltage is at least 25 V in the forward active mode.

Problem 6.24 An npn silicon bipolar transistor has a base doping of 10^{16} cm^{-3} and a heavily doped collector region. The neutral base width is 1.0 μm. What is the base collector reverse bias when punch through occurs?

Problem 6.25 The punch through voltage of a Ge pnp bipolar transistor is 20 V. The base doping is 10^{16} cm^{-3}, and the emitter and collector doping are 10^{18} cm^{-3}. Calculate the zero bias base width. If $\tau_B = 10^{-6}$ s, what is the α of the transistor at a 10 V reverse bias across the collector-base junction at 300 K? The hole diffusion coefficient in the base is 40 cm^{2}s^{-1}.

Problem 6.26 In a silicon npn transistor, the doping concentrations in the emitter and collector are $N_{de} = 10^{18}$ cm^{-3} and $N_{dc} = 5 \times 10^{15}$ cm^{-3}, respectively. The neutral base width is 0.6 μm at V_{BE} = 0.7 V and V_{CB} = 5 V. When V_{CB} is increased to 10 V, the minority carrier diffusion current in the base increases by 5%. Calculate the base doping and the Early voltage if $D_b = 20$ cm^2/s and $\tau_B = 5 \times 10^{-7}$s.

Problem 6.27 Consider a npn Si-BJT at 300 K with the following parameters:

$$\begin{aligned}
N_{de} &= 10^{18}\ \mathrm{cm}^{-3} \\
N_{ab} &= 10^{17}\ \mathrm{cm}^{-3} \\
N_{dc} &= 10^{16}\ \mathrm{cm}^{-3} \\
D_b &= 30.0\ \mathrm{cm}^2\mathrm{s}^{-1} \\
L_b &= 10.0\ \mu\mathrm{m} \\
W_b &= 1.0\ \mu\mathrm{m} \\
D_e &= 10\ \mathrm{cm}^2\mathrm{s}^{-1} \\
L_e &= 10.0\ \mu\mathrm{m} \\
\text{Emitter thickness} &= 1.0\ \mu\mathrm{m} \\
\text{Device area} &= 4.0 \times 10^{-6}\ \ \mathrm{cm}^2
\end{aligned}$$

Calculate the emitter efficiency and gain β when the EBJ is forward biased at 1.0 V and the BCJ is reverse biased at 5.0 V. Calculate the output conductance of the device defined by

$$g_o = \frac{\Delta I_C}{\Delta V_{CB}}$$

Problem 6.28 An important advance in Si bipolar transistors is the use of polysilicon emitters. If a normal ohmic contact is made to an emitter, the injected minority density goes to zero at the ohmic contact boundary. In polysilicon emitters, heavily doped polysilicon forms the contact to the emitter. The minority density does not go to zero at the polysilicon contact, but decreases to zero well inside it. This allows one to have very thin emitter contacts for high-speed operation. Discuss the disadvantage of such a contact over a normal ohmic contact in a thin emitter. (Consider the emitter efficiency and how it is affected by a thin emitter by using the discussions in chapter 5 on the narrow p-n diode.)

Problem 6.29 Consider an npn BJT with a base width of 0.5 μm and base doping of 10^{17} cm^{-3}. The hole mobility is 200 cm^2/V·s. An emitter stripe of 25 μm $\times$100 μm is placed to form the EBJ. If a base current of 100 μA passes in the device and the EBJ is forward biased at 0.7 V at the edge of the emitter, estimate the value of the forward bias of the EBJ at the middle of the emitter. Discuss the possible problems that the biasing difference could cause. (Assume that the base current is flowing through an area 100 μm $\times$0.5 μm.)

Problem 6.30 From our discussions of narrow p-n diodes, the importance of the boundary conditions imposed on the injected minority charge at the contact is quite obvious. We have used the condition that the minority charge density goes to zero at the contact. This is a reasonable approximation for the metal contact. One approach to defining the boundary conditions at any interface is through the concept of a recombination velocity. The recombination velocity v_{recom} is defined via the relation (say, for holes as minority charge)

$$J_p\left|_{boundary} = e\ v_{recom}\ \delta p\right|_{boundary}$$

where the current and the excess charge are evaluated at the boundary. Using the expression for minority current in terms of the diffusion coefficient and the charge density gradient, we have

$$D_p \frac{d(\delta p)}{dx} |_{boundary} = v_{recom} \; \delta p|_{boundary}$$

Consider the case where an excess hole density of $\delta p(x_1)$ is injected across a depletion region into an n-side. The boundary of the contact is at a position x_2. The distance $(x_2 - x_1) \ll L_p$ so that the hole concentration can be assumed to decrease linearly. Express $\delta p(x_2)$ in terms of the surface recombination velocity, $\delta p(x_1)$, and the diffusion coefficient D_p and $(x_2 - x_1)$.

Problem 6.31 Consider an npn bipolar transistor where the base is graded from $\mathrm{Al_{0.04}Ga_{0.96}As}$ at the emitter end to GaAs at the collector end. The emitter material is $\mathrm{Al_{0.22}Ga_{0.78}As}$ and it is graded to $\mathrm{Al_{0.04}Ga_{0.96}As}$ at the base.
(a) Sketch the detailed band diagram of the HBT in equilibrium, and under forward active bias. What is the quasi-electric field, $\mathcal{E}_\mathrm{B}$, in the base?
(b) Solve the drift-diffusion equation to obtain an expression for the base minority carrier concentration, n(x), in terms of the total current $\mathrm{J_C}$, $\mathcal{E}_\mathrm{B}$, and $\mathrm{W_B}$. Assume μ = 1000 $\mathrm{cm^2}$/Vs.

Problem 6.32 An npn HBT has a collector consisting of slabs of two different materials, A and B. The velocity in material A is v_s while that in material B is $\frac{v_s}{2}$. The thickness of these slabs is equal, ie. $W_A = W_B = \frac{W_C}{2}$, where W_C is the total collector width. Assume that the doping in the collector is N_D, and that the voltage V_{CB} depletes the entire collector region.
(a) Consider the case when slab A is adjacent to the base. Draw the charge, electric field and potential profiles in the collector when the Kirk threshold current is reached. What is the Kirk threshold of this transistor? How does it compare with a transistor whose collector is made up of material A throughout?
(b) Repeat the above but now with slab B adjacent to the base. Compare the two cases and explain the result.
(c) Assume that the breakdown field in material A is less than in material B. Which material should be placed adjacent to the base to maximize the breakdown voltage under high injection?

Problem 6.33 For some unknown reason, possibly dislocated assisted diffusion, the base doping in the bipolar transistor diffuses back into the collector forming a base profile shown in figure 6.26.
(a) Calculate the transit time across the depletion region of the base-collector junction for this transistor.
(b) Assuming that the collector remains fully depleted, and that the current density is measured by the collector current divided by the collector area, how is the Kirk current threshold for the transistor affected? Comment on your answer. Assume $v_s = 10^7 cm/s$ and $\mu = 1000 cm^2/V.s$ for electrons. Apply both the impulse function and charge control

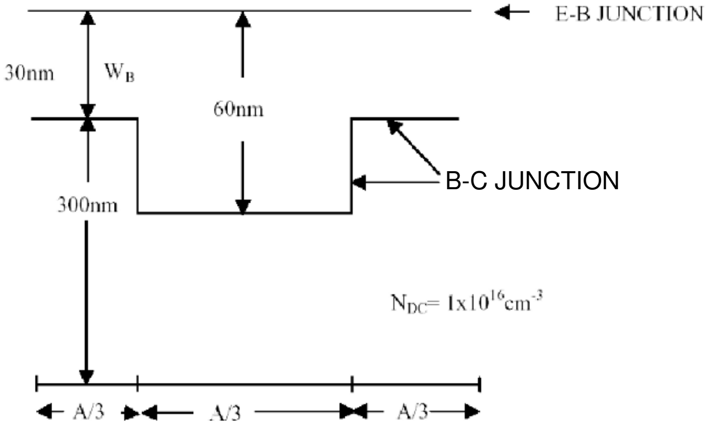


Figure 6.26: Figure for problem 6.33.

methods to arrive at the answer. For the impulse function, make sure the impulse charge is initiated at the base-emitter junction by the application of an impulse voltage. Think about how the charge density is distributed based on the fact that the base has a varying width.

Problem 6.34 An n-p-n AlGaAs-GaAs HBT is grown with the emitter-base junction graded from $\mathrm{Al}_{0.2}\mathrm{Ga}_{0.8}\mathrm{As}$ to GaAs over 0.19 μm. Assume that the emitter is doped 5×10^{16} cm^{-3}, and that the base is doped at 10^{18} cm^{-3}. Assume the conduction band offset between $\mathrm{Al}_{0.2}\mathrm{Ga}_{0.8}\mathrm{As}$ and GaAs to be 0.16 eV, and the bandgap of $\mathrm{Al}_{0.2}\mathrm{Ga}_{0.8}\mathrm{As}$ and GaAs to be 1.67 eV and 1.4 eV respectively.
(a) Draw the equilibrium band diagram of the emitter-base junction, indicating the band bending due to depletion charges and quasi-electric field. Calculate the depletion width of the junction.
(b) Calculate β for this device.
(c) Now, if a forward bias of 1 V is applied to this junction, what is the new depletion width? Calculate the conduction band profile and draw the band diagram for the device.
(d) Calculate the β when the emitter-base junction is forward biased.

Problem 6.35 Consider a GaAs npn BJT with the structure shown in figure 6.27. In this problem, we consider the effect of a non-zero lateral base resistance, R_B. The effective emitter base potential, $V_{BE}(x)$, drops along the lateral direction x (shown in the figure)

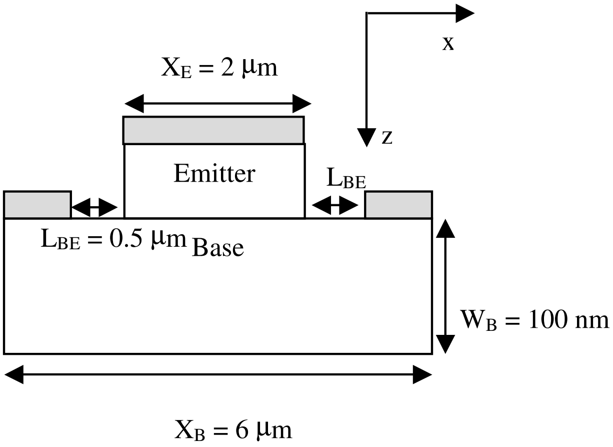


Figure 6.27: Figure for problem 6.35.

changing the current gain of the transistor. Note: You may ignore the vertical base resistance in this problem.
(a) Write the expression for the minority charge, $n(x, z)$, and integrated minority charge, $q(x) = \int n(x, z)dz$ (in cm^{-2}) in the base as a function of the varying emitter-base potential $V_{BE}(x)$.
(b) Derive the differential equation that relates the emitter base potential $V_{BE}(x)$ to $q(x)$ in terms of the base resistance, R_B, and recombination time, t_N. What are the boundary conditions?
(c) Using your results from (a) and (b), derive an expression for $n(x, z)$, $q(x)$, and the base current $I_B(x)$.
(d) What is the total emitter current and base current for this transistor? Find the expression for the current gain, β, in terms of R_B and t_N. Assume ideal emitter injection efficiency.

Problem 6.36 Consider an npn transistor where the base is open (figure 6.28). Assume that the β of the transistor is not impacted by recombination in the base. Show that the breakdown voltage, V$_{CEO}$, in this configuration is reduced from the normal breakdown voltage of the base-collector junction, V$_{CBO}$. Derive an analytical expression for V$_{CEO}$. State all your assumptions.

Problem 6.37 Consider a GaAs n-p-n transitor shown in figure 6.29. I make a mistake and during the growth and insert a 10nm thick quantum well in the center of a 100 nm base. The result of this mishap is to reduce the lifetime of the injected minority carriers to 0.1 ns

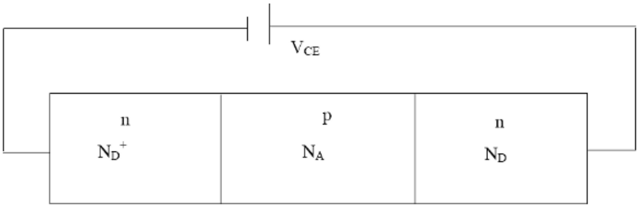


Figure 6.28: Figure for problem 6.36.

in the quantum well region compared to 1 ns in the rest of the base. The emitter is doped at $5 \times 10^{17} cm^{-3}$. The collector is doped at $5 \times 10^{16} cm^{-3}$. Assume that the elctron mobility is 5000 $\frac{cm^2}{Vs}$ and that the device is operate at room temperature. Calculate and sketch the electron current and charge profile in the structure at an emitter-base bias of 0.5 eV. Compare with the case of a homogeneous base. Next, calculate the early voltage of transistors with and without the quantum well in the base . Comment on all your solutions.

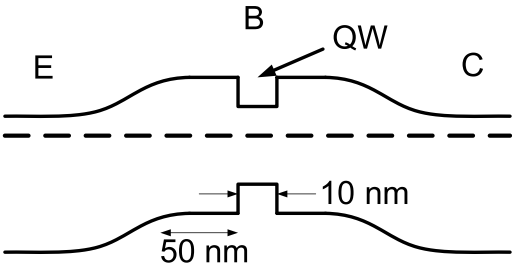


Figure 6.29:

Problem 6.38 Consider the transistors with two different collector doping profiles as shown in figure 6.30. Assume that the same base-collector bias, V_{BC}, is applied in both cases. You may also assume that the collector is fully depleted, and that the saturated velocity in both devices is v_{SAT}. Calculate the difference between the current densities at the Kirk threshold of transistors A and B. Give a physical explanation of your result using charge and electric field profiles.

Problem 6.39 The base-collector junction in a bipolar transistor has the structure shown in figure 6.31.

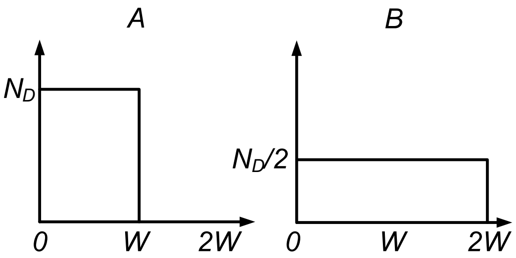


Figure 6.30: Figure for problem 6.38.

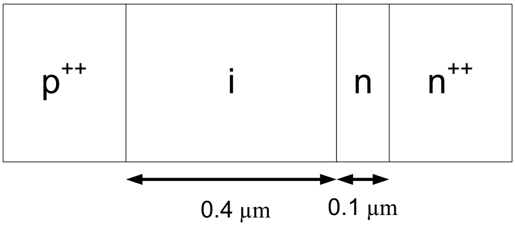


Figure 6.31: Figure for problem 6.39

1. Draw the band diagram of this base collector junction at zero bias. Assume that the base is heavily doped so that the entire voltage falls in the collector region. The built-in voltage is 1.4 eV.
2. Since the $\Gamma - L$ valley spacing in GaAs is 0.3 eV, we would like to make the voltage drop in the intrinsic collector adjacent to the base to be equal to that number. This would lead to high electron velocities without intervalley transfer. Is it possible to achieve this by adding a single sheet of acceptors or donors to this structure? Draw charge, electric field and energy band profiles to explain your answer. Calculate the acceptor or donor sheet charge density you would use.

6.8 DESIGN PROBLEMS

Problem 6.1 Consider a npn Si-BJT at 300 K with the following parameters:

$$\begin{aligned}
N_{de} &= 10^{18}\ \mathrm{cm}^{-3} \\
N_{ab} &= 10^{17}\ \mathrm{cm}^{-3} \\
N_{dc} &= 10^{16}\ \mathrm{cm}^{-3} \\
D_b &= 30.0\ \mathrm{cm}^2\mathrm{s}^{-1} \\
L_b &= 10.0\ \mu\mathrm{m} \\
W_b &= 1.0\ \mu\mathrm{m} \\
D_e &= 10\ \mathrm{cm}^2\mathrm{s}^{-1} \\
L_e &= 5.0\ \mu\mathrm{m} \\
\text{electron mobility in the emitter} &= 500\ \mathrm{cm}^2\ \mathrm{V}^{-1}\ \mathrm{s}^{-1} \\
\text{area} &= 5.0 \times 10^{-7}\ \mathrm{cm}^2
\end{aligned}$$

Calculate the emitter efficiency and gain β when the EBJ is forward biased at 1.0 V and the BCJ is reverse biased at (a): 5.0 V and (b) 10.0 V.

For high-speed operation, it is found that the BJT discussed above has too large an emitter resistance. The device designer wants to limit the emitter resistance (keeping the area unchanged) to 2.0 Ω. Calculate the emitter efficiency and β for the new device using the case (a) given above.

Problem 6.2 Consider a npn Si-BJT at 300 K with the following parameters:

$$\begin{aligned}
N_{de} &= 10^{18}\ \mathrm{cm}^{-3} \\
N_{ab} &= 10^{17}\ \mathrm{cm}^{-3} \\
N_{dc} &= 5 \times 10^{16}\ \mathrm{cm}^{-3} \\
D_b &= 30.0\ \mathrm{cm}^2\mathrm{s}^{-1} \\
L_b &= 15.0\ \mu\mathrm{m} \\
D_e &= 10.0\ \mathrm{cm}^2\mathrm{s}^{-1} \\
L_e &= 5.0\ \mu\mathrm{m}
\end{aligned}$$

Design the maximum base width, W_b, that will allow a current gain β of 100 when the EBJ is forward biased at 1.0 V and the BCJ is reverse biased at 5.0 V. You may make the following approximations:

- The reverse bias collector current is zero.
- W_b is much smaller than L_b.

Problem 6.3 Consider a npn Si-BJT at 300 K with the following parameters:

$$\begin{aligned}
N_{de} &= 10^{18}\ \mathrm{cm}^{-3} \\
N_{ab} &= 10^{17}\ \mathrm{cm}^{-3} \\
N_{dc} &= 10^{16}\ \mathrm{cm}^{-3} \\
D_b &= 30.0\ \mathrm{cm}^2\mathrm{s}^{-1} \\
L_b &= 10.0\ \mu\mathrm{m} \\
W_b &= 1.0\ \mu\mathrm{m} \\
D_e &= 10\ \mathrm{cm}^2\mathrm{s}^{-1} \\
L_e &= 10.0\ \mu\mathrm{m} \\
\text{emitter thickness} &= 1.0\ \mu\mathrm{m} \\
\text{device area} &= 4.0 \times 10^{-6}\ \mathrm{cm}
\end{aligned}$$

(a) Calculate the emitter efficiency and gain β when the EBJ is forward biased at 1.0 V and the BCJ is reverse biased at 5.0 V.
(b) Calculate the output conductance of the device defined by

$$g_o = \frac{\Delta I_C}{\Delta V_{CB}}$$

Problem 6.4 Consider an npn Si-BJT at 300 K with the following parameters:

$$\begin{aligned}
N_{de} &= 10^{18}\ \mathrm{cm}^{-3} \\
N_{ab} &= 10^{17}\ \mathrm{cm}^{-3} \\
N_{dc} &= 5 \times 10^{16}\ \mathrm{cm}^{-3} \\
D_b &= 20.0\ \mathrm{cm}^2\mathrm{s}^{-1} \\
L_b &= 15.0\ \mu\mathrm{m} \\
D_e &= 10.0\ \mathrm{cm}^2\mathrm{s}^{-1} \\
L_e &= 5.0\ \mu\mathrm{m} \\
\text{emitter dimensions} &= 100\ \mu\mathrm{m} \times 100\ \mu\mathrm{m}
\end{aligned}$$

(a) Calculate the base width, W_b, that will allow a current gain β of 200 when the EBJ is forward biased at 0.8 V and the BCJ is reverse biased at 5.0 V. Design the base width so that the gain goal is achieved and the base resistance is minimum.
(b) Estimate the base resistance. Note that the base hole current flows sideways into the device (figure 6.6). The hole mobility in the base is 300 cm^2/V·s.
You may make the following approximations :
- The reverse bias collector current is zero.
- W_b is much smaller than L_b.

6.9 FURTHER READING

- **General**
  - D. A. Neaman, Semiconductor Physics and Devices, Basic Principles (Irwin, Boston, 1997).
  - G. W. Neudeck, The Bipolar Junction Transistor (Vol. 3 of the Modular Series on Solid State Devices, Addison-Wesley, Reading, MA, 1989).
  - D. J. Roulston, Bipolar Semiconductor Devices (McGraw-Hill, New York, 1990).
  - B. G. Streetman and S. Bannerjee, Solid State Electronic Devices (Prentice-Hall, Englewood Cliffs, NJ, 2000).
  - D. A. Hodges and H. G. Jackson, Analysis and Design of Digital Integrated Circuits (McGraw-Hill, New York, 1988).

Chapter 7

TEMPORAL RESPONSE OF DIODES AND BIPOLAR TRANSISTORS

7.1 INTRODUCTION

In chapters 4-6, we studied the dc properties of diodes and bipolar transistors. In practice, these devices are used in circuits for both digital and analog applications, such as the circuit pictured in figure 7.1. In digital circuits, the devices will constantly be switched from the “on” (conducting) state to the “off” (non-conducting) state and back. The speed at which the circuit can process bits of data is largely determined by the switching speed of the devices.

In analog applications, the circuit is biased at some dc value, and then a small ac signal v_{in} is applied at the input. The input signal is amplified by the circuit, resulting in a signal v_{out} at the output. The gain of the devices in the circuit is frequency dependent and compresses at higher frequencies. Therefore, in order to design high frequency circuits, it is important to understand the frequency response of the devices.

In this chapter, we derive the frequency response of diodes and bipolar transistors. We address issues for both large-signal switching applications and small-signal high frequency applications. We will see that in many cases, there are trade-offs between achieving superior dc performance and being able to operate at higher frequencies.

7.2 MODULATION AND SWITCHING OF A P-N DIODE: AC RESPONSE

In chapter 5, we discussed the dc characteristics of the p-n diode. However, many applications of diodes will involve transient or ac properties of the diode. The transient properties of the diode

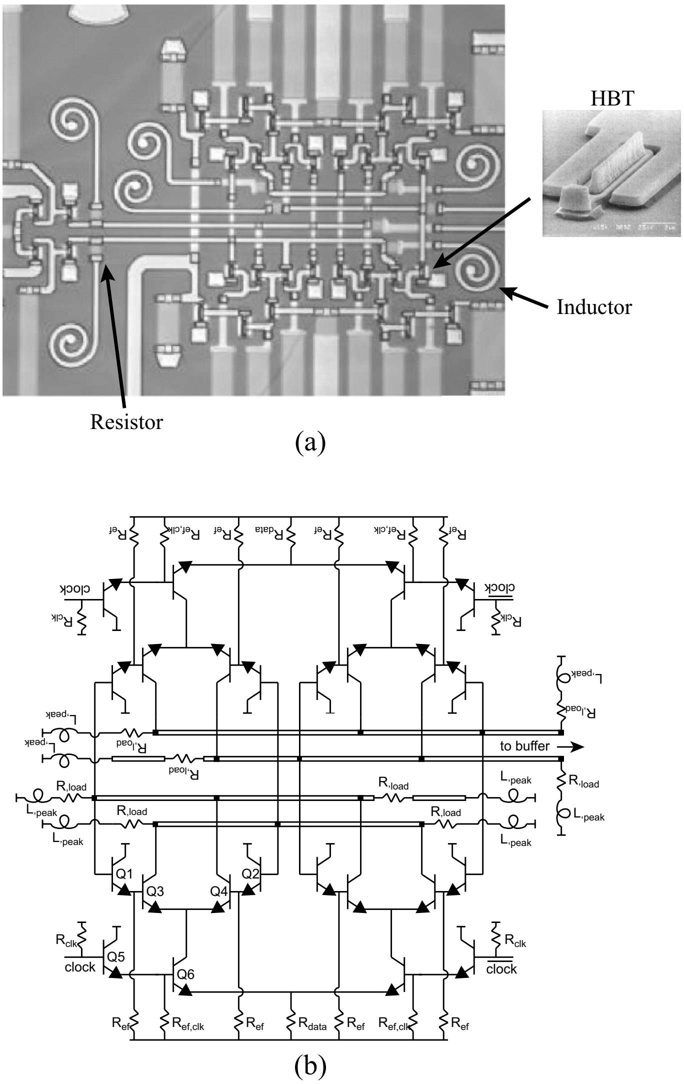


Figure 7.1: (a) Photograph of a 142 GHz master-slave latch, along with (b) the corresponding circuit diagram. The circuit is based on the InP HBT technology illustrated in figure 7.1 Figures courtesy of M. Rodwell and Z. Griffith, UCSB.

are usually not very appealing, especially for high-speed applications. This is one of that reasons that diodes have been replaced by transistors and Schottky diodes (to be discussed later) in many applications.

A homojunction p-n diode is a minority carrier device, i.e., it involves injection of electrons into a p-type region and holes into an $n-$type region. In forward-bias conditions where the diode is in a conducting state, the current is due to the minority charge injection. In figure 7.2a, we show the minority charge (hole) distribution in the n-side of a forward-biased p-n diode. If this diode is to be switched, this excess charge must be removed. The device time response, therefore, depends upon how fast one can alter the minority charge that has been injected. In figure 7.2b we show how the minority charge can be extracted. As noted in this figure, one can speed up the process either by introducing defects that speed up the recombination or by using very narrow diodes. Both these approaches have problems. A high defect density causes non-ideal diode behavior and increases reverse leakage and a narrow diode has a large reverse-bias current.

For the reverse-biased case, where no minority charge injection occurs, the device speed can be quite high and is dominated by the device RC time constant. Let us examine the response of the $p-n$ diode to large and small signals.

7.2.1 Small-Signal Equivalent Circuit of a p-n Diode

We will start by developing a model for the diode small-signal capacitance and resistance. The diode capacitance arises from two distinct regions of charge: i) The junction capacitance arises from the depletion region where there are regions of fixed positive and negative charge; and ii) The diffusion capacitance is due to the region outside the depletion region where minority carrier injection has introduced excess charges. The diffusion capacitance due to injected carriers dominates under forward-bias conditions. While in the reverse bias case the junction capacitance dominates. The small signal capacitance is in general defined by the relation

$$C = \left| \frac{dQ}{dV} \right| \tag{7.2.1}$$

It is important to note that by definition, capacitance is a lossless energy storage element. This implies that any charge which is stored in a capacitor must be reclaimable . Charge which is lost during modulation (for example through electron-hole recombination) is not reclaimed and therefore does not contribute to the capacitance defined in equation 7.2.1. We will see in this section that in p-n diodes, only a fraction of the stored charge is reclaimed during high frequency operation. This impacts the diode small-signal response.

We will now use the equations derived in chapter 4 to calculate the capacitance. The junction depletion width of the p-n diode is

$$W = \left[\frac{2\epsilon(V_{bi} - V)}{e} \left(\frac{N_a + N_d}{N_a N_d} \right) \right]^{1/2} \tag{7.2.2}$$

The depletion region charge is

$$|Q| = eA\, W_n N_d = eA\, W_p N_a \tag{7.2.3}$$

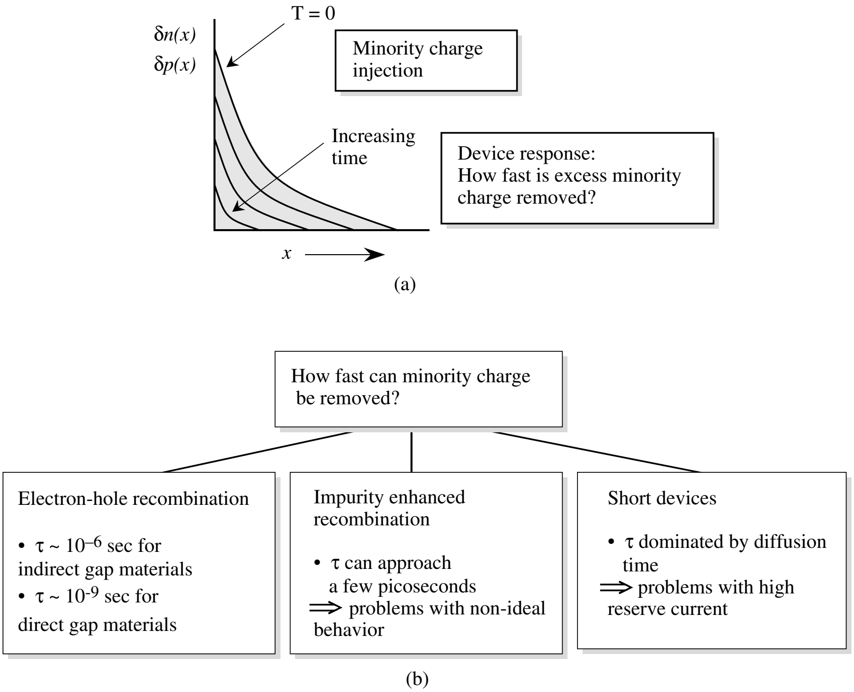

T = 0
δn(x)
δp(x)
Minority charge injection
Increasing time
Device response:
How fast is excess minority charge removed?
x
(a)
How fast can minority charge be removed?
Electron-hole recombination
• τ ~ 10⁻⁶ sec for indirect gap materials
• τ ~ 10⁻⁹ sec for direct gap materials
Impurity enhanced recombination
• τ can approach a few picoseconds
⟹ problems with non-ideal behavior
Short devices
• τ dominated by diffusion time
⟹ problems with high reserve current
(b)

Figure 7.2: (a) The minority hole distribution in a forward-biased p-n diode. If the diode is to be switched, the excess holes have to be extracted. (b) A schematic of what controls device response of minority-carrier-based devices. Three approaches used to speed up the device response are described.

where we showed earlier (see equation 4.2.22 through equation 4.2.24)

$$W_n = \frac{N_a}{N_a + N_d} W; \; W_p = \frac{N_d}{N_a + N_d} W \tag{7.2.4}$$

Thus

$$|Q| = \frac{eA \, N_a N_d}{N_d + N_a} W = A \left[2e\epsilon (V_{bi} - V) \frac{N_d N_a}{N_d + N_a} \right]^{1/2} \tag{7.2.5}$$

The small signal junction capacitance is then

$$\begin{aligned} C_j &= \left|\frac{dQ}{dV}\right| = \frac{A}{2}\left[\frac{2e\epsilon}{(V_{bi}-V)}\,\frac{N_aN_d}{N_a+N_d}\right]^{1/2} \\ &= \frac{A\epsilon}{W} = \frac{C_{jo}}{(1-\frac{V}{V_{bi}})^{1/2}} \end{aligned} \tag{7.2.6}$$

where C_{jo} is the capacitance at zero applied bias. Since the depletion width depends upon the applied bias, the diode capacitance can be tailored electronically. This voltage-dependent diode capacitor is called a varactor is useful for tuning frequency of a resonant cavity electronically.

In real diodes, the doping in the n-side and p-side gradually changes from $n-$type to $p-$type. In such cases, the depletion capacitance of the diode is written as

$$C_j = \frac{C_{jo}}{\left(1-\frac{V}{V_{bi}}\right)^m} \tag{7.2.7}$$

where m is a parameter called the grading parameter. For abrupt junctions, $m = 1/2$ as can be seen in equation 7.2.6. For linearly graded junctions, $m = 1/3$.

For the forward-biased diode, the injected charge density can be large and can dominate the capacitance. The injected hole charge is (see Eqn. 5.3.12 for the forward bias hole current; remember that charge is $I\tau_p$ and use $\tau_pD_p = L_p^2$; we also ignore 1 in the forward-bias state)

$$Q_p = I\tau_p = eA\,L_pp_n\,e^{eV/k_BT} \tag{7.2.8}$$

The diffusion capacitance is then

$$\frac{dQ_p}{dV} = \frac{e^2}{k_BT}A\,L_pp_n\,e^{eV/k_BT} = \frac{e}{k_BT}I\tau_p \tag{7.2.9}$$

Using the diode equation for small-signal ac response, the ac conductance of the diode is

$$\boxed{G_s = \frac{dI}{dV} = \frac{e}{k_BT}I(V)} \tag{7.2.10}$$

from the definition of the $I(V)$ function. We will show later that this expression only holds at low frequencies. At room temperature the conductance is (r_s is the diode resistance)

$$\boxed{G_s = \frac{1}{r_s} = \frac{I(\text{mA})}{25.86}\Omega^{-1}} \tag{7.2.11}$$

Consider now a p-n diode that is forward-biased at some voltage V_{dc}, as shown in figure 7.3a. If an ac signal is now applied to the diode, the current changes as shown schematically. The small signal equivalent circuit of the diode is shown in figure 7.3b and consists of the diode resistance r_s $(= G_s^{-1})$, the junction capacitance, and the diffusion capacitance. At first glance, it would appear that the diffusion capacitance in figure 7.3b is given by equation 7.2.9. However, this is not the case, since when an ac signal is applied, not all of the injected minority charge is reclaimed through the junction. Some of the charge simply recombines in the neutral region. In the forward-bias condition, the diffusion capacitance will dominate and we have the following relation between the current i_s and the applied voltage signal v_s:

$$i_s = G_s v_s + C_{diff} \frac{dv_s}{dt} \tag{7.2.12}$$

If we assume an input voltage with frequency ω $(v_s \sim v_s^o \ \exp\ (j\omega t))$, we get

$$i_s = G_s v_s + j\omega C_{diff} v_s \tag{7.2.13}$$

and the admittance of the diode becomes

$$y = \frac{i_s}{v_s} = G_s + j\omega C_{diff} \tag{7.2.14}$$

To find G_s and C_{diff} in equation 7.2.14. it is necessary to calculate the admittance y by solving the time-dependent continuity equation, and then G_s and C_{diff} can be extracted. We first solve the continuity equation to find the ac part of the injected charge distribution when a bias $V(t) = V_{dc} + v_s^o \ \exp\ (j\omega t)$ is applied to the diode. From this we can determine the ac part of the current and thus calculate the admittance. The general form for the time-dependent continuity equation is

$$\frac{\partial p}{\partial t} = -\frac{1}{e} \nabla \cdot \mathbf{J} + G - R \tag{7.2.15}$$

We assume here that we have a wide base diode (base $>> (D_p \tau_p)^{1/2}$), and we are applying a voltage $V(t) = V_{dc} + v_s^o \ \exp\ (j\omega t)$. The continuity equation then takes the form

$$eA \frac{\partial \left(\Delta p\right)}{\partial t} = -eA \frac{\Delta p}{\tau_p} - \frac{\partial I_p}{\partial x} \tag{7.2.16}$$

Assuming our current I_p is purely diffusion $(I_p = -eAD_p \frac{d(\Delta p)}{dx})$, the continuity equation becomes

$$\frac{\partial \left(\Delta p\right)}{\partial t} = -\frac{\Delta p}{\tau_p} + D_p \frac{\partial^2 \left(\Delta p\right)}{\partial x^2} \tag{7.2.17}$$

Under dc bias, the left hand side of this equation is zero, and our solution was given by

$$\left(\Delta p\right)_{dc} = \Delta p \left(0\right) e^{-x/L_p} \tag{7.2.18}$$

where $x = 0$ is at our depletion region edge and $L_p = (D_p \tau_p)^{1/2}$. When an ac signal is added, we assume a solution of the form

$$\Delta p = \left(\Delta p\right)_{dc} + \left(\Delta p\right)_{ac} = \Delta p \left(0\right) e^{-x/L_p} + \widetilde{\Delta p}(x) e^{j\omega t} \tag{7.2.19}$$

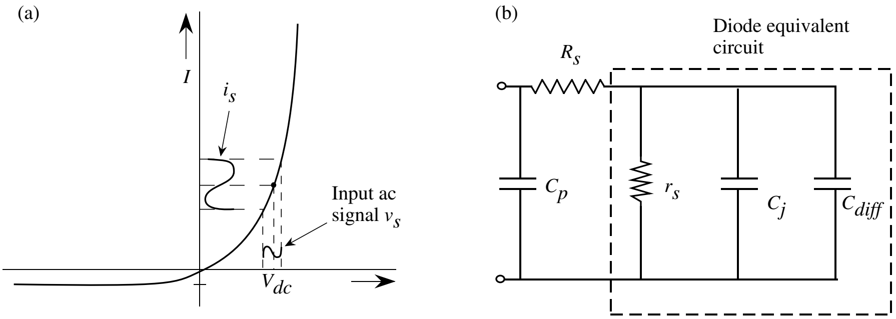


Figure 7.3: (a) A p-n diode is biased at a dc voltage V_{dc} and a small signal modulation is applied to it. (b) The equivalent circuit of a forward-biased diode

where $\widetilde{\Delta p}(x)$ is the amplitude of the ac hole concentration due to the small signal. Plugging this back into equation 7.2.16 and simplifying, we get

$$j\omega\widetilde{\Delta p}(x) = -\frac{\widetilde{\Delta p}(x)}{\tau_p} + D_p\frac{d^2\left(\widetilde{\Delta p}(x)\right)}{dx^2} \tag{7.2.20}$$

Equation 7.2.19 allows us to solve for the ac part of the injected charge distribution. The general solution to this equation is

$$\widetilde{\Delta p}(x) = C_1 e^{-x/\lambda} + C_2 e^{+x/\lambda} \tag{7.2.21}$$

where λ is the ac diffusion length and is given by

$$\lambda = \left[\frac{j\omega}{D_p} + \frac{1}{D_p\tau_p}\right]^{-1/2} \tag{7.2.22}$$

Applying appropriate boundary conditions ($\widetilde{\Delta p} = 0$ when $x \to \infty$ and $\widetilde{\Delta p} = \widetilde{\Delta p}(0)$ when $x = 0$) results in

$$C_2 = 0 \tag{7.2.23}$$
$$C_1 = \widetilde{\Delta p}(0) \tag{7.2.24}$$

and the ac injected charge distribution becomes

$$\boxed{(\Delta p)_{ac} = \left[\widetilde{\Delta p}(0)\exp\left(-x\sqrt{\frac{j\omega}{D_p} + \frac{1}{D_p\tau_p}}\right)\right]\exp\left(j\omega t\right)} \tag{7.2.25}$$

$\widetilde{\Delta p}(0)$ can be calculated by noting that the total injected charge at $x = 0$, $\Delta p_{tot}(0)$, is the sum of the dc and ac components $\left(\Delta p_{dc}(0) = p_{n0} \exp\left[\frac{eV_{dc}}{k_B T}\right]\right)$.

$$\begin{aligned}
\Delta p_{tot}(0) &= \Delta p_{dc}(0) + \widetilde{\Delta p}(0) = p_{n0} \exp\left(\frac{e\left(V_{dc} + v_s\right)}{k_B T}\right) \\
&= p_{n0} \exp\left(\frac{eV_{dc}}{k_B T}\right) \exp\left(\frac{ev_s}{k_B T}\right) = \Delta p_{dc}(0) \exp\left(\frac{ev_s}{k_B T}\right) \qquad (7.2.26)
\end{aligned}$$

Since we are assuming v_s to be small, $\exp\left(\frac{ev_s}{k_B T}\right) \simeq 1 + \frac{ev_s}{k_B T}$. Inserting this into equation 7.2.25 and solving for $\widetilde{\Delta p}(0)$ gives us

$$\Delta p_{tot}(0) = p_{n0} e^{\frac{eV_{dc}}{k_B T}} \left(1 + \frac{ev_s}{k_B T}\right) \qquad (7.2.27)$$

$$\boxed{\widetilde{\Delta p}(0) = \Delta p_{dc}(0) \cdot \frac{ev_s}{k_B T}} \qquad (7.2.28)$$

Notice that the time-independent part of the ac injected charge in equation 7.2.25 has the same form as the dc injected charge, only with a complex diffusion length, λ. It is interesting to note that λ is frequency dependent; when $\omega = 0$, $\lambda = L_p$, and as ω increases, λ decreases, since the injected charge can be reduced via reclamation through the junction during the negative swing of the ac voltage.

Using the results of equation 7.2.25 and equation 7.2.28, we can find the current i_s that results from our applied voltage signal v_s. The total current i_s is given by the diffusion current at $x = 0$.

$$\begin{aligned}
i_s &= -eAD_p \left.\frac{d\widetilde{\Delta p}(x)}{dx}\right|_{x=0} = eAD_p \widetilde{\Delta p}(0) \sqrt{\frac{j\omega}{D_p} + \frac{1}{D_p \tau_p}} \qquad (7.2.29) \\
&= \frac{eaD_p}{L_p} \Delta p_{dc}(0) \cdot \frac{ev_s}{k_B T} \sqrt{1 + j\omega\tau_p} \qquad (7.2.30) \\
&= \frac{eIv_s}{k_B T} \sqrt{1 + j\omega\tau_p} \qquad (7.2.31)
\end{aligned}$$

From this equation, we calculate the small signal admittance

$$y = \frac{i_s}{v_s} = \frac{eI}{k_B T} \sqrt{1 + j\omega\tau_p} \qquad (7.2.32)$$

The admittance, as well as the small signal parameters C_{diff} and G_s, take different forms at low frequencies ($\omega\tau_p < 1$) and at high frequencies ($\omega\tau_p > 1$). At low frequency, the admittance is

$$\boxed{y \simeq \frac{eI}{k_B T} \left[1 + \frac{j\omega\tau_p}{2}\right]} \qquad (7.2.33)$$

and G_s and C_{diff} are given by (see equation 7.2.14)

$$G_s = \frac{eI}{k_BT} \tag{7.2.34}$$

$$C_{diff} = \frac{e}{2k_BT} I\tau_p = \frac{1}{2}\left(\frac{dQ_p}{dV}\right) \tag{7.2.35}$$

While the diode conductance G_s is the same for low frequency ac response as its value at dc (equation 7.2.10), we see that the diffusion capacitance $C_{diff} = \frac{1}{2}\left(\frac{dQ_p}{dV}\right)$ (see equation 7.2.9), indicating that only half of the injected charge is reclaimed through the junction. The other half recombines in the neutral region. A similar analysis can be carried out for narrow base diodes to show that in that case, 2/3 of the injected charge is reclaimed through the junction. In general, the diffusion capacitance of the small-signal description can be written as

$$\boxed{C_{diff} = K\frac{e}{k_BT} I\tau_p} \tag{7.2.36}$$

where K is a factor which is 1/2 for long base diodes and 2/3 for narrow base devices.

At high frequencies, the admittance becomes

$$y \simeq \frac{eI}{k_BT}\sqrt{j\omega\tau_p} = \frac{eI}{k_BT}\sqrt{\frac{\omega\tau_p}{2}} + j\omega\frac{eI}{k_BT}\sqrt{\frac{\tau_p}{2\omega}} \tag{7.2.37}$$

and G_s and C_{diff} are given by

$$G_s = \frac{eI}{k_BT}\sqrt{\frac{\omega\tau_p}{2}} \tag{7.2.38}$$

$$C_{diff} = \frac{eI}{k_BT}\sqrt{\frac{\tau_p}{2\omega}} \tag{7.2.39}$$

We see that at high frequencies both the small signal resistance $r_s = G_s^{-1}$ and capacitance C_{diff} decrease with ω as $\frac{1}{\sqrt{\omega}}$.

In figure 7.3b we show the equivalent circuit of a packaged diode where we have the additional series resistance R_s associated with the diode n and $p-$type neutral regions and a capacitance C_p associated with the diode packaging. As discussed, at forward bias the diffusion capacitance dominates, while at reverse bias the junction capacitance is dominant.

7.2.2 Switching characteristics of diodes

In many approaches the diode is switched from the conducting state to its non-conducting state. Large-signal switching occurs in digital technology, in pulse shaping, and in optoelectronics. Accurate time responses of current to voltage switching are complex series solutions to the time-dependent semiconductor equations. However, simplified approaches give a good insight to the problem and will be discussed.

In the forward-biased state, minority charge is injected across the depletion region. In the reverse-bias state, the excess minority charge is below the equilibrium value. Thus in diode switching, minority charge has to be removed and injected, and the diode temporal response is controlled by the time it takes to inject and remove the minority charge.

To understand the time response of the diode, we use the relationship between the excess minority charge and the current in the diode. We will assume an asymmetrically doped diode $(p^+ - n)$ so that hole lifetime will limit the device response. The total charge Q_p injected into the n-region for a long diode is

$$Q_p = eA \int_{W_n}^{\infty} \delta p_n(x) dx \tag{7.2.40}$$

Using the relation between the charge and the voltage across the diode, we have

$$Q_p = eAL_p p_{no} \left(e^{eV/k_B T} - 1 \right) \tag{7.2.41}$$

In steady state the current is related to the charge by

$$I = \frac{Q_p}{\tau_p} \tag{7.2.42}$$

where τ_p is the hole recombination time. For a narrow diode the relevant time is the carrier extraction time from the neutral n-region of width $W_{\text{ln}} - W_n$, which is given by:

$$\tau_T = \frac{|W_{\text{ln}} - W_n|^2}{2D_p} \tag{7.2.43}$$

A change in density with time defines the current. This give the equation

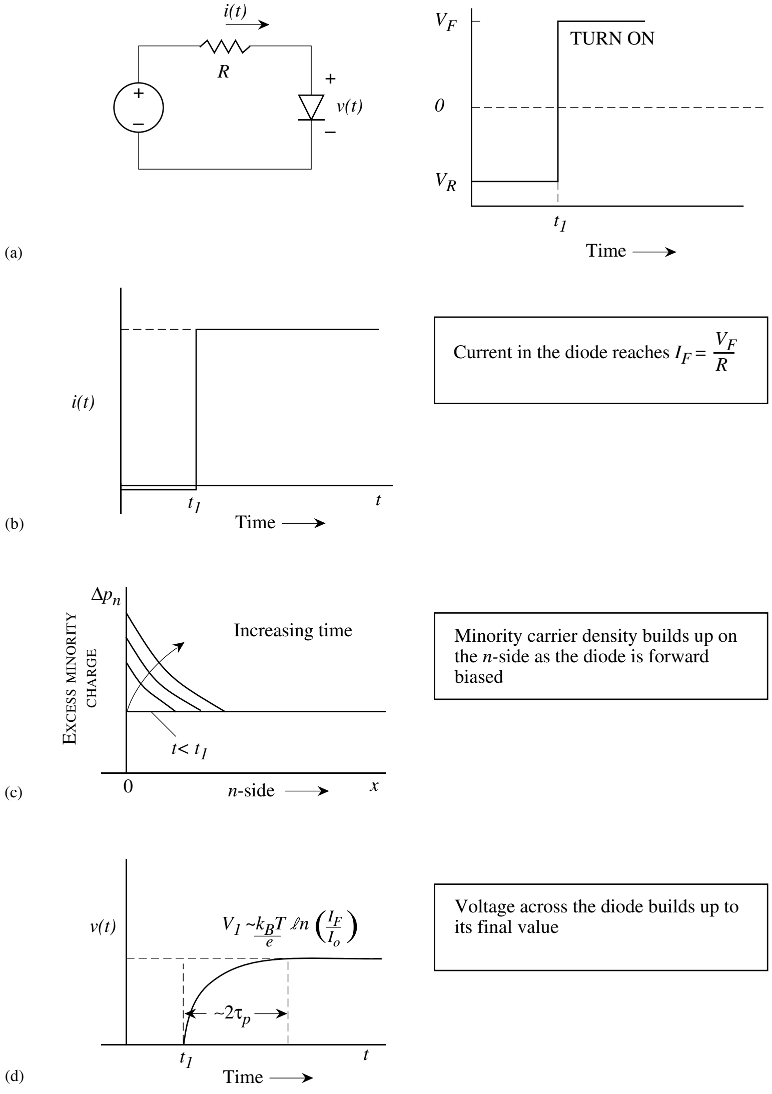
$$i(t) = \frac{Q_p}{\tau_p} + \frac{dQ_p}{dt} \tag{7.2.44}$$

where the first term is due to $e - h$ recombination and the second is due to the change in the minority charge with time.

Turn-ON Response

We will start by examining how a $p - n$ diode switches to its ON state. Consider the circuit of figure 7.4 where a diode is driven by a square wave pulse with the voltage switching between V_F and V_R. The voltage V_F is much larger than the voltage across the diode under forward-bias conditions. Let us consider how the diode responds when the voltage pulse switches to V_F. As shown in figure 7.4a, the voltage switches at $t = t_1$. Once the diode is forward biased, the current becomes

$$i(t) = \frac{V_F - V_1}{R} \tag{7.2.45}$$



Current in the diode reaches $I_F = \frac{V_F}{R}$

Minority carrier density builds up on the *n*-side as the diode is forward biased

Voltage across the diode builds up to its final value

Figure 7.4: Turn-ON characteristics of a p-n diode: (a) The voltage switches from V_R to V_F as shown at $t = t_1$. (b) Current response (c) Time evolution of the minority charge injected into the n-side. (d) Voltage across the $p - n$ diode.

where V_1 is the voltage across the diode and is related to the minority charge by the relation (see Eqn. 5.3.5)

$$V_1(t) = k_B T \ln\left(\frac{p(W_n)}{p_n}\right) \tag{7.2.46}$$

Since turn on voltage of the diode is small compared to V_F

$$i(t) \sim \frac{V_F}{R} \tag{7.2.47}$$

It is important to note that upon turn-on, the diode current reaches its peak value almost instantly, as shown in figure 7.4b.

The minority charge in the n-region increases gradually, and is controlled by diffusion, as shown in figure 7.4c. From equation 7.2.46 we see that the voltage across the diode also increases and saturates at

$$V_1 = k_B T \ln\left(\frac{I_F}{I_o}\right) \tag{7.2.48}$$

The time taken for the voltage to saturate to V_1 is approximately $2\tau_p$. The voltage across the diode starts from zero and grows to V_1 as shown in figure 7.4d.

Turn-OFF

We now discuss the turn-off behavior of the diode as shown in figure 7.5a. The voltage is switched from V_F to V_R at $t = t_2$. To understand the diode response to this turn-off, we note the relation between the excess hole density on the n-side and the voltage across the diode:

$$\delta p(W_n) = p_n \left[\exp\left(\frac{\mathrm{eV}}{k_B T}\right) - 1\right] \tag{7.2.49}$$

An important outcome of this equation is that as long as $\delta p(Wn)$ is positive, the voltage across the diode is essentially the forward bias voltage ($\sim$ 0.7 V). The diode current is

$$\begin{aligned} t < t_2 \quad &: \quad i(t) = I_F = \frac{V_F - V_1}{R} \\ t = t_2 \quad &: \quad i(t) = I_R = \frac{V_R - V_1}{R} \end{aligned} \tag{7.2.50}$$

Since the diode is in the forward-biased state before the diode is reverse biased, there is excess minority charge (holes) stored in the n-side. The diode response is controlled by the rate at which this charge is removed. If t_3 is the time at which the excess minority charge is extracted, then up to this time, the diode cannot be reverse biased (see equation 7.2.49). To examine the time response, let us examine the charge control equations

$$\begin{aligned} t < t_2 \quad &: \quad i(t) = I_F = \frac{Q_p}{\tau_p} \\ t = t_2 \quad &: \quad i(t) = I_R = \frac{Q_p}{\tau_p} + \frac{dQ_p}{dt} \end{aligned} \tag{7.2.51}$$

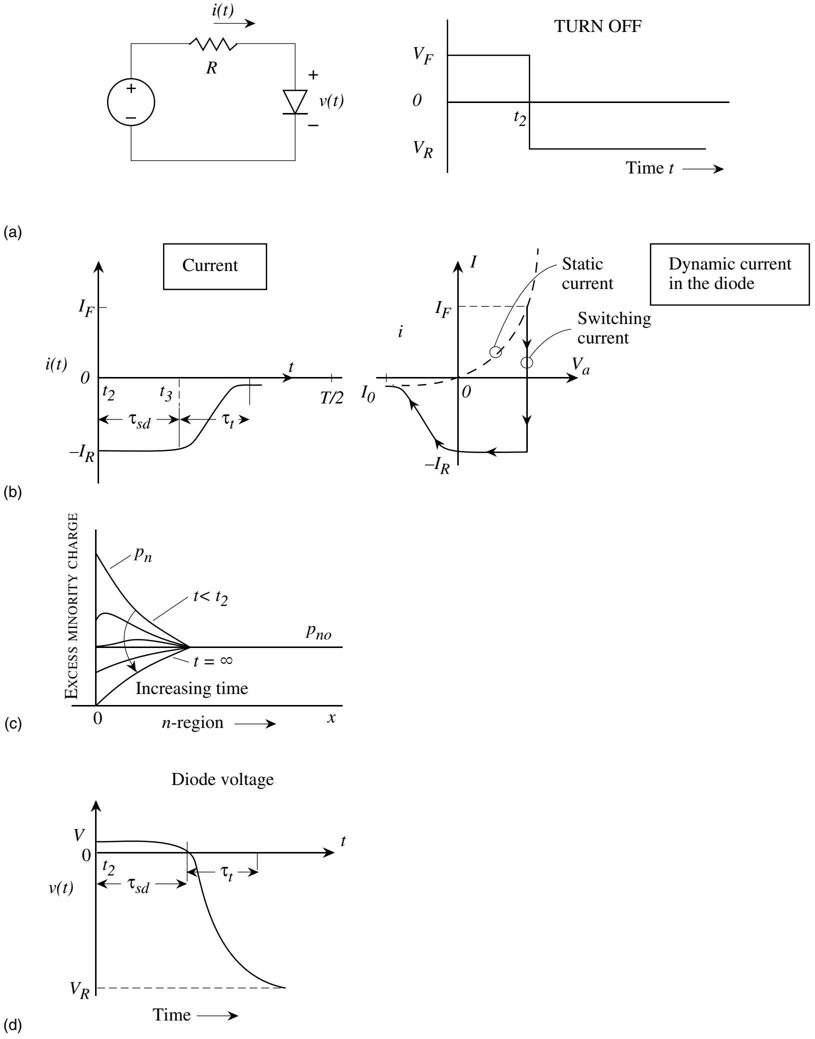


Figure 7.5: Turn-OFF characteristics of a $p - n$diode: (a) The external voltage switches from V_F to V_R at $t = t_2$. (b) A schematic of the current in the circuit and the current path (right) superimposed on the steady state I-V of a diode. (c) Minority charge distribution change. (d) Voltage across the diode.

The general solution of the equation is

$$Q_p(t) = i_R\tau_p + Ce^{-t/\tau_p} \tag{7.2.52}$$

To obtain the constant C, we note that at time just before t_2,

$$Q_p(t) = i_F\tau_p = i_R\tau_p + Ce^{-t_2/\tau_p}$$

or

$$C = \tau_p\left(i_F - i_R\right)e^{t_2/\tau_p} \tag{7.2.53}$$

The time dependence of the minority charge becomes

$$t > t_2 \ : \ Q_p(t) = \tau_p\left[i_R + \left(i_F - i_R\right)e^{(t_2-t)/\tau_p}\right] \tag{7.2.54}$$

At $t = t_3$, the entire excess minority charge is removed, i.e., $Q_p(t_3) = 0$. This gives us

$$i_R + \left(i_F - i_R\right)e^{-(t_3-t_2)/\tau_p} = 0$$

For the long diode, we get

$$t_3 - t_2 = \tau_p\ \ln\frac{i_F - i_R}{i_R} = \tau_{sd} \tag{7.2.55}$$

The time $(t_3 - t_2)$ it takes to remove the stored minority charge is called the storage delay time τ_{sd}. Until this time, the diode remains forward biased. For the short diode, the time τ_p is replaced by the transit time defined in equation 7.2.43. We have, for the short diode,

$$t_3 - t_2 = \tau_T\ \ln\frac{i_F - i_R}{i_R} = \tau_{sd} \tag{7.2.56}$$

Once the minority charge has been removed, the diode reverse biases in a time controlled by the circuit resistance and the average depletion capacitance of the diode. This time, known as the transition time, is

$$\tau_t \sim 2.3\ RC_j \tag{7.2.57}$$

where R is the resistance in the circuit and C_j is the average depletion capacitance.

The discussion of the turn-off process is represented schematically in figure 7.5.

7.3 Temporal Response of a Schottky Diode

In chapter 5 we have examined the Schottky diode. The key difference between the Schottky diode and the $p - n$ diode is that the Schottky diode is a majority carrier device and as a result minority carrier injection and extraction is not an issue. The small-signal equivalent circuit of a Schottky diode is shown in figure 7.6. One has the parallel combination of the resistance

$$R_d = \frac{dV}{dI} \tag{7.3.1}$$

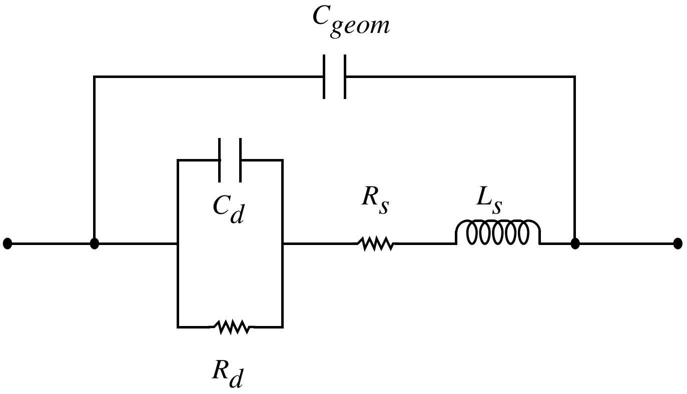


Figure 7.6: Equivalent circuit of a Schottky diode.

and the differential capacitance of the depletion region. The depletion capacitance has the form:

$$C_d = A\left[\frac{eN_d\epsilon}{2(V_{bi}-V)}\right]^{1/2} \tag{7.3.2}$$

As noted earlier, there is no diffusion capacitance. These circuit elements are in series with the series resistance R_s (which includes the contact resistance and the resistance of the neutral doped region of the semiconductor) and the parasitic inductance. Finally, one has to include the device geometry capacitance:

$$C_{geom} = \frac{\epsilon A}{L} \tag{7.3.3}$$

where L is the device length. The absence of the diffusion capacitance that dominates the forward-bias capacitance of a p-n diode allows a very fast response of the Schottky diode.

7.4 BIPOLAR JUNCTION TRANSISTORS: A CHARGE-CONTROL ANALYSIS

In our dc analysis of the bipolar transistor , we saw that when the device is under bias, current flows through each of the terminals, and a stored charge profile is established within the structure (figure 6.8). When an ac signal is applied, the stored charge and the current are modulated. However, the stored charge cannot be modulated instantaneously; once a signal is applied, a finite amount of time is required for the corresponding charge distribution to be established. Determining the switching behavior of a bipolar transistor essentially boils down to finding the relationship between the currents and the stored charge, the charge control model, and then determining the delays associated with modulation of stored charge. We shall discuss the small signal model of the bipolar transistor after the charge control model, since that allows the reader to better appreciate setting up the continuity equations for the minority carriers.

The charge-control model, presented in this section, establishes relationships between the currents and stored charge in the device. These relationships are quite useful for calculating delays. In section 7.5 the response of bipolar transistors to small signals is derived using the charge-control framework.

The two junctions of the BJT can be biased in several ways to produce four operating modes for the transistor, as was shown in figure 6.8. When the device is used for small-signal amplification, it remains biased in forward active mode. Hence, the analysis of the device in this mode is sufficient for deriving the response of the device to small signals (section 7.5). For large-signal applications, in addition to forward active mode, the device will also at times switch to saturation and cutoff modes. We will now briefly discuss behavior in all four modes and concentrate on the forward active mode in section 7.5.

Forward Active Mode

In this mode the emitter-base junction (EBJ) is forward biased, while the base-collector junction (BCJ) is reverse biased. We will use the subscript F to denote various terms in the forward active mode. The currents are given by the Ebers-Moll model discussed in section 6.3.3 ($eV_{CB} \gg k_BT$ in this mode):

$$\begin{aligned} I_E &= I_{ES} \exp\left(\frac{eV_{BE}}{k_BT}\right) + \alpha_R I_{CS} \\ I_C &= \alpha_F I_{ES} \exp\left(\frac{eV_{BE}}{k_BT}\right) + I_{CS} \end{aligned} \tag{7.4.1}$$

Here we assume that the emitter and collector current have the same direction. If we express $I_{ES} \exp\left(eV_{BE}/k_BT\right)$ in the second equation using the first equation, we can write

$$\begin{aligned} I_C &= \alpha_F \left(I_E - \alpha_R I_{CS}\right) + I_{CS} \\ &= \alpha_F I_E + I_{CS}\left(1 - \alpha_F \alpha_R\right) \end{aligned} \tag{7.4.2}$$

Using $I_E = I_B + I_C$, we have

$$I_C = \alpha_F I_B + \alpha_F I_C + I_{CS}\left(1 - \alpha_F \alpha_R\right) \tag{7.4.3}$$

or

$$\begin{aligned} I_C &= \frac{\alpha_F}{1-\alpha_F} I_B + \frac{I_{CS}\left(1 - \alpha_F \alpha_R\right)}{1-\alpha_F} \\ &= \beta_F I_B + \left(\beta_F + 1\right) I_{CS}\left(1 - \alpha_F \alpha_R\right) \end{aligned} \tag{7.4.4}$$

where

$$\beta_F = \frac{\alpha_F}{1-\alpha_F} \tag{7.4.5}$$

β_F represents the forward active current gain I_C/I_B for the transistor.

It is useful to examine the charge in the device in the forward active mode. In figure 6.8b we showed the minority charge injected in the emitter, base, and collector. The excess minority charge injected into the base region is given by

$$Q_F = \frac{eAW_{bn}n_{b0}}{2}\left(\exp\frac{eV_{BE}}{k_BT} - 1\right) \tag{7.4.6}$$

We can define the collector current in terms of the excess charge by defining a time constant τ_F which is the forward transit time of minority carriers through the base. We have

$$Q_F = \tau_F I_C \tag{7.4.7}$$

In other words, for a collector current I_C to be maintained, the excess minority charge in the base Q_F must be replaced every τ_F seconds. As discussed in chapter 3 , the forward transit time is

$$\tau_F = \frac{W_{bn}^2}{2D_b} \tag{7.4.8}$$

The base current I_B is due to recombination in the neutral base with the minority charge and hole injection into the emitter. These two effects can be summarized by a time constant τ_{BF} and we can write

$$I_B = \frac{Q_F}{\tau_{BF}} \tag{7.4.9}$$

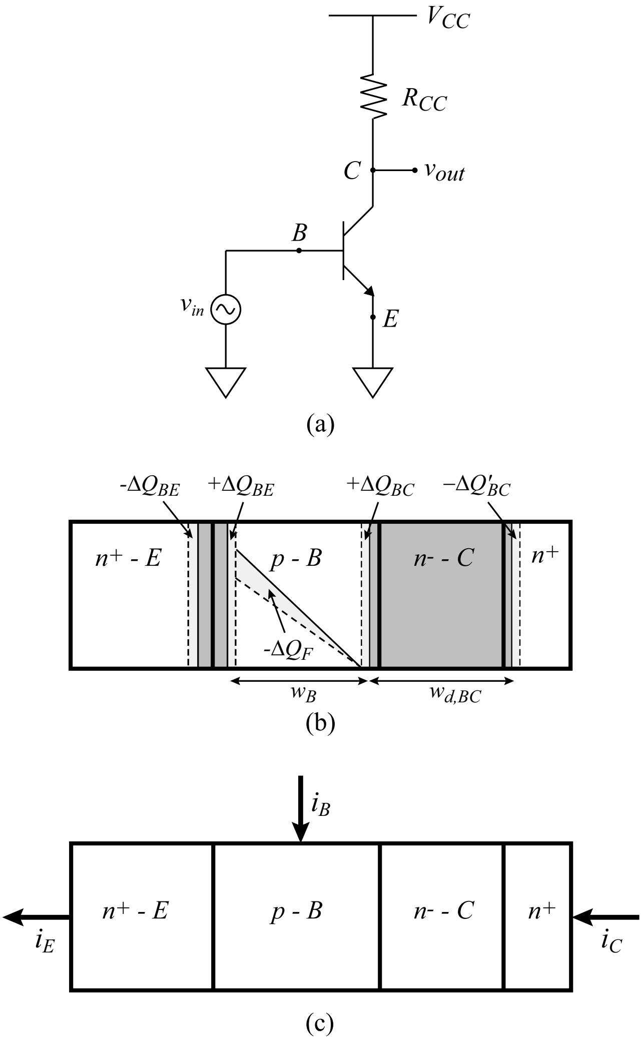
The current gain is then

$$\beta_F = \frac{I_C}{I_B} = \frac{\tau_{BF}}{\tau_F} \tag{7.4.10}$$

Now let us examine what happens when the junction voltages are modulated. Consider the transistor connected in a common emitter configuration shown in figure 7.7. When the emitter-base voltage is increased by ΔV_{BE}, the current <u>and</u> the stored charge in the device both change. The collector current increases by some amount ΔI_C, causing the collector voltage v_{out} to drop by an amount $\Delta v_{out} = \Delta I_C \cdot R_{CC}$. This <u>decreases</u> the reverse bias across the base-collector junction, causing the base-collector depletion region to become narrower, as illustrated in figure 7.7b. Additionally, because V_{BE} has increased, the emitter-base depletion region becomes narrower, and the injected minority charge in the base Q_F increases in magnitude. The variation in emitter-base and base-collector depletion widths implies a change in the amount of stored charge in each of the depletion regions, as indicated in figure 7.7b.

Figure 7.7c shows a schematic diagram of the current at each of the three terminals. Additional stored charge in the emitter, base, and collector regions must be supplied by the emitter, base, and collector currents. Including the current required to supply the additional stored charge, i_C, i_B, and i_E can be written as

$$\begin{aligned}
i_C &= \frac{Q_F}{\tau_F} - \frac{dQ_{BC}}{dt} \\
i_B &= \frac{Q_F}{\tau_{BF}} + \frac{dQ_F}{dt} + \frac{dQ_{BC}}{dt} + \frac{dQ_{BE}}{dt} \\
i_E &= i_C + i_B = Q_F\left(\frac{1}{\tau_F} + \frac{1}{\tau_{BF}}\right) + \frac{dQ_F}{dt} + \frac{dQ_{BE}}{dt}
\end{aligned} \tag{7.4.11}$$



(a)

(b)

(c)

Figure 7.7: Charge control model for forward active mode. (a) Bipolar transistor biased in common emitter configuration. (b) Charge components corresponding to a change in base-emitter voltage ΔV_{BE}. Because of charge injection into the base-collector depletion region, $\Delta Q_{BC} \neq \Delta Q^{'}_{BC}$ (c) Currents at each of the three transistor terminals.

This is valid at current densities much below the Kirk threshold density (section 6.6.1) where $\Delta Q_{BC} \approx \Delta Q^{'}_{BC}$ (see figure 7.7).

Reverse Active Mode

In the reverse active mode the EBJ is reverse biased while the BCJ is forward biased. Note that bipolar devices are asymmetrically doped, i.e., $N_{de} \gg N_{dc}$, and the reverse active mode has a poor current gain. The current and excess minority charge can be written, in analogy with the forward active mode case (note that the collector is now acting as the emitter):

$$\begin{aligned}
Q_R &= \frac{eAW_{bn}n_{b0}}{2}\left[\exp\left(\frac{eV_{BC}}{k_BT}\right)-1\right] \\
i_E &= \frac{-Q_R}{\tau_R}+\frac{dQ_{VE}}{dt} \\
i_B &= \frac{Q_R}{\tau_{BR}}+\frac{dQ_R}{dt}+\frac{dQ_{VC}}{dt}+\frac{dQ_{VE}}{dt} \\
i_C &= -Q_R\left(\frac{1}{\tau_R}+\frac{1}{\tau_{BR}}\right)-\frac{dQ_R}{dt}-\frac{dQ_{VC}}{dt}
\end{aligned} \qquad (7.4.12)$$

Cutoff Mode

In the cutoff mode, both junctions are reverse biased and we may write

$$\begin{aligned}
I_E &= -I_{ES}+\alpha_R I_{CS} \\
I_C &= -\alpha_F I_{ES}+I_{CS}
\end{aligned} \qquad (7.4.13)$$

In the cutoff mode the terminal currents are extremely small and there is an effective open circuit at the terminals.

Saturation Mode

In the saturation mode the EBJ and the BCJ are both forward biased. In this case a good approximation to the current-voltage equations is

$$\begin{aligned}
I_E &= I_{ES}\exp\left(\frac{eV_{BE}}{k_BT}\right)-\alpha_R I_{CS}\exp\left(\frac{eV_{BC}}{k_BT}\right) \\
I_C &= \alpha_F I_{FS}\exp\left(\frac{eV_{BE}}{k_BT}\right)-I_{CS}\exp\left(\frac{eV_{BC}}{k_BT}\right)
\end{aligned} \qquad (7.4.14)$$

In saturation there is charge injected into the base from the emitter (Q_F) and the collector (Q_R). The charge in the depletion region charge is negligible, since the junction voltages do not change much once the device is in saturation. The current-charge relations can be written as

$$\begin{aligned}
i_C &= \frac{Q_F}{\tau_F}-Q_R\left(\frac{1}{\tau_R}+\frac{1}{\tau_{BR}}\right)-\frac{dQ_R}{dt} \\
i_B &= \frac{Q_F}{\tau_{BF}}+\frac{Q_R}{\tau_{BR}}+\frac{d}{dt}\left(Q_F+Q_R\right)
\end{aligned} \qquad (7.4.15)$$

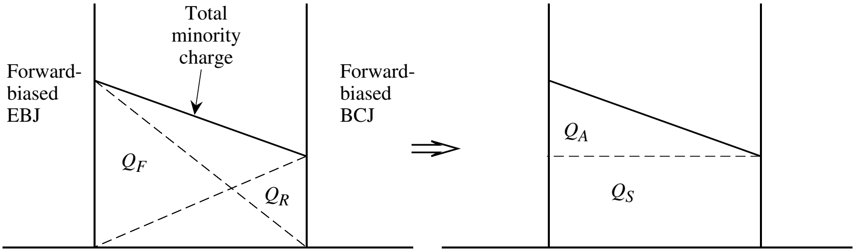


Figure 7.8: Minority charge in the base of a BJT in saturation mode. Charge is injected from the emitter and the collector into the base. The figure on the right shows a representation of the charge in terms of a uniform charge Q_S and charge Q_A.

The charges Q_F and Q_R are shown in figure 7.8. The total base charge may be written as shown on the right-hand side of figure 7.8:

$$\begin{aligned} Q_B &= Q_F + Q_R \\ &= Q_A + Q_S \end{aligned} \tag{7.4.16}$$

where Q_A represents the charge at the edge of saturation (EOS) and Q_S is the overdrive charge that drives the device into saturation. The charge Q_A can be written as

$$\begin{aligned} Q_A &= \tau_F I_{C(EOS)} = \tau_{BF} I_{B(EOS)} \\ \frac{I_{C(EOS)}}{I_{B(EOS)}} &= \beta_F \end{aligned} \tag{7.4.17}$$

The overdrive charge Q_S can be written as

$$Q_S = \tau_S I_{BS} \tag{7.4.18}$$

where I_{BS} is the base current over and above $I_{B(EOS)}$ that brings the device to the edge of saturation. The time τ_S is the weighted mean of τ_{BF} and τ_{BR}.

The static base current in saturation is

$$I_B = I_{B(EOS)} + I_{BS} \tag{7.4.19}$$

The instantaneous value of the base current is

$$i_B(t) = \frac{Q_A}{\tau_{BF}} + \frac{Q_S}{\tau_S} + \frac{dQ_S}{dt} \tag{7.4.20}$$

We can use the relation

$$\frac{Q_A}{\tau_{BF}} = \frac{\tau_F I_{C(EOS)}}{\tau_{BF}} = \frac{I_{C(EOS)}}{\beta_F} \tag{7.4.21}$$

so that we have

$$i_B(t) - \frac{I_{C(EOS)}}{\beta_F} = \frac{Q_S}{\tau_S} + \frac{dQ_S}{dt} \tag{7.4.22}$$

We will see later that in the switching characteristics of the BJT, the overdrive charge and the time constant τ_S appearing in the equation above play a critical role.

7.4.1 Junction Voltages at Saturation

Having discussed the various operating modes of the BJT, we will now obtain expressions for the junction voltages as the device goes into the saturation mode. These voltages are useful in studying the behavior of BJTs for logic elements. In figure 7.9 we show a simple model of the BJT in the saturation mode. Let us apply Kirchhoff's voltage law (KVL) to the voltage values:

$$\begin{aligned} V_{CE} &= V_{CB} + V_{BE} \\ &= -V_{BC} + V_{BE} \end{aligned} \tag{7.4.23}$$

Thus

$$V_{CE(sat)} = V_{BE(sat)} - V_{BC} \tag{7.4.24}$$

To obtain $V_{BE(sat)}$ we multiply the second of Eqns. 6.68 by α_R and subtract the resulting equation from the first of Eqns. 5.68. This gives

$$I_E - \alpha_R I_C = I_{ES}(1 - \alpha_F \alpha_R) e^{eV_{BE}/k_BT} \tag{7.4.25}$$

Using $I_E = I_B + I_C$, we find

$$I_B + I_C(1 - \alpha_R) = I_{ES}(1 - \alpha_F \alpha_R) e^{eV_{BE}/k_BT} \tag{7.4.26}$$

This gives for $V_{BE(sat)}$

$$V_{BE(sat)} = \frac{k_BT}{e} \ln \left[\frac{I_B + I_C(1 - \alpha_R)}{I_{EO}} \right] \tag{7.4.27}$$

where

$$I_{EO} = I_{ES}(1 - \alpha_F \alpha_R) \tag{7.4.28}$$

In a similar manner, the value of $V_{BC(sat)}$ is

$$V_{BC(sat)} = \frac{k_BT}{e} \ln \left[\frac{\alpha_F I_B - I_C(1 - \alpha_F)}{I_{CO}} \right] \tag{7.4.29}$$

with

$$I_{CO} = I_{CS}(1 - \alpha_F \alpha_R) \tag{7.4.30}$$

From these values of $V_{BE(sat)}$ and $V_{BC(sat)}$ we have

$$V_{CE(sat)} = \frac{k_BT}{e} \ln \left[\frac{I_B + I_C(1 - \alpha_R)}{\alpha_F I_B - I_C(1 - \alpha_F)} \cdot \frac{I_{CO}}{I_{EO}} \right] \tag{7.4.31}$$

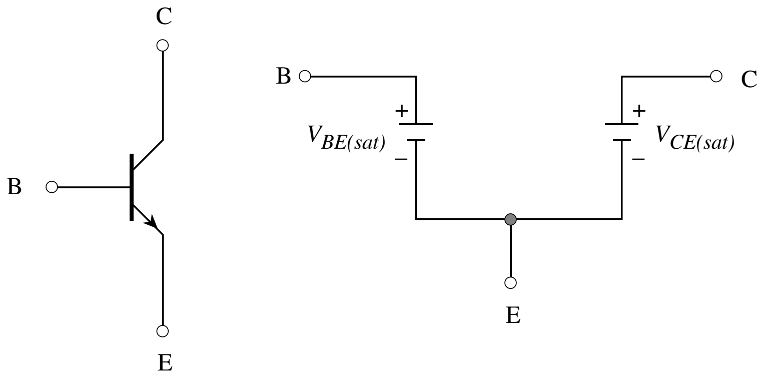


Figure 7.9: The BJT and a simple model for a device in the saturation mode.

Note that

$$\frac{I_{CO}}{I_{EO}} = \frac{I_{CS}}{I_{ES}} = \frac{\alpha_F}{\alpha_R} \tag{7.4.32}$$

The equation for $V_{CE(sat)}$, after some simple manipulation, can be written as

$$V_{CE(sat)} = \frac{k_B T}{e} \ln \left[\frac{\frac{1}{\alpha_R} + \frac{I_C}{I_B}\frac{1-\alpha_R}{\alpha_R}}{1 - \frac{I_C}{I_B}\frac{1-\alpha_F}{\alpha_F}} \right] \tag{7.4.33}$$

We finally substitute for the current gains $\beta_R = \alpha_R/(1-\alpha_R)$, $\beta_F = \alpha_F/(1-\alpha_F)$ to get

$$V_{CE(sat)} = \frac{k_B T}{e} \ln \left[\frac{\frac{1}{\alpha_R} + \frac{I_C}{I_B}\frac{1}{\beta_R}}{1 - \frac{I_C}{I_B}\frac{1}{\beta_F}} \right] \tag{7.4.34}$$

Typical values of $V_{CE(sat)}$ from the expression derived here are $\sim$ 50 mV. If one adds to this value the voltage drop across the neutral regions of the emitter and the collector, we find that $V_{CE(sat)}$ is $\sim$ 0.1 V. For silicon devices typical values for the various junction voltages, are

$$\begin{aligned} V_{BE(sat)} &\sim 0.8\text{ V} \\ V_{CE(sat)} &\sim 0.1\text{ V} \end{aligned} \tag{7.4.35}$$

7.5 HIGH-FREQUENCY BEHAVIOR OF A BJT

An important application of bipolar transistors is in the amplification of high-frequency small signals. For this application, the device is biased as shown in figure 7.7a, and a signal v_{in} is applied at the base terminal, resulting in an output voltage v_{out} at the collector terminal. The device remains in forward active mode at all times, so the charge control framework developed for this mode is sufficient to derive the small-signal response.

When the emitter current in the device is modulated by an amount Δi_E, the collector current does not respond immediately. The delay in establishing the change in collector current Δi_C is a result of the finite time required to modulate the various stored charge elements in the device. The total emitter to collector delay τ_{EC} is given by

$$\tau_{EC} = \tau_{BE} + \tau_B + \tau_{BC} \tag{7.5.1}$$

where τ_{BE} is the EBJ capacitance charging time, τ_B is the total delay in the quasi-neutral base region, and τ_{BC} is the delay associated with the base-collector capacitance (which includes the contribution due to change in mobile charge in the collector, which is equivalent to the collector transit delay). It will be shown later in section 7.5.3 that the current gain cutoff frequency f_τ of the device is given by

$$f_\tau = \frac{1}{2\pi\tau_{EC}} \tag{7.5.2}$$

This is the maximum frequency at which it is possible to achieve current gain in the device.

To calculate the delays in the device, we apply the following rule. The ratio of the change in stored charge to the change in current is the delay associated with the element. We are interested in the delay in setting the output current. We can write the ac portions of equation 7.4.11 as

$$\begin{aligned} |\Delta i_E| &= \frac{\Delta Q_F}{\tau_B} + \frac{\Delta Q_{BE}}{\tau_{BE}} \\ |\Delta i_C| &= \frac{\Delta Q_{BC}}{\tau_{BC}} \end{aligned} \tag{7.5.3}$$

The delay element τ_{BE} can be written as

$$\tau_{BE} = \frac{\Delta Q_{BE}}{\Delta I_E} = C_{BE}\left(\frac{\Delta V_{BE}}{\Delta I_E}\right) \cong C_{BE}\left(\frac{\Delta V_{BE}}{\Delta I_C}\right) \tag{7.5.4}$$

where

$$\left(\frac{\Delta V_{BE}}{\Delta I_C}\right)^{-1} = g_{m0} = (r_e)^{-1} \cong \frac{eI_C}{k_B T} \tag{7.5.5}$$

is the transconductance of the device. τ_{BE} is therefore given by

$$\boxed{\tau_{BE} = r_e C_{BE} = \left(\frac{k_B T}{eI_C}\right) C_{BE}} \tag{7.5.6}$$

The base delay τ_B is the time required to supply the additional charge ΔQ_F to the quasi-neutral base region. If we assume Shockley boundary conditions $(n_p(w_B) = 0)$, then τ_B can be written as

$$\tau_B = \frac{w_B^2}{2D_{nB}} \tag{7.5.7}$$

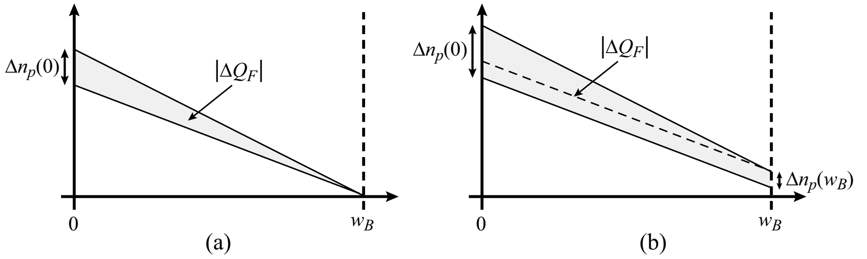


Figure 7.10: Charge profile in the base when device is modulated. (a) Approximate base charge profile with Shockley boundary conditions. (b) Base charge profile when velocity saturation in the collector is included.

where ΔQ_F for this case is shown in figure 7.10a. However, as we saw in the case of the Kirk effect, the carrier density cannot drop to zero at the collector side of the neutral base. Instead, $n_p(w_B)$ is given by

$$n_p(w_B) = \frac{I_C/A_E}{ev_{sat}} \tag{7.5.8}$$

where v_{sat} is the electron saturation velocity in the material and A_E is the emitter area. If we increase the current in the device by an amount ΔI_C, the electron density at the collector side of the base also rises, as indicated in figure 7.10b. This results in additional charge that must be supplied to the base region. In figure 7.10b, the charge above the dotted line is equal to the total charge supplied when Shockley boundary conditions are assumed (figure 7.10a). The charge below the dotted line is the additional charge due to velocity saturation in the collector. Including velocity saturation effects, τ_B can be written as

$$\boxed{\tau_B = \frac{w_B^2}{2D_{nB}} + \frac{w_B}{v_{sat}}} \tag{7.5.9}$$

In the base-collector depletion region, the change in stored charge ΔQ_{BC} results from two separate effects.

1. The base-collector depletion width changes due to the variation in the base-collector voltage, as was illustrated in figure 7.1b. We will refer to the associated change in space charge as ΔQ_{BC1}.

2. As discussed in the derivation of τ_B, because of velocity saturation, the mobile charge density in the base-collector depletion region must also increase when I_C increases (this is the origin of $\Delta n_p(w_B)$ in figure 7.10b). Because this increased mobile charge cannot result in a change in voltage across the depletion region (the voltage is fixed by the bias conditions), the base-collector depletion widths are adjusted by an additional amount to

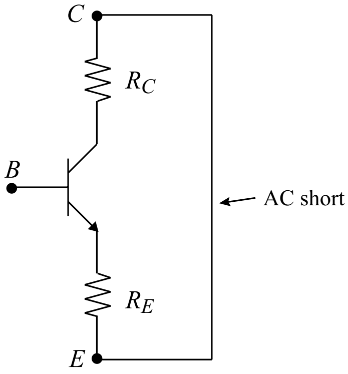


Figure 7.11: Biasing circuit for calculation of τ_{BC1}.

accommodate the mobile charge. Stated differently, each electron introduced into the base-collector depletion region must be imaged at the depletion edges. Because the induced charge due to this effect at both ends of the depletion region is positive, the total charge at the base end increases, while the charge at the collector end decreases in magnitude (becomes less negative). Hence referring to figure 7.7b,$\Delta Q_{BC} \neq \Delta Q'_{BC}$. Since the charge at the base end (ΔQ_{BC}) is the one which must be supplied at the input to induce a change in the output current, it is this charge that we are interested in calculating. We will call the change in the base-collector depletion charge associated with finite electron velocity in the collector, ΔQ_C.

The total change in charge in the base-collector depletion region $\Delta Q_{BC} = \Delta Q_{BC1} + \Delta Q_C$.

We will split the base-collector delay into two components, $\tau_{BC} = \tau_{BC1} + \tau_C$, where

$$\tau_{BC1} = \frac{\Delta Q_{BC1}}{\Delta I_C} \tag{7.5.10}$$

$$\tau_C = \frac{\Delta Q_C}{\Delta I_C} \tag{7.5.11}$$

τ_C is commonly referred to as the collector delay . To determine τ_{BC1}, we refer to the circuit shown in figure 7.11. Delay analysis is by convention carried out with the collector incrementally shorted to the emitter. Assuming a change in the base-emitter voltage ΔV_{BE} leads to a change in collector current ΔI_C, we can write the following expression for ΔV_{BC}.

$$\Delta V_{BC} = \Delta V_{BE} + \Delta I_C \left(R_E + R_C\right) \tag{7.5.12}$$

By definition,

$$\Delta Q_{BC1} = C_{BC} \Delta V_{BC} \tag{7.5.13}$$

Inserting the above equation and equation 7.5.12 into equation 7.5.10 gives us for τ_{BC1}

$$\begin{aligned}\tau_{BC1} &= \frac{C_{BC}\Delta V_{BC}}{\Delta I_C} = \frac{C_{BC}\left(\Delta V_{BE} + \Delta I_C\left(R_E + R_C\right)\right)}{\Delta I_C} \\ &= \left(\frac{\Delta V_{BE}}{\Delta I_C}\right) C_{BC} + \left(R_E + R_C\right) C_{BC}\end{aligned} \tag{7.5.14}$$

$$\boxed{\tau_{BC1} = \left(r_E + R_E + R_C\right) C_{BC}} \tag{7.5.15}$$

In calculating τ_C, we will assume that the electron velocity profile in the base-collector depletion region does not necessarily need to remain constant. This would, for example, be the case if the material composition in the collector was varied, such as in a double heterojunction bipolar transistor structure, or in the case of non-stationary transport in short collectors. The increased electron concentration in the collector $\Delta n(x)$ as a function of ΔI_C is then given by

$$\Delta n(x) = \frac{\Delta I_C}{A_E e v_e(x)} \tag{7.5.16}$$

where A_E is the emitter area and $v_e(x)$ is the electron velocity at a point x in the collector. figure 7.12 shows a schematic plot of $\Delta n(x)$ in the collector for an arbitrary velocity distribution $v_e(x)$.

We first need to calculate ΔQ_C. To do this, we find the induced charge $d\left(\Delta Q_C\right)$ at $x = 0$ caused by a sheet of charge $-e\Delta n(x)dx$ at a point x, and integrate from $x = 0$ to $x = w_C$, as illustrated in figure 7.13a (note that we assume $w_{d,BC} \simeq w_C$, since the base and subcollector are doped highly and the collector is typically fully depleted when the device is under bias). The electric field induced in the depletion region by each sheet charge element is shown in figure 7.13b. Using Gauss' Law, we can relate $d(\Delta Q_C)$ and $e\Delta n(x)dx$ to $d\mathcal{E}^+(x)$ and $d\mathcal{E}^-(x)$.

$$d\mathcal{E}^+(x) = \frac{d\left(\Delta Q_C\right)}{\epsilon A_E} \tag{7.5.17}$$

$$d\mathcal{E}^+(x) + d\mathcal{E}^-(x) = \frac{e\Delta n(x)dx}{\epsilon} \tag{7.5.18}$$

Also, since the change in voltage in the collector due to the induced charge must be zero, the area under $d\mathcal{E}^+(x)$ in figure 7.13b must equal the area above $d\mathcal{E}^-(x)$, or

$$x \cdot d\mathcal{E}^+(x) = \left(w_C - x\right) d\mathcal{E}^-(x) \tag{7.5.19}$$

Solving for $\Delta\mathcal{E}^-(x)$ in this equation gives us

$$d\mathcal{E}^-(x) = \left(\frac{x}{w_C - x}\right) d\mathcal{E}^+(x) \tag{7.5.20}$$

We can then substitute this result into equation 7.5.18 to get

$$d\mathcal{E}^+(x) + \left(\frac{x}{w_C - x}\right) d\mathcal{E}^+(x) = \frac{e\Delta n(x)dx}{\epsilon} \tag{7.5.21}$$

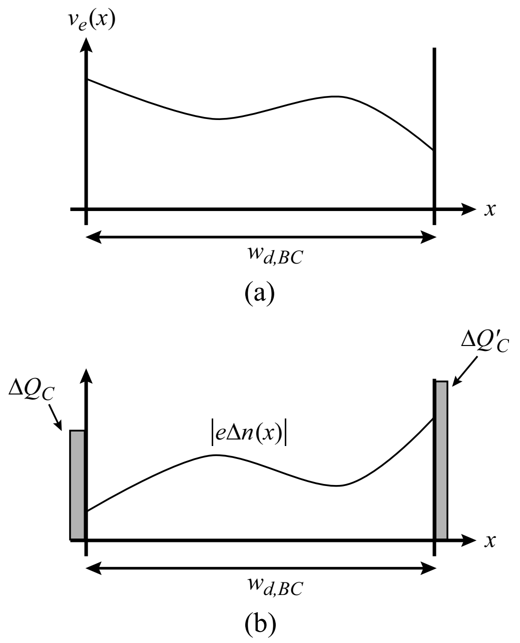


Figure 7.12: (a) Arbitrary velocity profile $v_e(x)$ in the base-collector depletion region. (b) Injected charge $\Delta n(x)$ corresponding to this velocity profile, along with the image charges at the depletion edges. $A_E \int_0^{w_{d,BC}} |e\Delta n(x)| dx = \Delta Q_C + \Delta Q_C^{'}$ (charge neutrality)

Solving for $d\mathcal{E}^+(x)$ and using equation 7.5.16 to substitute for $\Delta n(x)$ gives us

$$d\mathcal{E}^+(x) = \left(1 - \frac{x}{w_C}\right) \frac{\Delta I_C}{\epsilon A_E v_e(x)} dx \tag{7.5.22}$$

Substituting equation 7.5.17 into the equation above and integrating both sides, we can now solve for ΔQ_C.

$$d\left(\Delta Q_C\right) = \left(1 - \frac{x}{w_C}\right) \frac{\Delta I_C}{v_e(x)} dx \tag{7.5.23}$$

$$\Delta Q_C = \Delta I_C \int_0^{w_C} \left(1 - \frac{x}{w_C}\right) \frac{1}{v_e(x)} dx \tag{7.5.24}$$

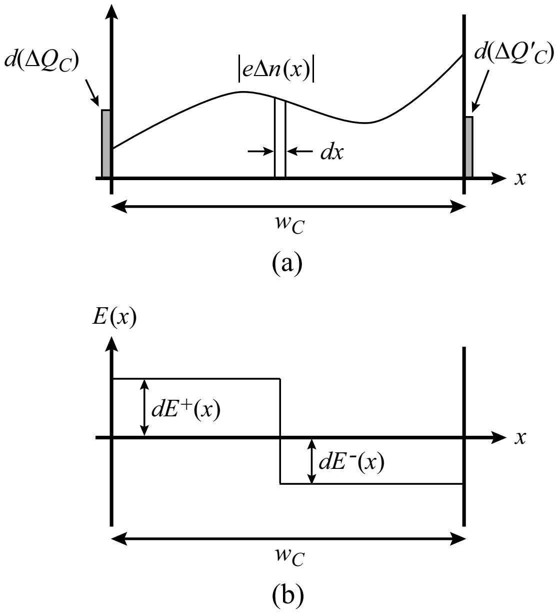


Figure 7.13: Induced (a) image charge and (b) electric field due to an injected sheet charge $-e\Delta n(x)dx$ at a point x.

Finally, we can solve for τ_C.

$$\boxed{\tau_C = \frac{\Delta Q_C}{\Delta I_C} = \int_0^{w_C} \left(1 - \frac{x}{w_C}\right) \frac{1}{v_e(x)} dx} \tag{7.5.25}$$

For the case of a constant electron velocity v_s, τ_C is given by

$$\tau_C = \frac{w_C}{2v_s} \tag{7.5.26}$$

The delay analysis for bipolar transistors presented here accurately describes the frequency limitations of the device and provides us with the tools required to design devices for high frequency operation. However, it does not give us any information about how the device will perform at frequencies less than f_τ. Since these transistors will ultimately be used in circuits, we need to be able to determine the frequency response of a circuit containing these devices. It is therefore necessary to derive a small-signal model of the device that can then be applied in circuit simulations. We will see in the next section that the discrete components of the bipolar equivalent circuit can be written in terms of the delays that we have derived.

Example 7.1 Consider an npn transistor with the following properties at 300 K:

Emitter current,	I_E	$=$	$1.5\ \mathrm{mA}$
EBJ capacitance,	C_{je}	$=$	$2pF$
Base width,	W_b	$=$	$0.4\ \mu\mathrm{m}$
Diffusion coefficient,	D_b	$=$	$60\ \mathrm{cm}^2/\mathrm{s}$
Width of collector depletion region,	W_{dc}	$=$	$2.0\ \mu\mathrm{m}$
Collector resistance,	r_C	$=$	30Ω
Total collector capacitance,	$(C_s + C_\mu)$	$=$	$0.4pF$
Saturated electron velocity,	v_s	$=$	$5 \times 10^6\ \mathrm{cm/s}$

Calculate the cutoff frequency of this transistor. How will the cutoff frequency change (i) if the emitter current level is doubled? (ii) if the base thickness is halved?

The emitter resistance $r_e^{'}$ is given by (see equation 7.5.11 for the resistance of a forward-biased diode)

$$r_e^{'} = \frac{dI_E}{dV_{BE}} \cong \frac{k_BT}{eI_E} = \frac{0.026}{1.5 \times 10^{-3}} = 17.3\ \Omega$$

This gives

$$\tau_e = r_e^{'} C_{je} = (17.3)(2 \times 10^{-12}) = 34.6\ \mathrm{ps}$$

The base transit time is

$$\tau_t = \frac{W_b^2}{2D_b} = \frac{(0.4 \times 10^{-4})^2}{2 \times 60} = 13.3\ \mathrm{ps}$$

The collector transit time is

$$\tau_C = \frac{W_{dc}}{2v_s} = \frac{(2.0 \times 10^{-4})}{1 \times 10^7} = 10\ \mathrm{ps}$$

The collector charging time is

$$\tau_c = r_c(C_\mu + C_s) = 30(0.4 \times 10^{-12}) = 12\ \mathrm{ps}$$

The total time is

$$\tau_{ec} = 34.6 + 13.3 + 10 + 12 = 69.9\ \mathrm{ps}$$

The cutoff frequency is

$$f_\tau = \frac{1}{2\pi\tau_{ec}} = \frac{1}{2\pi(69.9 \times 10^{-12}\ \mathrm{s})} = 2.3\ \mathrm{GHz}$$

If the emitter current is doubled (assuming no other change occurs), the time τ_e is reduced by half. This gives a cutoff frequency of 2.54 GHz. Similarly, if the base width is reduced by half, the base transit time becomes 3.3 ps and the cutoff frequency becomes 2.65 GHz. In this problem the dominant source of delay is the emitter junction.

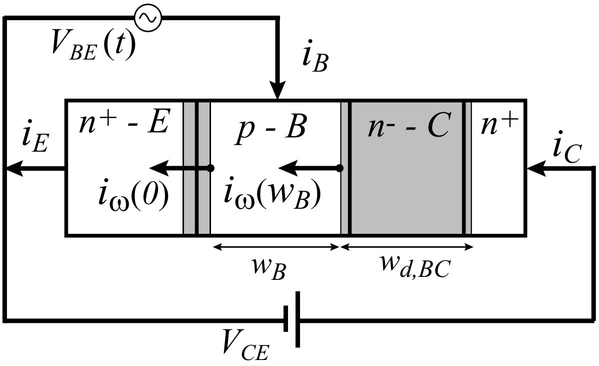


Figure 7.14: Current components in a bipolar transistor when a small signal v_{in} is applied. The direction of the arrows shows the direction of the electron flux.

7.5.1 Bipolar Transistor Small-Signal Equivalent Circuit

In figure 7.14, we show a schematic diagram of various current components in a bipolar transistor when a small signal $v_{in}(t)$ is applied. The total base-emitter voltage $V_{BE}(t)$ is given by

$$V_{BE}(t) = V_{dc} + v_{in}(t) \tag{7.5.27}$$

where we assume $v_{in}(t)$ to be of the form

$$v_{in}(t) = v_\omega \mathrm{e}^{j\omega t} \tag{7.5.28}$$

This generates a small-signal current i_E at the emitter. The current entering the base is denoted as $i_\omega(0)$. In general, $i_\omega(0)$ differs from i_E because of the delay in the emitter-base depletion region. Electrons then continue through the base, where they undergo a transit delay τ_B, resulting in a flux $i_\omega(w_B)$ leaving the base. Finally, the delay in the base-collector region results in an output current i_C at the collector. We are interested in determining the output current i_C as a function of the input voltage v_{in}.

We continue to make the assumption that the current in the base is purely diffusive and is therefore given by

$$i_\omega(x) = eA_E D_n \frac{\partial n_\omega(x)}{\partial x} \tag{7.5.29}$$

where $i_\omega(x)$ and $n_\omega(x)$ are the position-dependent amplitudes of the ac current and charge. Thus to determine $i_\omega(x)$, we must first calculate $n_\omega(x)$. In order to do this, it is necessary to solve the time-dependent continuity equation for electrons, which in the case of zero recombination takes the form

$$\frac{\partial n(x,t)}{\partial t} = -\frac{\partial}{\partial x}\left(\frac{I_e(x,t)}{-eA_E}\right) = D_n \frac{\partial^2 n(x,t)}{\partial x^2} \tag{7.5.30}$$

We assume solutions of the form

$$n(x,t) = n_{dc}(x) + n_\omega(x)\mathrm{e}^{j\omega t} \tag{7.5.31}$$

where $n_{dc}(x)$ is the dc component of the current calculated above and $n_\omega(x,t) = n_\omega(x)\mathrm{e}^{j\omega t}$. If we insert the ac part of equation 7.5.30 back into equation 7.5.29, the result can be written in the form

$$\frac{d^2 n_\omega(x)}{dx^2} = \frac{n_\omega(x)}{\lambda_e^2} \tag{7.5.32}$$

where λ_e is the frequency dependent diffusion length and is given by

$$\frac{1}{\lambda_e} = \sqrt{\frac{j\omega}{D_n}} = (1+j)\sqrt{\frac{\omega}{2D_n}} \tag{7.5.33}$$

Assuming $n_\omega(w_B) = 0$ (Shockley boundary conditions), the solution for $n_\omega(x)$ in equation 7.5.31 is given as

$$\boxed{n_\omega(x) = n_\omega(0)\frac{\sinh\left[\left(w_B - x\right)/\lambda_e\right]}{\sinh\left[w_B/\lambda_e\right]}} \tag{7.5.34}$$

where $n_\omega(0)$ is the amplitude of the the ac portion of the electron concentration at $x = 0$.

The value of $n_\omega(0)$ obviously depends on the magnitude of the ac voltage, which we have called v_ω. Again assuming Shockley boundary conditions, $n(0,t)$ can be written as

$$n(0,t) = n_{p0}\exp\left[\frac{eV_{BE}(t)}{k_BT}\right] = n_{p0}\exp\left(\frac{eV_{dc}}{k_BT}\right)\exp\left(\frac{ev_\omega \mathrm{e}^{j\omega t}}{k_BT}\right) \tag{7.5.35}$$

Assuming v_ω is small, we can linearize the second exponential in this equation, such that

$$\exp\left(\frac{ev_\omega \mathrm{e}^{j\omega t}}{k_BT}\right) \simeq 1 + \frac{ev_\omega}{k_BT}\mathrm{e}^{j\omega t} \tag{7.5.36}$$

We may then write equation 7.5.35 as

$$n(0,t) = n_{dc}(0) + n_\omega(0)\mathrm{e}^{j\omega t} \tag{7.5.37}$$

where

$$n_{dc}(0) = n_{p0}\exp\left(\frac{eV_{dc}}{k_BT}\right) \tag{7.5.38}$$

and

$$\boxed{n_\omega(0) = n_{p0}\exp\left(\frac{eV_{dc}}{k_BT}\right)\cdot\frac{ev_\omega}{k_BT} = n_{dc}(0)\frac{ev_\omega}{k_BT}} \tag{7.5.39}$$

Now that we know our ac charge distribution $n_\omega(x)$, we can calculate the amplitudes of the ac currents $i_\omega(0)$ and $i_\omega(w_B)$. If we insert $n_\omega(x)$ from equation 7.5.34 into equation 7.5.29 and evaluate the derivative at $x = 0$, we get

$$i_\omega(0) = -\frac{eA_E D_n n_\omega(0)}{\lambda_e}\coth\left[\frac{w_B}{\lambda_e}\right] \tag{7.5.40}$$

We see that the injected current amplitude is complex, indicating that the current has both a conductive (real) and capacitative (imaginary) part. If the frequency is sufficiently low such that $w_B << \sqrt{\omega/(2D_n)}$, the hyperbolic cotangent term may be expanded in the following manner:

$$\alpha \coth \alpha = 1 + \frac{\alpha^2}{3} + \text{H.O.T.} \tag{7.5.41}$$

This gives us for $i_\omega(0)$

$$i_\omega(0) = -\frac{eA_E D_n n_\omega(0)}{w_B}\left[1 + j\omega\frac{w_B^2}{3D_n}\right] \tag{7.5.42}$$

If we insert the expression for $n_\omega(0)$ from equation 7.5.38 into equation 7.5.41, we can express $i_\omega(0)$ in terms of our input signal v_ω

$$i_\omega(0) = -\frac{eA_E D_n n_{dc}(0)}{w_B}\frac{e}{k_B T}\left[1 + j\omega\frac{w_B^2}{3D_n}\right] v_\omega \tag{7.5.43}$$

or

$$\boxed{i_\omega(0) = -\frac{eI_E}{k_B T}\left[1 + j\omega\frac{w_B^2}{3D_n}\right] v_\omega} \tag{7.5.44}$$

where I_E is the dc emitter current. $i_\omega(0)$ may also be written in the form

$$\boxed{i_\omega(0) = -\left(G_s + j\omega C_{diff}\right) v_\omega} \tag{7.5.45}$$

where

$$G_s = \frac{1}{r_e} = \frac{eI_E}{k_B T} \tag{7.5.46}$$

is the emitter-base diode conductance , and

$$C_{diff} = \frac{2}{3}\frac{\partial Q_F}{\partial V_{BE}} = \frac{2}{3}C_B = \frac{eA_E n_{dc}(0) w_B}{3}\frac{e}{k_B T} \tag{7.5.47}$$

is the diffusion capacitance measured at the emitter terminal. As was discussed in section 7.2, the diffusion capacitance is 2/3 the value of the apparent diffusion capacitance (C_B), since for a short-base diode only 2/3 of the charge stored in the base is reclaimable. Finally, recognizing that C_B can be written related to the base transit time τ_B by

$$\tau_B = r_e \cdot C_B \tag{7.5.48}$$

we may express $i_\omega(0)$ as

$$\boxed{i_\omega(0) = -\frac{1}{r_e}\left(1 + j\omega\frac{2\tau_B}{3}\right) v_\omega} \tag{7.5.49}$$

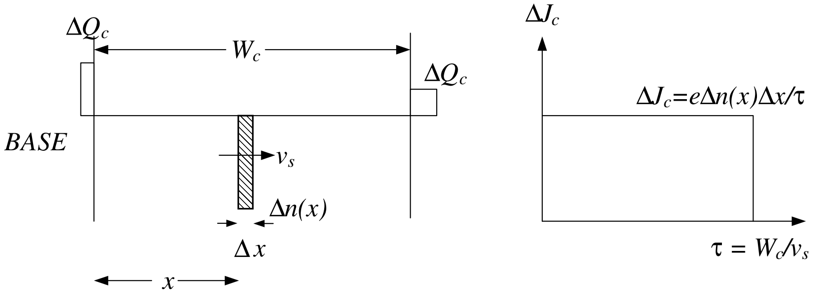


Figure 7.15: The induced charges and collector current versus time for a sheet of charge traveling at a constant velocity v_s.

Now, to calculate the flux leaving the base $i_\omega(w_B)$, we insert $n_\omega(x)$ from equation 7.5.34 into equation 7.5.29 and evaluate the derivative at $x = w_B$. If we Taylor expand the result and neglect higher order terms, we can express $i_\omega(w_B)$ as

$$\boxed{i_\omega(w_B) = -\frac{1}{r_e}\left(1 - j\omega\frac{\tau_B}{3}\right)v_\omega} \tag{7.5.50}$$

Note that the reactive part of $i_\omega(w_B)$ corresponds to a negative capacitance. This behavior results from the fact that any rise in current at the emitter end of the base appears at the collector end with a delay of one transit time τ_B, as can be seen by examination of equation 7.5.49 and equation 7.5.50.

Now all that remains is to calculate the collector current i_C. We showed in our time delay analysis (see figure 7.13) that electrons which are injected into the base-collector depletion region have a finite velocity and thus require a finite amount of time to transit this region. As the electrons travel through the depletion region, they induce image charges at the depletion edges, as illustrated in figure 7.13a. The collector current is equal to the time rate of change of the induced charge at the collector end of the depletion region.

7.5.2 Attenuation and Phase Shift of a Traveling Electron Wave

To analyze the delay introduced due to velocity saturation in the collector, we first derive the current, ΔJ_c induced by a sheet of charge of areal density $\Delta n(x) \cdot \Delta x$ traveling with a velocity v_s and a distance x from the edge of the base as shown in figure 7.15.

Following the analysis in section 7.5 (equation 7.5.16 through equation 7.5.26) and using

$$\frac{\Delta J_c}{v(x)} = e\Delta n(x) \cdot \Delta x \tag{7.5.51}$$

we see that the charge induced on the base is

$$\Delta Q_c = (1 - \frac{x}{W_c}) \cdot e\Delta n(x) \cdot \Delta x \tag{7.5.52}$$

and the image on the collector is by charge neutrality

$$\Delta Q_c^{'} = \Delta n(x)\Delta x - \Delta Q_c = e\Delta n(x)\Delta x \cdot \frac{x}{W_c} \tag{7.5.53}$$

The displacement current flowing in the external circuit, ΔJ_c is given by

$$\Delta J_c = \frac{d}{dt}\Delta Q_c^{'} = e\Delta n(x)\Delta x \cdot \frac{dx}{dt} \cdot \frac{1}{W_c} \tag{7.5.54}$$

Using $\frac{dx}{dt} = v_s$ we arrive at an important relationship also known as the Ramo-Shockley theorem

$$\Delta J_c = \frac{e\Delta n(x)\Delta x}{\tau} \tag{7.5.55}$$

The current carrying electrons in the collector can be assumed to comprise of several sheet charges of magnitude $\Delta n(x)\Delta x$. Hence the net induced current due to the electrons will be a sum (integral) of all the induced currents. The total current per unit area is therefore obtained by integration over all sheets:

$$J = -e \cdot (v/w) \cdot \int n(X) \cdot dX \tag{7.5.56}$$

We apply (equation 7.5.56) to a traveling electron wave of the form

$$n(x,t) = n_0 \exp\left[j\omega(t - x/v)\right] \tag{7.5.57}$$

It clearly corresponds to a wave of (angular) frequency ω traveling with a uniform speed v. The *convection* current in the plane $x = 0$ is evidently

$$i_\omega(w_B) = -en_0 v \cdot \exp\left(j\omega t\right) \tag{7.5.58}$$

this is the current density that would be flowing if the capacitor were infinitesimally thin and the transit time of the electrons through the capacitor were zero. With the help of equation 7.5.58 we may write equation 7.5.57 as

$$n(x,t) = -\left(\frac{i_\omega(w_B)}{ev}\right) \exp\left(-j\omega x/v\right) \tag{7.5.59}$$

If this is inserted into equation 7.5.56, and the integration executed, one finds readily

$$J = i_\omega(w_B)\frac{\exp(-j\omega t) - 1}{-j\omega t} \tag{7.5.60}$$

where we have introduced the electron transit time through the capacitor,

$$\tau = w/v \tag{7.5.61}$$

The expression equation 7.5.60 is easily transformed into the product of an amplitude factor and a phase factor:

$$J = i_\omega(w_B) \cdot \left[\frac{\sin(\omega\tau/2)}{\omega\tau/2}\right] \exp(-j\omega\tau/2) \tag{7.5.62}$$

The two factors following $i_\omega(w_B)$ indicate the attenuation and the phase shift of the current leaving the capacitor by the finite transit time through the capacitor.

We note first of all that the signal delay is only one-half the transit time of the electrons themselves. We also note that there is an attenuation, due to the destructive interference between different portions of the traveling wave. For $\omega = 2\pi/\tau$ the amplitude factor is zero, and no current is collected at all. This is the case when the wavelength $\lambda = 2\pi v/\omega$ of the traveling wave is equal to the capacitor plate separation w.

The two terms following $i_\omega(w_B)$ indicate that that the signal passing through the base-collector depletion capacitor has been both attenuated and phase shifted as a result of the finite transit time through this region.

Substituting for $i_\omega(w_B)$ we can express the output current i_C in terms of the input signal v_ω.

$$i_C = -\frac{v_\omega}{r_e}\left(1 - j\omega\frac{\tau_B}{3}\right) \cdot \frac{\sin(\omega\tau_C)}{\omega\tau_C} \cdot \exp(-j\omega\tau_C) \tag{7.5.63}$$

If the frequency ω is sufficiently small, this may be written as

$$\boxed{i_C = -\frac{v_\omega}{r_e} \cdot \frac{\sin(\omega\tau_C)}{\omega\tau_C} \cdot \exp\left[-j\omega\left(\frac{\tau_B}{3} + \tau_C\right)\right]} \tag{7.5.64}$$

The device transconductance, g_m, is defined as

$$g_m = \frac{\partial i_C}{\partial v_\omega} \tag{7.5.65}$$

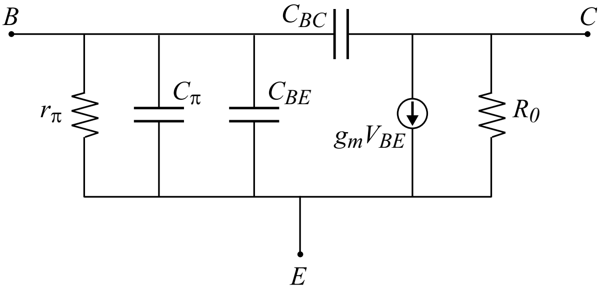
Inserting equation 7.5.65 into equation 7.5.63, we get for the bipolar transistor transconductance

$$\boxed{g_m = g_{m0} \cdot \frac{\sin(\omega\tau_C)}{\omega\tau_C} \cdot \exp\left[-j\omega\left(\frac{\tau_B}{3} + \tau_C\right)\right]} \tag{7.5.66}$$

where $g_{m0} = r_e^{-1}$ is the device transconductance at dc.

The base current i_B in figure 7.14 is simply the difference in i_E and i_C, or

$$i_B = i_E - i_C = i_\omega(0) - i_C \tag{7.5.67}$$



Element	Value
r_e	$\frac{k_B T}{I_E}$
r_π	$\beta_0 r_e$
C_π	$\frac{\tau_B+\tau_C}{r_e}$
R_0	$\frac{V_A}{I_C}$

Figure 7.16: Small-signal model of a bipolar transistor.

Inserting our expressions from , we get

$$i_B = -\frac{v_\omega}{r_e}\left[\left(1 + j\omega\frac{2\tau_B}{3}\right) - \left(1 - j\omega\frac{\tau_B}{3} - j\omega\tau_C\right)\right] \tag{7.5.68}$$

where we have assumed ω to be small so that

$$i_C \simeq -\frac{v_\omega}{r_e}\left[1 - j\omega\left(\frac{\tau_B}{3} + \tau_C\right)\right] \tag{7.5.69}$$

Simplifying equation 7.5.57 gives us for i_B

$$\boxed{i_B = -j\omega\frac{(\tau_B + \tau_C)}{r_e}v_\omega = -j\omega C_\pi v_\omega} \tag{7.5.70}$$

where

$$C_\pi = \frac{\tau_B + \tau_C}{r_e} \tag{7.5.71}$$

which illustrates how collector delay adds to the input capacitance.

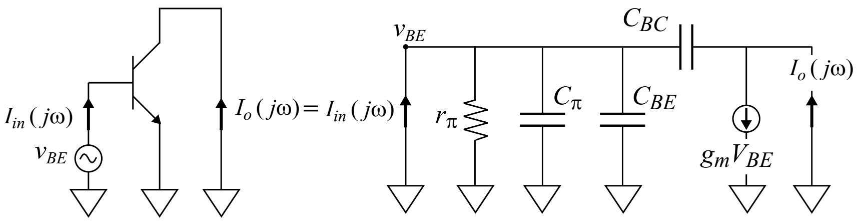


Figure 7.17: Bipolar equivalent circuit for calculating f_T.

Now that we have derived all of the small-signal currents in the device and expressed them in terms of conductive and capacitative components, it is relatively straightforward to construct a small-signal equivalent model. This model is shown in figure 7.16.

7.5.3 Small Signal Figures of Merit

Current gain cutoff frequency f_T

As stated earlier, the current gain cutoff frequency f_T is defined as the frequency at which the short circuit current gain becomes 1. We assumed earlier that f_T could be found by summing all the delays in the device (see equation 7.5.1 and equation 7.5.2). We will now show why this is the case.

The value of f_T is obtained by applying nodal analysis to the bipolar equivalent circuit for the termination shown in figure 7.17. The input capacitance $C_{in} = C_\pi + C_{BE}$. The frequency dependent current gain of the device $\beta(j\omega)$ is given by

$$\beta(j\omega) = \frac{I_o(j\omega)}{I_{in}(j\omega)} \tag{7.5.72}$$

We define the input impedance z_{in} as

$$z_{in} = r_\pi \left|\left| \frac{1}{j\omega C_{in}} = \frac{r_\pi}{1 + j\omega r_\pi C_{in}} \right.\right. \tag{7.5.73}$$

We can then write I_{in} and I_o as

$$I_{in}(j\omega) = \left[\frac{v_{BE}}{z_{in}} + j\omega v_{BE} C_{BC}\right] \tag{7.5.74}$$

$$I_o(j\omega) = v_{BE}\left[g_m - j\omega C_{BC}\right] \tag{7.5.75}$$

Using $g_m = r_e^{-1}$

$$\frac{I_o(j\omega)}{I_{in}(j\omega)} = \frac{z_{in}}{r_e}\left[\frac{1 - j\omega C_{BC} r_e}{1 + j\omega C_{BC} z_{in}}\right] \tag{7.5.76}$$

This expression reduces to

$$\frac{I_o\left(j\omega\right)}{I_{in}\left(j\omega\right)}=\frac{r_\pi}{r_e}\left[\frac{1-j\omega C_{BC}r_e}{1+j\omega r_\pi\left(C_{in}+C_{BC}\right)}\right] \tag{7.5.77}$$

Neglecting the zero introduced by r_eC_{BC}, since r_e and C_{BC} are both small, $\beta\left(j\omega\right)$ can be written as

$$\beta\left(j\omega\right)=\frac{\beta_0}{1+j\omega r_\pi\left(C_{in}+C_{BC}\right)}=\frac{\beta_0}{1+j\omega\beta_0 r_e\left(C_{in}+C_{BC}\right)} \tag{7.5.78}$$

or

$$\beta\left(j\omega\right)=\frac{\beta_0}{1+j\beta_0\left(\frac{\omega}{\omega_T}\right)} \tag{7.5.79}$$

where

$$\omega_T=\frac{1}{r_e\left(C_{in}+C_{BC}\right)} \tag{7.5.80}$$

It is readily seen that

$$\begin{aligned}\frac{1}{\omega_T} &= r_e\left(C_{in}+C_{BC}\right)=r_e\left[\left(C_\pi+C_{BE}\right)+C_{BC}\right]\\ &= r_e\left(C_{BE}+C_{BC}\right)+\tau_B+\tau_C\\ &= \tau_{EC}\end{aligned} \tag{7.5.81}$$

where τ_{EC} is the total delay from the emitter to the collector.

Let us examine our expression for $\beta\left(j\omega\right)$ in equation 7.5.79. When $\omega<\omega_T/\beta_0$, the denominator in equation 7.5.79 is approximately equal to 1, and $|\beta(j\omega)|$ is given by the dc current gain β_0. Once $\omega>\omega_T/\beta_0$, we can ignore the 1 in the denominator, and $\beta(j\omega)$ is approximately given by

$$\beta\left(j\omega\right)\simeq\frac{\omega_T}{j\omega} \tag{7.5.82}$$

So we see that at $\omega=\omega_T$, $|\beta|=1$, or the current gain is unity. ω_T is the transit frequency and determines the current gain cutoff frequency as $\omega_T=2\pi f_\tau$, or

$$\boxed{f_\tau=\frac{1}{2\pi\tau_{EC}}} \tag{7.5.83}$$

Maximum frequency of oscillation f_{max}

f_τ is a very important figure of merit because it is determined by the intrinsic delay in the device and is therefore related intimately to material parameters such as carrier velocity, lifetimes, etc. However, when used as an amplifier, in many cases the device can amplify power beyond f_τ, because often voltage gain can be achieved at frequencies higher than f_τ. The maximum frequency of operation beyond which the power gain is less than 1 is termed f_{max}. Beyond this frequency, the device dissipates more power than it outputs.

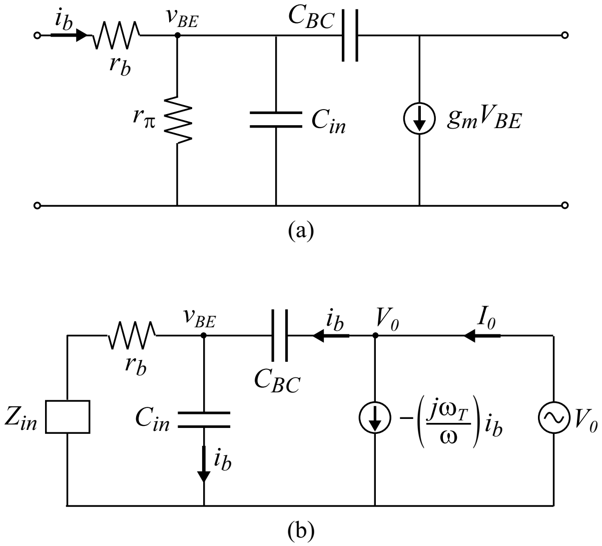


Figure 7.18: Equivalent circuit for determining (a) the input impedance and (b) the output impedance of a BJT biased in the common emitter configuration.

In well designed transistors, f_{max} is larger than f_T, though it is possible (but undesirable) to have $f_{max} < f_T$. Since an amplifier functions by delivering power to a load, the calculation of f_{max} is carried out under conditions of the load being conjugately matched to the output of the device. Let us calculate the input and output impedances of a BJT via its equivalent circuit in the common emitter configuration. For these calculations, we refer to the circuit diagrams in figure 7.18.

The input impedance (calculated by applying a test generator at the input and an open circuit at the output as in figure 7.18a) is seen to very rapidly approach r_b for

$$\omega > \frac{1}{r_\pi C_{in}}$$

The circuit used in the output impedance calculation is shown in figure 7.18b. Here, the input may be terminated with any impedance under the assumption that

$$\frac{1}{j\omega C_{in}} < r_b$$

so that the current flow through r_b is negligible. Applying a test voltage V_o, we can calculate the

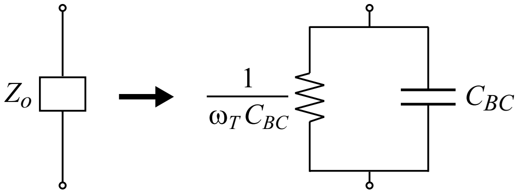


Figure 7.19: Equivalent circuit representation of the output impedance Z_o.

output impedance

$$Z_o = \frac{V_o}{I_o} \tag{7.5.84}$$

From nodal analysis, we get the following expression for I_o:

$$I_o = \frac{-j\omega_T}{\omega} \cdot i_B + i_B = \left[\frac{-j\omega_T}{\omega} + 1\right] i_B \tag{7.5.85}$$

We now assume $C_{BC} << C_{in}$, and since

$$i_B = j\omega C_{BC} \left(V_o - v_{BE}\right)$$

and

$$i_B = j\omega C_{in} v_{BE}$$

we can combine these expressions to show that

$$v_{BE} = \left(\frac{C_{BC}}{C_{in} + C_{BC}}\right) V_o$$

This shows that $v_{BE} << V_o$, and so $i_B \simeq j\omega C_{BC} V_o$.

We can now write I_o as a function of V_o.

$$\begin{aligned} I_o &= \left[\frac{-j\omega_T}{\omega} + 1\right] \cdot j\omega C_{BC} V_o \\ &= \left[\omega_T C_{BC} + j\omega C_{BC}\right] V_o \\ &= \left[G_s + jX\right] V_o \end{aligned} \tag{7.5.86}$$

This shows us that the output impedance can be expressed as a resistor $1/(\omega_T C_{BC})$ in parallel with a capacitor C_{BC}, as shown in figure 7.19. The dc output conductance

$$R_o^{-1} = \frac{I_C}{V_A}$$

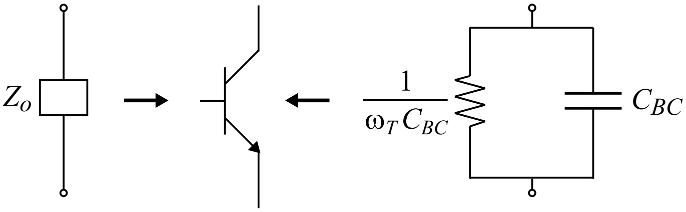


Figure 7.20: Equivalent circuit representation of the intrinsic device with the dominant extrinsic elements outside.

is also in parallel to the ac components in figure 7.19 but is typically very small compared to the ac conductance. The transistor can now be represented as an intrinsic device with the dominant extrinsic elements outside, as shown in figure 7.20.

Power gain is always calculated for the case of a conjugately matched load $Z_L = Z_o^*$ to enable maximum power transfer to the load. The conjugately matched load Z_L for the output impedance shown in figure 7.19 is illustrated in figure 7.21. Since the output current in the load is one-half of the short circuit current,

$$\frac{i_C}{i_B} = \frac{i_C(\text{short})}{2} \cdot \frac{1}{i_B} = \frac{\beta}{2} \tag{7.5.87}$$

The power gain can be written as

$$G = \frac{P_{load}}{P_{in}} = \frac{|i_C|^2 R_{load}}{|i_B|^2 R_{in}} = \left(\frac{i_C}{i_B}\right)^2 \cdot \frac{1}{\omega_T C_{BC}} \cdot \frac{1}{r_b} \tag{7.5.88}$$

or

$$G = \frac{|\beta|^2}{4} \cdot \frac{1}{\omega_T C_{BC}} \cdot \frac{1}{r_b} \tag{7.5.89}$$

Substituting $|\beta| = \omega_T/\omega$ (equation 7.5.82) and using $\omega = 2\pi f$, we get

$$G = \frac{f_\tau}{8\pi r_b C_{BC}} \cdot \frac{1}{f^2} \tag{7.5.90}$$

f_{max} is defined as the frequency at which $G \to 1$. This gives us

$$\boxed{f_{max} = \sqrt{\frac{f_\tau}{8\pi r_b C_{BC}}}} \tag{7.5.91}$$

where r_b is the total base resistance of the device including contact resistance, sheet resistance of the extrinsic base, and the intrinsic base resistance of the device.

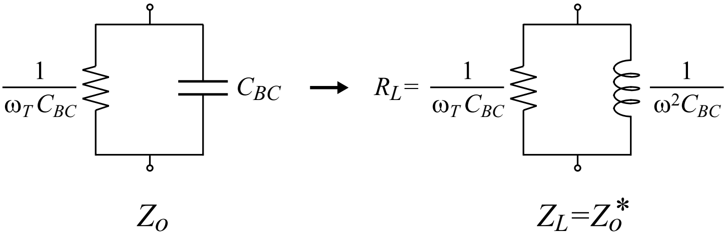

$\frac{1}{\omega_T C_{BC}}$
C_{BC}
$R_L = \frac{1}{\omega_T C_{BC}}$
$\frac{1}{\omega^2 C_{BC}}$
Z_o
$Z_L = Z_o^*$

Figure 7.21: Conjugately matched load Z_L (right) for the output impedance Z_o shown on the left.

7.6 BIPOLAR TRANSISTORS: A TECHNOLOGY ROADMAP

In this section we will discuss some of the important design considerations in the performance of bipolar devices. Bipolar devices must compete with the field effect transistor (FETs) and in many respects the two classes of families carry out similar functions. This puts a tremendous pressure on the BJT and HBT device designers to design the best devices in a given material system.

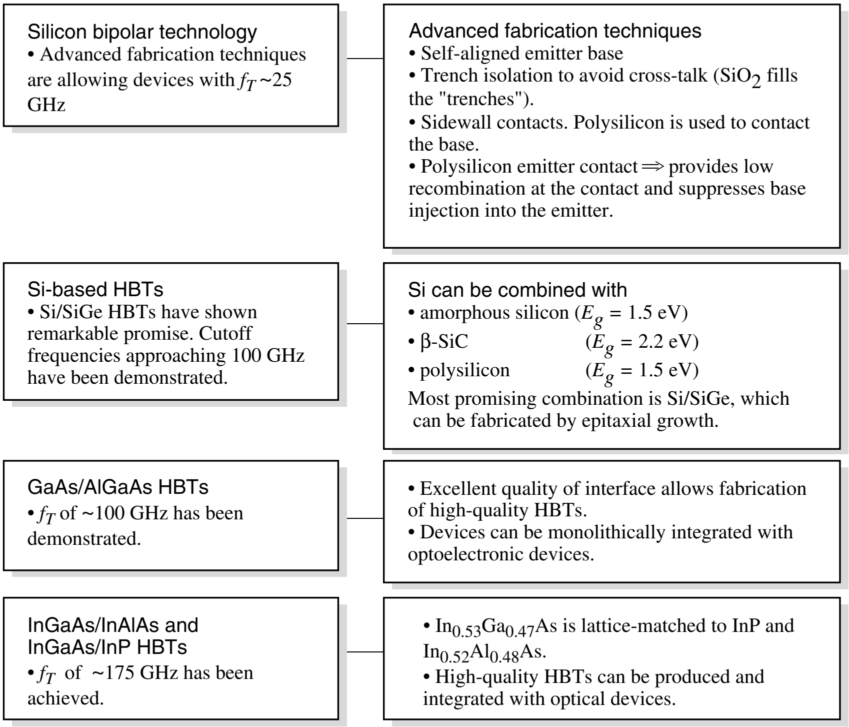
Bipolar devices are exploiting both fabrication techniques and new material systems to produce superior devices. A survey of the development of advanced devices was given in figure 7.22. We will now give a brief overview of these developments.

7.6.1 Si Bipolar Technology

In spite of the superior performance of HBTs, the Si bipolars continue to be the workhorse devices for both digital and some microwave applications. The advances in Si technology have come from two directions. The first direction relates to advanced fabrication technology and the second one relates to the use of polysilicon as a contact for the emitters.

The fabrication-technology-related advances in Si bipolars have resulted from: (i) self-aligned emitter and base contacts, which allow extremely precise placement of the base contact next to the emitter contact and thus reduce parasitic resistances; (ii) trench isolation, which allows very dense packing of the transistors without cross-talk. This involves etching narrow grooves around the transistor down to the substrate, lining them with SiO_2, and filling them with polysilicon. This greatly reduces the isolation capacitance; (iii) sidewall contact process, which dramatically reduces the extrinsic base collector capacitance. In this process polysilicon is used to contact the base and is isolated from the collector by a thick oxide. The device becomes essentially one-dimensional as a result and also becomes quite symmetric between the emitter and the collector.

The second source of improvements in Si bipolar devices is the use of polysilicon to contact the emitter. The advantages of polysilicon over metal contacts arise from the boundary



Silicon bipolar technology
- Advanced fabrication techniques are allowing devices with f_T ~25 GHz

Advanced fabrication techniques
- Self-aligned emitter base
- Trench isolation to avoid cross-talk (SiO_2 fills the "trenches").
- Sidewall contacts. Polysilicon is used to contact the base.
- Polysilicon emitter contact ⇒ provides low recombination at the contact and suppresses base injection into the emitter.

Si-based HBTs
- Si/SiGe HBTs have shown remarkable promise. Cutoff frequencies approaching 100 GHz have been demonstrated.

Si can be combined with
- amorphous silicon (E_g = 1.5 eV)
- β-SiC (E_g = 2.2 eV)
- polysilicon (E_g = 1.5 eV)

Most promising combination is Si/SiGe, which can be fabricated by epitaxial growth.

GaAs/AlGaAs HBTs
- f_T of ~100 GHz has been demonstrated.

- Excellent quality of interface allows fabrication of high-quality HBTs.
- Devices can be monolithically integrated with optoelectronic devices.

InGaAs/InAlAs and InGaAs/InP HBTs
- f_T of ~175 GHz has been achieved.

- $In_{0.53}Ga_{0.47}As$ is lattice-matched to InP and $In_{0.52}Al_{0.48}As$.
- High-quality HBTs can be produced and integrated with optical devices.

Figure 7.22: A survey of advanced bipolar devices.

conditions the contact places on the hole density injected into the emitter from the base. The boundary condition is very important for the thin emitters needed for high-frequency applications. The hole density goes to zero at a normal ohmic contact due to the very large recombination rate with the electrons. In the case of polysilicon the hole density goes to zero gradually so that the hole injection current is similar to that of a thick emitter. Due to this, the base injection into the emitter is strongly suppressed.

With advanced technology in use, Si BJTs have reached f_T values of $\sim$200 GHz.

7.6.2 Si-Based HBTs

Although Si BJTs are still workhorse devices for most applications, there is an increasing interest in Si HBTs for obvious reasons. Several wide-gap emitters have been proposed, although most still have technology-related problems. Among materials considered for emitters are: (i) amorphous Si, which has a large “effective bandgap” ($\sim$1.5 eV). The problems include poor-quality contacts to amorphous Si; (ii) β-SiC with bandgap of 2.2 eV. The material has a strong lattice mismatch with Si and it is not clear how reliable the technology will be; (iii) semi-insulating polycrystalline Si, which has a gap of 1.5 eV. High current gains have been reported for this system; (iv) use of III-V compounds like GaP. The main problem here is the cross-doping issue since Si dopes GaP while Ga and P dope Si.

A material system that appears to have a tremendous advantage and is still compatible with Si technology is the Si-SiGe system. The $\mathrm{Si}_{1-x}\mathrm{Ge}_x$ is an alloy with lattice constant that is mismatched from Si by $4x\%$. However, for very thin base regions n-Si/p-SiGe/n-Si HBTs can be fabricated with very high performance. The smaller gap of SiGe suppresses hole injection into the emitter. Devices operating up to 350 GHz have been reported in this material system.

7.6.3 GaAs/AlGaAs HBTs

In chapter 3 we discussed the bandstructure of GaAs and AlAs systems. The two semiconductors have excellent lattice matching ($\sim$0.14%) and high-quality GaAs/AlGaAs heterostructures can be grown. The bandgap of the alloy $\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{As}$ up to compositions of $x \sim 0.45$ is given by

$$E_g(x) = 1.42 + 1.247x$$

Above $x \sim 0.45$, the material becomes indirect and is usually not used for most device applications because of poor transport and optical properties.

GaAs material has a high bandgap and thus the intrinsic carrier concentration is quite low ($\sim 2.2 \times 10^6$ cm^{-3}) at room temperature. Thus the semi-insulating GaAs can have a very high resistivity ($\sim 5 \times 10^8$ Ω-cm), with the result that there is essentially negligible capacitance between the substrate and the interconnects or the collector. This is a serious problem for Si at high frequencies.

An important advantage of GaAs technology is that the electronic devices can be monolithically integrated with optoelectronic devices, leading to optoelectronic integrated circuits (OEICs), which are certainly not possible for Si technology (so far).

Another important advantage of GaAs technology is the ability to fabricate millimeter microwave integrated circuits (MMICs) in which the active and passive elements of the circuit are all made on the same chip. MMIC technology is quite advanced in GaAs while it is still primitive in Si.

In the GaAs/AlGaAs system, HBTs with f_T values around 200 GHz have been achieved, making this material system an important player in microwave technology.

7.6.4 InGaAs/InAlAs and InGaAs/InP HBTs

An important consideration in the development of any material technology is the substrate availability. One must have a high-quality substrate that is lattice-matched to the material and has very few defects. There are three main substrates that have reached a very high quality level: Si, GaAs, and InP. The material systems $\mathrm{In}_{0.53}\mathrm{Ga}_{0.47}\mathrm{As}$ ($E_g \sim 0.75$ eV) and $\mathrm{In}_{0.52}\mathrm{Al}_{0.48}\mathrm{As}$($E_g \cong$ 1.4 eV) are lattice-matched to InP. Thus the $\mathrm{In}_{0.53}\mathrm{Ga}_{0.47}\mathrm{As}/\mathrm{In}_{0.52}\mathrm{Al}_{0.48}\mathrm{As}$ and InGaAs/InP both

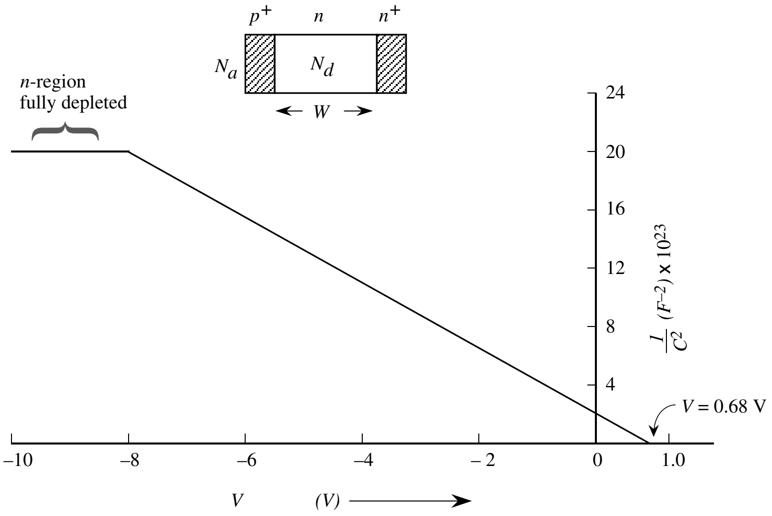


Figure 7.23: Plot of $\frac{1}{C^2}$ vs. V for problem 7.1.

can be exploited for high-performance HBTs. InGaAs has extremely attractive electronic properties and is therefore the material of choice for all high-speed/high-frequency applications. The InGaAs/InP HBTs have achieved f_T values of over 600 GHz.

7.7 PROBLEMS

- **Section 7.2**

Problem 7.1 The $\frac{1}{C^2}$ versus applied voltage relation in a silicon $p^+ - n - n^+$ junction diode is measured to have a form shown in figure 7.23. Calculate the thickness of the n-region, the built-in voltage, and the N_a and N_d concentrations in the p^+ and n regions. The diode area is 10^{-3} cm^2. Also calculate the width of the n-region.

Problem 7.2 Consider a long base p^+n diode that is biased to carry a forward current of 1 mA. The junction capacitance is 100 pF. If the minority carrier lifetime τ_p is 1μs, what is the admittance of the diode at 300 K for a 1 MHz signal?

Problem 7.3 A p^+-n silicon diode has an area of 10^{-2} cm^2. The measured junction capacitance is given by (at 300 K)

$$\frac{1}{C^2} = 5 \times 10^8 (2.5 - 4\text{ V})$$

where C is in units of μF and V is in volts. Calculate the built-in voltage and the depletion width at zero bias. What are the dopant concentrations of the diode?

Problem 7.4 In a long base n^+p diode, the slope of the C_{diff} versus I_F plot is $1.6 \times 10^{-5} F/A$. Calculate the electron lifetime, the stored charge, and the value of the diffusion capacitance at I_F = 1 mA.

Problem 7.5 Consider a Si p^+n diode with a long base. The diode is forward-biased (at 300 K) at a current of 2 mA. The hole lifetime in the n-region is 10^{-7} s. Assume that the depletion capacitance is negligible and calculate the diode impedance at the frequency of 100 KHz, 100 MHz, and 500 MHz.

Problem 7.6 Consider a diode with the junction capacitance of 16 pF at zero applied bias and 4 pF at full reverse bias. The minority carrier time is 2×10^{-8} s. If the diode is switched from a state of forward-bias with current of 2.0 mA to a reverse-bias voltage of 10 V applied through a 5kΩ resistance, estimate the response time of the transient.

Problem 7.7 Consider a Si p-n diode at room temperature with following parameters:

$$\begin{aligned} N_d &= N_a = 10^{17} \text{ cm}^{-3} \\ D_n &= 20 \text{ cm}^2/\text{s} \\ D_p &= 12 \text{ cm}^2/\text{s} \\ \tau_n &= \tau_p = 10^{-7} \text{ s} \end{aligned}$$

Calculate the reverse saturation current for a long ideal diode. Also estimate the storage delay time for the long diode. Now consider a narrow diode made from the structure given above. The thickness of the n-side region is 1.0 μm. The thickness of the p-side region is also 1.0 μm. Calculate the reverse saturation current in the narrow diode at a reverse bias of 2.0 volt. Also estimate the storage delay time for this diode.

Problem 7.8 Consider the p-n junction diode shown in figure 7.24. Assume $N_A = N_D = 10^{17} cm^{-3}$. Assume that the width of the n-region $L_1 << L_P$ and that the

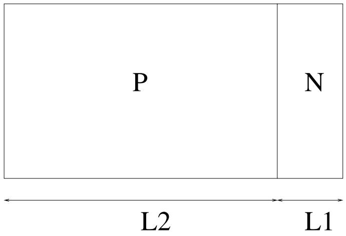


Figure 7.24: Figure for problem 7.8.

width of the p-region $L_2 >> L_n$. Calculate the depletion and diffusion capacitances of the diode. Obtain an expression for the ac resistance.

- **Section 7.3**

Problem 7.9 In the Schottky barrier, the electrons are injected across the barrier with energies equal to the barrier height. These electrons are very hot. Estimate the "temperature" of these electrons in a typical Si Schottky barrier with a barrier height of ϕ_b = 0.6 V. (Electron temperature, T_e, is defined by $\frac{3}{2}k_BT_e \sim \langle E_e \rangle$ where $\langle E_e \rangle$ is the average electron energy.)

Problem 7.10 An important consideration in the speed of Schottky barrier diodes is the time it takes hot electrons (see the previous problem) to lose their energy and achieve equilibrium thermal energy. In GaAs, electrons lose excess energy exponentially with a time constant of 1 ps. Consider a W-n-type GaAs Schottky diode with ϕ_b = 0.8 V. How far will electrons move in the GaAs before they lose 99% of their energy?

- **Section 7.4**

Problem 7.11 Consider an HBT with a base graded from InGaAs to GaAs so that the bandgap is narrow at the emitter and wide at the collector.

1. Draw the band diagram in the neutral base region of the device.
2. Write down the drift-diffusion equation governing the current in the base region assuming no recombination in the base. Assume a forward bias at the base-emitter junction and a reverse bias across the base-collector junction. What are the boundary conditions for this equation?
3. Solve the differential equation to get the minority charge profile ($n(x)$ versus x) as a function of injected current in the base.
4. Sketch (without actually calculating exact values) the minority charge profile with and without a reverse grade in the base for the same injected current density. Give physical arguments for your result.
5. How will the base transit time vary in these two cases? Why?

- **Section 7.5**

Problem 7.12 In a particular BJT, the base transit time forms 20% of the total delay time of the charge transport. The base width is 0.5 μm and the diffusion constant is $D_b = 25$ cm^{2}s. Calculate the cutoff frequency for the device.

Problem 7.13 A silicon npn bipolar transistor has a cutoff frequency at 300 K limited by base transit time. The cutoff frequency is 1 GHz. Estimate the base width if the base doping is 10^{16} cm^{-3}. The minority carrier mobility in the base is 500 cm^2/V·s.

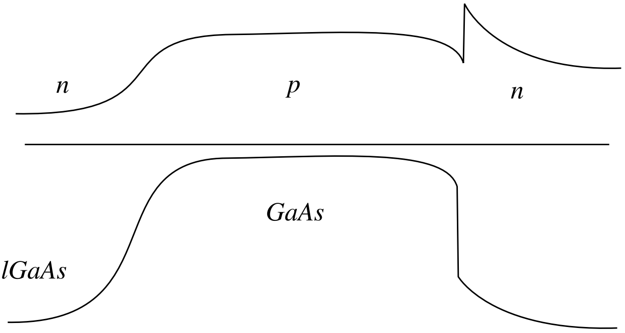


Figure 7.25: Band diagram for device in problem 7.14.

Problem 7.14 Consider a bipolar transistor where the wide bandgap collector is used, such that $\Delta E_c = 0.3eV$, as illustrated in figure 7.25. Calculate the additional delay introduced by the barrier for a current density of $10kA/cm^2$. Assume thermionic emission over the collector barrier. You may also assume that the notch is a quantum well of width 100 Å with infinite barriers when calculating the Fermi level in the notch.

Problem 7.15 Tired of making planar HBT's, I decide to make a cylindrical HBT as shown in figure 7.26. (a) Derive an expression for the transit time delay in the collector of this HBT.
(b) Calculate delays for $R_B = 1\mu$m and $R_C = 3\mu$m, and compare these delays with values for planar HBT's with the same base and collector thickness. Explain the difference. Assume that the electron velocity is saturated in the collector.
(c) Calculate the minority charge distribution in the base of the cylindrical HBT and compare it with the planar structure, assuming I_e is the same in both cases. Assume no recombination in the base. How is the delay affected relative to the planar HBT with the same base width?

Problem 7.16 Consider the HBT from prefxch07/6.36.
(a) Obtain an expression for the base transit time in this graded base. Compare it to an HBT with an ungraded base, but with the same collector current.
(b) What is the base transit time when the current density is 10 kA$\cdot$cm^{-2}. What will the base transit time at this current level be if the base is not graded? Assume μ = 1000 $\mathrm{cm}^2/(\mathrm{V}\cdot\mathrm{s})$, $v_{sat} = 10^7\mathrm{cm/s}$. You may assume that the electron velocity is saturated for electric fields greater than 2 kV/cm.

Problem 7.17 Consider two HBT structures, whose collector velocity profiles are shown in figure 7.27. Derive expressions for the collector transit delays in these two structures in terms of the saturated velocity v$_s$ and collector width, W$_C$. Now, calculate the base transit

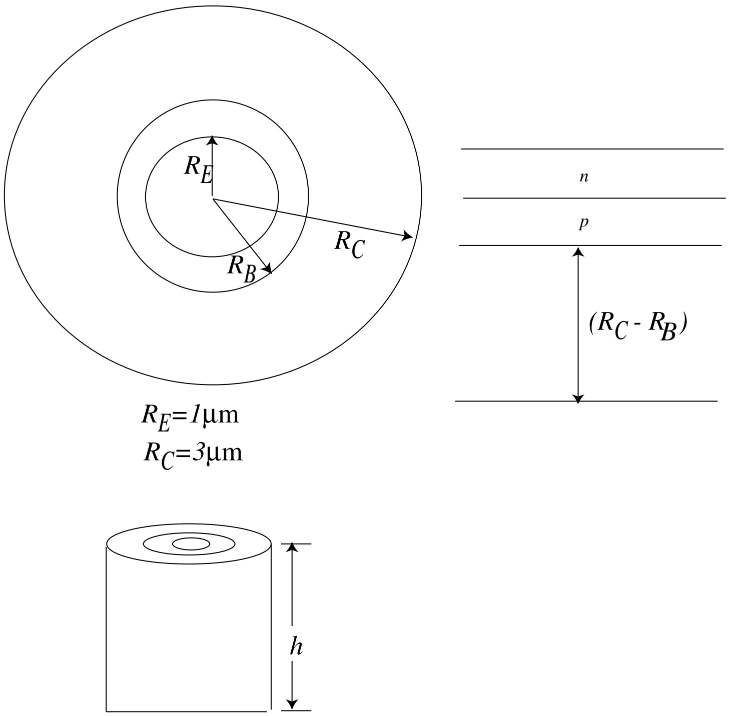


Figure 7.26: Figure for problem 7.15.

delay for each of these structures in terms of the base width, W$_B$, diffusion constant D$_n$, and carrier velocity v$_s$. Do not assume Shockley boundary conditions. Use charge control analysis to derive these delays.

Problem 7.18 To maintain a high breakdown voltage in the collector of an InP-based HBT, it is preferable to use an InP collector. In problem 7.14, you calculated the additional base delay introduced by a charge accumulation layer at the interface for a current density of $10kA/cm^2$.

1. To eliminate the above delay, I linearly grade the bandgap from InGaAs to InP across the base. Derive an expression for the new base delay. Based on physical arguments, sketch the expected minority charge profile. Compare this with a device whose base is not graded. Physically explain your result. Do *not* assume Shockley boundary conditions to calculate the delay. Instead, assume a saturated velocity v_s.

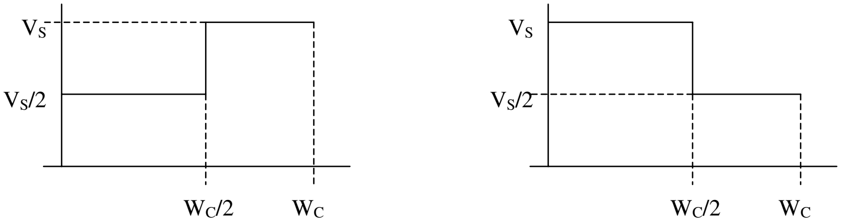


Figure 7.27: Figure for problem 7.17

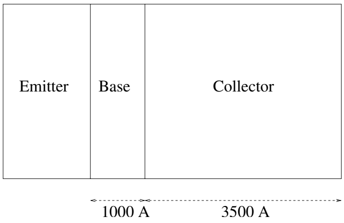


Figure 7.28: Figure for problem 7.18.

2. Next, I grade the InGaAs to InP in the collector, rather than in the base. Assume the doping in the collector to be $N_D = 10^{16} cm^{-3}$. What is the length of the parabolic grade necessary to eliminate the barrier at $V_{BC} = 0$ Why do you want the grading distance to be minimum?

3. How can you shorten the grading distance in the collector by half? What is the penalty you pay?

4. Calculate the collector delay time for the abrupt and graded cases.
   For the graded case, assume that the velocity profile is as shown in figure 7.29. For the abrupt case, assume a saturation velocity v_{s1}. In the figure, t_0 is the length of the grade calculated in part (b), $v_{s2} = 4.10^7 cm/s$, and $v_{s1} = 10^7 cm/s$. Use a collector width of 3500Å.

5. How do you expect the Kirk current threshold in the graded case to compare with the ungraded one? Assume the same velocity profiles as above.

Problem 7.19 The base-collector junction in a bipolar transistor has the structure shown in figure 7.30. A p-type doping sheet is added at the i-n junction to create a 0.3 eV drop across the intrinsic region. Calculate the density of this doping sheet. This is a ballistic

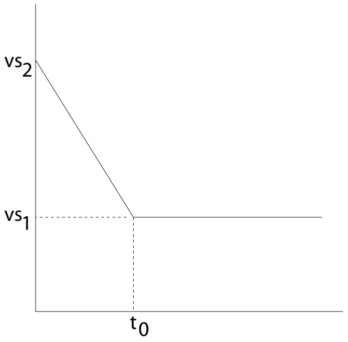


Figure 7.29: Figure for problem 7.18.

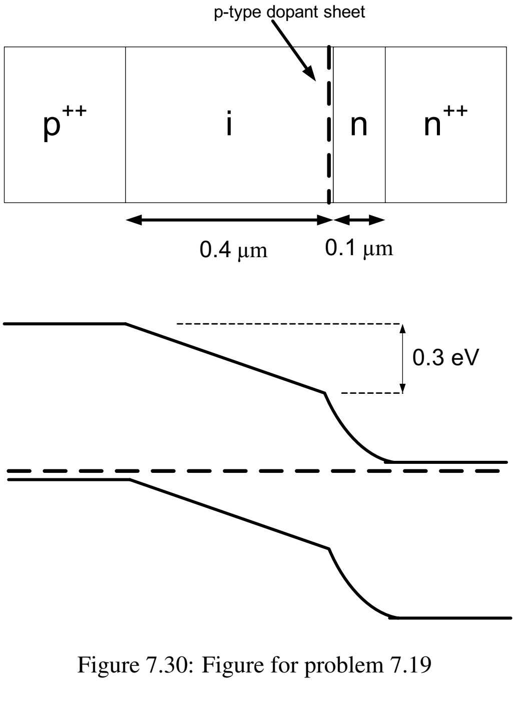


Figure 7.30: Figure for problem 7.19

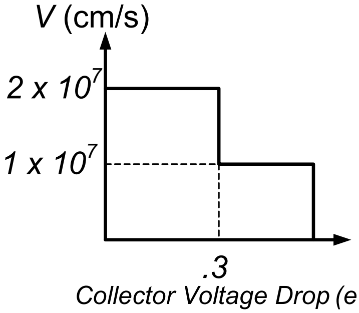


Figure 7.31: Figure for problem 7.19.

collector transistor. The velocity field profile for this material is given below. This is expressed in terms of the velocity versus voltage drop in collector. The reason for the sudden drop is the $\Gamma - L$ intervalley transfer of electrons in GaAs. Of course, this is an idealized profile to make the problem tractable.

1. Calculate the transit delay for this structure at $V_{CB} = 0V$.
2. I now apply a reverse bias of $V_{CB} = 1V$ on the collector-base junction. Calculate the transit delay at this bias. Assume that the depletion thickness is small in the n++ and p++ regions, and that the n-region is just fully depleted at $V_{CB} = 0V$.

7.8 DESIGN PROBLEMS

Problem 7.1 Design a p^+n Si diode that can be used in a digital system operating at 1 gigabit per second. Assume that the minority carrier lifetime is 10^7 s. Other parameters can be obtained from the text. Plot the I-V characteristics of this diode. At what applied reverse bias would the entire n-region be depleted in this diode?

Problem 7.2 A Si p^+n diode is to be used in the reverse bias state $(V_R = 5V)$ as a high-speed detector. Design the diode so it can operate up to a frequency of 5 GHz. Make reasonable assumptions for the material parameters.

Chapter 8

FIELD EFFECT TRANSISTORS

8.1 INTRODUCTION

In this and the next chapter we will examine the field effect transistor (FET) and Metal-Oxide-Semiconductor FETs (MOSFETs). These simple devices are majority carrier devices which are relatively simple to fabricate and are extremely versatile. FETs are now made from a wide variety of materials (Si, SiGe, GaAs, InGaAs, GaN, SiC, etc.). Figure 8.1 shows a GaAs-based Metal Semiconductor FET or MESFET.

The basic concept behind the FET is quite simple and is illustrated in figure 8.2. The device consists of an active channel through which electrons (or holes) flow from the source to the drain. The source and drain contacts are ohmic contacts. The conductivity of the channel is modulated by a potential applied to the gate. This results in the modulation of the charge density flowing in the channel. It is important to isolate the gate from the channel so that no current flows into the gate. The gate isolation is done in a variety of ways, leading to a number of different devices. In the MOSFET, the gate is isolated from the channel by an oxide. This is the basis of the silicon devices. In the metal-semiconductor FET or MESFET, the gate forms a Schottky barrier with the semiconductor and the gate current is small in the useful range of gate voltages. In the junction FET or JFET, a p-n junction is used in reverse bias to isolate the gate. Heterojunction field effect transistors (HFETs) or modulation doped FETs (MODFETs) use a large bandgap semiconductor to isolate the gate from the active channel. In this chapter we will examine the MESFET or JFET devices. In the next chapter we examine MOSFETs.

8.2 JFET AND MESFET: CHARGE CONTROL

The operation of JFETs and MESFETs are similar. The key difference is that in a JFET a $p^+ - n$ structure is used to create a barrier between the metal gate and the semiconductor while in a MESFET the Schottky barrier height is used. We have seen from chapter 5that the reverse current in a Schottky junction is much larger than that in a p-n junction. As a result the JFET is especially used in materials for which it is difficult to produce a large Schottky barrier.

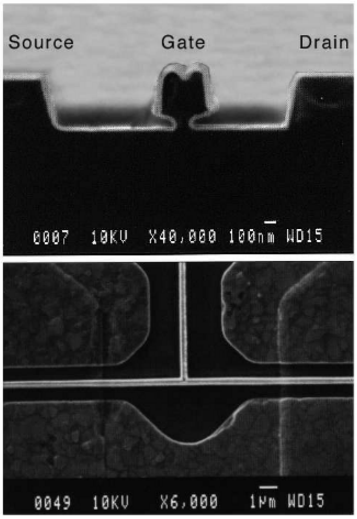


Figure 8.1: MESFETs and JFETs are important devices for high-speed, low-noise amplifiers, D/A and A/D converters, and much “front-end” processing where high speed is critical. These devices exploit materials, like GaAs, InP, and InGaAs, that have transport properties that are superior to Si. (Top) A cutoff cross-section of a 0.1 μm MESFET. (Bottom) Top view of the MESFET.

Let us examine a typical JFET or MESFET structure as shown in figure 8.3a. The device is based on a low-conductivity substrate on which an n-type region is grown to form an “active” conducting channel of thickness h. The gate is formed by a p^+ region (n^+ region for a $p-$type FET) or a Schottky barrier. The source and drain are ohmic contacts. In figure 8.3b we show a case where the active channel is partially depleted (say at zero gate bias). A negative bias on the gate reverse biases the gate-ohmic conductor junction (for an n-type device) and alters the width of the depletion region. This allows the gate to modulate the conductance of the device. In figure 8.3c we show a case where the channel is completely depleted.

Consider the cases shown in figure 8.4. In figure 8.4a we show the device with a small source-drain bias V_{DS} and no gate bias. As the gate bias is increased and the gate semiconductor junction is reverse biased, the current through the channel decreases until eventually the channel is “pinched off” and there are no free carriers in it. If the gate bias is fixed and the drain bias is increased, as shown in figure 8.5, the gate semiconductor junction near the drain becomes more reverse biased. Eventually, the channel is pinched off near the drain side. At this point the current cannot increase even if the drain voltage is increased. This is called the saturation region. Once the device reaches saturation, the current in the channel remains more or less unchanged.

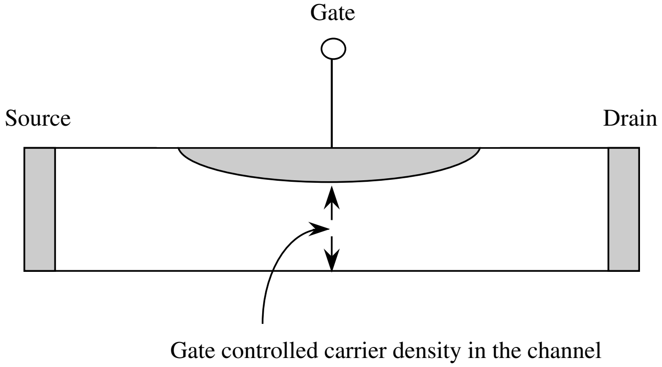


Figure 8.2: The physical principle behind the FET involves the use of a gate to alter the charge in a channel. Depending upon the method used for isolation of gate different FETs arise.

The thickness of the doped active channel is h as shown in figure 8.3 and V_{bi} is the built-in voltage. The gate bias required to pinch off the channel is simply given by the depletion approximation

$$h = \left\{ \frac{2\epsilon(V_{bi} - V_{GS})}{eN_d} \right\}^{1/2} \tag{8.2.1}$$

It is possible that the built-in voltage may by itself pinch the channel off. In an n-channel device, if the device is not pinched off by V_{bi}, then a negative gate bias will cause pinch-off. In a p-channel device, a positive bias is needed for pinch-off.

The pinch-off voltageV_p (called the intrinsic pinch-off voltage) is defined by

$$\boxed{V_p = \frac{eN_d h^2}{2\epsilon}} \tag{8.2.2}$$

and the gate bias needed for pinch-off for the n-channel device is

$$V_T = V_{bi} - V_p \tag{8.2.3}$$

where V_T is called the threshold voltage for the device. If the voltage V_p is smaller than the built-in potential V_{bi}, the device channel is completely depleted in the absence of a gate bias. A positive gate bias (for n-channel devices) can allow the channel to have free charge and be conducting. Such devices are said to be enhancement-mode devices. On the other hand, if V_p is larger than V_{bi}, the device has free charge in the channel at V_G = 0 since the channel is only partially depleted. A negative gate bias will then turn the device off, i.e., deplete the channel. Such devices are said to operate in the depletion mode. Electronic circuits may use enhancement- or depletion-mode devices or even combinations of them, depending upon the application.

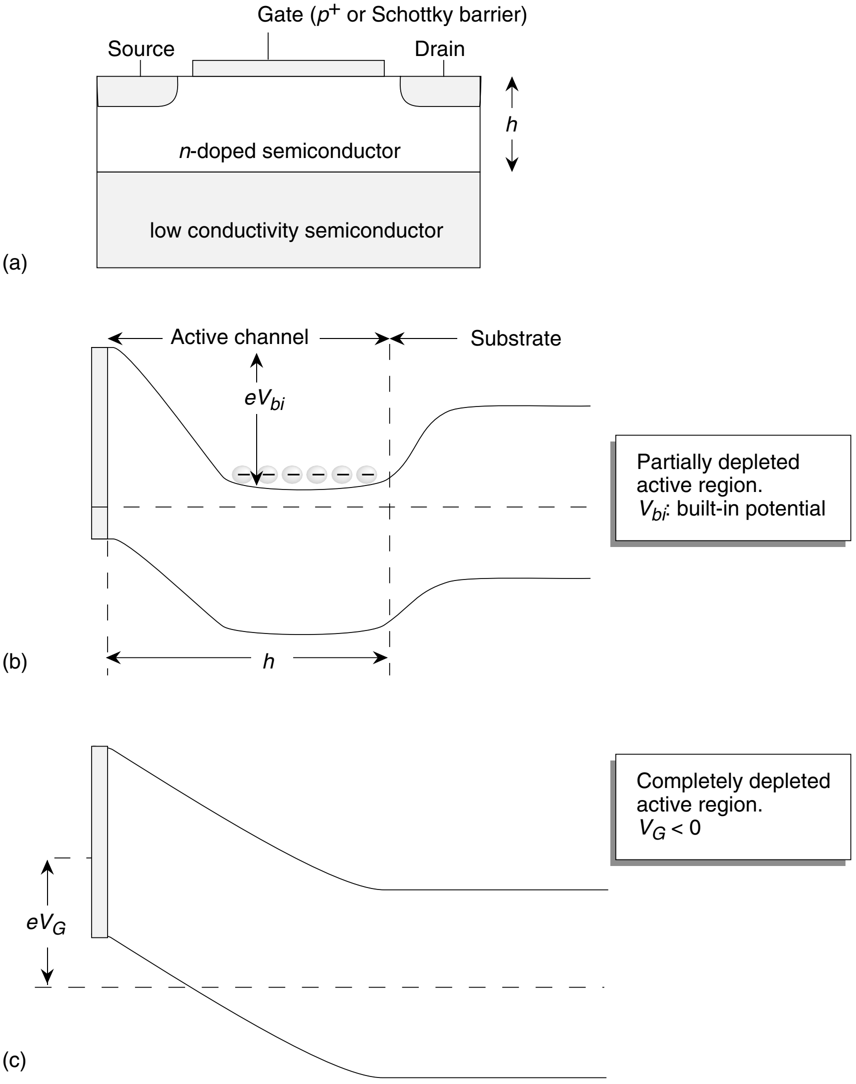

Gate (p^+ or Schottky barrier)
Source
Drain
h
n-doped semiconductor
low conductivity semiconductor
(a)
Active channel
Substrate
eV_{bi}
Partially depleted active region. V_{bi}: built-in potential
h
(b)
Completely depleted active region. $V_G < 0$
eV_G
(c)

Figure 8.3: (a) A schematic of a JFET or MESFET showing the source, drain, and gate. The channel width of this device is h; (b) the band profile when the applied gate bias is zero as is the source-drain bias; (c) the band profile with a negative gate bias so that the channel is depleted.

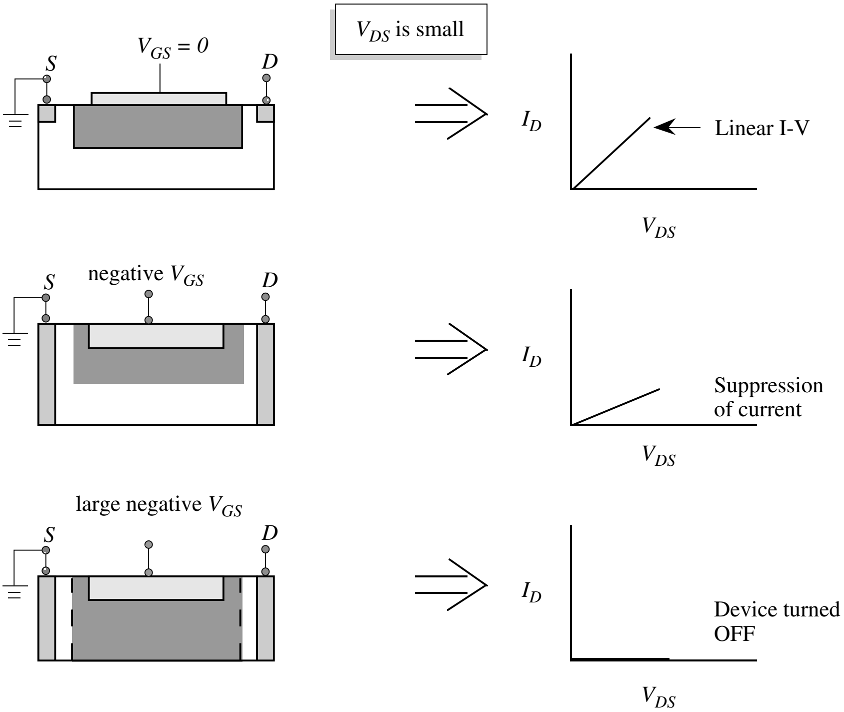


Figure 8.4: (a) Depletion width and channel in a JFET or MESFET under zero gate bias. The channel has a large opening. Such a device is called a depletion-mode device; (b) the device with a negative gate bias showing reduction in the channel opening and current; (c) the gate bias is large and negative and the channel is pinched off with current in the channel zero.

An important consideration in JFET or MESFET technologies is that the gate current be negligible. the requires that the gate be biased appropriately. This also requires a large built-in voltage or Schottky barrier height. For small gap semiconductors (e.g. InGaAs, InSb, etc.) this may not be possible.

Example 8.1 Consider an n-MESFET made from GaAs doped at 10^{17} cm^{-3}. Calculate the gate current density under normal operation if:

(i) the gate is made from a Schottky metal with a barrier $\phi_b = 0.8$ V;

(ii) the gate is made from a heavily doped p^+ GaAs.

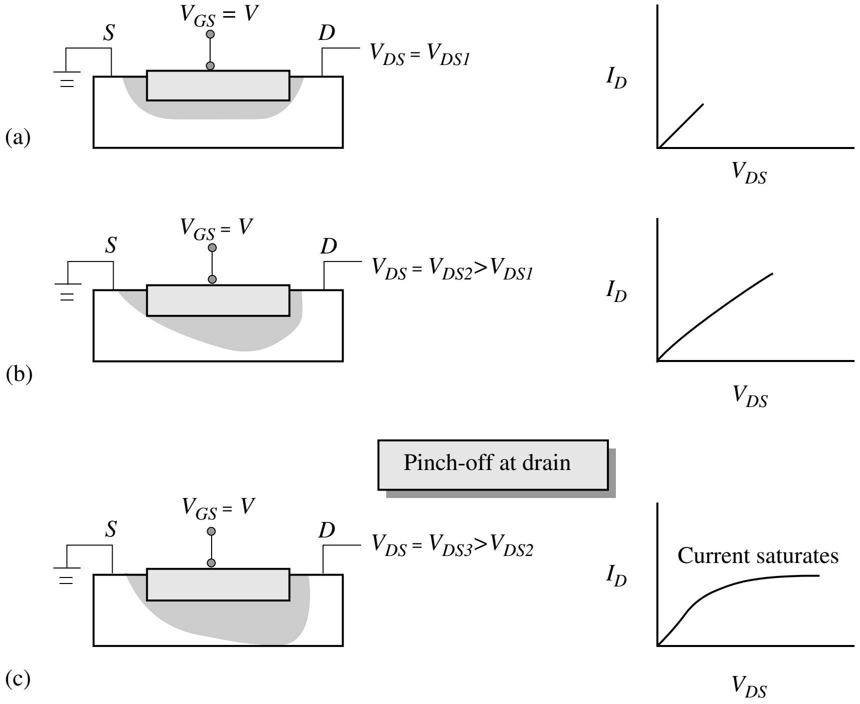


Figure 8.5: The effect of increased drain bias at a fixed gate bias. (a) the drain bias is small; (b) the drain bias is increased and the channel is constricted near the drain; (c) the drain bias is increased to the point that the channel is pinched off at the drain side. The drain current saturates as shown.

You may use the following parameters:

$$\begin{aligned} D_p &= 20 \text{ cm}^2/\text{s} \\ L_p &= 1.0\,\mu\text{m} \\ A^* &= 8 \text{ Acm}^{-2}K^{-2} \end{aligned}$$

The gate current under normal operation is just the reverse-bias current of the junction between the gate and the semiconductor. For the Schottky case we have (see chapter 5)

$$J_s = A^*T^2 \exp\left(-\frac{e\phi_b}{k_BT}\right)$$

This gives

$$J_s = 8 \times (300)^2 \times 4.34 \times 10^{-14} = 3.125 \times 10^{-8} \text{ A/cm}^2$$

For the p^+-gate we have from p-n diode theory (see chapter 4)

$$J_0 = \frac{eD_p p_n}{L_p}$$

This gives

$$J_0 = \frac{1.6 \times 10^{-19} \times 20 \times 3.38 \times 10^{-5}}{10^{-4}} = 1.08 \times 10^{-18} \text{ A/cm}^2$$

We see that the gate current is much smaller for the JFET case. However for the GaAs case considered here, the MESFET gate current is small enough for most applications.

8.3 CURRENT-VOLTAGE CHARACTERISTICS

The MESFET is one of the simplest three terminal devices to fabricate and to conceptually understand. The most common material used in MESFETs is GaAs. Other compound semiconductors can also be used although is common to use a HFET approach for most materials. It is important that one have a high resistivity substrate to avoid current flow through the substrate. This is usually done by impurity doping. These impurities create levels at midgap, pinning the Fermi level..

We will first present a very simple model for the current-voltage relation in the MESFET. Then we will describe a more accurate model. However, to obtain realistic results one needs to use computer simulation tools.

In figure 8.6 we show the device structure along with the band profile under the gate. In figure 8.7 we show the MESFET cross-section along with the depletion width under the gate region. In the absence of any bias, there is a uniform depletion region under the gate region, as shown in figure 8.7a. If the gate bias is made more negative, the depletion width spreads further into the active region until eventually the channel is completely depleted. Thus, as the gate bias is increased (to negative values), the total charge available for conduction decreases until the channel is pinched off.

If the gate bias is fixed and the drain voltage is increased toward positive values, current starts to flow in the channel. The depletion region now becomes larger near the drain side, as shown in figure 8.7b. As the drain voltage is increased, the depletion width toward the drain end starts to increase, since the potential difference between the gate and the drain end of the channel increases. The channel then starts to pinch off at the drain end. As this happens, the current starts to saturate. Once the drain voltage reaches a value $V_{DS}(sat)$ such that the channel pinches off at the drain end, the current remains essentially constant even as the drain voltage is increased.

8.3.1 The Ohmic Regime

As noted earlier we will start with a very simple model which, though not accurate for modern short gate devices, provides insight into the device operation. We will use the following

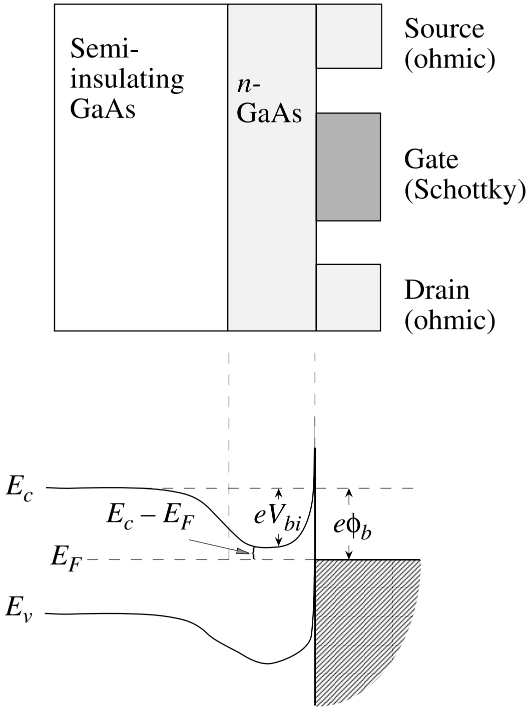


Figure 8.6: A schematic of a GaAs MESFET. Also shown are the energy band profile under the gate region and some important device parameters.

approximations:

- The mobility of the electrons is constant and independent of the electric field. Thus the velocity increases linearly with field.

- The gradual channel approximation introduced by Shockley is assumed. In the absence of any source-drain bias, the depletion width is simply given by the one- dimensional model we developed for the p-n diode. However, strictly speaking, when there is a source-drain bias, one has to solve a two-dimensional problem to find the depletion width and, subsequently, the current flow. In the gradual channel approximation, we assume that field in the direction from the gate to the substrate is much stronger than from the source to the drain, i.e., the potential varies “slowly” along the channel as compared to the potential variation in the direction from the gate to the substrate. Thus the depletion width, at a

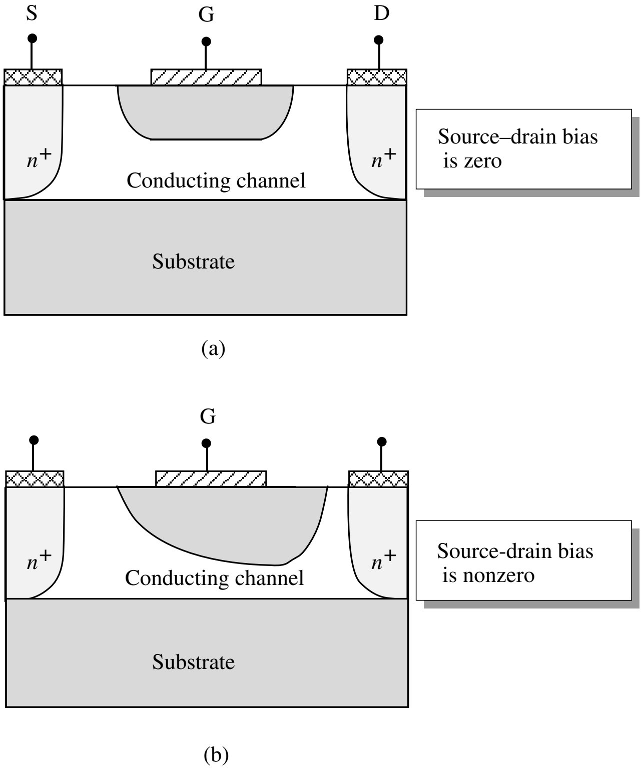


Figure 8.7: A schematic of a MESFET showing the depletion width under the gate. (a) In the absence of a source-drain bias, the depletion width is uniform and is controlled by the gate bias. (b) In the presence of a source-drain bias, the depletion width is greater in the drain side.

point x along the channel, is given by the potential at that point using the simple one-dimensional results.

Both the approximations given above are reasonable only if the channel fields are small. These approximations do not work for modern devices and we will discuss a better model later.

The current in the drain is given by (field = $-dV/dx$)

$$\begin{aligned} I_D &= \text{channel area} \times \text{(charge density)} \times \text{(mobility)} \times \text{(field)} \\ &= Z[h - W(x)]eN_d\,\mu_n\,\frac{dV}{dx} \end{aligned} \tag{8.3.1}$$

where $W(x)$ is the depletion width as shown in figure 8.8 and h is the channel thickness. Thus $h - W(x)$ is the channel opening. The depletion width at a point x is given in terms of the gate voltage V_{GS}, the built-in voltage V_{bi}, and the channel voltage $V(x)$ by the depletion equation

$$W(x) = \left[\frac{2\epsilon\,[V(x) + V_{bi} - V_{GS}]}{eN_d}\right]^{1/2} \tag{8.3.2}$$

To find I_D as a function of V_{DS} and V_{GS}, we substitute for $W(x)$ in equation 8.3.1 and integrate (I_D is constant throughout the channel) to get

$$I_D \int_0^L dx = e\mu_n N_d Z \int_0^{V_{DS}} \left[h - \left\{\frac{2\epsilon\,[V(x) + V_{bi} - V_{GS}]}{eN_d}\right\}^{1/2}\right] dV \tag{8.3.3}$$

which gives (after dividing by L)

$$I_D = \frac{e\mu_n N_d Z h}{L}\left\{V_{DS} - \frac{2\left[(V_{DS} + V_{bi} - V_{GS})^{3/2} - (V_{bi} - V_{GS})^{3/2}\right]}{3(eN_d h^2/2\epsilon)^{1/2}}\right\} \tag{8.3.4}$$

We denote by g_o the channel conductance when the channel is completely open,

$$\boxed{g_o = \frac{e\mu_n N_d Z h}{L}} \tag{8.3.5}$$

We have defined the pinch-off voltage V_p as

$$V_p = \frac{eN_d h^2}{2\epsilon} \tag{8.3.6}$$

In terms of V_p, the drain current versus drain voltage characteristics can be written as

$$I_D = g_o\left\{V_{DS} - \frac{2\left[(V_{DS} + V_{bi} - V_{GS})^{3/2} - (V_{bi} - V_{GS})^{3/2}\right]}{3V_p^{1/2}}\right\} \tag{8.3.7}$$

It must be remembered that this equation was derived under the condition that the gate and drain voltages are such that there is no pinch-off near the drain region, i.e.,

$$W(L) = \left\{2\epsilon\frac{V_{DS} + V_{bi} - V_{GS}}{eN_d}\right\}^{1/2} < h \tag{8.3.8}$$

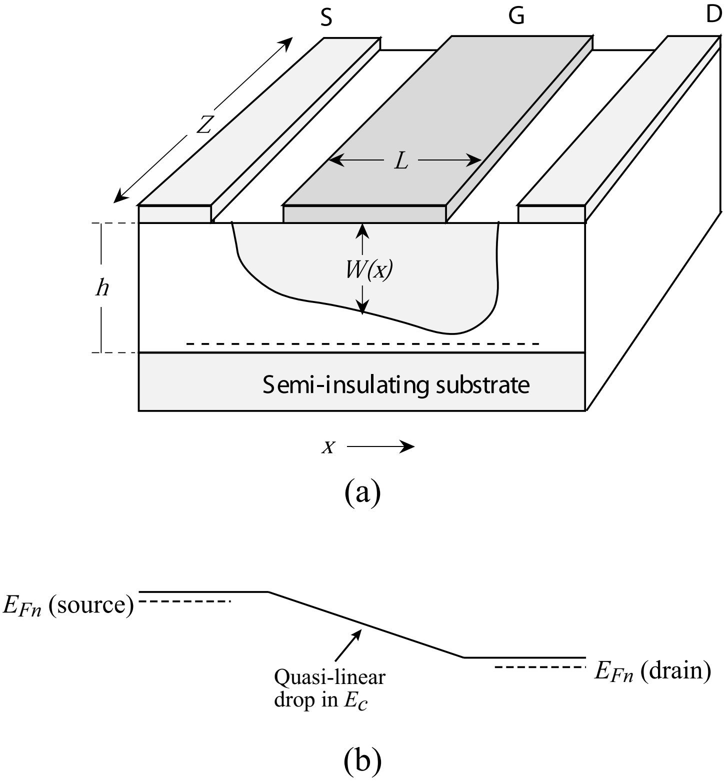


Figure 8.8: (a) A schematic of the MESFET with $V_D < V_{DS}(sat)$. The current flow occurs only in the undepleted region. The channel potential at any point x in the channel is $V(x)$. (b) Band diagram along the channel (dotted line in (a)).

We assume that to the first approximation, when pinch-off occurs, the drain current saturates. What happens once saturation occurs will be discussed in the following section. The drain voltage at which saturation occurs is

$$V_{DS}(sat) = V_p - V_{bi} + V_{GS} \tag{8.3.9}$$

and the saturated channel current becomes, from equation 8.3.7 (the drain current does not change in our simple model once $V_{DS} \geq V_{DS}(sat)$),

$$\boxed{I_D(sat) = g_o \left[\frac{V_p}{3} - V_{bi} + V_{GS} + \frac{2(V_{bi} - V_{GS})^{3/2}}{3V_p^{1/2}} \right]} \tag{8.3.10}$$

This expression will be reexamined with a better approximation later.

An important parameter of the device is the transconductance g_m, which defines the control of the gate on the drain current. From equation 8.3.7, the transconductance becomes

$$g_m = \left. \frac{dI_D}{dV_{GS}} \right|_{V_{DS}=constant} = g_o \frac{(V_{DS} + V_{bi} - V_{GS})^{1/2} - (V_{bi} - V_{GS})^{1/2}}{V_p^{1/2}} \tag{8.3.11}$$

From equation 8.3.5 and equation 8.3.11 we can see that the transconductance is improved by using a higher-mobility material as well as a shorter channel length. An improved transconductance means the gate has a greater control over the channel. This results in higher gain and high-frequency capabilities, as will be discussed later.

When the source-drain voltage is small, the expression for the current can be simplified by using

$$V_{DS} \ll V_{bi} - V_{GS} \tag{8.3.12}$$

Using the Taylor series, we then get from equation 8.3.7,

$$I_D = g_o \left[1 - \left(\frac{V_{bi} - V_{GS}}{V_p} \right)^{1/2} \right] V_{DS} \tag{8.3.13}$$

The device is ohmic in this regime, as shown in figure 8.9, with a transconductance

$$g_m = \frac{g_o V_{DS}}{2V_p^{1/2}(V_{bi} - V_{GS})^{1/2}} \tag{8.3.14}$$

In the saturation regime the transconductance is, from equation 8.3.10,

$$\boxed{g_m(sat) = g_o \left[1 - \left(\frac{V_{bi} - V_{GS}}{V_p} \right)^{1/2} \right]} \tag{8.3.15}$$

In the model discussed here (known as the Shockley model) the current cannot be calculated beyond pinch-off. At pinch-off, the channel width becomes zero so that the electron velocity must, in principle, go to infinity to maintain constant current. This, of course, does not happen. We will now discuss, using physical arguments, what happens in the saturation region.

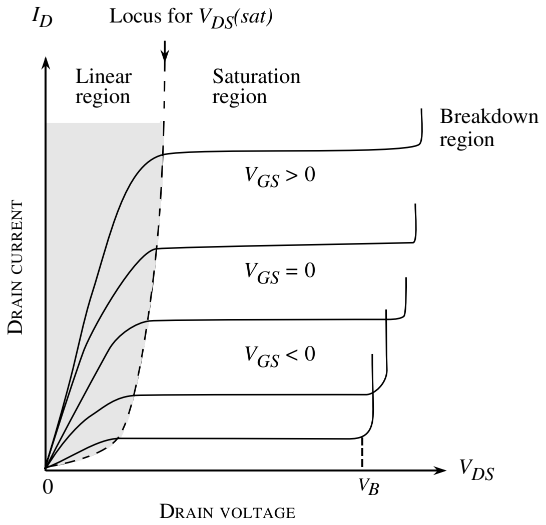


Figure 8.9: Typical I-V characteristics of an n-MESFET. In the Shockley model discussed in the text, it is assumed that once pinch-off of the channel occurs, the current saturates. In the figure, V_B is the breakdown voltage.

8.3.2 A Nearly Universal Model for FET Behavior : The Saturation Regime

The Shockley model is only valid for drain voltages smaller than $V_{DS}(sat)$. Consider again what happens when the gate voltage is held fixed and the drain voltage is increased toward positive values. As the drain voltage approaches $V_{DS}(sat)$, the drain end of the channel becomes very narrow, so the electric field in the direction of current flow must become large in this region in order for current continuity to be maintained. This clearly violates the assumption of a gradual channel that was used in the Shockley analysis. The current characteristics beyond pinch-off can be explained as follows.

Consider the two generic materials Si and GaAs. In materials like silicon the velocity-field relations are such that the velocity increases monotonically with the applied field and eventually saturates. In GaAs, the velocity peaks at a field $\mathcal{E}_p$ (~ 3 kV/cm) and then decreases and gradually saturates. Therefore, it is reasonable to assume that in a FET, once the drain voltage is very close to $V_{DS}(sat)$, the velocity of the electrons at the drain side of the channel saturates, as the channel on the drain side narrows approaching pinch-off; we denote the channel width on the drain side at pinch-off by the symbol δ.

$$\delta = \frac{I_D(sat)}{eN_d v_{sat} Z} \tag{8.3.16}$$

where v_{sat} is the electron saturation velocity in the material.

Figure 8.10a shows a schematic diagram of the FET depletion profile when $V_{DS} > V_{DS}(sat)$. Beyond pinch-off, the channel does not become any narrower near the drain, since this would imply a reduction in current and would have to be accompanied by a decrease in the electric field near the source. Rather, any additional drain voltage is supported by a lateral extension of the depletion region near the drain. The universality of the analysis alluded to in the title of this section comes from the similarity of the electrostatics that exists in all FETs to first order in the saturation regime.Once the device behavior is understood in the saturation region in say the JFET that is detailed below, the analysis can be readily extended to MOSFETs, HEMTs etc by merely changing materials and geometrical parameters but keeping the device physics the same. A very illustrative analysis of the JFET in saturation has been presented by Grebene and Ghandhi. The detail of their analysis leads to the physical understanding of the universality of FET electrostatics and I-V behavior and hence deserves consideration. It explains the basis of all FET design, namely the high aspect ration design, in an elegant, analytical manner. Following their analysis, it is useful to divide the channel into two separate regions in the direction of current flow, as shown in figure 8.10. In Region I, near the source, the electric field in the direction of current flow is small, so the gradual channel approximation is valid. In Region II, near the drain, the electric field in the direction of current flow is large, so carriers travel at their saturation velocity. Prior to pinch-off, Region I covers the entire channel, and the current characteristics are described by the Shockley model. $V_{DS} = V_{DS}(sat)$ represents the onset at which Region II appears. Beyond pinch-off, Region II continues to become longer. However, the field profile in Region I remains approximately constant, which implies that the current remains nearly constant even as V_{DS} is increased.

For the saturation region (Region II), a fundamentally different relationship between voltage and distance occurs. Here, the charge and electric field (x-component) distributions are shown schematically in figure 8.11. The voltage distribution $V(x, y)$ in the saturation region is determined by solving Poisson's equation.

$$\nabla^2 V(x,y) = -\frac{\rho(x,y)}{\epsilon} = -\frac{eN(y)}{\epsilon} \tag{8.3.17}$$

The solution to this partial differential equation can be divided into a homogeneous solution and a particular solution such that $V(x, y) = V_{hom}(x, y) + V_{par}(x, y)$, where

$$\begin{aligned} \nabla^2 V_{hom}(x,y) &= 0 \\ \nabla^2 V_{par}(x,y) &= -\frac{eN(y)}{\epsilon} \end{aligned} \tag{8.3.18}$$

The solution to the homogeneous part is the solution of a Laplacian is the part of the solution that is independent of the doping in the channel and is therefore independent of the particular form of the gating mechanism which defines the various categories of FETs, namely a junction gate for a JFET, an MOS capacitor for a MOSFET and a modulation doped doped structure for a HEMT. This provides the near-universality to the electrostatics of different FETs. The solution of the homogeneous part (the Laplacian) is assumed to be of the form

$$V_{hom}(x,y) = \sum_{n=1}^{\infty} A_n \sin(\alpha_n y) \sinh(\beta_n x) \tag{8.3.19}$$

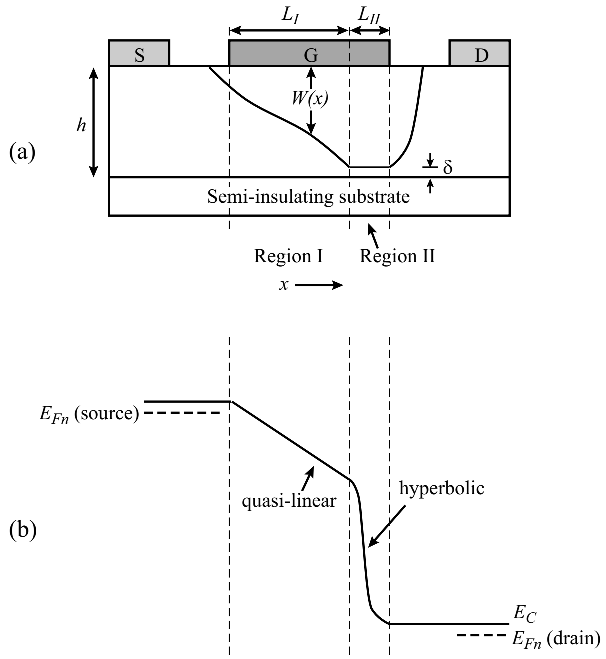


Figure 8.10: Schematic diagram of FET when $V_{DS} > V_{DS}(sat)$, drain end of channel is pinched off. (b) Band diagram across the channel of the device.

where the sine function represents the symmetric boundary conditions in the y-direction. This voltage distribution is caused by the positive charges on the drain electrode. The boundary conditions that have to be satisfied are

$$\begin{aligned} V_{hom}(x,0) &= 0 \\ V_{hom}(0,y) &= 0 \\ \frac{\partial V_{hom}(x,h)}{\partial y} &= 0 \\ \frac{\partial V_{hom}(0,h)}{\partial x} &= -\mathcal{E}_c \end{aligned} \tag{8.3.20}$$

The first two conditions are satisfied by the chosen functional form of V_{hom}. The third condition requires

$$\alpha_n = \beta_n = \frac{(2n-1)\,\pi}{2h} \tag{8.3.21}$$

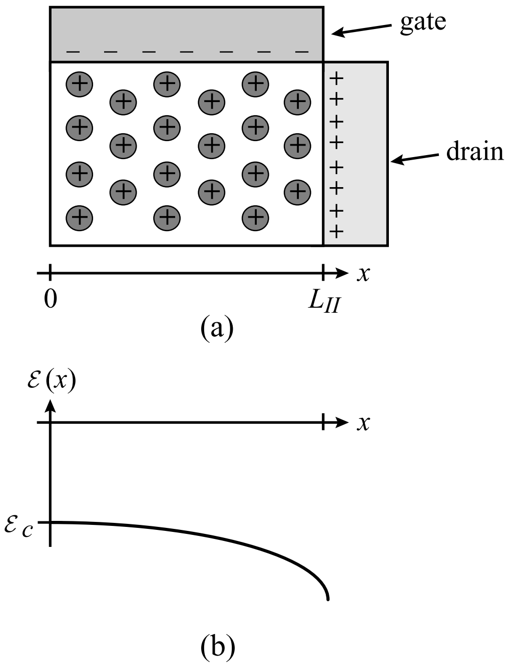


Figure 8.11: Schematic diagram of (a) the charge distribution and (b) the x-component of the electric field in the pinched-off region. The exact form of the electric field distribution is found by solving Laplace's equation.

The fourth condition, which reflects the assumption that pinch-off and velocity saturation occur simultaneously, leads to

$$\sum_{n=1}^{\infty} A_n\,(2n-1) = \frac{2h\mathcal{E}_c}{\pi} \tag{8.3.22}$$

For a physically meaningful solution, A_n must rapidly tend to zero for increasing values of n. Otherwise, the sinh function will lead to extremely high fields near the drain which are larger than the breakdown field of the semiconductor. We therefore retain only the first term in the series, which leads to

$$\boxed{V_{hom}(x,y) \cong \frac{2h\mathcal{E}_c}{\pi}\sin\left(\frac{\pi y}{2h}\right)\sinh\left(\frac{\pi x}{2h}\right)} \tag{8.3.23}$$

The particular solution, which is dependent on the gating structure and hence the type of FET under consideration, can be readily shown to be the gradual channel solution of the depletion region potential derived previously,

$$\boxed{V_{par}(x,y) = -V_G + \frac{e}{\epsilon}\left[\overline{\overline{N(y)}} - \overline{\overline{N(0)}} - y\overline{N(a)}\right]} \tag{8.3.24}$$

where the following convention has been used.

$$\overline{N(y)} \equiv \int N(y)dy \tag{8.3.25}$$

The above equation readily reduces to the equations in the previous section when N(y) is assumed to be constant, N_D. Hence the total voltage in Region II is given by

$$\boxed{V(x,y) = -V_G + \frac{e}{\epsilon}\left[\overline{\overline{N(y)}} - \overline{\overline{N(0)}} - y\overline{N(a)}\right] + \frac{2h\mathcal{E}_c}{\pi}\sin\left(\frac{\pi y}{2h}\right)\sinh\left(\frac{\pi x}{2h}\right)} \tag{8.3.26}$$

Along the line $y = h$, equation 8.3.26 reduces to

$$V(x,h) = V_{par} - V_G + \frac{2h\mathcal{E}_c}{\pi}\sinh\left(\frac{\pi x}{2h}\right) \tag{8.3.27}$$

The sine function in equation 8.3.26 reflects the inherent symmetry of the structure with a period $2h$ in the y-direction, whereas the hyperbolic function is the standard solution of Laplace's Equation in the non-symmetric direction. This hyperbolic dependence of V on x is critical to the operation of FETs in the saturation regime. The band diagram (and hence the voltage profile) of a FET biased in the saturation regime is shown in figure 8.10b.

As can be seen, the solution of Laplace's Equation leads to the electrostatic formation of a "collector" region, using terminology borrowed from bipolar transistors. The difference between this collector formed due to saturation/pinch-off is that voltage has an exponential dependence on length with applied voltage to conventional depletion regions that follow simple power laws of depletion depth with voltage. Voltages applied beyond $V_{DS}(sat)$, shown in figure 8.12 and labeled V_{dp}, are therefore absorbed efficiently within small extensions of Region II. The slope of the I-V curve in the saturation region is the output conductance g_d and is first explained quantitatively and then analytically.

Qualitative description of the output conductance

For a qualitative description of the output conductance, let us consider what is happening on the source side of the device. If the source electric field is $\mathcal{E}_s$, then the current in the device is

$$I_D = Ae\mu_n n_s \mathcal{E}_s$$

where $e\mu_n n_s = \sigma_s$ is the conductivity of the source and $A = W \cdot h$ is the cross-sectional area of the device. Under most conditions σ_s is not a function of drain bias V_D. Hence any increase in I_D can only result from an increase in $\mathcal{E}_s$. Therefore, a large increase in I_D with respect to

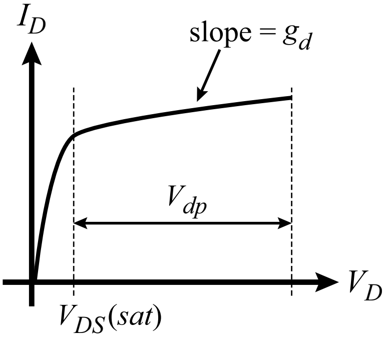


Figure 8.12: I_D vs. V_D for constant V_G.

V_D (or a large output conductance g_d) implies a large increase in $\mathcal{E}_s$, whereas insensitivity of I_D with respect to V_D (or a small g_d) implies a small increase in $\mathcal{E}_s$.

The first is the case before saturation where

$$\mathcal{E}_s \simeq \frac{V_D}{L}$$

$\mathcal{E}_s$ continues to increase until $V_D = V_{DS}(sat)$ and at a corresponding current

$$I_D(sat) = \sigma_s \frac{V_{DS}(sat)}{L}$$

In general

$$I_D(sat) = \sigma_s \frac{V_{DS}(sat)}{L_I}$$

where L_I is the length of Region I. Prior to saturation, $L = L_I$.

Once $V_D > V_{DS}(sat)$, the total channel voltage is split between Region I and Region II. The voltage drop across Region I remains close to $V_{DS}(sat)$, while the remaining voltage $V_{dp} = V_D - V_{DS}(sat)$ is dropped across Region II. In Region II, the hyperbolic relation of V to distance allows for large changes in V to be absorbed with only a small change in L_{II}. Hence the gradual channel length $L_I = L - L_{II}$ changes very slowly with drain bias, leading to a very small increase in $I_{DS}(sat)$ for $V_D > V_{DS}(sat)$, or a small output conductance in the saturation regime. This is critical to good device operation.

This analysis can also give a clear understanding of the square law behavior of $I_{DS}(sat) = K\left(V_G - V_T\right)^2$. The channel conductivity at the source σ_s can be written as

$$\sigma_s = e\mu_n n_s = e\mu_n C_G \left(V_G - V_T\right) \qquad (8.3.28)$$

where C_G is assumed to be constant (which is true for a MOSFET) and is the normalized capacitance of the gate. Therefore

$$\begin{aligned} I_{DS}(sat) &= \sigma_s \mathcal{E}_s Z = e\mu_n C_G \left(V_G - V_T\right) \cdot \frac{V_{DS}(sat)}{L} \cdot Z \\ &= e\mu_n C_G \left(V_G - V_T\right) \cdot \frac{\left(V_G - V_T\right)}{L} \cdot Z \\ &= \frac{e\mu_n C_G Z}{L} \left(V_G - V_T\right)^2 \end{aligned} \tag{8.3.29}$$

where Z is the device width.

This analysis assumed that the electric field along the length of the channel was uniform. A more rigorous analysis which allows for resistance and hence field variation leads to a factor of 2 in the expression, giving

$$I_{DS}(sat) = \frac{e\mu_n C_G Z}{2L} \left(V_G - V_T\right)^2 \tag{8.3.30}$$

Analytical derivation of the output conductance; high aspect ratio design

The output conductance g_d is given by

$$g_d = \left.\frac{\partial I_D}{\partial V_D}\right|_{V_G} \tag{8.3.31}$$

For a particular value of V_G, I_D is of the form

$$I_D = \frac{K(V_G)}{L_I} \tag{8.3.32}$$

Therefore

$$\frac{\partial I_D}{\partial V_D} = -\frac{K(V_G)}{L_I^2} \cdot \frac{\partial L_I}{\partial V_D} = -\frac{I_D}{L_I} \cdot \frac{\partial L_I}{\partial V_D} \tag{8.3.33}$$

Realizing that $L_I + L_{II} = L$ gives $\Delta L_I = -\Delta L_{II}$, equation 8.3.33 can be written as

$$\frac{\partial I_D}{\partial V_D} = \frac{I_D}{L_I} \left(\frac{\partial L_{II}}{\partial V_D}\right) \tag{8.3.34}$$

Also assuming $L_I >> L_{II}$, a reasonable assumption for long gate length devices ($L \geq 0.5\mu m$), L_I may be replaced by L, giving

$$g_d \simeq \frac{I_D}{L} \left(\frac{\partial L_{II}}{\partial V_D}\right) \tag{8.3.35}$$

To evaluate $(\partial L_{II}/\partial V_D)$ let us consider equation 8.3.26 again. We can write

$$V(x,h) = V_{par} - V_G + \frac{2h\mathcal{E}_c}{\pi} \sinh\left(\frac{\pi x}{2h}\right) \tag{8.3.36}$$

and

$$V(L_{II}, h) = V_D \tag{8.3.37}$$

The voltage drop in the pinched region (Region II) is $V_D - (V_G - V_T)$ or $V_D - V_{DS}(sat) = V_{dp}$. Therefore

$$L_{II} = \frac{2h}{\pi} \sinh^{-1}\left[\frac{V_{dp}\pi}{2\mathcal{E}_c h}\right] \tag{8.3.38}$$

and

$$\frac{\partial L_{II}}{\partial V_D} = \frac{1}{\mathcal{E}_c}\left[1 + \left(\frac{V_{dp}\pi}{2\mathcal{E}_c h}\right)^2\right]^{-1/2} \tag{8.3.39}$$

For the case of interest, where the device is biased well into saturation, the second term in the bracket becomes large, giving

$$r_d = \frac{1}{g_d} = \frac{\pi V_{dp}}{2I_D}\left(\frac{L}{h}\right) \tag{8.3.40}$$

This is a very important result. It says that to maintain a high value of r_d, which is necessary for high voltage gain, it is essential to maintain a high aspect ratio of the gate length to channel thickness (L/h). Typically, (L/h) should be at least 10. Of course, as the current in the channel decreases, r_d increases, but this benefit is largely negated by a similar decrease in the device g_m. An increase in r_d with V_{dp} is based on the increase in the saturated region which further isolated the drain potential from the source, reducing g_d.

8.4 HFETs: INTRODUCTION

In the previous sections we have discussed the MESFET (or JFET) devices. In the MESFET the gate is insulated from the channel by a barrier created by either a Schottky barrier (or a p^+n junction). The charge in the channel is provided by dopants in the channel. The dopants, while providing charge, also cause scattering and reduce mobility. The question arises: Can we have channel charge but avoid dopant scattering? This is possible in the Si MOSFET where charge can be induced by inversion. However, the MOSFET charge has to contend with interface roughness scattering. In the Si/SiO$_2$ case the interface is between a high quality semiconductor and a non-epitaxial layer and the interface scattering can greatly reduce mobility. Thus while room temperature electron mobility in pure Si is $\sim$ 1300 cm^2/V.s, it is only half this value in the NMOS channel. It would be ideal to have a heterostructure grown epitaxially where band inversion could occur. However, so far this has been difficult, though advances continue to be made. It is possible to make heterostructure devices where mobile charge is donated by dopants or other fixed charges.

The most widely used heterostructure FET utilizes the modulation doping concept. The device is called modulation doped field effect transistor (MODFET) or high electron mobility transistor (HEMT) or 2-dimensional gate field effect transistor (TEGFET), etc. It has also been shown that polar charges created at interfaces by piezoelectric and/or Spontaneous polarization can also be exploited to create free charge. This approach has become quite dominant in nitride based devices. Heterostructure field effect transistors (HFETs) offer many advantages over MESFETs

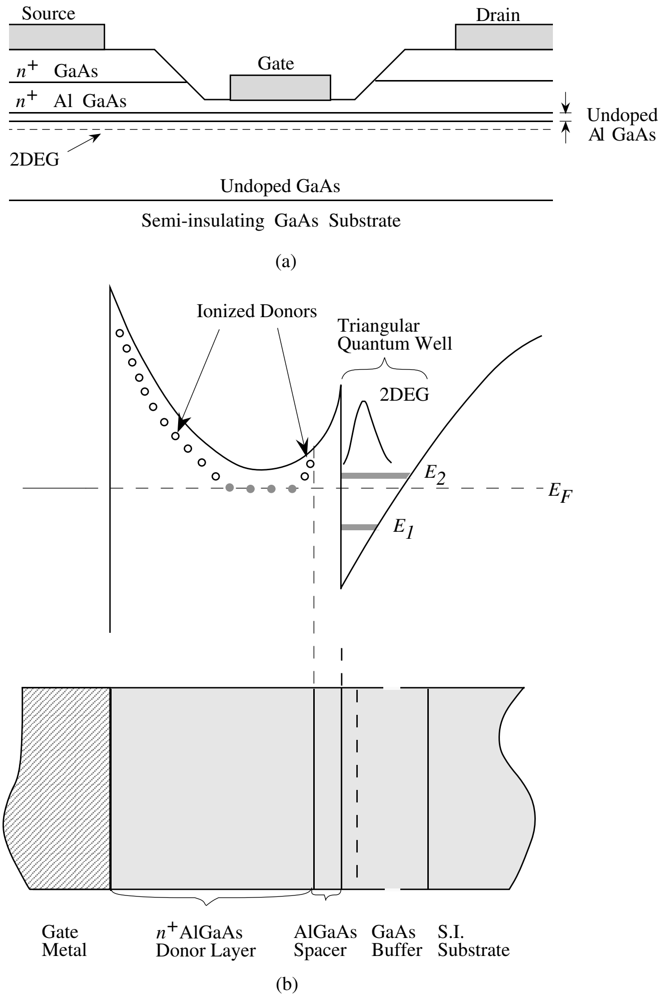


Figure 8.13: (a) A schematic of a GaAs/AlGaAs MODFET. In the figure shown, the gate is “recessed” to have a better control over the 2-dimensional electron gas (2DEG). (b) The band profile of an n-MODFET showing bandbending leading to a triangular quantum well at the GaAs/AlGaAs interface.

or MOSFETs. A typical modulation doped device structure is shown in figure 8.13a. We show a structure fabricated by epitaxial techniques such as MBE or MOCVD and using the recessed gate technology. For the AlGaAs/GaAs structure the substrate is semi-insulating GaAs on which an undoped GaAs layer is grown. A heterostructure is formed by depositing AlGaAs which is left undoped to provide a "spacer" region. The remaining barrier material is doped strongly. Finally, a heavily doped GaAs cap layer is deposited on which the ohmic source contacts are deposited. The cap layer is etched off and the Schottky gate is deposited on the high barrier material.

The electrons from the donor atoms in the high barrier material spill over into the low bandgap material conduction band creating a dipole layer. As a result, the band bends as shown in figure 8.13b to produce a quantum well in which the electrons are trapped. The quantum well has a triangular form and the electrons have 2-dimensional properties; i.e., they are free to move in the plane of the device but are confined in the device growth direction. As a result the density of states of the electrons have the usual 2-dimensional features. The term 2-dimensional electron gas (2DEG) is used to describe the electron system .

The key motivations for HFETs are:

- High Mobility Due to Suppression of Ionized Impurity Scattering: We have earlier discussed the effect of ionized impurity scattering on mobility. In the HFET, due to the physical separation of the dopants from the free electrons, the mobility is greatly improved. For example, in a GaAs MESFET channel, doped at 5 × 10^{17} cm^{-3}, the room temperature mobility is $\sim$ 4000 cm$^2V^{-1}$s^{-1}. In a MODFET channel with equivalent charge density the mobility is essentially limited by phonon scattering to $\sim$ 8000 cm$^2V^{-1}s^{-1}$. The effects are even more dramatic at low temperatures.

  The improved mobility allows the device to have a very low resistance between the source and the gate region (low access resistance or source resistance). The high field transport in the MODFET channel is, however, not too much better than the MESFET channel since at high fields, transport is governed primarily by phonon (lattice vibration) scattering.

- Superior Low Temperature Performance: We had noted in Chapter 3 the carrier freezeout effect that occurs in doped semiconductors at low temperatures. In a MODFET channel, this effect is avoided since the electrons are in a region of energy below the donor levels in the high bandgap material. Thus a high carrier density can be maintained at very low temperature exploiting the low temperature improvement in transport. Extremely low noise, high gain microwave devices are exploiting this low temperature feature for special applications such as deep space signal reception.

- Use of Superior Materials in the Channel: In the MODFET, the active channel in which the transport takes place need only be $\sim$ 200 Å. Thus one can use a very high mobility material system in the channel. Normally materials like InAs or InSb which have very high mobilities cannot be used as MESFETs since it is difficult to process these narrow bandgap materials which are very "soft" and defect prone. However, when only a narrow region is used, the device can be quite robust. On GaAs substrates one can use In$_x$Ga$_{1-x}$As channels for active regions while on InP one can use In$_{0.53+x}$Ga$_{0.47-x}$As as active channel materials. In the two cases above, if $x \neq 0$, the channel is under strain, resulting in pseudomorphic MODFETs.

- High Sheet Charge Density: The charge density in the 2-dimensional HFET channel depends upon the doping density in the large bandgap material or on polar charge at the interface and the conduction band discontinuity at the channel-barrier interface. By using materials with large conduction band discontinuities, a very high sheet charge density ($\gtrsim 10^{13}$ cm^{-2}) can be introduced. This results in a very large device transconductance and device performance.

In figure 8.14, we show an SEM image of a state-of-the-art InP-based HFET along with its layer structure, I-V characteristics, and an integrated circuit composed of these devices. A careful examination of the gate in the SEM image shows that the gate is recessed; the advantages of this are described later in this chapter. The T-gate structure, which is characteristic of all modern high speed FETs, is desirable because it is possible to achieve a very small intrinsic gate length (in this case 0.1 μm) while still maintaining a bulkier gate metal, which reduces the lateral gate resistance. Also evident is the dielectric passivation layer which covers the device. This prevents undesirable charging of surface states as well as protects the device from contaminants that may be present in the ambient environment.

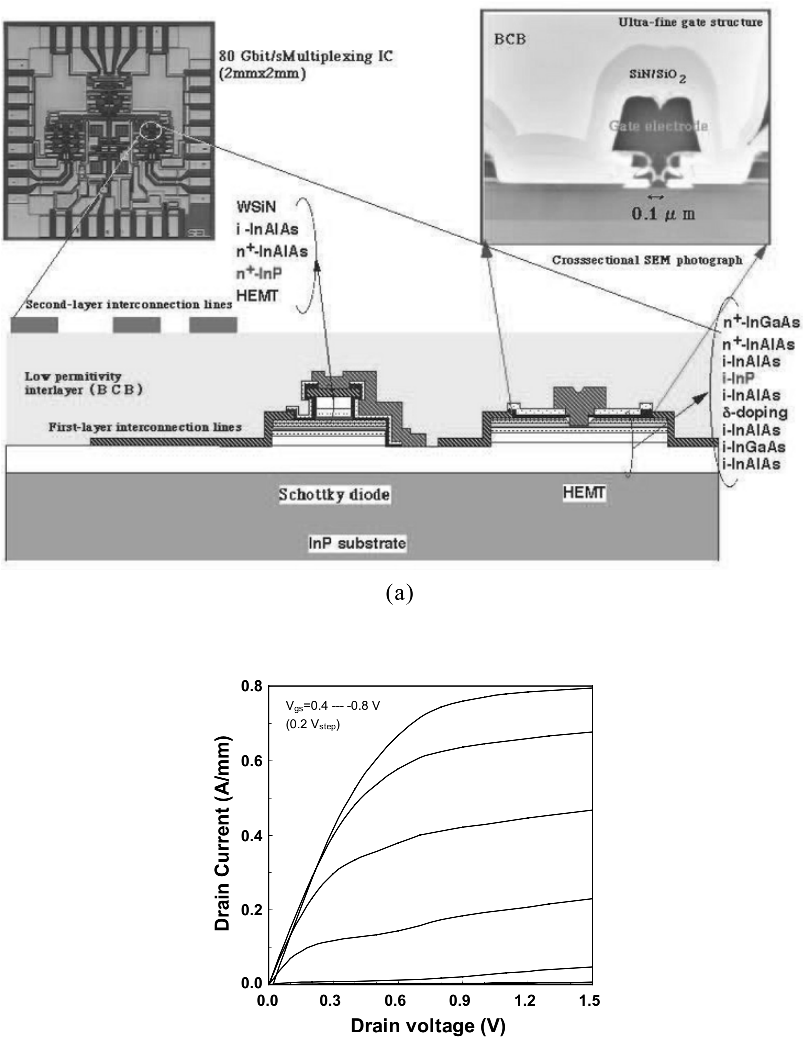
In the I-V characteristics, we see that the current saturates at a very low voltage, indicative of the low contact resistance, access resistance, and channel resistance that can be achieved with this technology. However, one can see that the current does not completely saturate. This non-zero output conductance results from short channel effects. Additionally, in the I-V curves, the device is only biased to 1.5 V, since the breakdown voltage for InP devices with such short gate lengths is typically $\sim$ 3 V. In GaN-based HFET technology, much higher breakdown voltages can be achieved due to the wide bandgaps of the materials in the nitride system.

In this chapter we will examine some important issues in HFETs. In particular we will examine how polar charge can be exploited to create free electron or hole gas. Such undoped HFETs have become very important due their use in the large bandgap AlGaN/GaN technology.

8.5 CHARGE CONTROL MODEL FOR THE MODFET

In a MODFET, electrons are introduced into the channel via doping of a region which is spatially separated from the channel, as shown in figure 8.13 and figure 8.15. In this way, the electron mobility is not degraded by ionized impurity scattering. A number of different doping schemes are possible for this device. The entire barrier material, with the exception of a thin spacer layer near the channel, can be doped, resulting in the structure that was shown in figure 8.13. Alternatively, one can dope a very thin ($\sim$ 10 Å) layer of barrier material separated from the channel by an undoped spacer layer. This scheme, known as δ-doping, is illustrated in figure 8.15a.

While the continuous doping scheme is in practice easier to implement, δ-doping is preferable because the maximum amount of charge that can be induced in the channel is higher. Additionally, δ-doping reduces the risk of inducing a parasitic channel within the barrier material. For the MODFET charge control model introduced in this chapter, we assume a δ-doped layer with an areal donor density N_d [cm^{-2}] separated from the heterointerface by a distance d_s, as shown in figure 8.15a.



(a)

(b)

Figure 8.14: (a) Cross sectional SEM image of a state-of-the-art InP-based HFET device. Also shown is a diagram indicating the device layer structure, as well as a microscope image of a circuit made up of these devices. (b) I-V characteristics of the device. Images courtesy of T. Enoki, NTT.

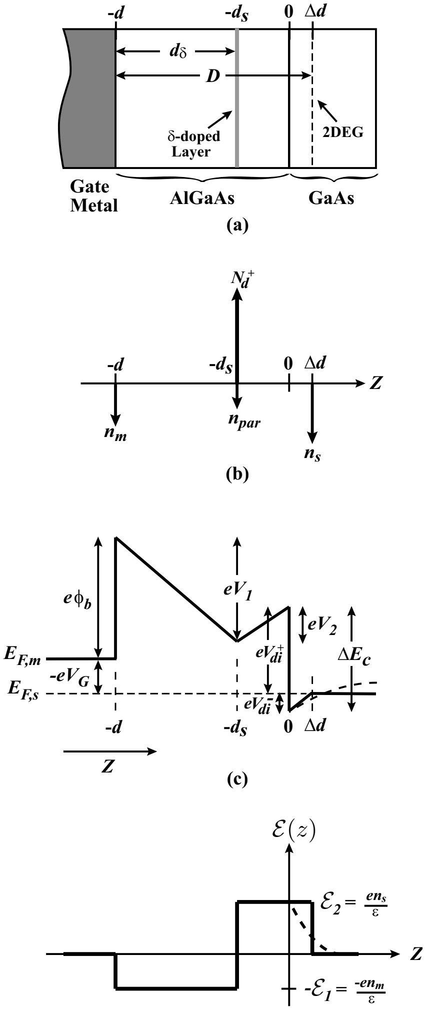


Figure 8.15: (a) Schematic diagram of a δ-doped AlGaAs/GaAs MODFET, along with (b) the charge distribution, (c) the band diagram, and (d) the electric field distribution in the structure.

The charge distribution in the system is determined by electrostatics and can be varied by applying a voltage to the gate. In general, electrons from donors in the barrier region can end up in one of three places:

1. In the channel. We will call this charge n_s.
2. On the gate. We will call this charge n_m.
3. Inside the barrier material, where they create a parasitic channel. We call this charge n_{par}.

All charges are expressed in units [cm^{-2}]. We treat the distributed 2DEG as if it were a perfect 2-dimensional sheet placed a distance Δd from the heterointerface, where Δd is simply the centroid of the 2DEG charge distribution. The resulting charge distribution, band diagram, and electric field profile in the system is are shown in figure 8.15. For the purpose of this discussion, let us assume the heterojunction is between AlGaAs and GaAs. In this analysis we use the result of Kroemer that the capacitance of a Schottky barrier on a semiconductor with an arbitrary charge distribution is

$$C = \frac{\Delta Q}{\Delta V} = \frac{\epsilon}{< x >}$$

where $< x >$ is the centroid of incremental displaced electron distribution, ΔQ, caused by ΔV.

Note that when the charge centroid approximation is used, the electric field in the GaAs is terminated at the centroid of the charge distribution. The actual electric field in the GaAs, indicated by the dashed line in figure 8.15c, is gradually terminated by the 2DEG following Gauss' Law, where

$$\frac{\partial \mathcal{E}}{\partial z} = -\frac{en(z)}{\epsilon} \tag{8.5.1}$$

and $n(z)$ is the local volume electron concentration [cm^{-3}]. Similarly, the band diagram has been drawn as a solid line for the charge centroid approximation and as a dashed line for the true behavior.

Charge neutrality states that the total charge in the system must be zero, or

$$N_d^+ = n_m + n_{par} + n_s \tag{8.5.2}$$

For the purpose of MODFET operation, it is desirable that $n_{par} = 0$, since electrons in the barrier region create a low mobility parallel parasitic current path. n_{par} can become significant when a large forward bias is applied to the gate or if N_d is very large. For the remainder of this discussion, we will assume that the device is biased such that n_{par} is negligible.

In HFETs with channels having electrons with a low electron effective mass and hence a low density of states, it is important to consider the variation in eV_{di}^- (see figure 8.15c) as a function of the charge n_s in the channel. Clearly, an increase in n_s also requires an increase in eV_{di}^-. This is an undesirable effect because

1. It decreases the channel confinement potential and hence sets a limit on the maximum current available.

2. It effectively acts as a voltage divider between the gate and source as the Fermi level in the channel $E_{F,ch}$ is raised relative to the Fermi level in the source $E_{F,s}$, and hence only part of the applied gate-source voltage V_{GS} is used for charge control. Thus the intrinsic source-gate bias $V_{GS,int}$ is related to the applied source gate voltage V_{GS} by

$$V_{GS,int} = V_{GS} - (E_{F,ch} - E_{F,s}) \tag{8.5.3}$$

This voltage division (or reduced charge control) can also be represented by a displacement of the centroid of the 2DEG Δd away from the heterointerface (see figure 8.15c), effectively increasing the gate to channel distance to $d + \Delta d$ and reducing the gate capacitance C_G to

$$C_G = \frac{\epsilon A}{d + \Delta d}$$

This is sometimes referred to as gate capacitance reduction due to a quantum capacitance associated with motion of the fermi level. We now calculate an analytic expression for Δd. We first assume that the 2DEG forms a triangular potential well, as shown in figure 8.16. The sub-band energies are well known to be

$$E_i \simeq \left(\frac{\hbar^2}{2m^*}\right)^{1/3} \left[\frac{3}{2} e\mathcal{E}_2 \pi \left(i + \frac{3}{4}\right)\right]^{2/3} \tag{8.5.4}$$

We assume that the electric field $\mathcal{E}_2$ is generated by only the 2DEG charge n_s, yielding

$$\begin{aligned} E_i &\simeq \left(\frac{\hbar^2}{2m^*}\right)^{1/3} \left[\frac{3}{2} e\pi \left(i + \frac{3}{4}\right)\right]^{2/3} \left(\frac{en_s}{\epsilon}\right)^{2/3} \\ &= \gamma_i n_s^{2/3} \end{aligned} \tag{8.5.5}$$

where $i = 0, 1, 2, ..., n$. The coefficients γ_i are material dependent, explicitly related to the density of states effective mass m^*. Typical values for GaAs are

$$\gamma_0 = 2.5 \times 10^{-12} \text{ eV} \cdot \text{cm}^{4/3}$$
$$\gamma_1 = 3.2 \times 10^{-12} \text{ eV} \cdot \text{cm}^{4/3}$$

The 2DEG concentration is related to the position of the Fermi level via the Fermi-Dirac distribution

$$n_s = D_s \frac{k_B T}{e} \sum_{i=0}^{n} \ln \left[1 + \exp\left(\frac{e(E_F - E_i)}{k_B T}\right)\right] \tag{8.5.6}$$

where D_s is the 2D density of states

$$D_s = \frac{em^*}{\pi \hbar^2} \tag{8.5.7}$$

Assuming only the first sub-band is dominant, we can write

$$n_s = D_s \frac{k_B T}{e} \ln \left[1 + \exp\left(\frac{e(E_F - E_0)}{k_B T}\right)\right] \tag{8.5.8}$$

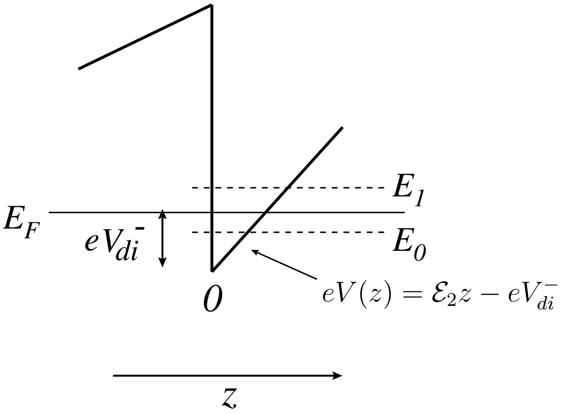


Figure 8.16: Band structure of the HFET channel region represented as a triangular potential well.

For the case of $(E_F - E_0)/k_BT \geq 1$, we get

$$n_s \cdot \frac{e}{k_BT} \cdot \frac{1}{D_s} = \frac{E_F - E_0}{k_BT} \tag{8.5.9}$$

or

$$E_F - E_0 = \frac{e}{D_s} n_s = \frac{\pi\hbar^2}{m^*} n_s \tag{8.5.10}$$

For $E_F - E_0 \approx eV_{di}^-$, we get

$$V_{di}^- = \left(\frac{\pi\hbar^2}{em^*}\right) = an_s \tag{8.5.11}$$

This tells us that eV_{di}^-, the amount that the conduction band drops below the Fermi energy at the heterointerface, increases linearly with n_s. The coefficient a in equation 8.5.11 is clearly material dependent since it varies with m^*. By examining the band diagram and the electric field profile near the channel, we can calculate Δd.

$$V_{di}^- = an_s = \mathcal{E}_2 \cdot \Delta d = \frac{en_s}{\epsilon} \cdot \Delta d \tag{8.5.12}$$

$$\boxed{\Delta d = \frac{\epsilon a}{e}} \tag{8.5.13}$$

Typical values of Δd are 80 Å for the AlGaAs/GaAs system, 50 Å for the AlInAs/GaInAs, and 20 Å for the AlGaN/GaN system and the Si/SiO$_2$2 system. When calculating the band diagram of a HEMT, one of two boundary conditions are typically used:

1. The electric field in the buffer (or bulk) is zero.
2. The voltages in the system are specified

The first condition allows voltages in the system to adjust and the second allows charges and hence fields to adjust. Both conditions should not be applied simultaneously. We can now use the band diagram in figure 8.15c to calculate the charge in the 2DEG as a function of gate bias V_G. The methodology is to follow the energy bands from the Fermi level in the metal to that in the GaAs and set the difference equal to the gate bias V_G. After dividing by the electron charge e, we get the following equation:

$$-V_G + \phi_b - V_1 + V_2 - \frac{\Delta E_c}{e} + V_{di}^- = 0 \tag{8.5.14}$$

V_1 and V_2 are found by solving Poisson's equation and are given by

$$V_1 = \frac{en_m\left(d - d_s\right)}{\epsilon} \tag{8.5.15}$$

$$V_2 = \frac{en_s d_s}{\epsilon} \tag{8.5.16}$$

Substituting the relationships from equation 8.5.6 and equation 8.5.7 into equation 8.5.5 and rearranging terms, we get

$$\frac{en_s\left(d + \Delta d\right)}{\epsilon} - \frac{eN_d^+\left(d - d_s\right)}{\epsilon} - \left[V_G - \left(\phi_b - \Delta E_c/e\right)\right] = 0 \tag{8.5.17}$$

From figure 8.15a, we see that $d - d_s = d_\delta$ is the distance between the gate and the δ-doped layer, and $d + \Delta d = D$ is the distance between the gate and the 2DEG. Solving for n_s gives us

$$\boxed{n_s(V_G) = \frac{eN_d^+ d_\delta + \epsilon\left[V_G - \left(\phi_b - \Delta E_c/e\right)\right]}{eD}} \tag{8.5.18}$$

The term $N_d^+(d_\delta/D)$ in our expression for n_s depicts what is known as the Lever Rule for charge sharing. To illustrate its impact, consider the special case where $\phi_b - V_G = \Delta E_c/e$. When the δ-doped layer is half way between the gate and the channel ($d_\delta = D/2$), the charge is shared equally between the gate metal and the 2DEG ($n_s = n_m = N_d^+/2$). When the δ-doped layer is brought closer to the metal, more of the charge is imaged on the gate; as $d_\delta \rightarrow 0$, $n_m \rightarrow N_d^+$. Similarly, as the δ-doped layer is brought closer to the channel, more of the charge is imaged in the 2DEG. The change in the charge in the 2DEG can be related to the gate voltage as

$$\begin{aligned}\Delta n_s &= n_s\left(V_g = 0\right) - n_s\left(V_g\right)\\ &= \frac{N_d d_z - \epsilon/e\Delta\left(0\right)}{D} - \frac{N_d d_z - \epsilon/e\Delta\left(V_g\right)}{D}\end{aligned}$$

or

$$e\Delta n_s = \frac{\epsilon}{D}\cdot\left[\Delta\left(V_g\right) - \Delta\left(0\right)\right]$$

therefore

$$e\Delta n_s = \frac{\epsilon}{D} \cdot (\pm V_g)$$

This is the charge control equation of the gate capacitor where

$$\Delta Q_{2DEG}\ \left[\mathrm{C} \cdot \mathrm{cm}^{-2}\right] = C_g \cdot \Delta V_g \quad \left[\mathrm{F} \cdot \mathrm{V} \cdot \mathrm{cm}^{-2}\right]$$

This is to be expected since we are indeed dealing with a capacitor where the depleted AlGaAs layer is the dielectric and the two plates of the capacitor are the gate metal and the centroid of the 2DEG separated by a distance D. By examining equation 8.5.9, we see that at a given gate voltage, n_s increases linearly with d_δ. Thus, moving the δ-doped layer closer to the AlGaAs/GaAs heterojunction causes more of the induced charge to be imaged in the channel rather than on the gate. This also illustrates why a δ-doped structure is preferable to continuous doping; in the δ-doped structure, the centroid of the donor charge distribution is much closer to the 2DEG, resulting in more charge being induced in the channel. However, moving the doped layer too close to the heterointerface causes a degradation in channel mobility, since ionized impurity scattering increases.

The pinch-off voltage V_p in a MODFET is the gate voltage required to deplete the channel of carriers. To find V_p, we set n_s in equation 8.5.9 equal to zero and solve for the gate voltage. This gives us

$$V_p = -\frac{eN_d^+ d_\delta}{\epsilon} + (\phi_b - \Delta E_c/e) \tag{8.5.19}$$

figure 8.17a shows the band diagram of a MODFET biased at pinch-off. Here, $n_m = -N_d^+$ and $n_s = 0$, so the the only region with a non-zero electric field is between the gate and the δ-doped layer.

In figure 8.17b, we show a MODFET with a large forward bias on the gate. If we bias the device at pinch-off (figure 8.17a) and then increase the voltage on the gate, charge is transferred from the gate (n_m) to both the 2DEG (n_s) and the barrier (n_{par} increases and N_d^+ decreases, since some of the electrons end up in the conduction band and some fill empty donor states). Initially, almost all of the charge from the gate is transferred to the channel, and the change in n_{par} and N_d^+ remains small. However, as V_G becomes large, the conduction band in the AlGaAs begins to approach the Fermi level, implying that the electron concentration in the barrier must be increasing (see figure 8.17b). Hence, if the gate voltage is further increased, charge is transferred from the gate into both the 2DEG and the barrier. This is obviously not the biasing required for good MODFET performance. The device operates between the two limits given by figure 8.17a and figure 8.17b.

8.5.1 Modulation Efficiency

We have seen that in general, modulating the gate voltage causes charge to be transferred from the gate to both the 2DEG and the barrier region. Even under optimal MODFET bias conditions, n_{par} and $N_d^0 = N_d - N_d^+$ (the density of occupied donors in the AlGaAs) are typically negligible, but they are not zero, so increasing V_G will still cause a small change in the charge density in the AlGaAs. The concept of modulation efficiency was introduced by Foisy et al to describe

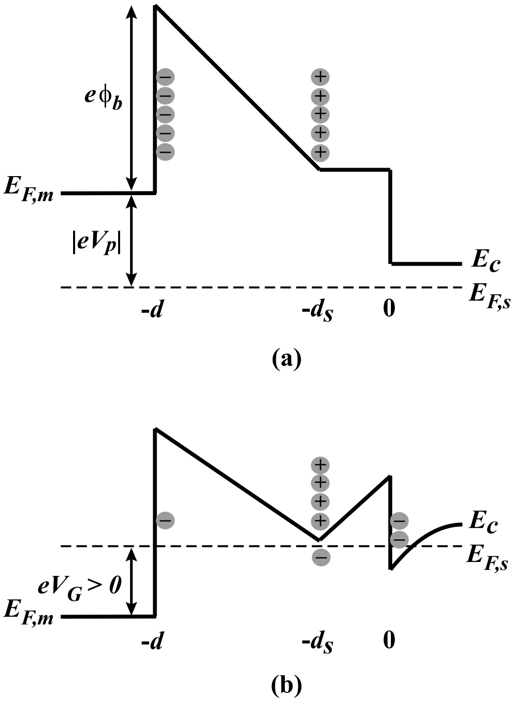


Figure 8.17: A schematic diagram of a MODFET band profile under conditions where (a) a negative gate bias is applied to completely deplete the 2DEG, and (b) a large positive gate bias is applied, such that the gate loses control over the 2DEG.

charge transfer in the situation where a change in V_G does not exclusively result in a change in the 2DEG concentration. We define the modulation efficiency (ME) to be

$$ME(V_G) = \frac{e\Delta n_s}{\Delta V_G} \cdot \frac{1}{C_G} \tag{8.5.20}$$

where $C_G = \epsilon/D$ (see figure 8.15a) is the gate-channel capacitance. The denominator of equation 8.5.11 $C_G \cdot \Delta V_G$ represents the ideal induced charge in the 2DEG.

In general, the change in the charge density on the gate is

$$|\Delta n_m| = \Delta n_s + \Delta n_{par} + \Delta N_d^0 \tag{8.5.21}$$

where ΔN_d^0 is the change in density of occupied donors in the AlGaAs, the superscript emphasizing that this is the change in the concentration of neutral donor atoms. For simplicity, we will assume $\Delta n_{par} = \int_{-d}^{0} n(\text{AlGaAs})dx \to 0$. From equation 8.5.9, we have

$$n_s(V_G) = \frac{\left(N_d - N_d^0\right) d_\delta - \epsilon/e\left[V_G - (\phi_b - \Delta E_c/e)\right]}{D} \tag{8.5.22}$$

From this expression, we can solve for $\frac{\Delta n_s}{\Delta V_G}$:

$$\frac{\Delta n_s}{\Delta V_G} = \frac{1}{D}\left[-\frac{\Delta N_d^0}{\Delta V_G} \cdot d_\delta + \frac{\epsilon}{e}\right] \tag{8.5.23}$$

Inserting this into equation 8.5.11 gives us the following expression for the modulation efficiency:

$$ME(V_G) = 1 - \frac{e}{\epsilon}\frac{\Delta N_d^0}{\Delta V_G} \cdot d_\delta = 1 - \frac{C_{p,eff}}{C_p} \tag{8.5.24}$$

where

$$C_{p,eff} = \frac{e\Delta N_d^0}{\Delta V_G}$$
$$C_p = \frac{\epsilon}{d_\delta}$$

Again, if the change in charge density in the AlGaAs ΔN_d^0 is negligible, then $C_{p,eff} \to 0$ and $ME \to 1$. Finding an expression for $C_{p,eff}$ requires solving for the conduction band occupancy in the AlGaAs as a function of V_G. This must be done numerically and is left as a problem for the reader.

Example 8.1 Consider an n-type GaAs/Al$_{0.3}$Ga$_{0.7}$As MODFET at 300 K with the following parameters:

Schottky barrier height,	ϕ_b	$=$	0.9 V
Barrier doping,	N_d	$=$	10^{18} cm^{-3}
Conduction band discontinuity,	ΔE_c	$=$	0.24 eV
Dielectric constant of the barrier,	ϵ_b	$=$	12.2
Spacer layer thickness,	d_s	$=$	30 Å
Barrier thickness,	d	$=$	350 Å

Calculate the 2DEG concentration at V_G = 0 and V_G = −0.5 V.

The parameter V_{p2} of this structure is given by

$$\begin{aligned} V_{p2} &= \frac{eN_d}{\epsilon_b}(d - ds)^2 = \frac{\left(1.6 \times 10^{-19}\ \text{C}\right)\left(10^{18}\ \text{cm}^{-3}\right)\left(320 \times 10^{-8}\ \text{cm}\right)^2}{12.2\,(8.85 \times 10^{-14}\ \text{F/cm})} \\ &= 1.52\ \text{V} \end{aligned}$$

The threshold voltage V_{off} is given by

$$V_{off} = 0.9 - 0.24 - 1.52 = -0.86\ \text{V}$$

The device is thus a depletion mode MODFET. The 2DEG carrier concentration is given by

$$\begin{aligned} n_s\left(V_G=0\right) &= \frac{12.2\left(8.85\times10^{-14}\ \mathrm{F/cm}\right)\left(0.86\ \mathrm{V}\right)}{\left(1.6\times10^{-19}\ \mathrm{C}\right)\left(350\times10^{-8}cm\right)} = 1.66\times10^{12}\ \mathrm{cm}^{-2} \\ n_s\left(V_G=-0.5\right) &= \frac{12.2\left(8.85\times10^{-14}\right)\left(0.36\right)}{\left(1.6\times10^{-19}\right)\left(350\times10^{-8}\right)} = 6.94\times10^{11}\ \mathrm{cm}^{-2} \end{aligned}$$

8.6 POLAR MATERIALS AND STRUCTURES

8.6.1 Polar Materials

An emerging class of materials is the (Al,Ga,In)N-based system for use in both optoelectronics and electronics. These materials are fundamentally different from conventional cubic semiconductors in that they exist normally in the wurtzite phase and exhibit strong polarization in the <0001> direction (also known as the C-direction). Before studying HFETs fabricated from these materials, it is necessary to first understand the effects that these polarization fields have on the electronic properties of the material.

Figure 8.18a shows the ball and stick model of GaN in the C^+ orientation (Ga face on top) and the associated polarization in the crystal. In the classical model, these polarization charges exist on each unit cell. The sum of the internal polarization within the crystal is zero, as shown in figure 8.18b, leaving $\pm\ Q_\pi$ charge at each end of the crystal forming a dipole . Since an unscreened dipole will result in a non-sustainable dipole moment, nature will always provide for a screening dipole by placing equal and opposite charges at or close to the charges of polarization dipole, as shown in figure 8.18c. Let us consider some numbers to see how large the polarization dipole moment is.

The spontaneous polarization charge density in GaN $n_\pi \sim 10^{13}\ \mathrm{cm}^{-2}$. This leads to an electric field

$$\mathcal{E}_\pi = \frac{Q_\pi}{\epsilon} = \frac{en_\pi}{\epsilon} \simeq 1.6\ \mathrm{MV/cm} \tag{8.6.1}$$

In a crystal of thickness $d = 1\ \mu m$, the voltage across the material that results from this dipole charge is

$$V_\pi = \mathcal{E}_\pi \cdot d = 160\ \mathrm{V} \tag{8.6.2}$$

which is not sustainable. Hence a screening dipole is essential. This raises the question of what is the nature of the charges that form the screening dipole. They could arise from counter ions from the atmosphere (such as H^+ and OH^-). This is probably the case for bulk polar materials used in the ceramic industry (such as ZnO for varistors and piezoelectric sensors). However, this is probably not the case for epitaxial GaN thin films, since these films can be created in an atmosphere free of counter ions, such as in an MBE reactor. This begs the question of whether screening is possible without external counter ions. The following discussion addresses this issue.

Consider a lightly doped n-type GaN sample in the initial stages of growth, shown in figure 8.19a. Due to the lack of availability of GaN substrates, currently GaN is typically grown

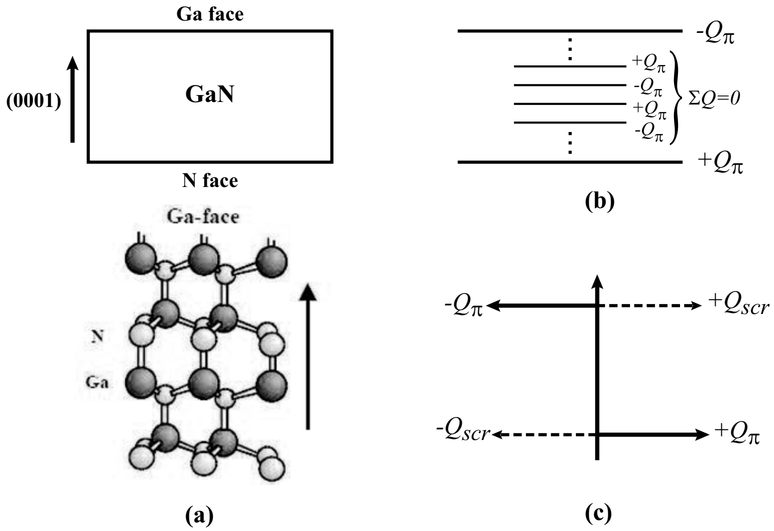


Figure 8.18: (a) Stick-ball representation of Wurtzite GaN crystal structure. (b) Classical model of polarization charge in a polar material such as GaN. (c) Crystal will draw in charge to screen the polarization dipole - From M. J. Murphy et al, MRS Internet J. Nitride Semicond. Res. 4S1, G8.4(1999)

heteroepitaxially (on sapphire, Si, or SiC substrates). The material at the substrate / thin film interface is highly defective and therefore capable of trapping mobile charges. We will assume that the effect of the background n-type doping on the electric field profile within the material is negligible compared to the electric field generated by the polarization charges. We will also ignore the effects of surface states on the electrical properties of the material. Both of these effects will be considered later.

In the absence of surface states, as the material becomes thicker, the electric field in the material (given by the slope of the conduction and valence band) will remain constant until the valence band crosses the Fermi level, as shown in figure 8.19b. The thickness of the film d_{cr} at which this occurs is given simply by

$$d_{cr} = \frac{E_g}{e\mathcal{E}_\pi} = \frac{3.4\,\mathrm{eV}}{1.6\,\mathrm{MeV/cm}} \simeq 215\ \text{Å} \tag{8.6.3}$$

where $E_g = 3.4$ eV is the bandgap of GaN. Once $d > d_{cr}$, holes begin to accumulate at the surface (created by generation across the gap), leading to an equal electron concentration which drifts to the substrate-epi interface (the GaN N-face), creating a screening dipole. This is illustrated in figure 8.19c. The magnitude of the screening charge Q_{scr} increases continuously

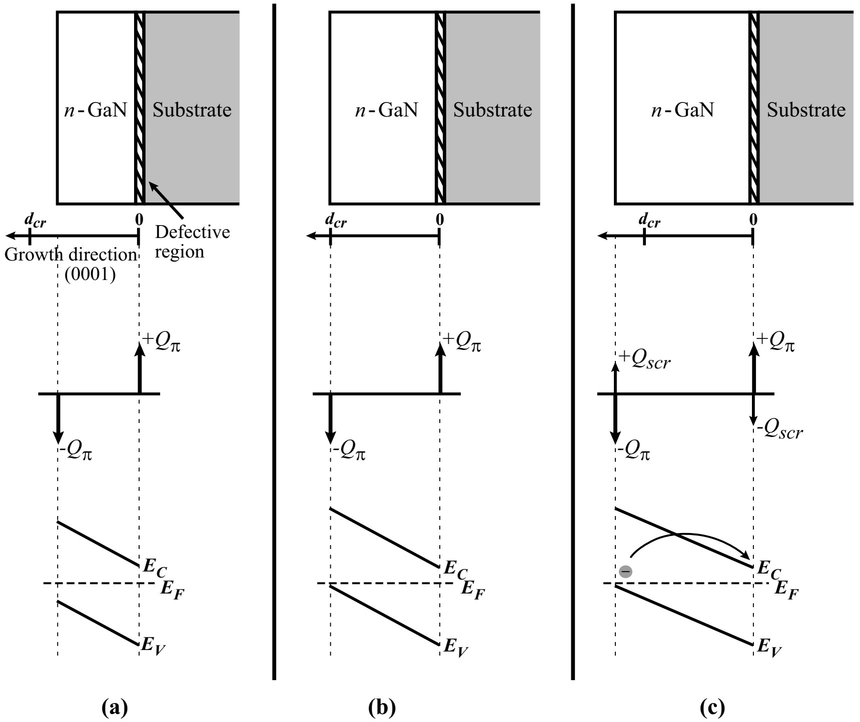


Figure 8.19: Schematic diagram of an n-type GaN sample along with charge profile and band diagram (a) during the initial stages of growth, (b) for $d = d_{cr}$, and (c) for $d > d_{cr}$.

with epitaxial layer thickness. The evolution of the screening charge with distance is obtained by recognizing that the maximum voltage across the structure is the bandgap of the material, or

$$\frac{1}{e}E_g = |\mathcal{E}| \cdot d = \left(\frac{Q_\pi - Q_{scr}}{\epsilon}\right) d \tag{8.6.4}$$

$$Q_{scr} = Q_\pi - \frac{\epsilon E_g}{ed} \tag{8.6.5}$$

As $d \rightarrow \infty$, $Q_{scr} \rightarrow Q_\pi$, or in other words for very thick samples the polarization dipole is fully screened.

If we now assume that there exists a surface donor state, a very similar situation develops, except that instead of holes providing the positive screening charge, ionized surface donors do. These states pin the Fermi level at the surface to create a built-in voltage equal to the donor

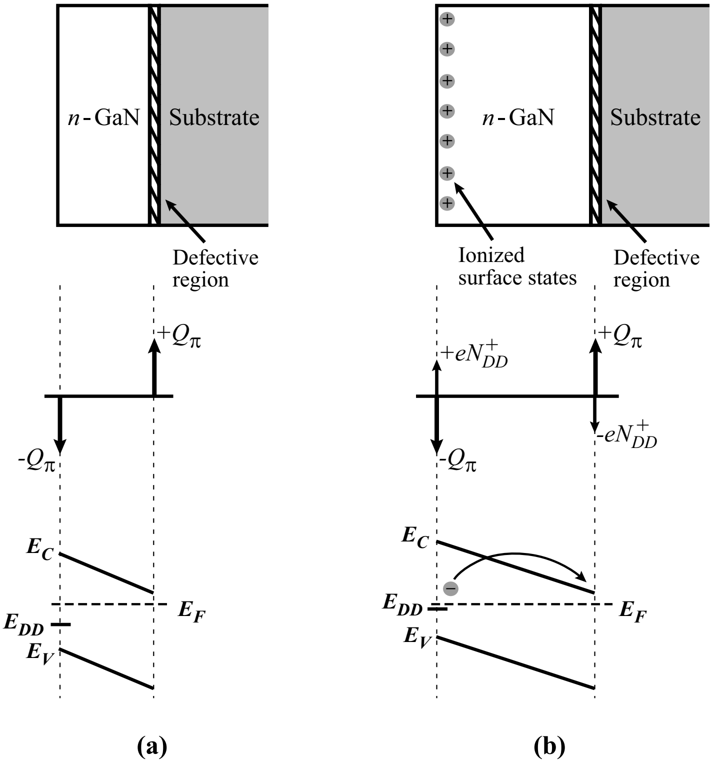


Figure 8.20: Schematic diagram of an n-type GaN sample along with charge profile and band diagram when the effects of surface states are taken into account. (a) Very thin GaN, for which surface states are not ionized. (b) Once GaN is thick enough such that E_{DD} is very close to E_F at the GaN surface, surface donors become ionized and polarization charge is screened.

depth $(E_C - E_{DD})/e$, as illustrated in figure 8.20. As the epitaxial thickness increases, the donor level E_{DD} approaches the Fermi level E_F at the GaN surface, and the screening charge N_{DD}^+ increases as given by the Fermi-Dirac occupancy probability

$$N_{DD}^+ = [1 - f(E_{DD}(0))]\, N_{DD} \tag{8.6.6}$$

$$= \left[\frac{\exp\left(\frac{E_{DD}(0)-E_F}{k_BT}\right)}{1+\exp\left(\frac{E_{DD}(0)-E_F}{k_BT}\right)}\right] \cdot N_{DD} \tag{8.6.7}$$

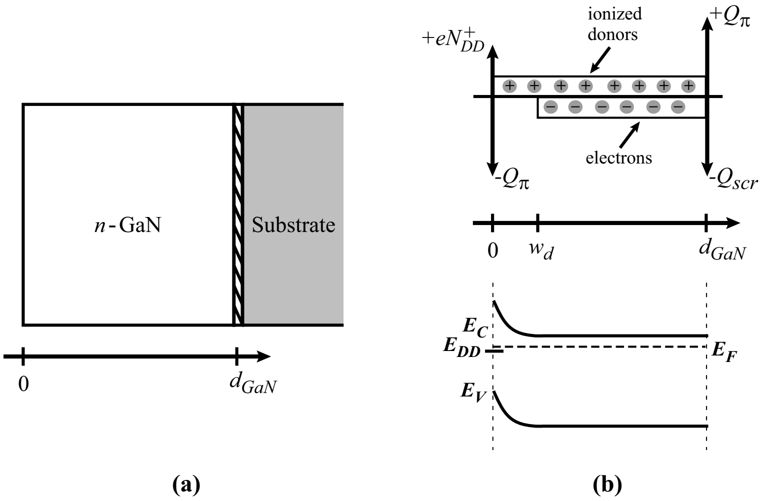


Figure 8.21: Schematic diagram of (a) a thick n-type GaN sample along with (b) the corresponding charge profile and band diagram when surface states and the donor charge are taken into account.

If $eN_{DD} > Q_{\pi}$, then Q_{π} will be fully screened when $(E_C - E_{DD}) - (E_C - E_F)$ is very close to zero, or in other words when E_{DD} is very close to E_F. Analogous to the previous case of screening via holes,

$$N_{DD}^{+} = Q_{\pi} - \frac{\epsilon E_{DD}}{ed} \tag{8.6.8}$$

where E_{DD}/e is now the built-in voltage as opposed to E_g/e.

We have ignored the effects of the donor charge in the analysis until now because we were seeking to understand the formation of the screening dipole. Figure 8.21 shows the band diagram of a thick n-type GaN film. Experimental evidence has shown that the surface of GaN indeed has a neutral level, E_{DD}, which is currently assumed to be the position of the surface donor. Since the GaN is considered to be very thick, the energy bands must be flat (zero electric field). The surface negative polarization charge $-Q_{\pi}$ is balanced by the sum of the positively charged ionized surface states N_{DD}^{+} and the areal density of charges in the depletion region $N_d \cdot w$, or

$$Q_{\pi} = N_{DD}^{+} + N_d \cdot w \tag{8.6.9}$$

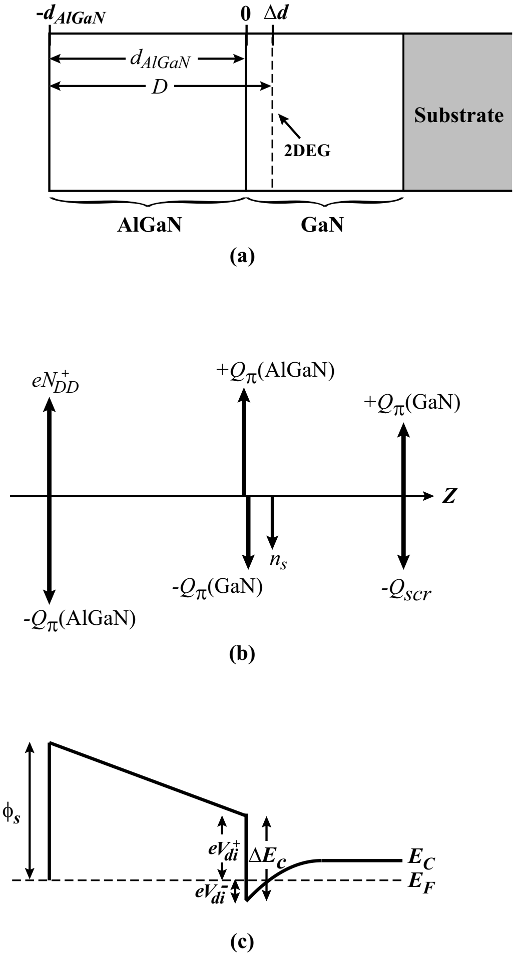


Figure 8.22: Polar heterostructures can generate a 2DEG which is used as the channel region of an HFET. (a) Typical AlGaN/GaN heterostructure used in polar HFET technology, along with (b) the charge distribution and (c) the band diagram of the structure.

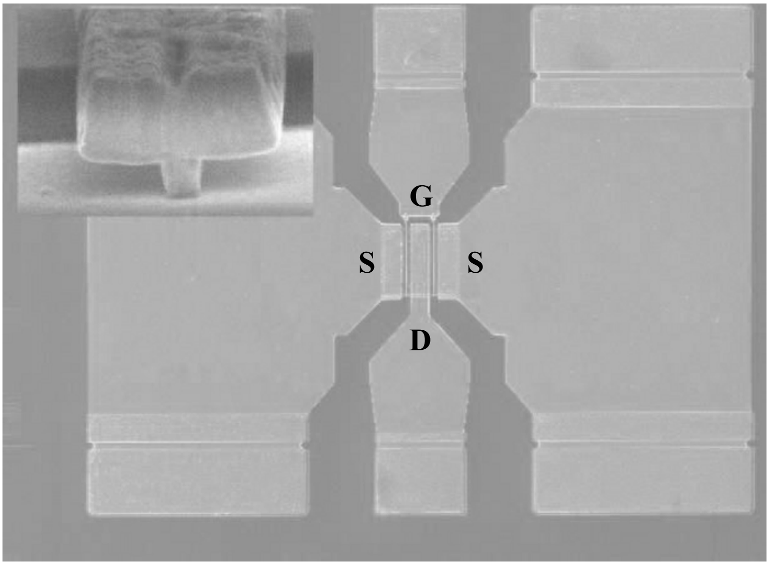


Figure 8.23: Top view of an AlGaN/GaN HFET structure with 2 gate fingers. Pictured in the inset is a close up of the 0.12 μm T-gate. Picture courtesy of Ilesanmi Adesida.

8.6.2 Polar HFET Structures

Now that we have described how charge is distributed within polar materials, we are ready to show how polarization fields can be used to generate a 2DEG in polar heterostructures. Consider the AlGaN/GaN structure illustrated in figure 8.22. The charge density distribution is shown along with the band diagram. For sufficiently thick AlGaN layers, the surface potential $e\phi_s$ is pinned by the surface donor and is approximately equal to the donor depth E_{DD}. Due to the lattice mismatch between AlGaN and GaN, the thin AlGaN cap is under tensile strain. Hence the total polarization charge at the AlGaN surface $-Q_\pi(\mathrm{AlGaN})$ is the sum of the spontaneous and piezoelectric contributions from the AlGaN. In addition to the negative polarization charge, there will also be a positive charge at the surface N_{DD}^+ resulting from the ionized surface donors.

At the AlGaN/GaN interface, the net polarization charge $Q_\pi(\mathrm{net})$ is the sum of the polarization contributions from the AlGaN and the GaN, or

$$Q_\pi(\mathrm{net}) = Q_\pi(\mathrm{AlGaN}) - Q_\pi(\mathrm{GaN}) \tag{8.6.10}$$

$Q_\pi(\mathrm{net})$ is a positive number for Ga-face polarity because of the higher polarization in the AlGaN relative to GaN. From the band diagram, we can see that there must also exist a distribution of electrons in the GaN near the AlGaN/GaN heterointerface. Again, we have drawn the distributed charge as a 2 dimensional sheet charge of density en_s a distance Δd from the heterointerface, where Δd is the centroid of the charge distribution.

To find our 2DEG carrier density n_s, we follow the same methodology as in our MODFET analysis (see equation 8.5.5). Doing this, we get

$$\phi_s - V_1 - \frac{\Delta E_c}{e} + V_{di}^- = 0 \tag{8.6.11}$$

where V_{di}^- was given in equation 8.5.3 and V_1 is given by

$$V_1 = \frac{[Q_\pi(\mathrm{net}) - en_s]\, d_{AlGaN}}{\epsilon} \tag{8.6.12}$$

Substituting these values into equation 8.6.11 and setting $d_{AlGaN} + \Delta d = D$ gives us for n_s

$$\boxed{n_s = \frac{Q_\pi(\mathrm{net}) \cdot d_{AlGaN} - \epsilon\,(\phi_s - \Delta E_c/e)}{eD}} \tag{8.6.13}$$

which is the same expression as that derived for conventional HFETs with $d_s = 0$ (i.e. with the donor sheet at the heterointerface). This is reassuring, since in the case of AlGaN/GaN heterostructures, the positive sheet charge that induces the 2DEG is the net polarization charge at the heterointerface. The physical difference between conventional and polar HFETs is simply the origin of the electrons in the 2DEG. In conventional HFETs, the channel electrons are provided by a donor sheet, while in GaN-based HFETs, they come from ionized surface donor states.

In an HFET structure, we place a gate metal on top of the AlGaN layer and apply a gate voltage to modulate the charge in the 2DEG. The only difference between the HFET charge control analysis and the one presented here for an AlGaN/GaN heterostructure is that in the HFET, the potential barrier at the metal/AlGaN interface is given by $\phi_b - V_G$, where ϕ_b is the metal-semiconductor barrier height and V_G is the applied gate voltage. Thus, for the AlGaN/GaN HFET, the 2DEG sheet charge density as a function of gate voltage can be written as

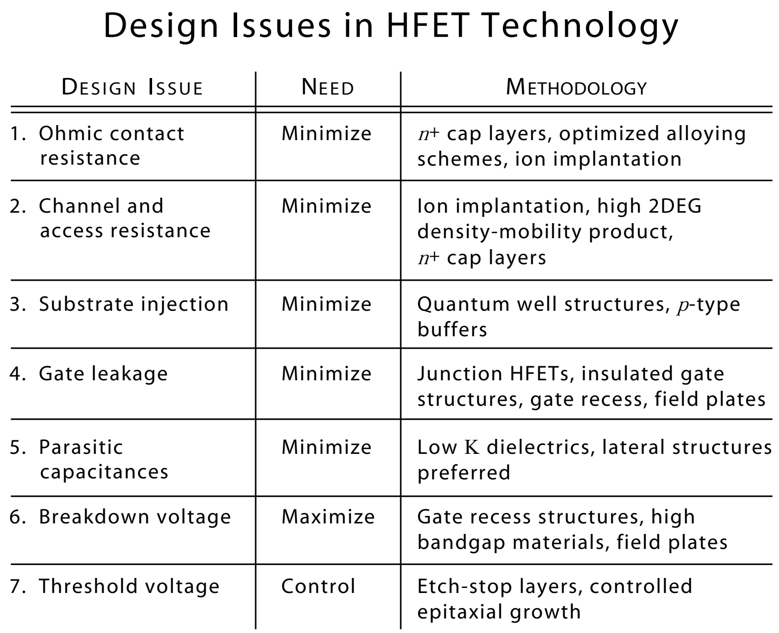
$$\boxed{n_s(V_G) = \frac{Q_\pi(\mathrm{net}) \cdot d_{AlGaN} + \epsilon\,[V_G - (\phi_b - \Delta E_c/e)]}{eD}} \tag{8.6.14}$$

8.7 DESIGN ISSUES IN HFETS

In addition to the issue of aspect ratio discussed for MESFETs and JFETs in Chapter 8, there are several other design issues to be considered in HFETs. They are summarized in table 8.1. We will discuss a number of techniques which are employed in modern HFET processes that address these issues.

8.7.1 n^+ Cap Layers

n^+ cap layers are used to reduce the contact resistance as well as the access resistance in the device. The schematic of an AlGaAs/GaAs HFET with an n^+ GaAs cap layer is shown in



Design Issues in HFET Technology

Design Issue	Need	Methodology
1. Ohmic contact resistance	Minimize	n+ cap layers, optimized alloying schemes, ion implantation
2. Channel and access resistance	Minimize	Ion implantation, high 2DEG density-mobility product, n+ cap layers
3. Substrate injection	Minimize	Quantum well structures, p-type buffers
4. Gate leakage	Minimize	Junction HFETs, insulated gate structures, gate recess, field plates
5. Parasitic capacitances	Minimize	Low K dielectrics, lateral structures preferred
6. Breakdown voltage	Maximize	Gate recess structures, high bandgap materials, field plates
7. Threshold voltage	Control	Etch-stop layers, controlled epitaxial growth

Table 8.1: Overview of technology issues that must be addressed in HFET design.

figure 8.24 along with a band diagram. The access resistance of the HFET is comprised of the sheet resistance of the n^+ cap layer R_{n^+}, the sheet resistance of the 2DEG R_{2DEG}, and the interchannel resistance posed by the barrier to electron flow between the cap and the channel R_{int} (see figure 8.24b). The total resistance can be modeled as a distributed network of all of these components, as shown schematically in figure 8.24a.

Solutions to reducing R_{int} are to reduce the barrier to electron flow from the n^+ layer to the 2DEG channel, and to reduce the barrier to tunneling. The first is best achieved by increasing the doping in the n^+ layer so that it is very degenerate, causing E_F to rise above E_C. The second is achieved by adding n^+ doping in the AlGaAs layer nearest to the surface, which in turn enhances the surface electric field and thereby tunneling. A schematic diagram illustrating the benefits of both these solutions is shown in figure 8.25.

8.7.2 Maximizing 2DEG Conductivity

The 2DEG conductivity in MODFET structures σ is given by

$$\sigma = e\mu_n n_s \tag{8.7.1}$$

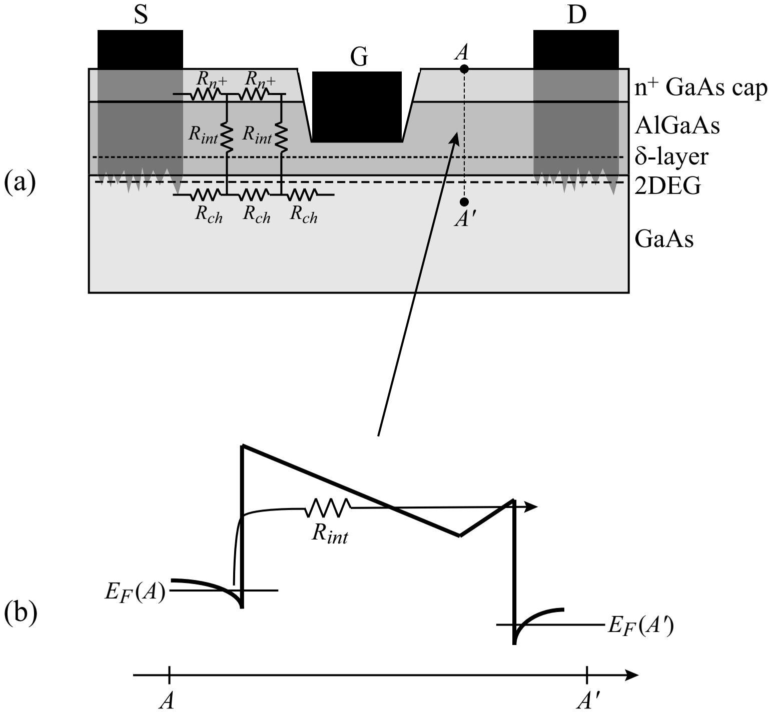


Figure 8.24: (a) Schematic of an AlGaAs/GaAs HFET with an n^+ cap layer and a recessed gate. Also shown are the various resistive components that make up the source access resistance. (b) Band diagram across the structure.

The 2DEG conductivity is at a maximum when the $\mu_n \cdot n_s$ product is maximized. From our discussion of the Lever Rule in section 8.5, it is clear that the 2DEG density n_s increases as the δ-doping sheet is brought closer to the heterointerface. However, decreasing the spacer distance d_s also causes the electron mobility μ_n to decrease because of the increase in remote ionized impurity scattering. It is therefore clear that the 2DEG conductivity will have a maximum at a value d_s(optimum) which must be determined for each material system and doping level. Typical values are 5 nm for the AlInAs/GaInAs system, 3 nm for the AlGaAs/InGaAs system, and 2 nm for the AlGaAs/GaAs system.

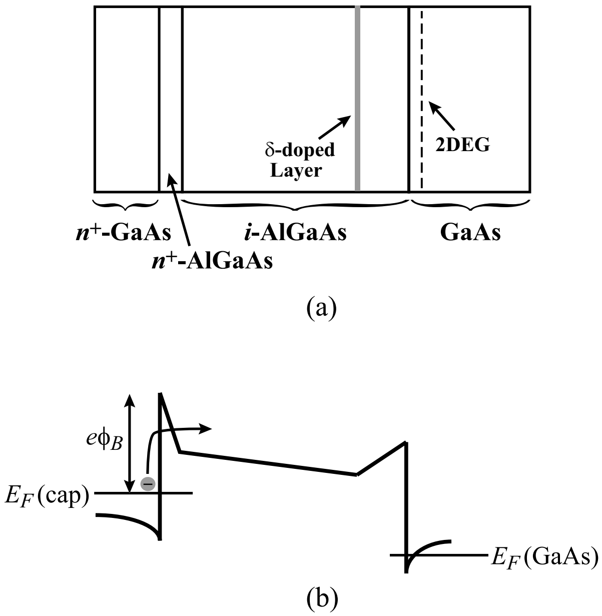


Figure 8.25: (a) Layer structure and (b) band diagram of an AlGaAs/GaAs HFET with a highly degenerate n^+ GaAs cap layer directly above a thin n^+ AlGaAs layer. The high doping in the GaAs cap layer reduces the barrier $e\phi_B$ that electrons must overcome, and the n^+ AlGaAs layer increases the probability of electrons tunneling through a portion of the barrier.

8.7.3 Back-barriers to Substrate Injection

Control of channel charge is the essence of FET operation. If electrons travel through the path labeled I_s in figure 8.26a, then they are effectively controlled by the gate. Electrons traveling along the path labeled I_{par} within the substrate and far from the gate are not effectively modulated and are parasitic currents leading to both reduced output resistance (hence low power gain) and low current gain. To keep electrons from being injected into the substrate, we need to present a barrier to current flow, as shown in figure 8.26b. This can be done by introducing a fully depleted p-type layer or a wider bandgap buffer. The band diagrams for each of these are shown in figure 8.26c and figure 8.26d.

The p-type buffer introduces negative space charge to the region immediately below the channel, thus increasing the electrostatic barrier to electron injection into the buffer by a maximum amount

$$\Delta\phi_b \simeq \frac{eN_a d_{bar}^2}{2\epsilon} \tag{8.7.2}$$

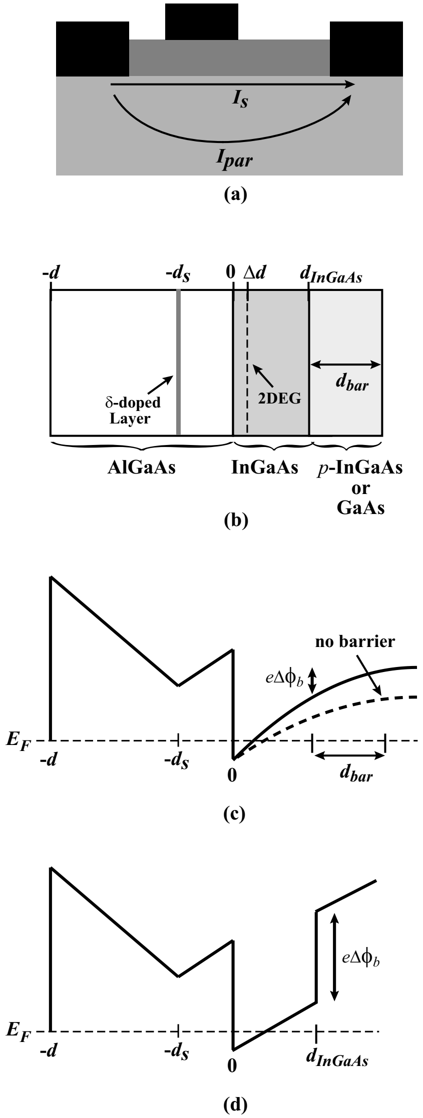


Figure 8.26: (a) In standard MODFET structures, a small parasitic leakage current I_{par} flows through the substrate. (b) MODFET structure which incorporates a back barrier to prevent current injection through the substrate. (c) Band diagram for a p-type barrier. (d) Band diagram for a barrier composed of a wide bandgap buffer.

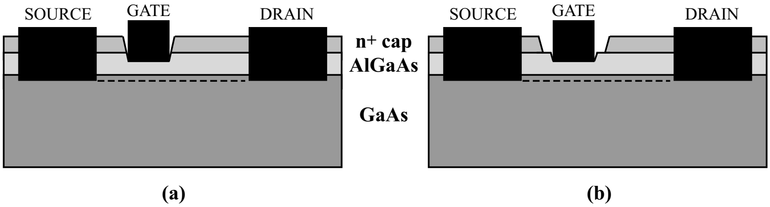


Figure 8.27: HFET structures with (a) a single gate recess and (b) a double recess.

The wide bandgap buffer approach provides a barrier increase of

$$\Delta\phi_b = \frac{\Delta E_C}{e} \tag{8.7.3}$$

The latter confinement scheme is common in AlGaAs/InGaAs/GaAs pseudomorphic HFETs and is the preferred design for GaAs-based HFET structures.

8.7.4 Gate Recess Design

Recessed gate structures are required when n^+ cap layers are employed, and they can also be designed to improve gate leakage and breakdown characteristics as well as to control the device threshold voltage. Designing the gate recess is one of the more important issues in HFET design. Recess structures can generally be placed into two categories: single recess (figure 8.27a) and double recess (figure 8.27b) structures.

The single recess is designed so that the recess and the gate metal are both defined through a single opening in the photoresist such that the recess width is approximately equal to the gate length L_g. The advantage of this process is that the source and drain access resistances are minimized, so the transit delay is determined dominantly by the gate length, as the high field region in the structure is effectively terminated by the source and drain cap layers. The major disadvantage of this scheme is that the lack of depletion field extension beyond the gate increases the electric field at the drain edge of the gate, thus increasing gate leakage and decreasing the breakdown voltage.

The double recess design, shown in figure 8.27b, allows one to trade off transit delay versus gate leakage and breakdown. By utilizing the first recess of length L_R to etch through the n^+ cap layer and the second recess to simultaneously define the gate length and threshold voltage.

Single recess structures are used for small signal analog applications such as low noise amplifiers and in digital circuits, whereas double recess designs are used in large signal analog applications such as power amplifiers.

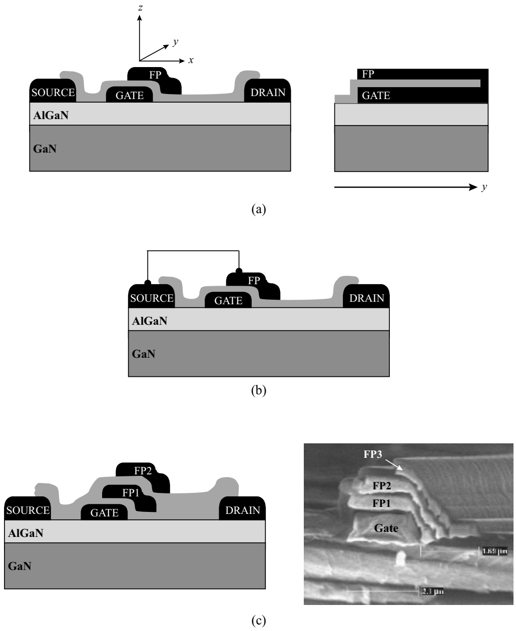


Figure 8.28: Various field plate configurations. (a) Gate-terminated field plate. (b)Source-terminated field plate. (c) Multiple field plate structure. SEM image courtesy of Y. Dora,UCSB.

8.7.5 Field Plates

One can actively control the gate extension beyond the drain edge of the gate and thereby reduce the peak electric field by using field plate structures. This is advantageous for applications such as high voltage switching and high power amplifiers, in which very high breakdown voltages are necessary. For this reason, field plates have become especially popular for HFETs in the GaN-based material system. There are a number of methods of implementing field plates, a few of which are shown in figure 8.28. One can have a dielectric-assisted extension of the gate toward the drain (i.e. a gate-terminated field plate). The gate extension effectively modulates the channel beyond the primary gate, thereby spreading the electric field between two peaks, one at the gate edge and the other at the edge of the termination, as shown in figure 8.32. The penalty for this approach is the enhanced gate-drain feedback capacitance C_{GD}.

Field shaping can also be achieved by utilizing a field plate connected to the source, as shown in figure 8.28b. Here, image charges on the plate result in an enhanced drain-source capacitance

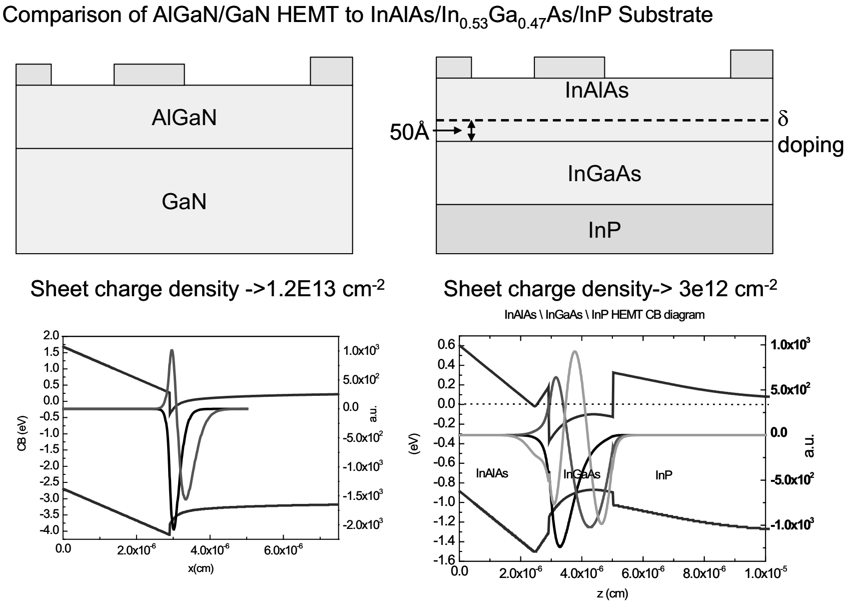


Figure 8.29: Schematic structure and band diagrams of AlInAs/GaInAs and AlGaN/GaN HFETs.

C_{DS} and an enhanced gate-source capacitance C_{GS}, but a reduced C_{GD} because of the screening of the gate from the drain by the source-connected field plate. One can effectively trade off the capacitances based on the geometry of the gate-connected and source-connected field plates, thus mapping out a design space of gain and breakdown voltage.

8.7.6 Comparison of two disparate material systems: AlInAs/GaInAs and AlGaN/GaN

It is instructive to compare the behavior of two families of HFET devices which could in some ways be considered to be at opposite ends of compound semiconductor space. One is the AlInAs/GaInAs/InP system, where the In composition in the GaInAs channel can be increased beyond the 53% value required to achieve lattice matching to come close to the 6.1 Å lattice constant of InAs. The bandgap of course decreases from 0.74 eV in the lattice matched case toward the bandgap of InAs ($\sim$ 0.36 eV). The mobility in the 2DEG can increase from 9,000 cm$^2/V$·s to over 15,000 cm$^2/V$·s. The effective mass of the electron decreases from $0.47m_0$ to $0.25m_0$.

At the other extreme is the AlGaN/GaN system, where E_G in the channel is 3.4 eV. and the effective mass is 0.2 m_0. Though the channel can be modified in a pseudomorphic fashion by adding In, the advantages are not obvious. Figure 8.29 shows the band structure of the AlGaN/GaN HFET and the AlInAs/GaInAs HFET. One feature to note is that the AlInAs/GaInAs HFET is modulation doped whereas the AlGaN/GaN HFET achieves its 2DEG as a result of polarization. The second is that the 2DEG concentration is only 3×10^{12} cm^{-2} in the AlInAs/GaInAs HFET, as opposed to 1.2×10^{13} cm^{-2} in the AlGaN/GaN HFET. The reason is that beyond an electron concentration of that order, the conduction band in the AlInAs touches the Fermi level, drastically reducing the modulation efficiency. The low scattering rates in GaInAs because of the small electron effective mass and the large separation between the Γ and L valleys results in large electron velocity overshoot in channels which are much smaller than the mean free path.

It is imperative to include velocity overshoot in calculating current-voltage ($I - V$) curves of InGaAs HEMTs where an average velocity of over $4{\times}10^7$ $\mathrm{cm/s}$ is easily attained for gate lengths of $0.1\mu\mathrm{m}$. In comparison, the large effective mass of electrons in GaN, the high phonon energy, and the strong coupling between electrons and phonons increases the scattering rate by over an order of magnitude compared to InGaAs (10^{13} s^{-1} vs. 10^{12} s^{-1} in bulk materials). Hence, the probability of overshoot is much lower in this case.

Figure 8.30 shows that the GaN HFET has a very small fraction of the 0.1 μm long channel exhibiting velocity overshoot, whereas the InGaAs HFET exhibits it over the full channel. Initial estimates suggest that velocity overshoot will become important at gate lengths of 20 nm or less in the GaN system. The difference in the non-stationary electron transport behavior is the primary reason why the InGaAs HFET shows excellent f_τ behavior with decreasing gate length, as shown in figure 8.30. The current state-of-the-art is an f_τ of over 560 GHz at a gate length of 30 nm. On the other hand, AlGaN/GaN HFETs have achieved an f_τ value of 163 GHz at 90 nm. The power performance of a state-of-the-art AlGaN/GaN HFET, which has high breakdown voltage because of the large E_g is shown in figure 8.31a.

8.7.7 Non-idealities in state-of-the-art transistors

The performance of state-of-the-art HEMTs is strongly affected by gate modulation efficiency, electron confinement in the channel and small signal access resistances. This section will show several examples of how these parameters affect the performance of the transistors. It will also describe some techniques that allow a higher performance by overcoming these limitations. We use AlGaN/GaN HEMTs as the vehicle for demonstration.

As shown in previous sections, a good aspect ratio between the gate length and the gate-to-channel distance is critical to obtain a good modulation of the channel electrons by the gate. This is especially important in high frequency devices where a poor gate modulation efficiency degrades f_τ. To illustrate this problem, figure 8.33 shows the f_τ of AlGaN/GaN HEMTs for different gate aspect ratios. There is a clear increase in f_τ and $f_{\max}$ as the aspect ratio increases.

However, a good aspect ratio is not enough to allow a good modulation of the channel electrons by the gate. A poor carrier confinement in the channel can also degrade the performance of transistors, even with good gate aspect ratios. Figure 8.35a shows the transconductance as a

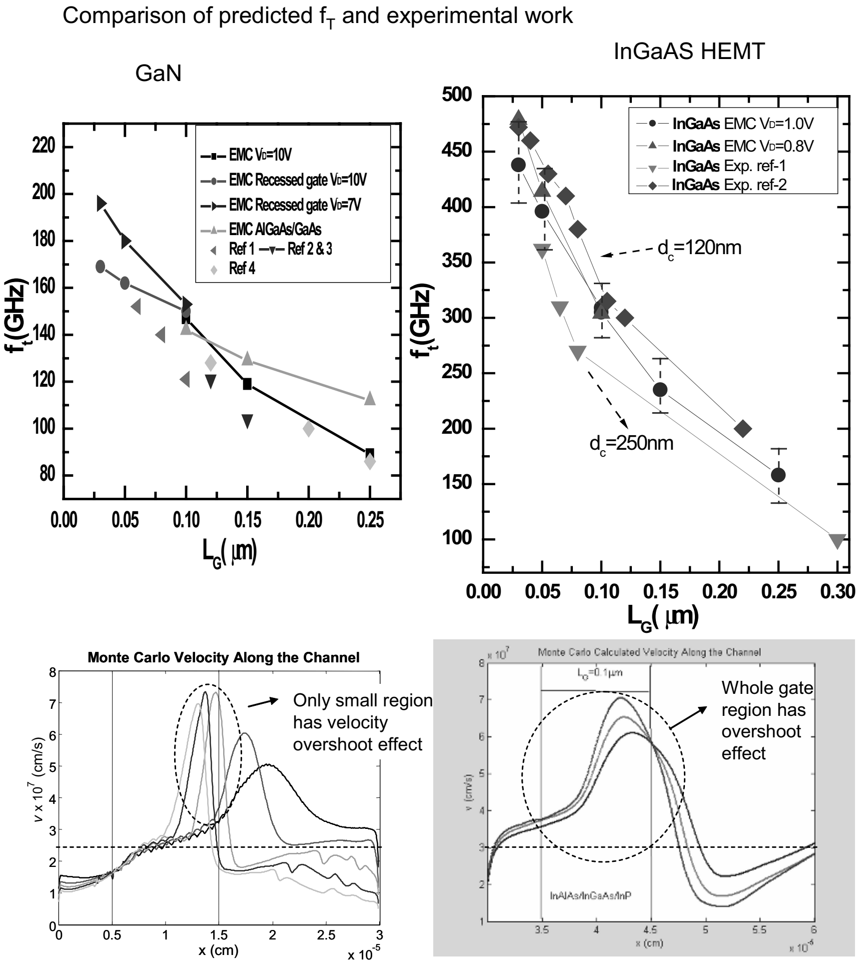


Figure 8.30: f_T vs. L_G and velocity field profiles along the channel for both GaN and InGaAs HFETs.

function of gate voltage for different drain voltages in an AlGaN/GaN HEMT. At low drain voltages, the gate can easily modulate the electrons in the channel and a good pinch-off voltage is obtained for a gate voltage of -5 V. However, as the drain voltage increases, the pinch-off degrades significantly, shifting to lower VGS voltages and becoming softer. These problems are the consequence of the poor electron confinement typical of single heterojunction devices

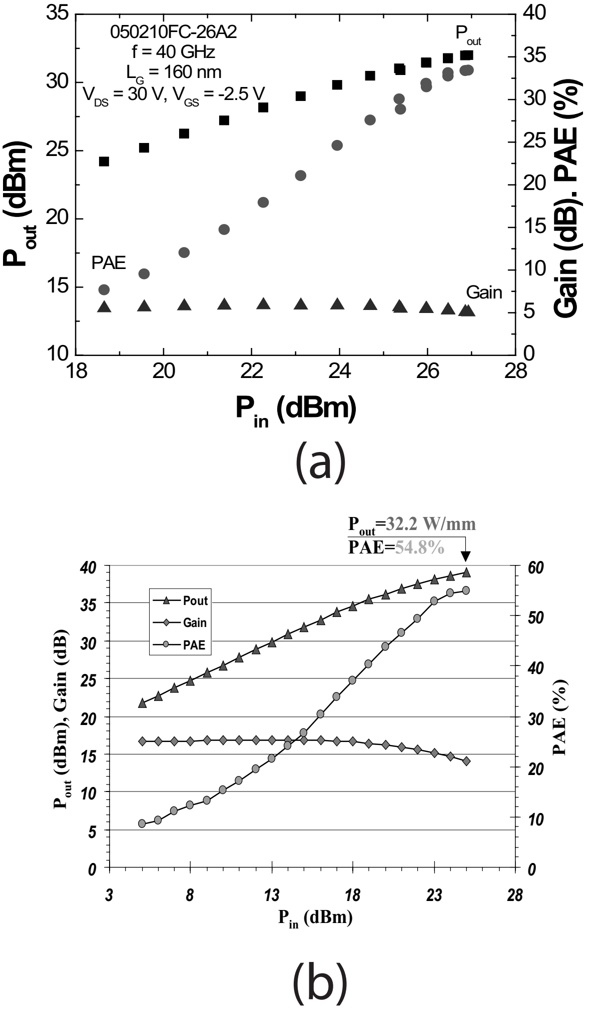


Figure 8.31: (a) Power performance of an AlGaN/GaN HFET at 40 GHz. The maximum power output of this device $P_{out} > 10.5$ W/mm with a PAE of 33%. Figure courtesy of T. Palacios, UCSB.(b) Record power densities have been achieved by employing field plates in AlGaN/GaN technology. Shown here are power measurements taken at 4 GHz of a 246 μm wide device biased at $V_{DS} = 120$ V. The maximum output power density $P_{out} = 32.2$ W/mm with a PAE of 54.8%. Figure courtesy of Y.-F. Wu, Cree Inc.

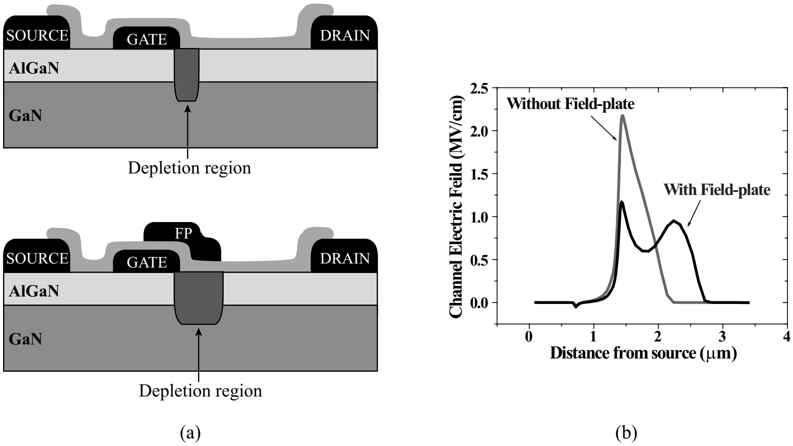


Figure 8.32: (a) Schematic diagrams of HFET structures with and without gate-terminated field plates. In the field plated device, the depletion region extends over a larger lateral distance. (b) Electric field profiles within the depletion region along the channel of both devices.

like the AlGaN/GaN HEMT. While there is a significant potential barrier at the AlGaN/GaN heterointerface, there is no barrier between the channel and the buffer. Therefore, it is easy for hot electrons to get injected into the buffer, which increases the gate to channel distance and degrades the performance.

Especially in high frequency devices with very short channels, it is important to increase the confinement of the channel by providing a potential barrier between the channel and the buffer. This barrier can be formed by the conduction band discontinuity between a wide bandgap semiconductor buffer and a narrow bandgap channel. This is the approached normally followed in AlGaAs/GaAs/AlGaAs transistors. The channel confinement can also be increased by doping the buffer $p-$type, which generates an electric field in the buffer in a direction that opposes the injection of hot electrons from the channel. An additional option in nitride-based devices is to use ultra thin InGaN backbarrier layers as shown in figure 8.36. In this device, the difference in the polarization coefficients between the GaN buffer and the InGaN backbarrier induces two sheets of fixed charge at the GaN/InGaN interfaces. These polarization induced charges generate an electric field in the InGaN layer which lowers the conduction band in the GaN channel with respect to the GaN buffer. This creates an effective conduction band discontinuity which provides a barrier for the flow of electrons into the buffer as shown in the band diagram in figure 8.36b. The improved confinement provided by the InGaN back barrier allows much better gate modulation at high drain voltages as shown in the transconductance measurements of figure 8.35b. In these improved devices, there is no degradation in the quality of the pinch-off as the drain voltage increases, although there is still a shift in the pinch-off voltage.

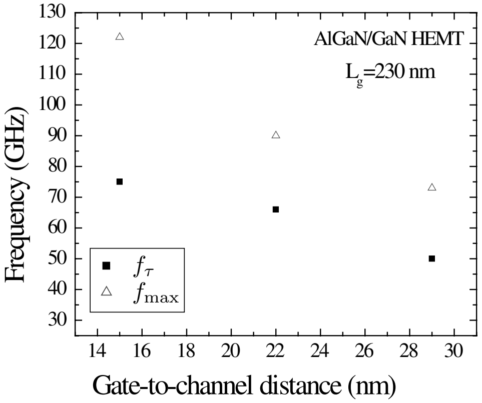


Figure 8.33: Effect of the gate-to-channel distance in the frequency performance of an AlGaN/GaN HEMT with a gate length of 230 nm.

The higher channel confinement of double heterojunction devices is also beneficial in increasing the output resistance of the transistors. Figure 8.34 compares the output resistance as a function of gate length for several standard AlGaN/GaN HEMTs and some AlGaN/GaN HEMTs with InGaN backbarrier. Almost a 50% increase in the output resistance can be measured in the devices with higher channel confinement. This increase in output resistance also causes an increase in the f_{max} of the devices, as predicted by equation 8.8.12, and shown in figure 8.37.

Another interesting non-ideality in the behavior of many transistors is the decrease of gm with drain current. From equation 8.8.2, the transconductance of a HEMT operating in the saturated mode should be independent of the drain current level. However, this is normally not the case. As shown in figure 8.38 for an AlGaN/GaN HEMT, g_m decreases as current increases once that the maximum g_m has been reached. This kind of behavior has been observed in many different transistor technologies, including Si MOSFETs, AlGaAs/GaAs MODFETs and AlGaN/GaN HEMTs. The cause of this decrease is different in each technology. In Si MOSFETs, simulations have related this decrease in performance with roughness at the Si/SiO2 interface. On the other hand, in AlGaAs/GaAs HEMTs, as the drain current increases, there is a reduction in the modulation efficiency of the gate due to the capture of channel electrons by the ionized donors in the AlGaAs barrier, which reduced g_m. Finally, in GaN technology, the reason for this decrease is related to the increase in small signal source access resistance due to a reduction in the electron mobility at higher electric fields in this material system . Other studies have also proposed the emission of hot phonons and the subsequent reduction of the electron velocity as a possible cause for this decrease in performance.

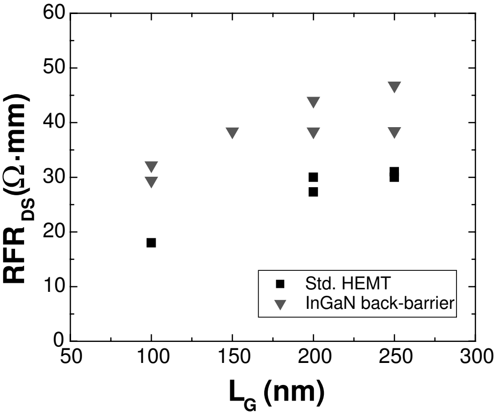


Figure 8.34: RF output resistance in standard AlGaN/GaN HEMTs and in HEMTs with enhanced confinement due to the use of an InGaN back-barrier.

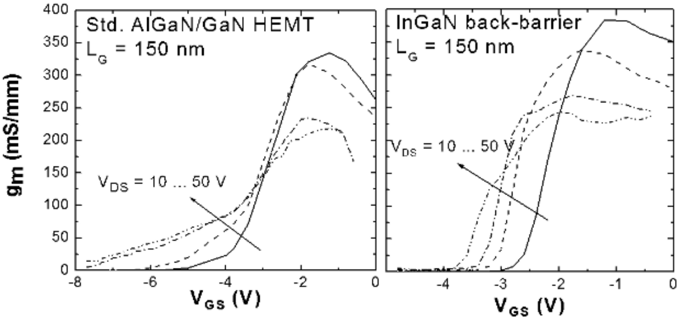


Figure 8.35: Change in g_m and pinch off with V_{DS} in a standard AlGaN/GaN HEMT and in a HEMT with an InGaN back-barrier.

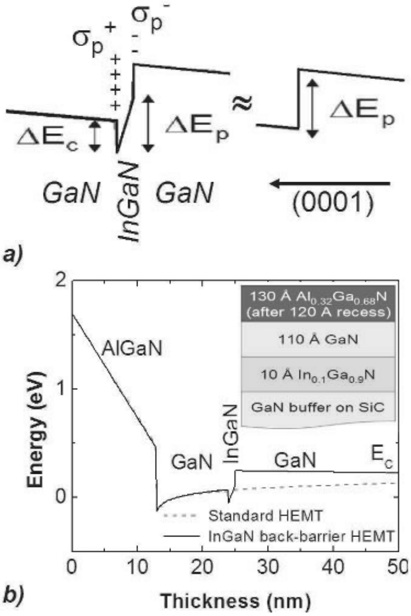


Figure 8.36: a) Effect of the insertion of an ultra-thin layer of InGaN in the conduction band diagram of a GaN buffer. Due to the extremely thin InGaN layer, the conduction band discontinuity, ΔE_c, of one side of the heterostructure is canceled by the ΔE_C in the other side and it can be neglected, resulting in an effective band discontinuity equal to ΔE_p. In the figure, the polarization-induced sheet charges at the heterointerfaces are also shown. b) Schematic and conduction band diagram of the basic InGaN back-barrier sample used in this work.

In conclusion, in this section we have reviewed several of the problems limiting the performance of real HEMTs as well as some of the solutions normally adopted to overcome them. Some solutions, like to keep a good gate aspect ratio, are common to every semiconductor family; others, like the use of InGaN back-barriers, are specific to some materials. Therefore, to fabricate high performance transistors is fundamental not only to understand the physics of the device but also to know the particularities of each material system, its limitations and advanced properties.

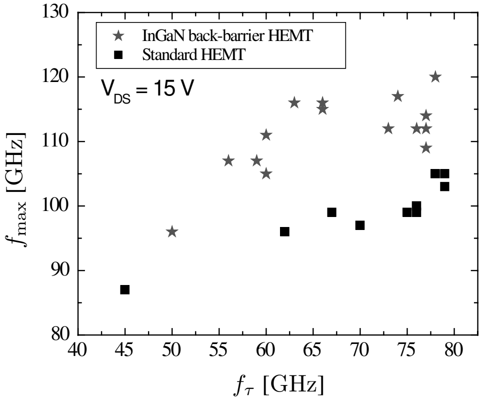


Figure 8.37: Effect of the InGaN back-barrier on the power gain of AlGaN/GaN HEMTs. Each data point represents a different transistor. The variation in f_T is due to different gate lengths, which vary from 0.1 to 0.4 μmfor the measured devices.

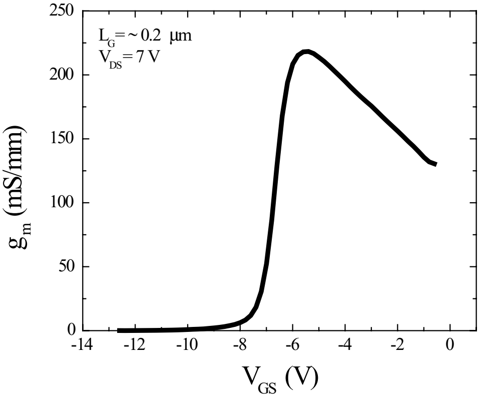


Figure 8.38: Decrease of $g_{\rm m}$ as the drain current increases (i.e. V_{GS} increases) in an AlGaN/GaN HEMT.

8.8 SMALL AND LARGE SIGNAL ISSUES AND FIGURES OF MERIT

It is important to understand the behavior of devices at higher frequencies both in small and large signal operation. The former refers to applications such as low noise amplifiers in receivers whereas the latter to applications such as power amplifiers used in transmitters.In this context several figures of merit have been defined to characterize device performance. It is important to recognize that frequencies of merit are in general a function of the application or equivalently a function of the input and output networks that the device is connected to. In the following sections we will study this in more detail and present in this introduction a short synopsis of the treatment. The most important figure of merit is the current gain cut-off frequency, f_T, which is proportional to the inverse of the electron transit time across the device. The output termination of the device when f_T is calculated is always an AC short circuit and hence reflects the device behavior independent of the circuit. The f_T is the primary indicator of the average electron velocity through the transistor and detailed analysis can extract electron velocity in regions of the transistor. The power gain cutoff frequency of the device $f_{\max}$ is evaluated with the output of the device presented with the complex conjugate of its output impedance to maximize power transfer. This again is predominantly dependent on the device as the termination is determined uniquely by the device characteristic. In other instances, like in large signal amplifiers driving 50 ohms, the load line is what determines the termination and hence another figure of merit, f_{lsg} , the large signal figure of merit is used. In the discussion of the bipolar device high frequency response we had to discuss minority carrier injection and removal. The FET is a majority carrier device. The device performance is essentially controlled by carrier transit time effects. Thus lithographic techniques defining the gate length and carrier mobility and velocity figure strongly in device response.

8.8.1 Small-Signal Characteristics

The equivalent circuit of a MESFET and the source of the various terms are shown in figure 8.39. A change of charge ΔQ on the gate produces the change ΔQ in the channel (assuming charge neutrality). If Δt is the time taken by the device to respond to this change, the change in the current in the channel is

$$\delta I_D = \frac{\delta Q}{\Delta t} \tag{8.8.1}$$

where I_D is the current flowing between the source and the drain. The time Δt can be interpreted as the average transit time t_{tr} for the electrons to move through the device. The transistor transconductance can be related to the transit time. The transistor intrinsic transconductance is

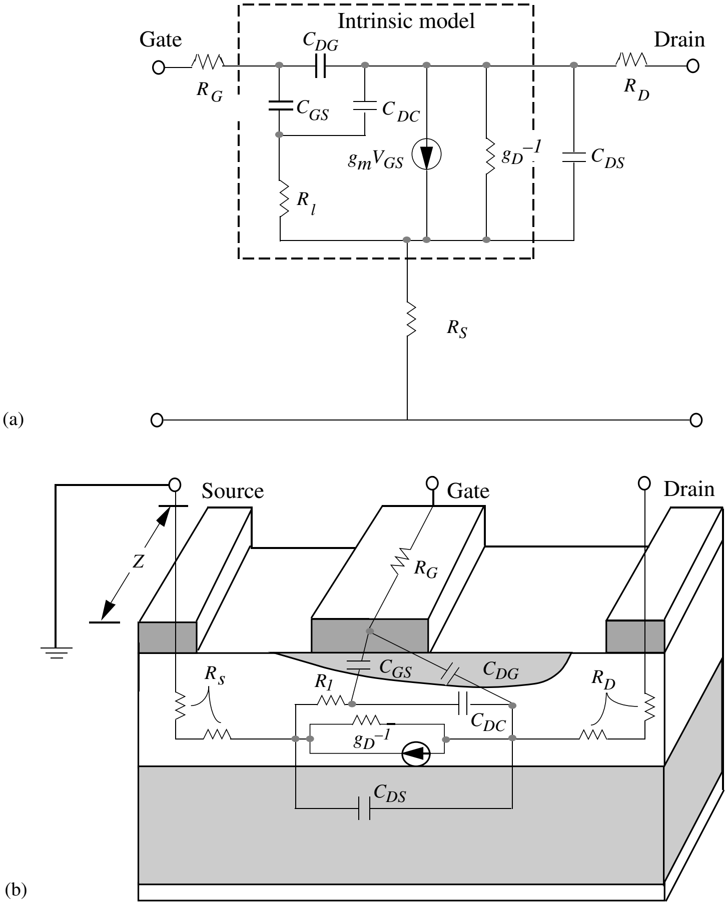


Figure 8.39: (a) Equivalent circuit of a MESFET. (b) Cross-section of a MESFET indicating the origins of the elements.

given by

$$\begin{aligned} g_m &= \left.\frac{\partial I_D}{\partial V_G}\right|_{V_D} \\ &= \left.\frac{\partial I_D}{\partial Q}\right|_{V_D} \left.\frac{\partial Q}{\partial V_G}\right|_{V_D} \\ &= \frac{C_G}{\Delta t} = \frac{C_G}{t_{tr}} \end{aligned} \tag{8.8.2}$$

where C_G is the gate-to-channel capacitance and describes the relationship between the gate voltage and the gate charge. The intrinsic transconductance is thus inversely proportional to the carrier transit time. The gate capacitance can be characterized by the gate-source capacitance C_{GS} and the gate-drain capacitance C_{DG} shown in figure 8.39b.

We can also define the output conductance g_D, which describes the effect of the drain bias on the drain current as

$$g_D = \left.\frac{\partial I_D}{\partial V_{DS}}\right|_{V_{GS}} \tag{8.8.3}$$

In addition to the intrinsic circuit elements discussed above, important extrinsic parasitic elements are the gate resistance R_G, the drain resistance R_D, and the source resistance R_S , which represents the series resistance of the ohmic contact and the channel region between the source and the gate. Also, we have the drain-to-substrate and drain-to-channel capacitances C_{DS} and C_{DC} respectively. These parameters lead to a simplified circuit model for the FET device shown in figure 8.39. This figure shows the equivalent circuit based on the physical origin of the circuit elements discussed above. An important characterization parameter is the forward current gain cutoff frequency f_τ, which is measured with the output short-circuited. The parameter f_τ defines the maximum frequency at which the current gain becomes unity. Figure 8.40 shows a simplified AC equivalent circuit where the input resistances are condensed into R_i and the output is represented by $R_{DS} = 1/g_D$.

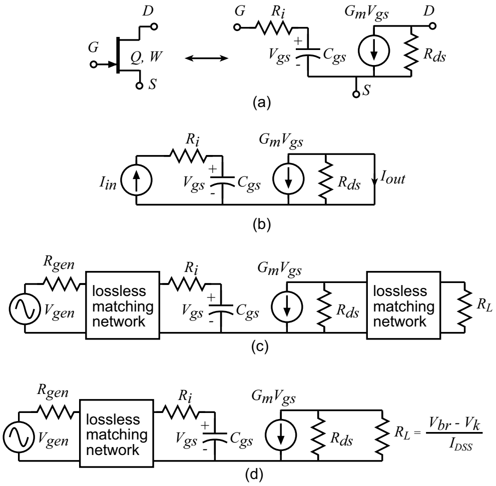


Figure 8.40: (a)Simplified A.C. π-model for the transistor used in this analysis, (b) definition of short circuit current gain and (c) definition of maximum available power gain (d) definition of large signal power gain for load line match.

If the capacitance charging time is the limiting factor, at the cutoff frequency the gate current I_{in} is equal to the magnitude of the output channel current $g_m V_{GS}$. The input current is the current due to the gate capacitor, and for a small-signal sinusoidal signal we have

$$I_{in} = j\omega C_G V_{GS} \tag{8.8.4}$$

Equating this to $g_m V_{GS}$ at $\omega = 2\pi f_\tau$, we get for the cutoff frequency

$$\boxed{f_\tau = \frac{g_m}{2\pi C_G} = \frac{1}{2\pi t_{tr}}} \tag{8.8.5}$$

where t_{tr} represents the transit time of the electrons through the channel. The frequency response is therefore improved by using materials with better transport properties and shorter channel lengths. If we assume that the carriers are moving at a saturated velocity, the transit time t_{tr} is simply

$$t_{tr} = \Delta t = \frac{L}{v_s} \tag{8.8.6}$$

and the cutoff frequency becomes

$$\boxed{f_\tau = \frac{v_s}{2\pi L}} \tag{8.8.7}$$

It may be noted that the source resistance R_S has an effect of reducing the effective transconductance of the device. In the presence of a source resistance the gate bias is $V_{GS}^{'}$ since a part of the input voltage drops across the resistance R_S. The drain current is

$$I_D = g_m V_{GS}^{'} \tag{8.8.8}$$

Also we have

$$V_{GS} = V_{GS}^{'} + \left(g_m V_{GS}^{'}\right) R_S = (1 + g_m R_S) V_{GS}^{'} \tag{8.8.9}$$

The drain current now becomes

$$I_D = \frac{g_m V_{GS}}{1 + g_\mathrm{m} R_S} = g_m^{'} V_{GS} \tag{8.8.10}$$

where $g_m^{'}$ is the extrinsic transconductance and is smaller than g_m.

The transistor provides the maximum power gain when both the input and output are conjugate matched to the generator and load impedance respectively figure 8.40(c)). This maximum available power gain (MAG) is given by,

$$MAG = \frac{P_{load}}{P_{av,gen}} = \frac{G_m^2 R_{ds}}{16\pi^2 f^2 C_{gs}^2 R_i} \equiv \left(\frac{f_\mathrm{max}}{f}\right)^2 , \tag{8.8.11}$$

where

$$f_\mathrm{max} = \frac{f_\tau}{2\sqrt{R_i/R_{ds}}} , \tag{8.8.12}$$

is the frequency at which the power gain becomes unity, also called the power gain cut-off frequency. In power amplifiers a load line match is usually provided at the output (equation 8.8.19) as in figure 8.40d, rather than a match for the maximum power gain as in figure 8.40c. The large signal power gain (LSG) is then given by (for the case $R_L \gg R_{ds}$),

$$LSG = \frac{P_{load}}{P_{av,gen}} = \left(\frac{V_{br} - V_k}{V_p}\right) \frac{G_m}{4\pi^2 f^2 C_{gs}^2 R_i} \equiv \left(\frac{f_{\text{lsg}}}{f}\right)^2 \tag{8.8.13}$$

where

$$f_{\text{lsg}} = \sqrt{\frac{V_{br} - V_k}{I_{DSS}}} \frac{f_\tau}{\sqrt{R_i}} \tag{8.8.14}$$

Here the large signal power gain cut-off frequency (f_{lsg}) is the frequency at which the power gain becomes unity for a load line match.

With the transistor parameters scaling with the device periphery (W) as $I_{DSS} \propto W$, $C_{gs} \propto W$, $G_m \propto W$, $R_i \propto 1/W$ and $R_{ds} \propto 1/W$, f_τ, $f_{\max}$ and f_{lsg} are independent of the device periphery.

8.8.2 Power-frequency limit

An important limitation called the power-frequency (pf^2) limit relates to the inherent limit on the breakdown voltage a high frequency device technology can achieve. This limits the output power one can obtain from a given device technology. The pf^2 limit, well-known in microwave power transistor design, imposes particularly severe performance limits on broadband microwave power amplifiers.

In high frequency transistors, whether HEMT or HBT, there is a high-field drift region separating the control region (the HEMT channel, the HBT base) from the output terminal. In HEMTs it is the extension of the gate depletion region laterally toward the drain contact, while in a HBT this drift region is the collector depletion layer. If the length of this region is D_{drift}, and the semiconductor breakdown electric field is $\mathcal{E}_{max}$, then the transistor breakdown voltage is,

$$V_{br} = \mathcal{E}_{max} D_{drift} \tag{8.8.15}$$

This drift layer introduces space-charge transit time, τ_{sct}. If the electron velocity is v_{sat}, then the space charge transit time

$$\tau_{\text{sct}} = \frac{D_{drift}}{2v_{sat}} \tag{8.8.16}$$

and (ignoring all other transit delays) the unity current-gain cutoff frequency is

$$f_\tau \leq \frac{v_{sat}}{\pi D_{drift}} \tag{8.8.17}$$

Combining equation 8.8.15 and equation 8.8.17, we get

$$f_\tau V_{br} \leq \frac{\mathcal{E}_{max} v_{sat}}{\pi} \tag{8.8.18}$$

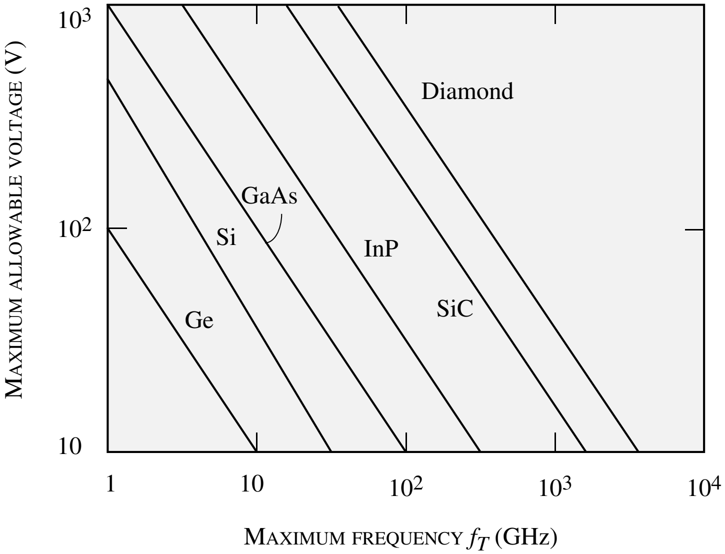


Figure 8.41: Plot of the maximum operating voltage for transistors made of selected semiconductors as a function of estimated f_T. The f_T estimates are based on the steady-state velocity-field curve for each material. (After M. W. Geis, N. N. Efremow, and D. D. Rathman, "Summary Abstract: Device Applications of Diamond," J. Vac. Sci. Technol. A6, 1953 (1988).)

which is purely dependent on the material parameters. So, the transistor f_T and V_{br} have to be traded against each other, with extended drift regions giving high breakdown voltages but low f_T and thin drift regions giving low breakdown voltages but high f_T.

8.8.3 Classes of operation of transistor power amplifiers and necessary device characteristics

The configuration of the uses of transistor amplifiers in say transmitter(power) applications are called classes and determined by one or more of the following criterion:

1. where the device is biased

2. what load-line the device sees

3. whether the active device is operated as an amplifier or a switch

Though the number of classes in existence is far too many to be described here in detail, we will briefly describe the class-A, and class-AB/B operations and highlight their performance with

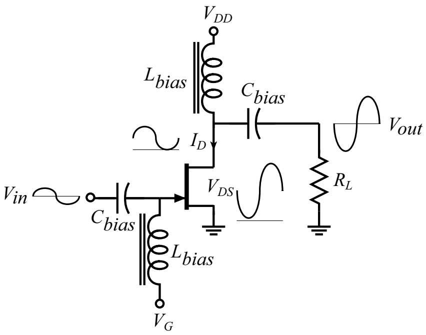


Figure 8.42: Circuit schematic of a simple class-A power amplifier.

respect to efficiency and bandwidth and associated device requirements . We will conclude the section by referring to power amplifiers where the devices is used as a switch. These classes offer the highest efficiency of operation but are the most stringent on device requirements .

Class-A : Least restrictive on device characteristics

Figure 8.42 shows the circuit schematic of a simple class-A power amplifier. This class of amplifiers is used for highly linear applications and can be used for both narrow and large bandwidth applications. For narrow band applications, a tuning network might be added at the output to terminate the harmonics created due to the variation in device transconductance. In this class of power amplifiers the device is biased normally-on, at about half the peak-peak output current and half the peak-peak output voltage (figure 8.43).

The load-line in class-A operation is linear at low frequencies and primarily determined by the load resistance (R_L). To obtain the maximum power from the device, the load-line is chosen so that the device operates between the maximum allowed drain to source voltage (the breakdown voltage, V_{br}) at one extreme and the maximum allowed drain current (the saturation current, I_{DSS}) at the other extreme. This requires that the optimum load resistance $R_{L,opt}$ be,

$$R_{L,opt} = \frac{(V_{br} - V_k)}{I_{DSS}} \tag{8.8.19}$$

This ensures that device provides the maximum output power obtainable, given by

$$P_{out,max} = \frac{1}{2} \cdot \frac{1}{2} (V_{br} - V_k) \cdot \frac{1}{2} I_{DSS}$$

where the first term of $1/2$ comes from time averaging. Therefore,

$$P_{out,max} \leq \frac{(V_{br} - V_k) I_{DSS}}{8} \equiv \frac{(V_{br} - V_k)^2}{8 R_{L,opt}} \tag{8.8.20}$$

Choosing this load-line minimizes the total device periphery (and hence the die area) required for a given RF output power. This also provides the best bandwidth. Larger device periphery results in larger device input and output capacitances which degrade the bandwidth. Oversizing is done if the device on-resistance in the linear region (R_{on}) is large (i.e. V_k/V_{br} is large). Then by operating at $I_{d,max} < I_{DSS}$, the $I_d^2 R_{on}$ losses are reduced and the efficiency is improved. However this is achieved at the cost of reduced bandwidth.

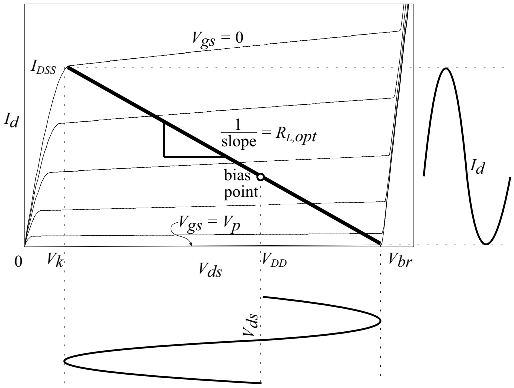


Figure 8.43: Optimum load-line for class-A operation

The maximum output power obtainable is roughly half the DC power, which means that the theoretical maximum drain efficiency (ratio of the output RF power to the DC power) is 50%. Since the device is normally on, a constant DC power of nearly twice the peak RF output power is dissipated at all times. This might degrade the performance of high power amplifiers with time. But the advantages of class-A operation include broadband operation and high linearity.

Reflecting on the discussion above the following device requirements can be extracted. To minimize the losses in the linear region or on-state of the device the channel conductivity or the product of sheet charge and electron mobility should be maximized. This allows the device periphery for a certain value of allowed on resistance to be minimized which in turn reduces the device capacitances and hence reduces the amount of circuit inductance required to tune the device. This device requirement is applicable to all classes of operation, a universal requirement. It is intuitive clear that the presence of large tuning elements result in LC networks which are inherently narrow band centered around their resonance frequency and are undesirable in broadband applications. The output power is a function a the product of available current and voltage. The current is typically proportional to the channel conductivity and electron velocity. This is compatible to the requirement of low on resistance but typically materials that have high mobility and electron velocity have low bandgaps such as, Si, GaInAs and InAs. The one remarkable exception is GaN which has a large bandgap, high electron mobility, and high electron velocity enabling large currents and large voltages simultaneously and is hence the subject of intensive research. Another extremely important figure of merit for power transistors after output power capability is efficiency. Why is this important? Power amplifiers transmit power. Depending on the application these powers can be large. 120W per amplifier and over 1kW for a base station is typical for cellular phone applications. The requirements for RADAR are even larger. Imagine an amplifier that has an efficiency of say 50% (the best one can do in class A operation). Then approximately 1 kW is wasted as heat for 1 kW of transmitted power. This is not only is wasteful but poses a severe challenge in the packaging of the device as the heat has to be removed from the chip. If the temperature of the chip rises then the mobility of the materials drop as discussed in chapter 3 and the resistances rise which in turn heat up the chip even more. To prevent this from resulting in a catastrophic failure of the device, adequate thermal management (cooling) has to be in place. The average efficiency of an amplifier operating under GSM modulation schemes (a popular scheme in wireless transmission the world over) is closer to 18%. It should not be a surprse to the reader that forced air cooling is required for many base stations. The efficiency of amplifiers is therefore becoming as important if not more than the requirement for high power. There are two definitions of efficiency; the Drain Efficiency (DE) and the power added efficiency (PAE) and are explained below.

1. Drain efficiency ($DE\%$) or D.C. to RF conversion efficiency is defined as the ratio of R.F. output power (P_{out}) to the D.C. power drawn from the drain supply ($P_{DC,D}$) expressed as a percentage ;

$$DE\% = \frac{P_{out}}{P_{DC,D}} \cdot 100\% \; . \tag{8.8.21}$$

   Drain efficiency represents what fraction of the D.C. power is converted into R.F. output power.

2. Power Added Efficiency ($PAE\%$) is defined as the ratio of the difference in the R.F. output to input power, to the total D.C. power drawn from all bias supplies (P_{DC}).

$$PAE\% = \frac{P_{out} - P_{in}}{P_{DC}} \cdot 100\% \equiv \frac{P_{out}}{P_{DC}} \left(1 - \frac{1}{G}\right) \cdot 100\% \; , \tag{8.8.22}$$

   where P_{in} is the RF input power, and G is the power gain. As PAE also accounts for the input R.F. drive power required for the amplifier it is representative of how the power amplifier output stage is going to impact the overall system efficiency.

The PAE is the more important of these figures because it includes the amount of input power required to achieve the desirable output power. In many instances the amount of gain and output power simultaneously required for the system may not be achievable in a single amplifier stage. An example is the emerging need for mm-wave imaging (cameras operating at 94 GHz). Here gain of over 30 dB is required which requires multi-stage amplifiers. In this instance a high input drive (low power gain) would imply that the efficiency of the driver stage is also going to significantly affect the overall efficiency. So at least a power gain of 10 (10 dB) is required to obtain high PAE (say up to 45% for the class-A case). This ensures that the system efficiency is primarily determined by the efficiency of the output stage and the driver stages do not significantly affect the overall efficiency.

Class-AB and B: Requires devices with excellent pinch-off characteristics and preferably complementary devices for push-pull architectures

In class-AB (class-B) amplifiers the device is biased close to pinch-off (at pinch-off) so that the device operates as a amplifier for half the cycle and remains cut-off for the other half of the cycle. In tuned class-AB/B amplifiers sinusoidal output swings are obtained using a resonator at the output (figure 8.44) tuned to the fundamental frequency. The drain voltage and current waveforms are sinusoidal and half-sinusoidal respectively and the drain is biased at roughly half the peak-peak RF output voltage swing (figure 8.45). Under these conditions the LC resonant circuit is charged during the conducting portion of the device cycle and discharges into the load when the device is off providing the sinusoidal outputs desired. It is also apparent from figure 8.45 that for the same device periphery as the class-A case the optimum load is now reduced by a factor of 2, but the net fundamental output power is the same. Since the device is off when the voltage across it is high, lower D.C. consumption and hence higher efficiency up to 78.6% is expected. However, this configuration is inherently narrow-band because the series (parallel) resonator that is designed to be a short (open) at a fundamental frequency acts close to a open (short) at the second harmonic, and so even bandwidths of 2:1 are hard to realize. This bandwidth limitation is mitigated by employing a push-pull architecture where sinusoidal output swings are obtained using two devices, each operating for half the cycle and combining the output currents.

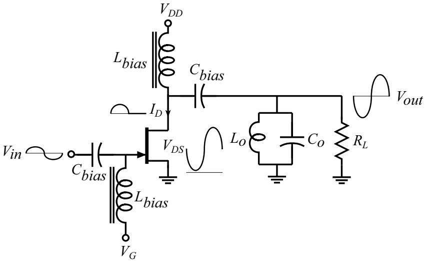


Figure 8.44: Circuit schematic of a simple class-B tuned power amplifier

In push-pull class-AB/B configuration. (Figure 8.46) This configuration could be made relatively broadband but requires broadband transformers or complementary devices. The most prevalent of all complementary devices is the Si-CMOS structure which is the dominant technology in the world today. The importance of having complementary devices is apparent by comparing Figure 8.46a and b, where figure 8.46a shows the case of a technology that does not have a complementary architecture and figure 8.46b the Si CMOS case. The complexity and the size penalty in the former is obvious as one has to generate out of phase signals at the input using transformer and add the outputs also using transformers. As mentioned before these transformers have typically narrow band and hence limit bandwidth. As in the tuned class-B case, higher efficiency up to 78.6% is obtained due to reduced D.C. consumption. The device requirements

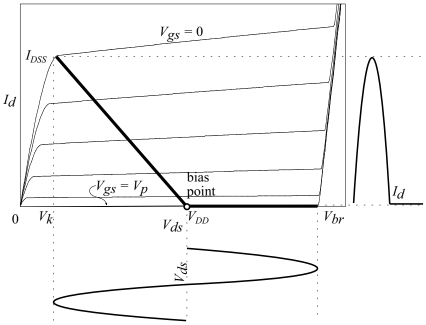


Figure 8.45: Device bias point and load-line for class-B operation.

for high performance Class B amplifiers is high gain near pinch-off conditions as this is the bias condition of the device is biased near pinch-off. this requires that the device leakage near pinch-off be minimal. Devices that have quantum well channels, p-type buffer layers, wide bandgap buffer layers or other means of enhanced charge control are required.

Higher Classes of Operation; Most stringent on device requirements

Class-C amplifiers are similar to the class-B tuned power amplifier with the device biased deep into cut-off region, so that the conduction angle (the fraction of the cycle that the device is on) for a sinusoidal waveform is less than 180°. Higher efficiencies are obtained by lowering the conduction angle, up to a theoretical limit of 100% with 0° conduction angle. Class-D,E are switched mode power amplifiers, where the device is operated as a switch. The load networks are chosen to minimize the current and voltage waveform overlap across the device, resulting in higher efficiency. But, again as in the case of class-B tuned power amplifiers, these classes as well as other classes like F, G, H etc. use a resonator at the output to obtain the fundamental power and are of no significance in broadband amplifiers. The requirements for devices operating as switches are the most extreme because since the waveforms are nearly square waves, the device has to have have gain at least to the third harmonic of the fundamental to be able to reconstruct the waveforms with reasonable accuracy. In addition the requirements of high gain near pinch-off, low leakage and high subthreshold slope, similar to to the case of Class B amplifiers are still desirable.

8.9 Implications on device technology and circuits

Having understood the limitations on decade bandwidth high power amplifiers, the following implications on the choice of circuits and device technology are apparent:

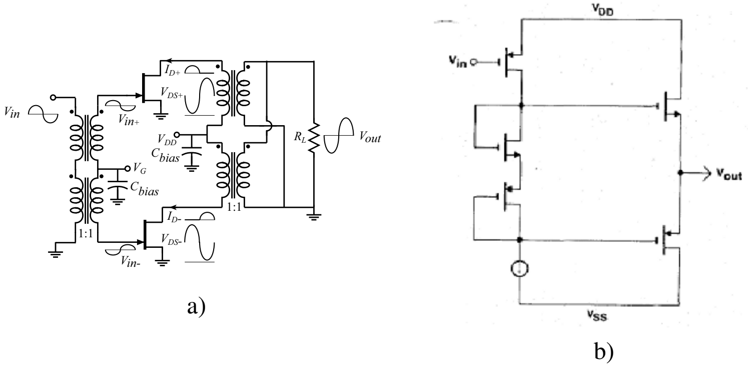


Figure 8.46: Push-pull amplifiers realized using (a) non-complementary devices, and b) complementary (CMOS) devices

Table 8.2: Typical power obtainable from various device technologies driving a $Z_o = 50\Omega$ load.

device technology	typical V_{br} (V)	typical V_k (V)	typical I_{DSS} (mA/mm)	device periphery (mm)	typical $P_{out,max}$ (W)
GaAs MESFET	20	1	300	600	200
§InP PHEMT	12	1	500	0.45	0.3
†GaN HEMT	150	5	1000	57.6	500

§ HRL, † CREE

1. Class-A mode of operation is desired when requirements of linearity and bandwidth have to be simultaneously satisfied
2. Push-pull class - AB/B operation is attractive if complementary devices. are available such as CMOS
3. designs must be for at least 10 dB gain to ensure high PAE.
4. circuits must use a device technology with high $f_\tau V_{br}$ product.

8.10 PROBLEMS

- **Section 8.2**

Problem 8.1 Discuss the reasons why one needs a large Schottky barrier value for the gate in a MESFET.

Problem 8.2 By drawing the band profile of a MESFET, discuss the restrictions on the gate bias values that can be allowed.

Problem 8.3 Consider an n-channel Si JFET at 300 K with the following parameters:

$$\begin{array}{lcl} p^+\text{-doping}, & N_a & = & 5 \times 10^{18}\ \text{cm}^{-3} \\ n\text{-doping}, & N_d & = & 10^{17}\ \text{cm}^{-3} \\ \text{Channel thickness}, & h & = & 0.5\ \mu\text{m} \end{array}$$

(a) Calculate the internal pinch-off for the device. (b) Calculate the gate bias required to make the width of the undepleted channel 0.25 μm.

Problem 8.4 Consider a GaAs JFET with the same characteristics as those of the Si device in problem 8.3. Repeat the calculations for this GaAs device.

Problem 8.5 An n-type In$_{0.53}$Ga$_{0.47}$As epitaxial layer doped at 10^{16} cm^{-3} is to be used as a channel in a FET. A decision is to be made whether the JFET or MESFET technology is to be used for the device. In the JFET technology a p^+ region can be made with a doping of 5 × 10^{17} cm^{-3}. In the MESFET technology a Schottky barrier with a height of 0.4 V is available. Which technology will you use? Give reasons considering gate isolation issues. ($R^* = 5$ Acm$^{-2}K^{-2}$; $D_p = 20$ cm^2/s; $D_n = 50$ cm^2/s; $L_n = 5$ μm; $L_p = 5$ μm.)

Problem 8.6 Consider a p-channel Si JFET with the following parameters:

$$\begin{array}{lcl} p\text{-doping}, & N_a & = & 5 \times 10^{16}\ \text{cm}^{-3} \\ n^+\text{-doping}, & N_d & = & 5 \times 10^{18}\ \text{cm}^{-3} \\ \text{Channel depth}, & h & = & 0.25\ \mu\text{m} \end{array}$$

(a) Calculate the internal pinch off for the device as well as the gate bias needed for pinch off.
(b) Calculate the width of the undepleted channel for gate biases of V_{GS} = 1 V and V_{GS} = 2 V for V_{DS} = 0.

Problem 8.7 Design an AlInAs/GaInAs HEMT for maximum g_m such that charge in the channel is 3 × $10^{12}cm^{-2}$. Assume the doping in the AlInAs is 5 × $10^{12}cm^{-2}$. Assume the surface barrier is 0.8 V and $\Delta E_C = 0.55eV$. Also, assume that the substrate is GaInAs and is doped p-type such that $E_F = E_V$ in the substrate. Assume the buffer is $0.5\mu m$ thick. Also assume the minimum spacer allowed is 2 nm. What is the current available from the device. At zero gate bias assuimng a gate length of $1\mu m$. The velocity-field wave is shown in figure 8.48.

- **Section 8.3**

Problem 8.8 Consider an n-channel GaAs MESFET with the following parameters:

$$\begin{array}{lcl} \text{Schottky barrier height}, & \phi_b & = & 0.8\ \text{V} \\ \text{Channel doping}, & N_d & = & 5 \times 10^{16}\ \text{cm}^{-3} \\ \text{Channel width}, & h & = & 0.8\ \mu\text{m} \end{array}$$

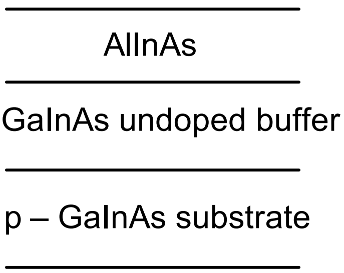


Figure 8.47: Figure for problem 8.7.

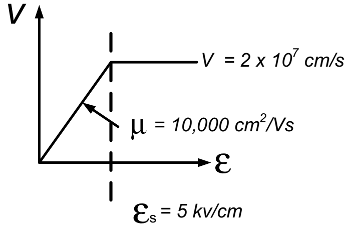


Figure 8.48: Figure for problem 8.7.

Calculate the minimum width of the undepleted channel (near the drain side) with V_{GS} = 0.5 V when (a) V_{DS} = 0.0 V; (b) V_{DS} = 1.0 V; (c) V_{DS} = 2.0 V; (d) V_{DS} = 10 V.

Problem 8.9 (a) An n-type GaAs MESFET is to be designed so that the device is just turned off at a gate voltage of V_{GS} = 0 V. The Schottky barrier height ϕ_b is 0.8 V and the channel thickness is 0.2 μm. Calculate the channel doping required. To calculate the depletion region thickness (only) you may assume that $V_{bi} \cong \phi_b$ - 0.1 V.
(b) If a gate bias of 0.2 V is applied, calculate the gate current.
(c) What is the saturation drain current when the gate bias is 0.2 V? Compare the gate current with the drain current.

Mobility,	μ_n	$=$	$5000\ \mathrm{cm}^2/\ \mathrm{V}\cdot\mathrm{s}$
Gate length,	L	$=$	$2.0\ \mu\mathrm{m}$
Gate width,	Z	$=$	$20.0\ \mu\mathrm{m}$
Channel width,	h	$=$	$0.2\ \mu\mathrm{m}$

Problem 8.10 In the text we used the constant-mobility model to obtain the relation between the drain current and the gate and drain voltages below pinch-off. Obtain a result for the depletion region $h - h(x)$ as a function of x (distance from source to drain) for a gate bias V_{GS} and a drain bias V_{DS}. Use the symbols used in the text for other device parameters.

Problem 8.11 Consider the design of two n-channel GaAs MESFETs with the following parameters:

$$\begin{array}{llcll} \text{Schottky barrier height,} & \phi_b & = & 0.8\text{ V}; & 0.8\text{ V} \\ \text{Channel doping,} & N_d & = & 2\times 10^{16}\text{ cm}^{-3}; & 2\times 10^{17}\text{ cm}^{-3} \end{array}$$

The first sequence belongs to one device and the second sequence to the other. Calculate the depths of the channel needed for each device so that the devices are just turned off in the absence of any gate bias.

Problem 8.12 Consider an n-channel GaAs MESFET at 300 K with the following parameters:

$$\begin{array}{llcl} \text{Schottky barrier height,} & \phi_b & = & 0.8\text{ V} \\ \text{Channel thickness,} & h & = & 0.25\ \mu\text{m} \end{array}$$

Calculate the channel doping needed so that the device turns off at a gate bias of V_{GS} = V_T = 0.5 V.

Problem 8.13 Consider an n-channel Si MESFET at 300 K with the following known parameters:

$$\begin{array}{llcl} \text{Barrier height,} & \phi_b & = & 0.7\text{ V} \\ \text{Channel doping,} & N_d & = & 10^{16}\text{ cm}^{-3} \end{array}$$

It is found that when a gate bias of V_{GS} = -0.3 V is applied (V_{DS} = 0), the channel is just fully depleted. Calculate the channel depth h for the device.

Problem 8.14 Consider a GaAs n-channel MESFET at 300 K with the following parameters:

$$\begin{array}{llcl} \text{Schottky barrier height,} & \phi_b & = & 0.8\text{ V} \\ \text{Electron mobility,} & \mu_n & = & 6000\text{ cm}^2/\text{ V}\cdot\text{s} \\ \text{Channel width,} & Z & = & 25\ \mu\text{m} \\ \text{Channel length,} & L & = & 1.0\ \mu\text{m} \\ \text{Channel depth,} & h & = & 0.25\ \mu\text{m} \\ \text{Channel doping,} & N_d & = & 1.0\times 10^{17}\text{ cm}^{-3} \end{array}$$

(a) Calculate the gate bias $V_{GS} = V_T$ needed for the device to just turn off.
(b) Calculate $V_D(sat)$ for gate biases of V_{GS} = -1.5 V and V_{GS} = -3.0 V.
(c) Calculate the saturation drain current for the cases considered in part b.

Problem 8.15 Consider an n-channel GaAs MESFET at 300 K with the parameters of problem 8.14. Calculate the transconductance of the device in the saturation region for the gate biases V_{GS} = -1.5 V and V_{GS} = -2.0 V. Express the results in terms of mS/mm.

Problem 8.16 Consider an n-channel Si MESFET at 300 K. The following parameters define the MESFET:

Schottky barrier height,	ϕ_b	$=$	0.8 V
Channel mobility,	μ_n	$=$	$1000\ \mathrm{cm}^2/\ \mathrm{V}\cdot\mathrm{s}$
Channel doping,	N_d	$=$	$5\times10^{16}\ \mathrm{cm}^{-3}$
Channel length,	L	$=$	$1.5\ \mu\mathrm{m}$
Channel depth,	h	$=$	$0.25\ \mu\mathrm{m}$
Gate width,	Z	$=$	$25\ \mu\mathrm{m}$

(a) Calculate the turn-on voltage V_T for the structure.
(b) Calculate $V_{DS}(sat)$ at a gate bias of V_{GS} = 0. Also calculate the device transconductance.
(c) If the device turn-on voltage is to be V_T = -2.0 V, calculate the additional doping needed for the channel.

Problem 8.17 In a MESFET, as the gate length shrinks, the channel doping has to be increased. Discuss the reasons for this.

Problem 8.18 Derive and plot the I-V curves for a GaAs MESFET with $N_D = 5\times10^{17}cm^{-3}$, and a channel thickness of 50 nm. Assume a two-region mobility model, with a saturated velocity $v_{sat} = 2\times10^7\frac{cm}{s}$. Plot these curves for a gate length of $1\mu m$ and $10\mu m$, with maximum drain voltage, $V_{DS} = 2V$, and maximum gate voltage, $V_{GS} = 0V$. Assume the electron mobility in the doped GaAs to be $5000\frac{cm^2}{Vs}$, and a Schottky barrier height of 1 eV for the gate metal. Normalize the current to unit with(mA/mm).

- **Section 8.5**

Problem 8.19 Consider an $Al_{0.3}Ga_{0.7}N$/GaN HEMT structure. Assume that the Schottky barrier is 1.7 eV on AlGaN and 0.9 eV on GaN.
(a) How does the sheet charge at the AlGaN/GaN junction vary with the thickness of the AlGaN barrier? Plot the sheet charge n_s for AlGaN thickness up to 40 nm.
(b) Plot the band diagram of an AlGaN/GaN HEMT with a 30 nm AlGaN cap at zero gate bias, and at pinch-off. What is the pinch-off voltage?
(c) Now, a 5 nm layer of GaN is added *above* the AlGaN barrier. Calculate and plot the band diagram of this structure at zero bias and at pinch-off. What is the effective Schottky-barrier height in these two cases? Do you expect the gate leakage of this diode to be different from the AlGaN/GaN structure? Why (not)?

Problem 8.20 I grow an AlGaN on GaN HEMT (Device A) where the net polarization charge, $Q_{\pi,NET}$ = 1.5 x 10^{13} cm^{-2}. The spontaneous and piezoelectric polarizations (due

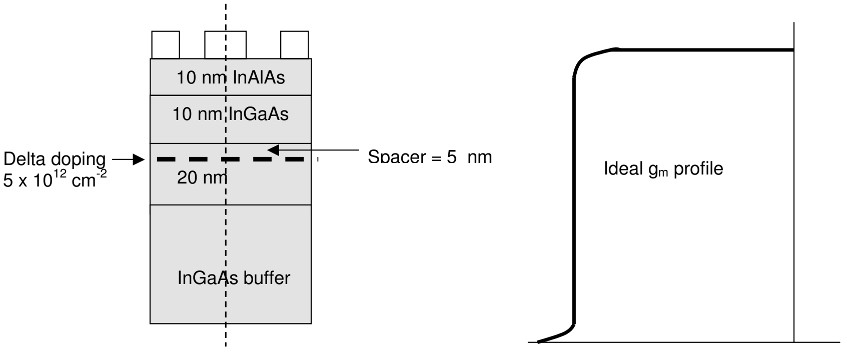


Figure 8.49: Figure for problem 8.21.

to the strain in the AlGaN) contribute 1 x 10^{13} cm^{-2} and 5 $\times$ 10^{12}cm^{-2} electrons to this charge, respectively.
(a) Draw the band diagram of this structure assuming that the surface pinning is 1.8 eV and the conduction band discontinuity of 0.7 eV.
(b) When measuring device A, I find the output conductance is high. I therefore grow a different device on a relaxed AlGaN buffer to reduce substrate injection and grow the strained 10 nm GaN QW followed by the same AlGaN cap I grew before (ie. 200 $\AA$). Draw the band diagram of device B by calculating and showing the relevant voltages and changes in the system. How much electron charge is available? How would you expect the output conductance to change?
(c) What is the main problem in device B? Suggest a qualitative solution to this problem.

Problem 8.21 I make a HEMT as shown in figure 8.49 and get a g$_m$ versus V$_{GS}$ curve that deviates from the ideal one. Draw the charge, electric field and energy band profiles for this structure along the line AA'. What is the transconductance curve you measure and why? The electron velocity in the structure is 2 x 10^7 cm^{-2}. Assume ΔE$_C$ = 0.5 eV, Schottky barrier height, ϕ_B = 0.8 eV, and Δd = 5 nm, where Δd is the mean distance between the electron gas and hetero-interface. You may also assume that the transistor operates in the fully saturated region.

Problem 8.22 Consider a GaAs n-channel MESFET operating under conditions such that one can assume that the field in the channel has a constant value of 5.0 kV/ cm^{-1}. The channel length is 2.0 μm. Calculate the transit time for an electron to traverse the channel if one assumes a constant mobility of 7500 cm^2/V·s. What would the time be if the correct velocity-field relations plotted in chapter B were used?

Problem 8.23 Consider a 1.0 μm channel length n-channel Si MESFET operating under the condition that the average field in the channel is 15 kV/cm. Assume the electric field in the channel is constant at this value. Calculate the electron transit time assuming a constant mobility of 1000 cm^2/V·s and using the velocity-field relations for Si given in the text.

Problem 8.24 Consider two n-channel GaAs MESFETs operating at a source-drain bias of 2.0 V. Assume that the electric field in the channel is constant and has a value of V_{DS}/L where L = 1.0 μm for one device and 5 μm for the second. Calculate the transit time for electrons in the two devices using two models for transit: (a) constant-mobility model with μ = 6000 cm^2/V·s; (b) correct velocity-field relations for the velocity. Use the curves given in chapter B for the velocity field. Note that the discrepancy in the two models is larger for the shorter channel device.

Problem 8.25 Consider an n-channel GaAs MESFET with the following parameters:

Schottky barrier height,	ϕ_b	$=$	0.8 V
Channel doping,	N_d	$=$	5×10^{16} cm^{-3}
Channel depth,	h	$=$	$0.5\ \mu$m
Channel mobility,	μ_n	$=$	5000 cm^2/ V · s
Channel length,	L	$=$	$1.5\ \mu$m
Channel width,	Z	$=$	$20.0\ \mu$m

Calculate the value of $V_{DS}(sat)$ at V_{GS} = 0. Also calculate the output resistance of the channel at V_{DS} = $V_{DS}(sat)$ + 2.0 V.

Problem 8.26 Consider an n-channel GaAs with the same parameters as the device in problem 8.25 except for the channel length. A maximum value of V_{DS} is 10.0 V for the device, and it is required that the effective channel length L' at V_{GS} = 0 and the maximum drain voltage should be no less than 90% of the actual channel length L. What is the smallest channel length L that satisfies this requirement?

Problem 8.27 Consider the nominal AlGaAs-GaAs ($\Delta E_C = 0.25eV$) HEMT structure shown in figure 8.50. The sheet charge in the channel is $1 \times 10^{12} cm^{-2}$.

(a) Calculate the sheet doping in the donor layer required to achieve that. Show clearly the electric field distribution and the resultant band diagram of the structure.

(b) I wish to now have a flat quantum wll (as opposed to a triangular quantum well) holding the same sheet charge density. First, clearly state the design methodology to achieve this. Next, proceed with the quantitative analysis.

(c) Explain why I would want a flat quantum well. Are there any disadvantages?

(d) Calculate the g_m vs. V_{gs} curve for the transistor assuming that $v_s(GaAs) = 1 \times 10^7 \frac{cm}{s}$ and $v_s(AlGaAs) = 2 \times 10^7 \frac{cm}{s}$. Use 3-d density of states in the AlGaAs for your calculation.

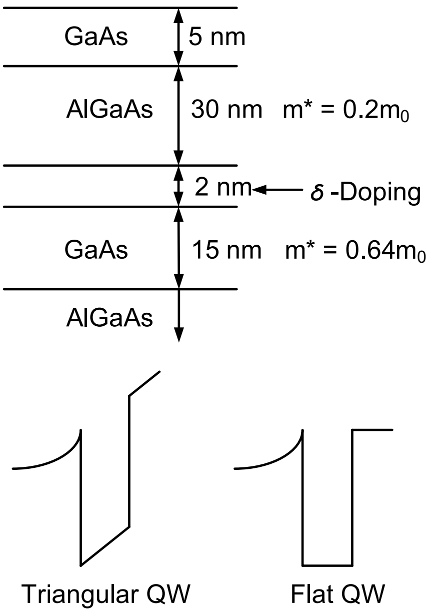


Figure 8.50: Figure for problem 8.27.

Problem 8.28 Consider the AlInAs/GaInAs HEMT, shown in figure 8.51 where the AlInAs is delta-doped with Si to the level of $5 \times 10^{12} cm^{-2}$. The spacer layer thickness is 5nm. You may assume that the Schottky barrier height is determined by Fermi level pinning of the surface and is 1 eV. Next, I consider the same structure grown on p^+ GaInAs $(E_{Fp} \approx E_V)$, where the thickness of the buffer is $1\mu m$ to enable threshold voltage adjustment. What is the sheet charge in this structure compared to the structure grown on undoped GaInAs? Last, but not least consider a forward bias of 0.8 eV applied to the conventional HEMT structure. Assuming an effective mass of $0.5m_0$. Assume tunneling as the transport mechanism. Calculate the position of the Fermi level around the donor by balancing the current in with that out and linking the resident electron concentration to E_F. Use $E_g(GaInAs) = 0.7eV$ and $E_g(AlInAs) = 1.4eV$, and $\Delta E_C = 0.5eV$.

- **Section 8.8**

Problem 8.29 In this problem we will consider the effect of the source resistance on the device transconductance. Consider an n-channel GaAs MESFET with the following

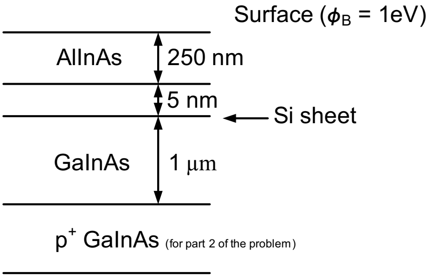


Figure 8.51: Figure for problem 8.28.

parameters:

Schottky barrier height,	ϕ_b	$=$	0.8 V
Gate length,	L	$=$	$3.0\ \mu$m
Channel mobility,	μ_n	$=$	$6000\ \mathrm{cm}^2/\ \mathrm{V}\cdot\mathrm{s}$
Channel doping,	N_d	$=$	$5\times10^{16}\ \mathrm{cm}^{-3}$
Channel depth,	h	$=$	$0.5\ \mu$m
Gate width,	Z	$=$	$25\ \mu$m

Calculate the intrinsic transconductance of the device. If the source-to-gate separation is 0.5 μm, calculate the value of the extrinsic transconductance.

Problem 8.30 Calculate the maximum cutoff frequency for the ideal device of problem 8.24 (with the source resistance assumed equal to zero). Calculate the degradation in the cutoff frequency due to the effect of the source series resistance.

Problem 8.31 Consider an n-type GaAs MESFET at 300 K with the following parameters:

Schottky barrier height,	ϕ_b	$=$	0.8 V
Channel doping,	N_d	$=$	$10^{17}\ \mathrm{cm}^{-3}$
Channel mobility,	μ_n	$=$	$6000\ \mathrm{cm}^2/\ \mathrm{V}\cdot\mathrm{s}$
Channel depth,	h	$=$	$0.2\ \mu$m
Channel width,	Z	$=$	$2.0\ \mu$m
Channel length,	L	$=$	$1.0\ \mu$m

Calculate the maximum cutoff frequency using the constant-mobility model and the saturation velocity model.

Problem 8.32 An important effect in short-channel FETs made from high-mobility materials like GaAs and InGaAs is the "velocity overshoot effect." The average time for scattering τ_{sc} in such materials is ~1.0 ps. If the electron transit time is less than 1 ps, the electron moves "ballistically," i.e., without scattering. Consider a FET in which the average electric field is 20 kV/cm. Electrons are injected at the source with thermal velocities and move in the average electric field toward the drain. Estimate the gate length at which the velocity overshoot effect will become important for Si, GaAs, and InAs. Assume that the average scattering time is 1 ps for all three materials. Assume electron effective masses of 0.26 $m_0, 0.067$ m_0, and 0.02 m_0, respectively.

8.11 DESIGN PROBLEMS

Problem 8.1 Consider an n-MESFET made from GaAs operating in an ON state. Sketch schematically (i.e., only semi-quantitatively) the electric field in the channel below the gate going from the source to the drain for the following cases:
(a) the device is in the linear regime, i.e., the drain bias is very small.
(b) the device is under a high drain bias (i.e., $V_D \sim V_D(sat)$).
Give reasons for your results.

Problem 8.2 A field-effect transistor is to be made from the high-speed material n-InGaAs. The doping is 10^{17} cm^{-3}. The bandgap of the material is 0.8 eV and the maximum Schottky barrier height possible is 0.4 eV. In the device the maximum gate leakage current density allowed is 10^{-2} Acm^{-2}. Discuss how you would design the FET using the MESFET and JFET approach.

$$\begin{aligned} R^* &= 4.7 \text{ Acm}^{-2}\text{K}^{-2} \\ D_p &= 25 \text{ cm}^2/\text{s} \\ L_p &= 1.5\ \mu\text{m} \\ n_i &= 2 \times 10^{11} \text{ cm}^{-3} \end{aligned}$$

Discuss the limitations on the gate bias for the MESFET and the JFET.

Problem 8.3 An n-MESFET is made from GaAs doped at 10^{17} cm^{-3}. The gate width Z is 50.0 μm and the gate length is 2.0 μm and the channel thickness h is 0.25 μm.

To characterize the gate properties, the gate semiconductor current is measured and is found to have the value (at 300 K)

$$I_G = 3.12 \times 10^{-14}[\exp(eV/k_BT) - 1] \text{ A}$$

where V is the bias between the gate and the semiconductor. The mobility in the semiconductor is measured to be 4000 cm^2/V·s.
(a) Calculate the threshold voltage V_T for the device.
(b) Calculate the transconductance at saturation when the gate bias is $V_{GS} = -2.0$ V.

Problem 8.4 Consider a GaAs MESFET with a gold Schottky barrier of barrier height 0.8 V. The n-channel doping is 10^{17} cm^{-3} and the channel thickness is 0.25 μm. Calculate the 300 K threshold voltage for the MESFET.

Problem 8.5 Consider the device in problem 8.4. Calculate the maximum channel thickness at which the device is OFF when no gate bias is applied, i.e., the device is an enhancement MESFET.

Problem 8.6 Consider a GaAs MESFET with the following parameters:

Schottky barrier height	=	0.8 V
Channel doping	=	10^{17} cm^{-3}
Channel depth	=	0.06 μm

Calculate the gate bias needed to open up the MESFET channel.

Problem 8.7 Consider a GaAs MESFET with the following parameters:

Channel mobility,	μ_n	=	6000 cm^2/ V $\cdot$ s
Schottky barrier height,	ϕ_b	=	0.8 V
Channel depth,	h	=	0.25 μm
Channel doping,	N_d	=	5×10^{16} cm^{-3}
Channel length,	L	=	2.0 μm
Gate width,	Z	=	25 μm

Calculate the 300 K saturation current when a gate bias of 0.0 V and -1.0 V is applied to the MESFET. Also calculate the transconductance of the device at these biases.

8.12 FURTHER READING

- **General**
  - D. A. Neaman, Semiconductor Physics and Devices (Irwin, Boston, MA, 1997).
  - R. F. Pierret, Field Effect Devices (Vol. 4 of the Modular Series on Solid State Devices, Addison-Wesley, Reading, MA, 1990).
  - M. Shur, Physics of Semiconductor Devices (Prentice-Hall, Englewood Cliffs, NJ, 1990).
  - S. M. Sze, Physics of Semiconductor Devices (Wiley, New York, 1981).

Chapter 9

FIELD EFFECT TRANSISTORS: MOSFET

9.1 INTRODUCTION

The basic principles of the field effect transistor have been discussed in chapter 8. A key requirement for a FET is zero or negligible gate leakage current. To ensure this one needs some kind of barrier for electron (hole) from the gate to the source, channel and drain. In the devices in chapter 8 this barrier is provided by a Schottky barrier (or $p^+ - n$ built-in voltage.) or a large bandgap semiconductor. Of course one may ask: Why not use an insulator to isolate the gate from the channel? Obviously an insulator would be an ideal choice but so far only on Si has it been possible to grow a high quality and reliable insulator. This has led MOSFET technology to become so dominant. In many ways the MOSFET is an ideal device since a large gate

bias can be applied to "invert" the bands and induce electron (or holes) in a channel without the concern of gate leakage. An example of a MOSFET today is shown in cross-section in figure 9.2. Over the last several years steady progress has been made on using the MOSFET concept with other semiconductors, notably GaAs. Indeed GaAs NMOSFETs have been demonstrated with channel mobilities much higher than those in NMOS FET based on Si. However, widespread use of such devices is still not near.

In this chapter we will first discuss the MOS capacitor and examine how mobile charge is induced in the the MOS structure by "inversion." It is important to note that in a MOSFET, unlike the MESFET or JFET, channel charge is induced electrostatically by the gate by using the gate as a capacitor with the gate metal electrode and the semiconductor being the other plate of the capacitor without the need for doping, however the addition of dopants in the channel provides additional control on the charge. Once we discuss the MOS capacitor we will examine the operation of the MOSFET.

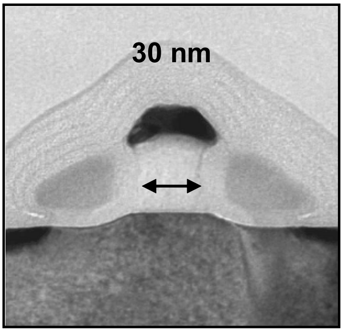


Figure 9.1: SEM cross-sectional image of a state-of-the-art MOSFET with a physical gate length of 30nm. Figure courtesy of R. Chau, Intel.

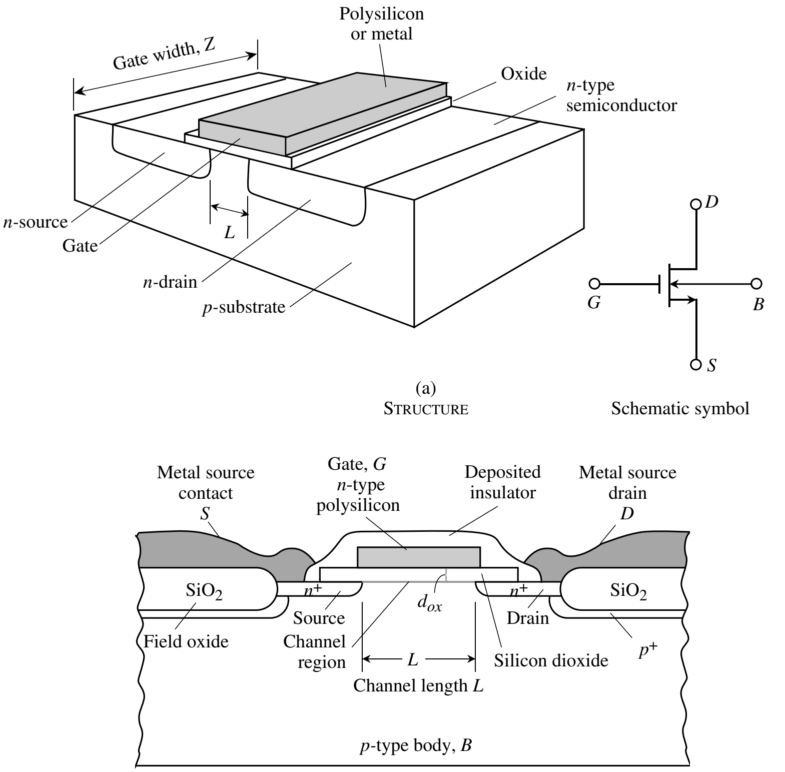
9.2 MOSFET: DEVICES AND IMPACT

MOSFETs can be made so that the current from the source to drain is carried by electrons (NMOS), by holes (PMOS), or in the case of complementary MOSFET (CMOS), by electrons and by holes in two devices. In figure 9.2 we show a schematic of an NMOS device. The structure starts with a p-type substrate. We will see later that a voltage applied to the gate "inverts" the polarity of the carriers and produces electrons near the oxide-semiconductor interface.In figure 9.3 we show the well known "Moore's Law" and its impact on technology. It is well known that the advances shown in figure 9.3 have been possible because of the Si MOSFET devices.

CMOS Technology

CMOS technology has become the most widely used technology, finding use in wireless, microprocessors, memories, and a host of other applications. The chief attraction is low power dissipation. Since both NMOS and PMOS transistors are to be fabricated on the same substrate, additional steps are needed compared to the NMOS case discussed earlier. The cross-section of a typical CMOS structure is shown in figure 9.4a. As can be seen, the NMOS transistor is fabricated within a p-type well that is implanted or diffused into the n-substrate. The p-well acts as the body or substrate for the NMOS. In addition to creating the p-well, one needs to do an n^+ implant for the source and drain of the transistor. In figure 9.4b the symbolic representation of the CMOS transistor is shown.

It is critical in MOSFETs to follow a voltage convention to make sure that errors are avoided in calculating critical parameters such as threshold voltage. Consider two materials 1 and 2, shown in figure 9.5 with work functions ϕ_1 and ϕ_2 which form a junction. We always reference voltages with respect to the material 2. The electrochemical potential of material 1 with respect to material 2 is $\phi_1 - \phi_2$. Hence the built-in voltage of this structure, which by definition is the



(a)
STRUCTURE

Schematic symbol

(b)
CROSS-SECTIONAL VIEW

Figure 9.2: (a) A schematic of an NMOS device along with a symbol for the device. The contact B denotes the body or substrate of the device. (b) A cross-section of the NMOS. Modern devices involve considerably more complexities.

voltage required to align the two Fermi levels, is therefore equal to

$$V_{bi} = -(\phi_1 - \phi_2) \tag{9.2.1}$$

The applied voltage necessary to create flat bands in the junction is $V_{fb} = -V_{bi}$.

Now let us consider an MOS capacitor. Figure 9.6c shows the device band diagrams with zero bias across an MOS structure and $V = V_{fb}$ applied to material 1 with respect to material 2.

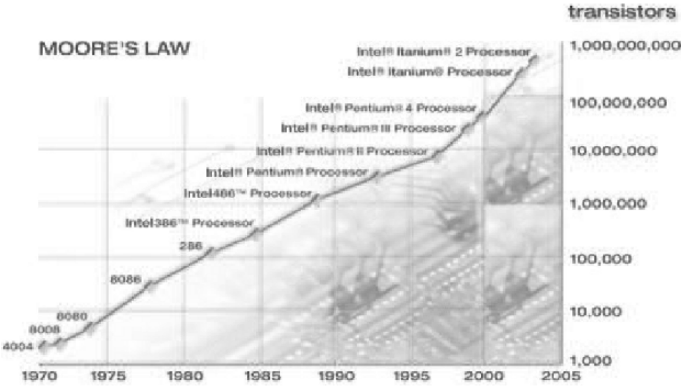


Figure 9.3: Illustration of Moore's Law.

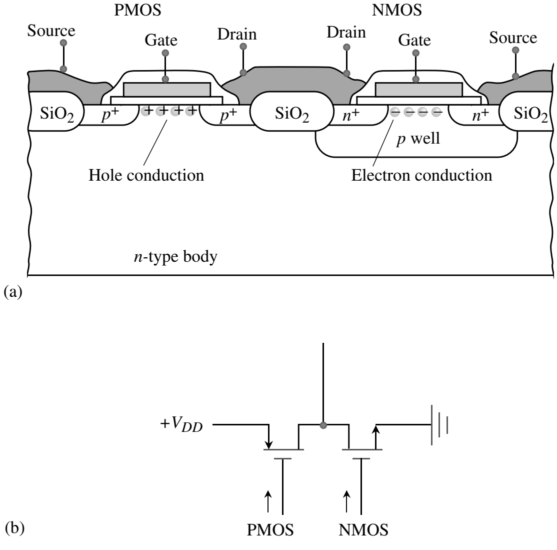


Figure 9.4: (a) A cross-section of a CMOS device. (b) Symbol representing the CMOS.

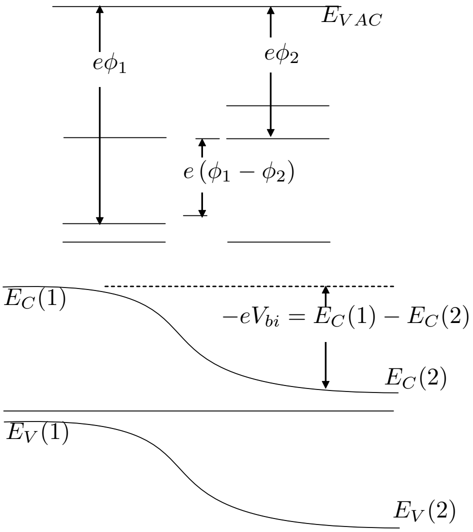


Figure 9.5: Above: Band line-up before junction formation. Below: Band line-up after junction formation

Following our convention

$$V_{bi} = -(\phi_m - \phi_s) = -\phi_{ms}$$

In the case shown ϕ_{ms} is negative and therefore V_{bi} is a positive number. Hence from $V_{fb} = -V_{bi}$ we get $V_{fb} = \phi_{ms}$. When applied to the case shown we see that V_{fb} is negative

9.3 METAL-OXIDE-SEMICONDUCTOR CAPACITOR

We have noted several times in this book that Si technology is so far unique in that a high-quality oxide SiO_2 that can be formed on Si. The Si-SiO_2 interface perfection has been the reason why field-effect devices are suitable for many applications. Their higher areal density, better switching characteristics and lower power dissipation have made them the dominant device in electronic systems and the engine driving Moore's law.

To understand the operation of the MOSFET we first need to examine the MOS capacitor, whose structure and band diagrams are shown in figure 9.6. An oxide layer is grown on top of a p-type semiconductor and a metal contact is placed on the oxide. In general, the insulator could be any large bandgap material. The main purpose of the oxide layer is to provide isolation between the metal and the semiconductor.

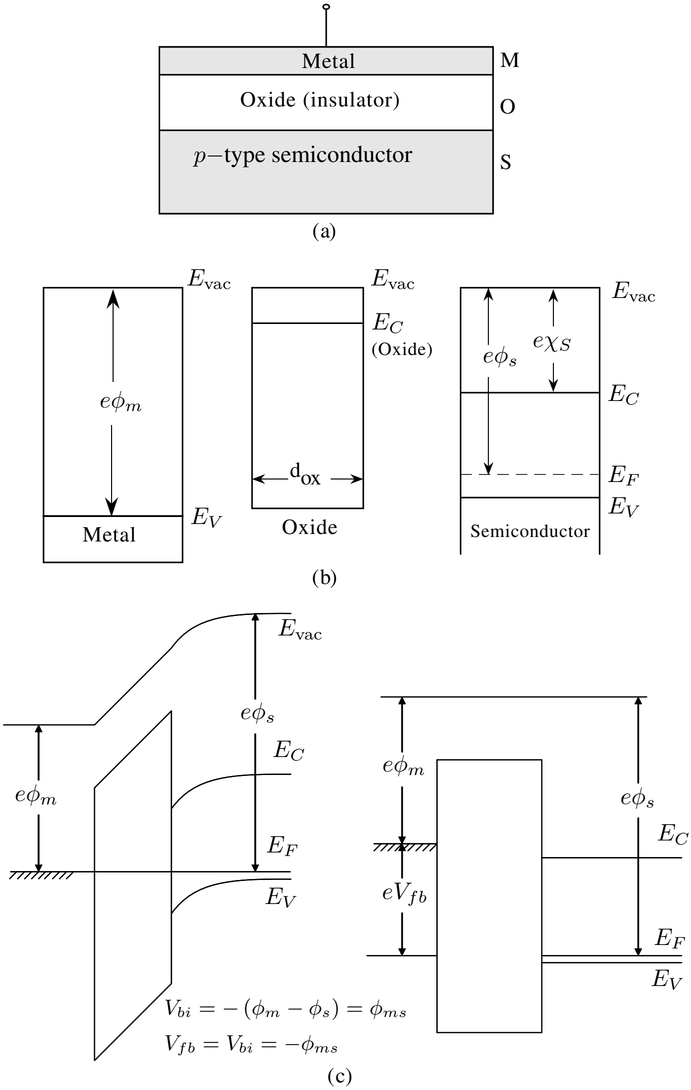


Figure 9.6: (a) A schematic of an MOS capacitor. (b) Band profiles of the isolated metal, oxide, and semiconductor. Shown are the metal work function, semiconductor work function, and electron affinity. (c) Band profile of the MOS structure in equilibrium and in flatband.

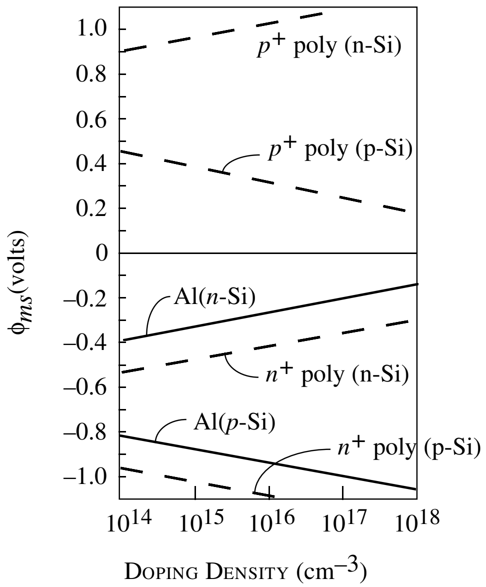


Figure 9.7: Metal-semiconductor work function difference for some important gate metals used in MOS devices. Note the signs of ϕ_{ms} for three different gate types for NMOS and PMOS.

In figure 9.7 we show the values of ϕ_{ms} for several different metals as a function of doping density. Starting from the flat band position, there are three important regimes of biasing in the MOS capacitor, as shown in figure 9.8.

(i) Hole Accumulation: If a negative bias is applied between the metal and the semiconductor, the valence bands are bent to come closer to the Fermi level, causing an accumulation of holes at the interface as shown in figure 9.8a. The difference between the Fermi level in the metal and the semiconductor is the applied bias.

(ii) Depletion: If a positive bias is applied to the metal with respect to the semiconductor, the Fermi level in the metal is lowered by an amount eV with respect to the semiconductor, causing the valence band to move away from the semiconductor Fermi level near the interface. As a result the hole density near the interface falls below the bulk value in the p-type semiconductor as shown in figure 9.8b. So, $n \sim p \sim 0$.

(iii) Inversion: If the positive bias on the metal side is increased further, the conduction band at the oxide-semiconductor region comes close to the Fermi level in the semiconductor. This reverses the mobile charges from holes to electrons at the interface and the electron density at the interface starts to increase. If the positive bias is increased until E_c comes quite close to the electron quasi Fermi level near the interface, the electron density becomes very high and the

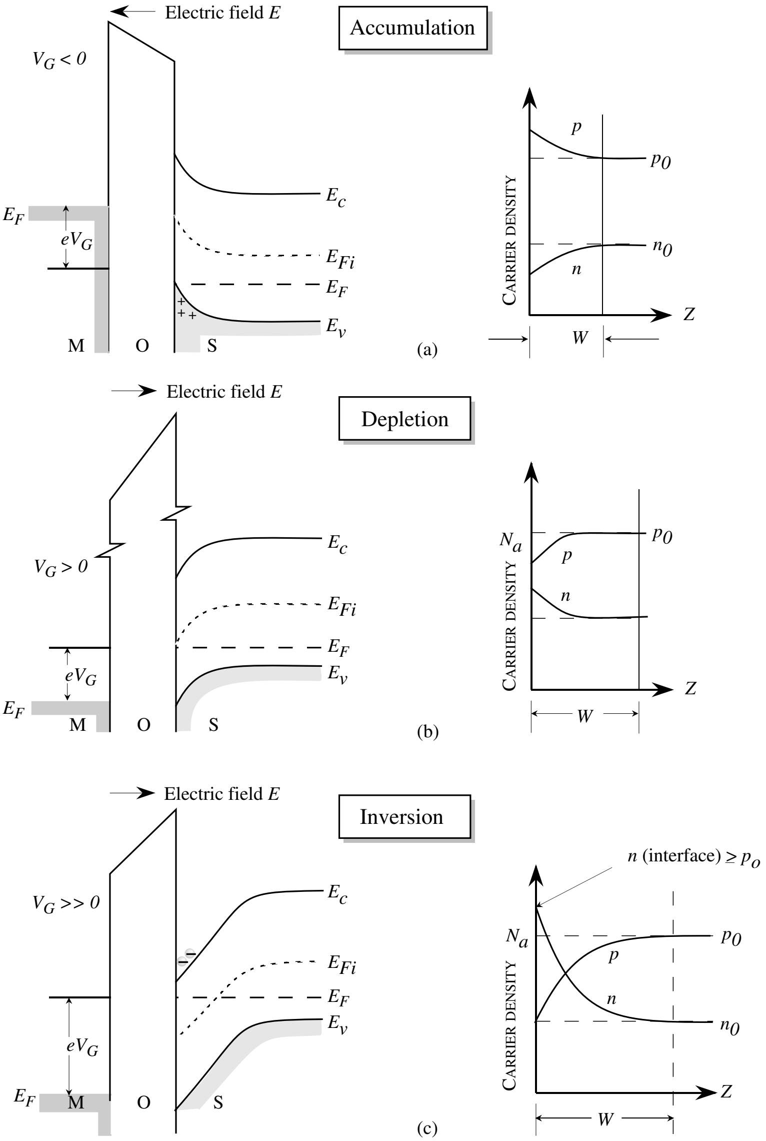


Figure 9.8: Effects of applied voltage on interface charge density in the ideal MOS capacitor: (a) negative voltage causes hole accumulation in the p-type semiconductor; (b) positive voltage depletes holes from the semiconductor surface; and (c) a larger positive voltage causes inversion—an "n-type" layer at the semiconductor surface.

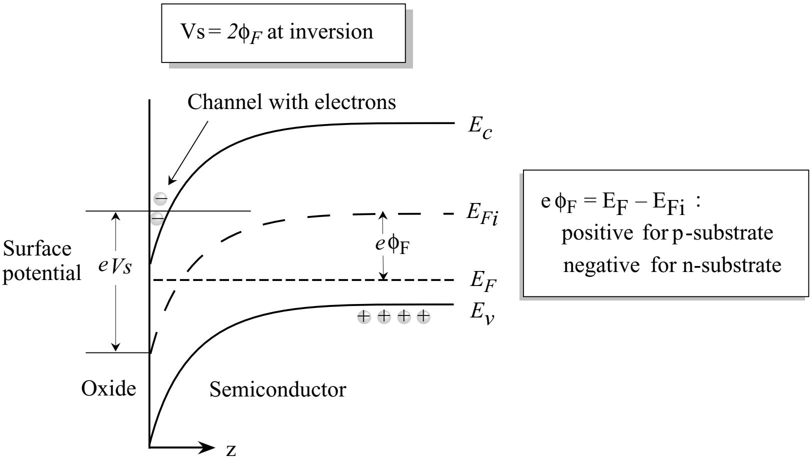


Figure 9.9: Band bending of the semiconductor in the inversion mode. The interface potential is ψ_s. A simple criterion for inversion is that $\psi_s = 2\phi_F$. The electron density changes monotonically near inversion.

semiconductor near the interface has electrical properties of an n-type semiconductor. This is shown in figure 9.8c. The device can be switched between depletion (OFF) and inversion (ON) and as a result current flow can be modulated by a gate bias.

Due to the importance of the inversion regime in the MOSFET, let us examine it in quantitative detail. In figure 9.9 we show the band bending of the semiconductor on the onset of strong inversion. The band bending is described by the quantity $e\psi$, which measures the position of the intrinsic Fermi level with respect to the bulk intrinsic Fermi level. The surface band bending at the oxide-semiconductor interface is described in terms of the potential $e\psi_s$ as shown in figure 9.9.

The onset of inversion is a gradual process as a function of gate bias. We will first use the criterion that strong inversion occurs when the electron concentration at the interface is equal to the bulk p-type concentration. Thus the intrinsic level E_{Fi} should be at a position $e\phi_F$ below the Fermi level at the interface. Thus the surface band bending is given by

$$\boxed{\psi_s(inv) = 2\phi_F} \tag{9.3.1}$$

Note that for an NMOS FET, the substrate is p-type and ϕ_F is positive and a positive bias ψ_s is needed to cause inversion. For a PMOS FET the substrate is $n-$type and ϕ_F is positive. A negative bias is needed to cause inversion. From Chapter 2, using Boltzmann statistics,

$$\phi_F = \frac{k_BT}{e} \ln \frac{p}{n_i} \sim \frac{k_BT}{e} \ln \frac{N_a}{n_i} \tag{9.3.2}$$

where N_a is the acceptor density and n_i is the intrinsic carrier concentration. The strong inversion criterion then becomes

$$\psi_s(inv) = 2\frac{k_BT}{e} \ln \frac{N_a}{n_i} \tag{9.3.3}$$

Later we will develop a model for sub-threshold current based on a more gradual transition in the electron density. At the onset of strong inversion there is an electron charge density of $\sim 10^{11}$ cm^{-2} at the surface so that the interface region's conductivity is high. Let us now evaluate the charge in the semiconductor channel. The electron concentration is approximately given by the Boltzmann distribution. In the bulk region, this concentration is

$$n_{p0} = n_i \; \exp \; (E_F - E_{Fi})/k_BT = n_i \; \exp \; \left(\frac{e\phi_F}{k_BT}\right) \tag{9.3.4}$$

We are interested in calculating the carrier concentration in the semiconductor near the Si-SiO$_2$ interface.

A detailed overview of the charge, electric field, and potential in the inversion regime is shown in figure 9.10. The areal charge density on the metal Q_m is balanced by the channel depletion charge Q_d and the inversion charge Q_n. We are interested in calculating the gate voltage needed to cause inversion in the channel. This voltage is called the threshold voltage.

The total surface charge density is related to the surface field by Gauss' law and is

$$|Q_s| = \epsilon_s \, |\mathcal{E}_s| \tag{9.3.5}$$

This charge Q_s is the total surface charge density at the semiconductor-oxide interface region and includes the induced free charge (in inversion) and the background ionic charge. The charge Q_s goes to zero when the bands are flat.

We can relate the gate voltage to the surface potential ψ_s by using the continuity of the electric displacement across the oxide-semiconductor interface ($\mathcal{E}_s$ and $\mathcal{E}_{ox}$ are the electric fields in the semiconductor and the oxide at the interface):

$$\epsilon_s\mathcal{E}_s = \epsilon_{ox}\mathcal{E}_{ox} \tag{9.3.6}$$

The voltage between the gate and the semiconductor is best understood by starting from the flat band condition such that

$$V_{GS} - V_{fb} = \Delta V_{ox} + \psi_s$$

or the applied voltage difference from flat-band is the sum of the change the oxide voltage, ΔV_{ox} and ψ_s. (Note: In the absence of additional fixed charges and traps in the system, V_{ox} at flat-band is zero and $V_{fb} = \phi_{ms}$.) In general

$$V_{GS} - V_{fb} = \Delta V_{ox} + \psi_s \tag{9.3.7}$$

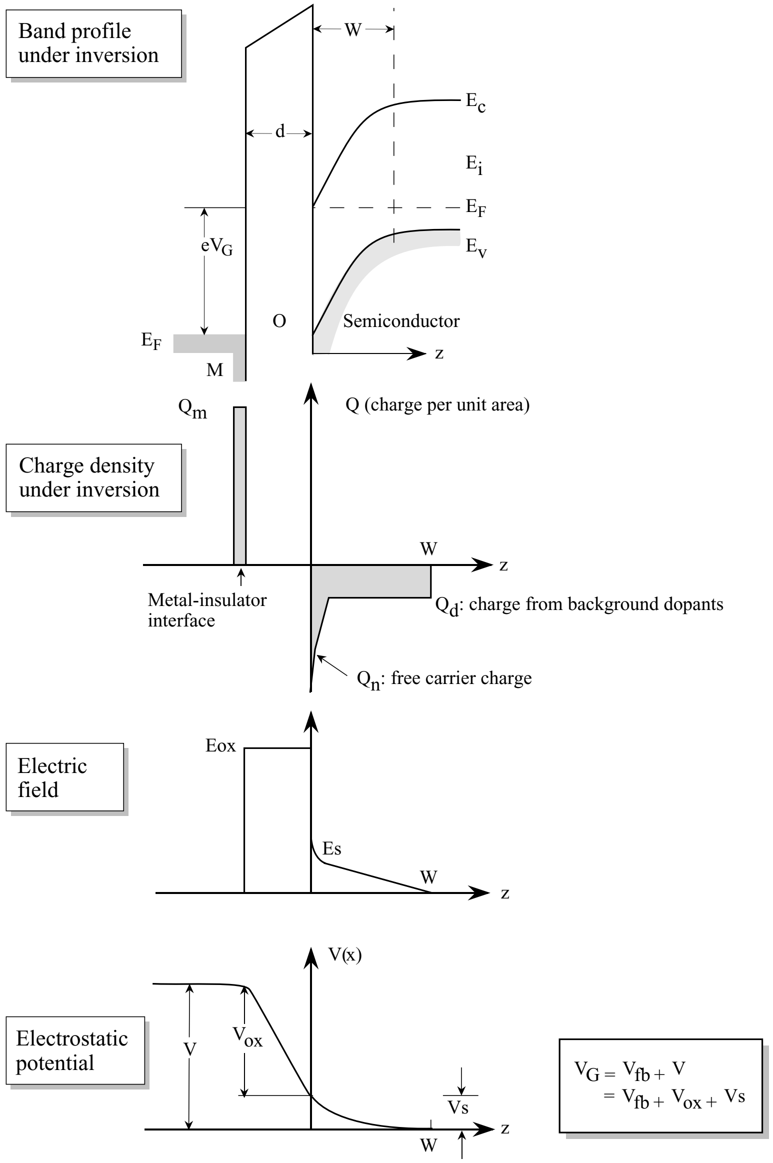


Figure 9.10: A schematic of the distributions of charge, electric field, and electrostatic potential in the ideal MOS capacitor in inversion. Once inversion begins, the depletion width W does not increase further because of the high mobile electron density at the interface region.

Also

$$\begin{aligned} \Delta V_{ox} &= \Delta \mathcal{E}_{ox} \cdot d_{ox} = \frac{\epsilon_s \Delta \mathcal{E}_s d_{ox}}{\epsilon_{ox}} \\ &= \frac{\epsilon_s \Delta \mathcal{E}_s}{C_{ox}} \end{aligned} \tag{9.3.8}$$

where C_{ox} is the oxide capacitance per unit area (= ϵ_{ox}/d_{ox}). Thus

$$V_{GS} = V_{fb} + \psi_s + \frac{\epsilon_s \Delta \mathcal{E}_s}{C_{ox}} = V_{fb} + \psi_s + \frac{Q_s}{C_{ox}} \tag{9.3.9}$$

Let us evaluate the threshold voltage V_T applied to the gate at which strong inversion starts in the channel. A reasonable approximation when inversion just occurs, the charge in the channel is essentially due to the depletion charge (= eN_aW) since the total free charge is still small. This is because even though the maximum mobile charge at inversion is equal to the bulk charge concentration, N_a, its concentration drops off exponentially with band bending and hence the areal charge density is much smaller than the depletion charge eN_aW. Using the relation between the depletion width W and the potential V_s,

$$W = \left(\frac{2\epsilon_s \left| \psi_s \right|}{eN_a} \right)^{1/2} \tag{9.3.10}$$

the areal charge density ($Q_s = eN_aW$) becomes (using $\psi_s(inv) = 2\phi_F$)

$$Q_s = \left(2\epsilon_s eN_a \left| \psi_s \right| \right)^{1/2} = \left(4\epsilon_s eN_a \left| \phi_F \right| \right)^{1/2} \tag{9.3.11}$$

This gives, from equation 9.3.10,

$$V_T = V_{GS} \left(\psi_s = +2\phi_F \right) = V_{fb} + 2\phi_F + \left(4e\epsilon_s N_a \left| \phi_F \right| \right)^{1/2} \frac{1}{C_{ox}} \tag{9.3.12}$$

Once the inversion condition is satisfied, the depletion width does not change since the large density of free carriers induced after inversion starts prevent further depletion as all additional applied voltage is dropped across the oxide since small changes in semiconductor band bending cause exponential increases in the inversion charge. The maximum depletion width is given by using $\psi_s = +2\phi_F$ in equation 9.3.11 as

$$\boxed{W_{max} = \left(\frac{4\epsilon_s \left| \phi_F \right|}{eN_a} \right)^{1/2}} \tag{9.3.13}$$

Using the above equation and equation 9.3.6, the field at the surface at the onset of strong inversion is

$$\mathcal{E}_s = \left(\frac{4eN_a \left| \phi_F \right|}{\epsilon_s} \right)^{1/2} \tag{9.3.14}$$

If the body is at a bias V_{SB} with respect to the inversion region, then the surface potential needed to cause inversion becomes $+2\phi_F + V_{SB}$. Replacing this value for ψ_s in Eqns. 9.3.12 and 9.3.14, we get, for the threshold voltage,

$$V_T = V_{fb} + 2\phi_F + \left(2e\epsilon_s N_a \left|2\phi_F + V_{SB}\right|\,\right)^{1/2} \frac{1}{C_{ox}} \tag{9.3.15}$$

In the Si/SiO$_2$ interface region there are often traps or charge centers. Since Si and SiO$_2$ have quite different lattice structures. These centers can cause a shift in the threshold voltage. Let $N_t(x)$ be the position-dependent trap density in the MOS device in the oxide region. The traps will have additional charge, which will cause a voltage drop across the insulator. The voltage drop will cause a shift in the flat-band voltage and hence the threshold voltage that is given by Gauss' law and the superposition principle as

$$\Delta V_{fb}(oxidecharge) = \Delta V_T = \frac{-e}{C_{ox}} \int_o^{d_{ox}} \frac{zN_t(z)}{d_{ox}} dz \tag{9.3.16}$$

Note that the value of the integral is the centroid of the charge distribution. Variations in V_T can have serious consequences for the device turn-on Note that that the effect of the interface trap charge on the threshold voltage depends upon where the charge is spatially located. It has the least effect if it is near the gate ($z = 0$), and has the maximum effect if it is at the Si-SiO$_2$ interface ($z = d_{ox}$). If Q_{ss} is the effective fixed charge density per unit area at the oxide-semiconductor interface, the potential drop will occur across the oxide and the flat-band voltage changes from its ideal value ϕ_{MS} to $\phi_{MS} - Q_{SS}/C_{ox}$ or

$$\Delta V_{fb}(interfacecharge) = V_{ox}(@FB) = \frac{-Q_{ss}}{C_{ox}}$$

Adding the voltage shift due to interface charge, the threshold voltage expression becomes

$$V_T = V_{fb} + 2\phi_F + \left[2e\epsilon_s N_a \left|-2\phi_F + V_{SB}\right|^{1/2}\right] \frac{1}{C_{ox}} \tag{9.3.17}$$

where $V_{fb} = \phi_{MS} - Q_{SS}/C_{ox} - \frac{e}{C_{ox}} \int_o^{d_{ox}} \frac{zN_t(z)}{d_{ox}} dz$ and defining a parameter, γ, known as the body factor as

$$\gamma = \frac{1}{C_{ox}} \sqrt{2e\epsilon_s N_a} \tag{9.3.18}$$

we can write the equation for the threshold voltage as

$$V_T = V_{TO} + \gamma \left(\sqrt{|2\phi_F + V_{SB}|} - \sqrt{2\,|\phi_F|}\right) \tag{9.3.19}$$

where V_{TO} is the threshold voltage when $V_{SB} = 0$. The expressions given above are valid for NMOS or PMOS. Of course, N_a has to be replaced by substrate doping N_d in the case of a PMOS. The signs for various terms in the threshold voltage equation for NMOS and PMOS are provided in table 9.1.

Parameter	NMOS	PMOS
Substrate	p-type	n-type
ϕ_{ms}		
Al-gate	–	–
n$^+$ Si-gate	–	–
p$^+$ Si-gate	+	+
ϕ_F	+	–
Q_{ox}	+	+
γ	+	–
C_{ox}	+	+
Source-to-body voltage V_{SB}	+	–

Table 9.1: Signs for various terms in the threshold voltage equation for a MOSFET.

Example 9.1 Assume that the inversion in an MOS capacitor occurs when the surface potential is twice the value of $e\phi_F$. What is the maximum depletion width at room temperature of a structure where the p-type silicon is doped at $N_a = 10^{16}\text{cm}^{-3}$

At room temperature, the intrinsic carrier concentration is n_i = 1.5 $\times 10^{10}\text{cm}^{-3}$ for Si. Thus, we have for the potential ϕ_F,

$$\begin{aligned}\phi_F &= \frac{k_BT}{e} \ln \frac{N_a}{n_i} = (0.026\text{eV}) \ln \left(\frac{10^{16}}{1.5 \times 10^{10}}\right) \\ &= 0.347 \text{ V}\end{aligned}$$

The corresponding space charge width is

$$\begin{aligned}W &= \left[\frac{4\epsilon_s \left|\phi_F\right|}{eN_a}\right]^{1/2} = \left[\frac{4 \times 11.9 \times (8.85 \times 10^{-14})(0.347)}{1.6 \times 10^{-19} \times 10^{16}}\right]^{1/2} \\ &= 0.30\,\mu\text{m}\end{aligned}$$

Example 9.2 Consider an aluminum-SiO$_2$-Si MOS device. The work function of Al is 4.1 eV, the electron affinity for SiO$_2$ is 0.9 eV, and that of Si is 4.15 eV. Calculate the potential V_{fb} if the Si doping is $N_a = 10^{14}\text{cm}^{-3}$.

The potential V_{fb} is given by

$$eV_{fb} = e\phi_m - (e\chi_s + (E_c - E_F))$$

The position of the Fermi level is

$$E_F = E_{Fi} + k_BT \ \ln \ \frac{N_a}{n_i}$$

below the conduction band. Also, $E_{Fi} = E_g/2$ where for Si, E_g = 1.11 eV. Using T = 300 K, we get

$$E_F = 0.555 + 0.026 \ \ln \ \left(\frac{10^{14}}{1.5 \times 10^{10}}\right) = 0.783 \text{ eV}$$

below the conduction band. Thus

$$V_{fb} = 4.1 - (4.15 + 0.783) = -0.833 \text{ V}$$

Example 9.3 Consider a p-type silicon doped to $3 \times 10^{16}\text{cm}^{-3}$. The SiO_2 has a thickness of 500 Å. An n^+ polysilicon gate is deposited to form the MOS capacitor. The work function difference $V_{fb} = -1.13$ eV for the system; temperature = 300 K. Calculate the threshold voltage if there is no oxide charge and if there is an oxide charge of 10^{11}cm^{-2}.

The position of the Fermi level is given by (measured from the intrinsic Fermi level)

$$\phi_F = 0.026 \ \ln \ \left(\frac{3 \times 10^{16}}{1.5 \times 10^{10}}\right) = 0.376 \text{ V}$$

Under the assumption that the charge Q_s is simple N_aW where W is the maximum depletion width, we get

$$\begin{aligned} Q_s &= (4\epsilon_s e N_a \left|\phi_F\right|)^{1/2} \\ &= \left(4 \times (11.9) \times (8.85 \times 10^{-14} \text{ F/cm}) \ (1.6 \times 10^{-19} \text{ C})(3 \times 10^{16} \text{ cm}^{-3})(0.376 \text{ V})\right)^{1/2} \\ &= 8.64 \times 10^{-8} \text{ C cm}^{-2} \end{aligned}$$

In the absence of any oxide charge, the threshold voltage is

$$\begin{aligned} V_T &= -1.13 + 2(0.376) + \left(8.64 \times 10^{-8}\right)\left(\frac{500 \times 10^{-8}}{3.9(8.85 \times 10^{-14})}\right) \\ &= 0.874 \text{ V} \end{aligned}$$

In the case where the oxide has trap charges, the threshold voltage is shifted by

$$\begin{aligned} \Delta V_T &= \left(10^{11}\right)\left(1.6 \times 10^{-19}\right)\left(\frac{500 \times 10^{-8}}{3.9 \times 8.85 \times 10^{-14}}\right) \\ &= -0.23 \text{ V} \end{aligned}$$

It can be seen from this example that oxide charge can cause a significant shift in the threshold voltage of an MOS device.

Example 9.4 Consider an n-MOSFET made from Si-doped p-type at $N_a = 5 \times 10^{16}$ cm^{-3} at 300 K. The other parameters for the device are the following:

$$\begin{aligned} \phi_{MS} &= -0.5 \text{ V} \\ \mu_n &= 600 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1} \\ \mu_p &= 200 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1} \end{aligned}$$

The inversion condition is $\psi_s = 2\phi_F$. Assume that the electrons induced under inversion are in a region 200 Å wide near the Si/SiO$_2$ interface.

(i) Calculate the channel conductivity near the Si-SiO$_2$ interface under flat band condition and at inversion.

(ii) Calculate the threshold voltage.

(i) Assuming that all of the acceptors are ionized, we have at flat band

$$p = N_a = 5 \times 10^{16} \text{ cm}^{-3}$$

This gives

$$\sigma(fb) = (5 \times 10^{16} \text{ cm}^{-3})(1.6 \times 10^{-19} \text{ C})(200 \text{ cm}^2/\text{ V} \cdot \text{s}) = 1.6 \ (\Omega \text{ cm})^{-1}$$

At inversion with $\psi_s = 2\phi_F$ we have

$$n(\text{interface}) = p(\text{bulk}) = 5 \times 10^{16} \text{ cm}^{-3}$$

This gives (near the interface)

$$\sigma(inv) = (5 \times 10^{16} \text{ cm}^{-3})(1.6 \times 10^{-19} C)(600 \text{ cm}^2/\text{ V} \cdot \text{s}) = 4.8 \ (\Omega \text{ cm})^{-1}$$

(ii) To calculate the threshold voltage we need ϕ_F. This is given by

$$\phi_F = \frac{k_B T}{e} \ln(\frac{p}{p_i}) = +0.39 \text{ V}$$

Using the parameters given and the equation for the threshold voltage we get

$$V_T = -0.5 + 0.78 + 1.637 \text{ V} = 1.93 \text{ V}$$

9.4 CAPACITANCE-VOLTAGE CHARACTERISTICS OF THE MOS STRUCTURE

The study of capacitance-voltage characteristics of a MOSFET provides valuable information on threshold voltage, oxide thickness, trap density, etc. In the C-V measurement, a dc bias V is applied to the gate, and a small ac signal ($\sim$ 5-10 mV) is applied to obtain the capacitance at the bias applied.

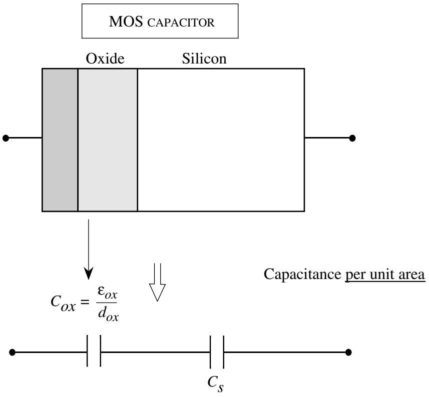

MOS CAPACITOR
Oxide
Silicon
Capacitance per unit area
$C_{ox} = \frac{\varepsilon_{ox}}{d_{ox}}$
C_s

Figure 9.11: A simple equivalent capacitance model for the MOS capacitor.

As shown in figure 9.11, the capacitance of the MOS structure is the series combination of the oxide capacitance C_{ox} and the semiconductor capacitance C_s. The semiconductor capacitance per unit area is, by definition

$$C_s = \frac{dQ_s}{dV_s} \tag{9.4.1}$$

and the capacitance of the MOS capacitor is

$$C_{mos} = \frac{C_{ox}C_s}{C_{ox} + C_s} \tag{9.4.2}$$

In the accumulation region (negative V_{GS}), the holes accumulate at the surface and C_s is much larger than C_{ox}. This is because a small change in bias causes a large change in Q_s in the accumulation regime. The MOS capacitance becomes

$$C_{mos} \cong C_{ox} = \frac{\epsilon_{ox}}{d_{ox}} \tag{9.4.3}$$

As the gate voltage becomes positive and the channel is depleted of holes, the depletion capacitance becomes important. The depletion capacitance is simply given by ϵ_s/W, and the total capacitance

$$C_{mos} = \frac{C_{ox}}{1 + \frac{C_{ox}}{C_s}} = \frac{\epsilon_{ox}}{d_{ox} + \frac{\epsilon_{ox}W}{\epsilon_s}} \tag{9.4.4}$$

With greater bias, the value of C_{mos} decreases, as shown in figure 9.12. At the strong inversion condition, the depletion width reaches its maximum value W_{max}. At this point there is essentially negligible free carrier density. The minimum capacitance takes the value

$$C_{mos}(min) = \frac{\epsilon_{ox}}{d_{ox} + \frac{\epsilon_{ox}W_{max}}{\epsilon_s}} \tag{9.4.5}$$

where W_{max} is defined by equation 9.3.13. At the onset of strong inversion, free electrons begin to collect in the inversion channel and the depletion width remains unchanged with bias. The low frequency capacitance of the semiconductor again increases since a small change in ψ_s causes a large change in Q_s. The capacitance of the MOS device thus returns toward the value of C_{ox}:

$$C_{mos}(inv) = C_{ox} = \frac{\epsilon_{ox}}{d_{ox}} \tag{9.4.6}$$

Another important point on the C-V characteristics is the point where the bands become flat. The flat band capacitance of the MOS device is (see example 9.5)

$$C_{mos}(fb) = \frac{\epsilon_{ox}}{d_{ox} + \frac{\epsilon_{ox}}{\epsilon_s}\sqrt{\frac{k_BT}{e}\,\frac{\epsilon_s}{eN_a}}} \tag{9.4.7}$$

One must now ask as to where the electrons come from when the device is in inversion. The excess electrons needed are introduced into the channel by e-h generation, by thermal generation processes, or by diffusion of the minority carriers from the p-type substrate. Since the generation process takes a finite time, the inversion sheet charge can follow the voltage only if the voltage variations are slow. If the variations are fast, the capacitance due to the free electrons makes no contribution and the capacitance is dominated by the depletion capacitance. Thus, under high-frequency measurements, the capacitance does not show a turnaround and remains at the value $C_{mos}(min)$, as shown in figure 9.12. The capacitance in the inversion regime starts to decrease even at frequencies of 10 Hz and at 10^4 Hz it reaches the low value of $C_{mos}(min)$. In the MOSFET this is not an issue since electrons can be rapidly supplied by the ohmic contacts. The presence of the fixed charge simply causes a voltage drop across the oxide given by

$$\Delta V_{fb} = \Delta V = \frac{-Q_{ss}}{C_{ox}} \tag{9.4.8}$$

where Q_{ss} is the fixed charge density (cm^{-2}) in the oxide. As a result, if Q_{ss} is positive the entire C-V curve shifts to a more negative value. Since the charge Q_{ss} is independent of the gate bias, the entire C-V curve shifts as shown schematically in figure 9.13a. The value of Q_{ss} can be obtained by measuring the shift as compared with the calculated ideal curve. Such measurements are very important for characterizing the quality of MOS devices.

The interface charge, Q_{is} , has a somewhat different effect on the C-V characteristics. In an ideal system, there are no allowed electron states in the bandgap of a semiconductor. However, since the Si-SiO$_2$ interface is not ideal, a certain density of interface states are produced that lie in the bandgap region.

In contrast to the fixed charge, electrons can flow into and out of these interface states depending upon the position of the Fermi level. The character of the interface states is defined as "acceptor-like" and "donor-like." An acceptor state is neutral if the Fermi level is below the state (i.e., the state is unoccupied) and becomes negatively charged if the Fermi level is above it (i.e., the state is occupied). The donor state is neutral if the Fermi level is above it (i.e., the state is occupied) and positively charged when it is empty. As a result, when the position of the Fermi level is altered, the charge at the interface changes.

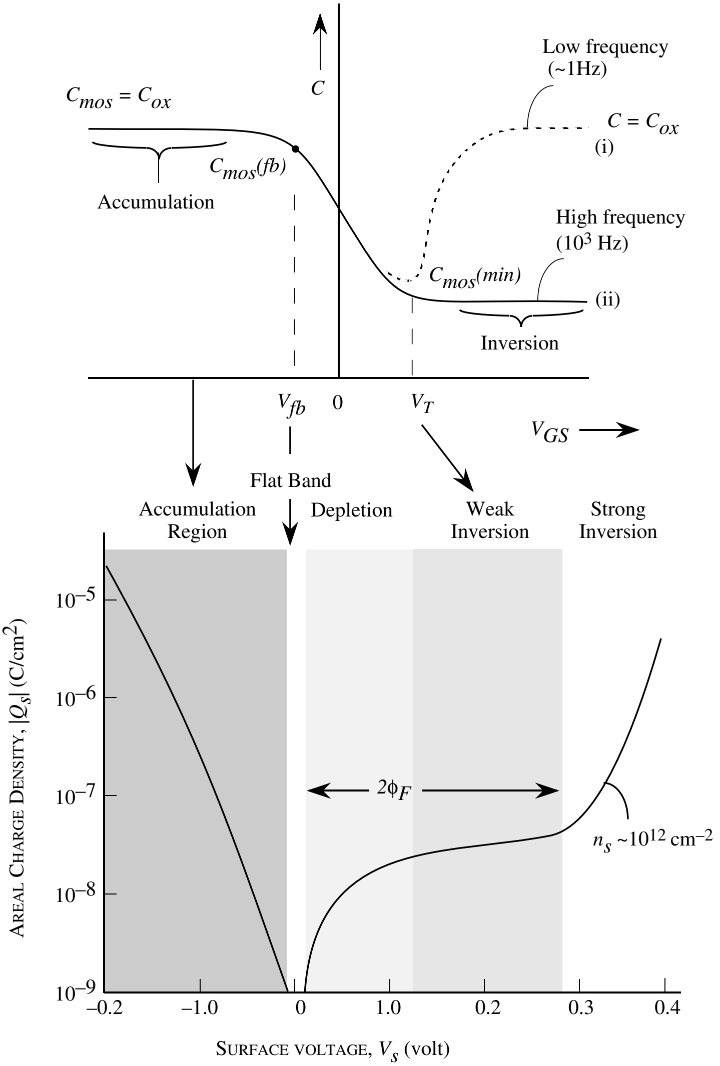


Figure 9.12: (a) A typical dependence of MOS capacitance on voltage. Curve (i) is for low frequencies and curve (ii) is for high frequencies. Also shown are the various important regions in the capacitance-voltage relations. (b) The charge density $|Q_s|$ is shown schematically as a function of the surface potential V_s.

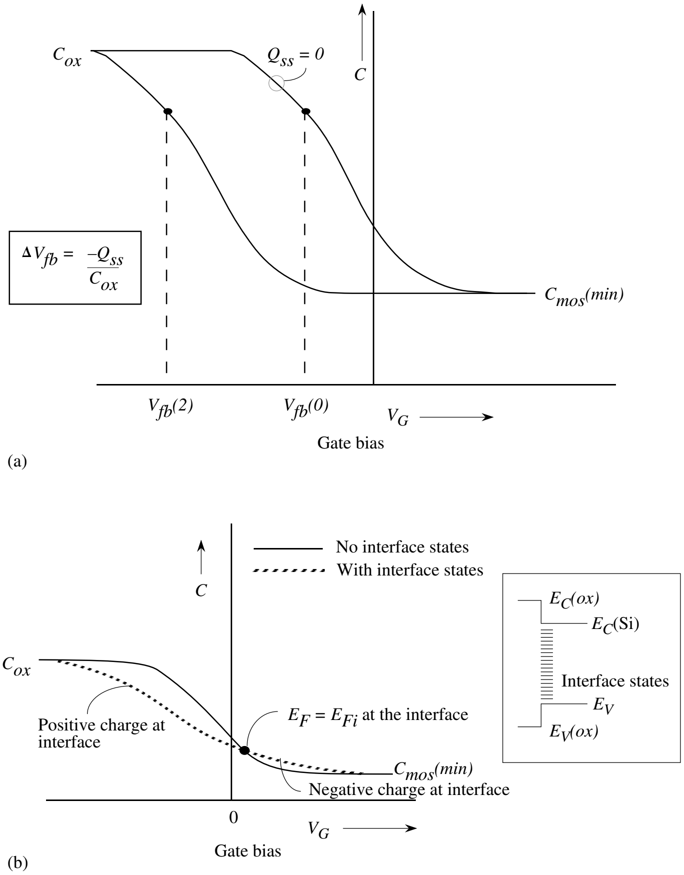


Figure 9.13: (a) A schematic plot of the high-frequency capacitance voltage of MOS capacitors with different values of the fixed oxide charge. (b) Interface states cause a smearing out the C-V curves.

When the interface charge is positive, the C-V curve shifts toward negative voltages, while when it is negative, the curve shifts toward positive voltages. This is shown schematically in figure 9.13b. The C-V curve is thus “smeared out” due to the presence of interface states. In modern high-quality MOS structures, the interface state density is maintained below 10^{10} cm^{-2}, so that the effect is negligible.

Example 9.5 Derive the relation for the semiconductor capacitance per unit area of the MOS capacitor at the flat band condition. The charge density near the flat band is

$$\delta\rho(z) = \frac{e\delta V(z) p_o}{k_B T} = \frac{e N_a \delta V(z)}{k_B T}$$

The Poisson equation then gives us

$$\frac{d^2 \delta V(z)}{dz^2} = -\frac{e N_a \delta V(z)}{k_B T}$$

The solution is a simple exponentially decaying function:

$$\delta V(z) = \delta V_s \ \exp\ (-bz)$$

where

$$b = \sqrt{\frac{e N_a}{k_B T}}$$

The charge density is now

$$\delta\rho(z) = \frac{e N_a}{k_B T}$$

The areal charge density is obtained by integrating this from the interface into the bulk, with the result

$$\mid \delta Q_s \mid = \int_o^\infty \rho(z) dz = \frac{e N_a}{b\, k_B T} \delta V_s$$

The capacitance is now

$$C_s = \frac{\delta Q_s}{\delta V_s} = \frac{e N_a}{b\, k_B T}$$

which gives the result given in equation 9.4.7 when the value of b is used.

Example 9.6 Consider a MOS capacitor made on a p-type substrate with doping of $10^{16}\mathrm{cm}^{-3}$. The SiO_2 thickness is 500 Å and the metal gate is made from aluminum. Calculate the oxide capacitance, the capacitance at the flat band, and the minimum capacitance at threshold.

The oxide capacitance is simply given by

$$C_{ox} = \frac{\epsilon_{ox}}{d_{ox}} = \frac{3.9 \times 8.85 \times 10^{-14}}{500 \times 10^{-8}} = 6.9 \times 10^{-8} \ \mathrm{F/cm}^2$$

To find the minimum capacitance, we need to find the maximum depletion width at the threshold voltage. The value of ϕ_F is given by

$$\begin{aligned} \phi_F &= 0.026 \ \mathrm{V} \ \ln\left(\frac{N_a}{n_i}\right) = 0.026 \ \ln\left(\frac{10^{16}}{1.5 \times 10^{10}}\right) \\ &= 0.347 \ \mathrm{V} \end{aligned}$$

The maximum depletion width (assuming $V_s = -2\phi_F$) is

$$W_{max} = \left(\frac{4\epsilon\left|\phi_F\right|}{eN_a}\right)^{1/2} = \left\{\frac{4(11.9 \times 8.85 \times 10^{-14})(0.347)}{1.6 \times 10^{-19} \times 10^{16}}\right\} = 0.3 \times 10^{-4} \text{ cm}$$

The minimum capacitance is now

$$\begin{aligned} C_{min} &= \frac{C_{ox}C_s}{C_{ox}+C_s} = \left(\frac{\epsilon_{ox}}{d_{ox} + \frac{\epsilon_{ox}}{\epsilon_s}W_{max}}\right) \\ &= 2.3 \times 10^{-8} \text{ F/cm}^2 \end{aligned}$$

The capacitance under flat band conditions is

$$\begin{aligned} C_{fb} &= \frac{\epsilon_{ox}}{d_{ox} + \frac{\epsilon_{ox}}{\epsilon_s}\sqrt{\left(\frac{k_BT}{e}\right)\left(\frac{\epsilon_s}{eN_a}\right)}} \\ &= \frac{3.9 \times (8.85 \times 10^{-14})}{(500 \times 10^{-8}) + \frac{3.9}{11.9}\sqrt{\frac{0.026 \times 11.7 \times 8.85 \times 10^{-14}}{1.6 \times 10^{-19} \times 10^{16}}}} \\ &= 5.42 \times 10^{-8} \text{ F/cm}^2 \end{aligned}$$

It is interesting to note that C_{fb} is $\sim 80\%$ of C_{ox} and C_{min} is $\sim 33\%$ of C_{ox}.

9.5 MOSFET OPERATION

With some important differences the MOSFET behaves in a manner similar to the MESFETs and HFETs discussed in chapter 8. A key difference is of course the electron density created by inversion. In figure 9.14 we show the basic NMOSFET structure.

9.5.1 Current-Voltage Characteristics

The full three-dimensional analysis of the MOSFET requires complex numerical techniques. However, we will present a simplified approach that gives a good semi-quantitative understanding of the current-voltage characteristics of the device.

Qualitatively, we can see how the MOSFET I-V characteristics behave. When a bias is applied between the source and the drain, current flows in the channel near the Si-SiO$_2$ interface if a channel exists. The charge density in the channel is controlled by the gate bias as well as the source-drain bias. The gate bias can thus modulate the current flow in the channel, as discussed for the MESFET or JFET case . For a simple model we assume that the mobility is constant. We also use the gradual channel approximation. In the analysis discussed here we will assume that the source is grounded and all voltages are referred to the source. Using the gradual channel approximation for the induced charge in the channel, we can treat the charge-voltage problem

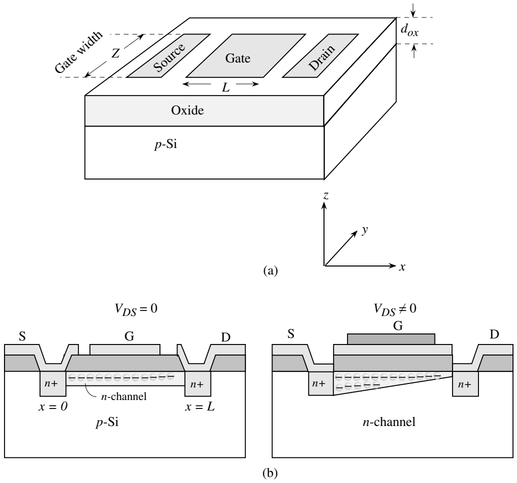


Figure 9.14: a) A schematic of the MOSFET structure. b) a cross-section of the NMOSFET.

as a one-dimensional problem. The induced charge per unit area, once we are in the inversion region, is

$$Q_s = C_{ox} \left[V_{GS} - V_T - V_c(x) \right] \tag{9.5.1}$$

We know that

$$\begin{aligned} V_c(x) &= 0 \quad \text{at the source} \\ &= V_{DS} \quad \text{at the drain} \end{aligned} \tag{9.5.2}$$

We also assume that the body bias is zero. The case of finite body bias will be discussed later. The current is given by (current = surface charge density × mobility × electric field × gate width)

$$I_D = Q_s \mu_n \frac{dV_c(x)}{dx} Z \tag{9.5.3}$$

where Z is the width of the device. The current I_D is constant at any cross-section of the channel. The above equation may be rewritten as

$$I_D dx = Q_s \, \mu_n \, dV_c(x) Z \tag{9.5.4}$$

The integration of this equation from the source ($x = 0$) to the drain ($x = L$) after using the value of Q_s gives ($V_c(L) = V_{DS}$)

$$I_D = \frac{\mu_n Z C_{ox}}{L}\left[\left\{V_{GS} - V_T - \frac{V_{DS}}{2}\right\}\right] V_{DS} \tag{9.5.5}$$

Let us define parameters k and k' to define the prefactor in the equation above:

$$k = \frac{\mu Z C_{ox}}{L} = \frac{k' Z}{L} \tag{9.5.6}$$

From equation 9.5.1 we see that for a sufficiently high drain bias, the channel mobile charge becomes zero (the channel is said to have pinched off) at the drain side. This defines the saturation drain voltage $V_{DS}(sat)$, i.e.,

$$Q_s(V_{DS}) = Q_s(V_{DS}(sat)) = 0$$

The pinch-off occurs at the drain end of the channel.

$$V_{DS}\left(Q_s(x = L) = 0\right) = V_{DS}(sat) = V_{GS} - V_T \tag{9.5.7}$$

Our derivation of the current is valid only up to pinch-off. Beyond pinch-off as discussed in chapter 7 the current essentially remains constant except for a small increase related to a decrease in effective channel length. Other factors that cause increase in drain current beyond pinch-off such as lowering of the threshold voltage and substrate injection are considered later.

Linear or Ohmic Region

In the case where the drain bias V_{DS} is less than $V_{DS}(sat)$

$$V_{DS} < V_{DS}(sat) = V_{GS} - V_T \tag{9.5.8}$$

For very small drain bias values, the current increases linearly with the drain bias, since the quadratic term in V_{DS} in equation 9.5.6 can be ignored. The current in this linear regime is

$$\boxed{I_D = k\left[(V_{GS} - V_T)V_{DS}\right]} \tag{9.5.9}$$

where V_T is the gate voltage required to "turn on" the transistor by creating strong inversion.

Saturation Region

The analysis discussed above is valid up to the point where the drain bias causes the channel to pinch off at the drain end. The saturation current now becomes, after substituting for $V_{DS}(sat)$ in equation 9.5.5,

$$\boxed{\begin{aligned} I_D(sat) &= k\left\{(V_{GS} - V_T)^2 - \tfrac{(V_{GS}-V_T)^2}{2}\right\} \\ &= \tfrac{k}{2}(V_{GS} - V_T)^2 \end{aligned}} \tag{9.5.10}$$

Thus once saturation starts, the drain current has a square-law dependence upon the gate bias similar to all FETs.

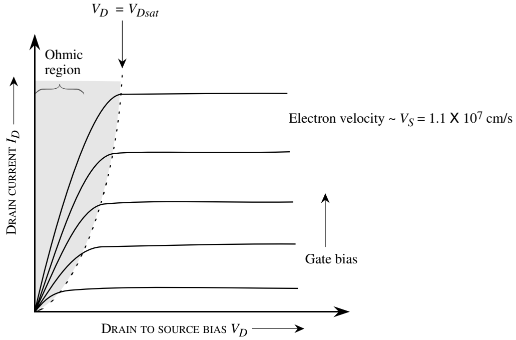


Figure 9.15: A schematic of the I-V characteristics of a MOSFET. In the ohmic region the current increases linearly with the drain bias for a fixed gate bias.

Material and Device Parameters

Important material and device parameters can be extracted from the I-V characteristics of the MOSFET. At low drain bias we can ignore the quadratic term in V_{DS}. The drain current is given by

$$I_D = \frac{Z\mu_n C_{ox}}{L}(V_{GS} - V_T)V_{DS} \tag{9.5.11}$$

so that the extrapolation of the low drain bias current points gives the threshold voltage V_T. This is shown schematically in figure 9.16. Also, if the drain current is measured at two different values of V_{GS} while keeping V_{DS} fixed, the mobility in the channel can be determined, since

$$I_{D2} - I_{D1} = \frac{Z\mu_n C_{ox}}{L}(V_{GS2} - V_{GS1})V_{DS} \tag{9.5.12}$$

where I_{D1} and I_{D2} are the currents at gate biases of V_{GS1} and V_{GS2}. Since Z, L and C_{ox} are known, the inversion channel mobility can be obtained. It is worth noting that the mobility in a MOSFET channel is usually much smaller than the mobility in bulk silicon. This is because of the strong scattering that occurs due to the roughness of the Si-SiO$_2$ interface. Typical MOSFET electron mobilities are $\sim$ 600 cm^2/V·s while typical electron mobilities in bulk silicon are $\sim$ 1300 cm^2/V·s.

The performance of the MOSFET as a device is defined via two important parameters, the drain conductance (output conductance) and the transconductance.

The drain conductance is defined as

$$g_D = \left.\frac{\partial I_D}{\partial V_{DS}}\right|_{V_{GS}=\text{constant}} \tag{9.5.13}$$

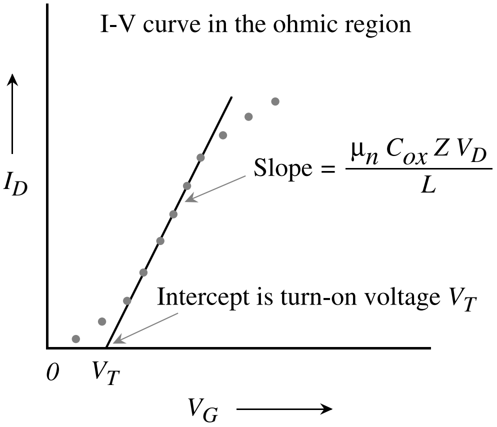


Figure 9.16: A schematic showing how the basic parameters V_T and mobility can be obtained from the $I_D - V_{GS}$ curves in the ohmic region of the MOSFET.

At low drain biases we get from equation 9.5.9 for the ohmic region

$$g_D = \frac{Z\mu_n C_{ox}}{L}(V_{GS} - V_T) \tag{9.5.14}$$

In the saturation region in our simple model, the drain conductance is zero. In real devices g_D is not zero at saturation, as discussed in section 9.6.2. The transconductance of the MOSFET is closely linked to the speed of the device and is given by

$$g_m = \left.\frac{\partial I_D}{\partial V_{GS}}\right|_{V_{DS}=\text{constant}} \tag{9.5.15}$$

In saturation we have

$$\boxed{g_m = \frac{Z\mu_n C_{ox}}{L}(V_{GS} - V_T)} \tag{9.5.16}$$

A high-transconductance device is produced if the channel length is small and channel mobility is high. The transconductance represents the control of the gate on the channel current and is usually quoted in millisiemens per millimeter (mS/mm) to remove the dependence on the gate width Z.

9.5.2 Substrate Bias Effects

In the analysis above we have assumed that the substrate bias is the same as the source bias. In MOSFET circuits, the source-to-substrate (or body) bias V_{SB} is an additional variable that can be exploited. In figure 9.17 we show an n-channel MOSFET showing the source-to-body bias, which is chosen to be zero or positive to reverse bias the source-to-substrate junction.

In the absence of V_{SB}, the inversion condition occurs when V_s, the surface potential, is equal to $-2\phi_F$ as shown in figure 9.17b. In case V_{SB} is positive, the surface voltage required for inversion is increased as shown in figure 9.17c by an amount V_{SB}, since the body is at a higher energy level.

When $V_{SB} > 0$, the depletion width is no longer W_{max} but is increased to absorb the added potential V_{SB}. As noted previously the body bias alters the threshold voltage. The change in the threshold voltage is given by

$$\Delta V_T = \frac{\sqrt{2e\epsilon_s N_a}}{C_{ox}}\left[\sqrt{|2\phi_F + V_{SB}|} - \sqrt{|2\phi_F|}\right] \tag{9.5.17}$$

To ensure a positive shift in the threshold voltage, V_{SB} must be positive for the NMOS.

The threshold voltage of a MOSFET can also be modified by altering the doping density in the silicon region as well. This can be done by ion implantation so that an added dose of acceptors (or donors) is introduced. This changes the value of the depletion charge and consequently the threshold voltage is altered.

Example 9.7 Consider a n-channel MOSFET at 300 K with the following parameters:

Channel length,	L	$=$	$1.5\ \mu$m
Channel width,	Z	$=$	$25.0\ \mu$m
Channel mobility,	μ_n	$=$	600 cm^2/ V $\cdot$ s
Channel doping,	N_a	$=$	1×10^{16} cm^{-3}
Oxide thickness,	d_{ox}	$=$	$500\ \mathring{A}$
Oxide charge,	Q_{ss}	$=$	10^{11} cm^{-2}
Metal-semiconductor work function difference,	ϕ_{ms}	$=$	-1.13 V

Calculate the saturation current of the device at a gate bias of 5 V.

The Fermi level position for the device is given by

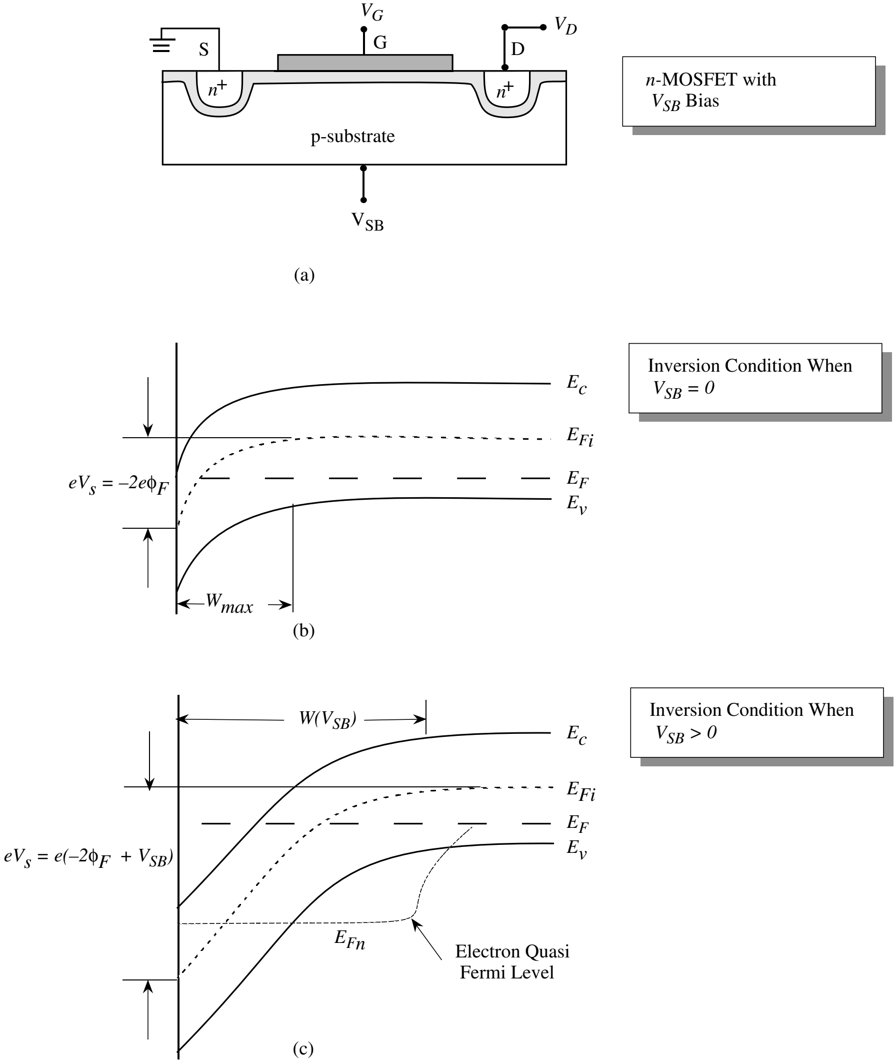
$$\phi_F = 0.026\ \ln\left(\frac{1 \times 10^{16}}{1.5 \times 10^{10}}\right) = 0.348 \text{ V}$$

The flat band voltage is

$$V_{fb} = \phi_{ms} - \frac{Q_{ss}}{C_{ox}} = -1.13 - 0.23 = -1.35 \text{ V}$$

The threshold voltage is given by equation 9.3.12 as

$$\begin{aligned} V_T &= -1.35 + 0.696 \\ &+ \frac{\left[4(1.6 \times 10^{-19})(11.9)(8.84 \times 10^{-14})(10^{16})(0.348)\right]^{1/2}(500 \times 10^{-8})}{3.9(8.85 \times 10^{-14})} \\ &= 0.04 \text{ V} \end{aligned}$$


V_G
G
S
D
V_D
n^+
n^+
p-substrate
V_{SB}
n-MOSFET with V_{SB} Bias
E_c
E_{Fi}
E_F
E_v
$eV_s = -2e\phi_F$
W_{max}
Inversion Condition When $V_{SB} = 0$
$W(V_{SB})$
E_c
E_{Fi}
E_F
E_v
$eV_s = e(-2\phi_F + V_{SB})$
E_{Fn}
Electron Quasi Fermi Level
Inversion Condition When $V_{SB} > 0$

(a)

(b)

(c)

Figure 9.17: (a) An n-MOSFET showing the voltage between the source and the body of the transistor; (b) band profiles of the MOSFET with V_{SB} = 0 at the inversion condition; (c) band profile of the MOSFET when $V_{SB} > 0$. The depletion width increases when $V_{SB} > 0$.

The saturation current is now, from equation 9.5.11,

$$\begin{aligned} I_D(sat) &= \frac{(25\times10^{-4})(600)}{2(1.5\times10^{-4})}\,\frac{(3.9)(8.85\times10^{-14})}{500\times10^{-8}}\,[5.0-0.04]^2 \\ &= 8.5\,\mathrm{mA} \end{aligned}$$

Example 9.8 Consider a silicon NMOS device at 300 K characterized by $\phi_{ms}=0, N_a=4\times10^{14}$ cm$^{-3}, d_{ox}=200$ Å$, L=1.0\ \mu$m$, Z=10\ \mu$m. Calculate the drain current for a gate voltage of V_{GS} = 5 V and drain voltage of 4 V. The electron mobility in the channel is 700 cm^2/V·s.

We start by calculating the threshold voltage. The potential ϕ_F is given by

$$\phi_F=(0.026)\ \ln\left(\frac{4\times10^{14}}{1.5\times10^{10}}\right)=0.264\ \mathrm{V}$$

The threshold voltage is, from equation 9.3.12,

$$\begin{aligned} V_T &= 0.528+\frac{\left[4(1.6\times10^{-19})(11.9\times8.85\times10^{-14})(4\times10^{14})(0.264)\right]^{1/2}}{3.9\times8.85\times10^{-14}}\cdot\left(2\times10^{-6}\right) \\ &= 0.58\ \mathrm{V} \end{aligned}$$

The saturation voltage for a gate bias of 5 V is, from equation 9.5.7,

$$V_{DS}(sat)=4.42\ \mathrm{V}$$

The saturation current is now, from equation 9.5.11,

$$I_D(sat)=11.8\ \mathrm{mA}$$

Example 9.9 Consider an n-channel MOSFET with gate width Z = 10 μm, gate length L = 2 μm and oxide capacitance $C_{ox}=10^{-7}$F/cm^2. In the linear region, the drain current is found to have the following values at V_{DS} = 0.1 V:

$$\begin{aligned} V_{GS} &= 1.5V, I_D=50\ \mu\mathrm{A} \\ &= 2.5V, I_D=80\ \mu\mathrm{A} \end{aligned}$$

The intercept of the I_D-V_{GS} curve is at -0.16 V, which is the threshold voltage.

Example 9.10 Consider an n-channel MOSFET with a substrate doping of
$N_a=2\ \times\ 10^{16}$
cm^{-3} at 300 K. The SiO$_2$ thickness is 500 Å and a source-body bias of 1.0 V is applied. Calculate the shift in the threshold voltage arising from the body bias.

The potential ϕ_F is given by

$$\phi_F=0.026\ \ln\left(\frac{N_a}{n_i}\right)=0.026\ \ln\left(\frac{2\times10^{16}}{1.5\times10^{10}}\right)=0.367\ \mathrm{V}$$

The oxide areal capacitance is

$$C_{ox} = \frac{\epsilon_{ox}}{d_{ox}} = \frac{3.9(8.84 \times 10^{-14})}{500 \times 10^{-8}} = 6.9 \times 10^{-8}\ \mathrm{F/cm}^2$$

The change in the threshold voltage is

$$\begin{aligned} \Delta V_T &= \frac{\left[2(1.6 \times 10^{-19})(11.9)(8.84 \times 10^{-14})(2 \times 10^{16})\right]^{1/2}}{6.9 \times 10^{-8}} \\ &\quad \cdot \left\{[2(0.367) + 1.0]^{1/2} - [2(0.367)]^{1/2}\right\} \\ &= 0.54\ \mathrm{V} \end{aligned}$$

9.5.3 Depletion and Enhancement MOSFETs

In the discussions of the MOSFET so far, we saw that as the gate voltage is increased, at some positive value V_T, inversion occurs and the device starts conducting or turns ON. This, of course, is not the only configuration in which the device can operate. It is possible to design devices that are ON when no gate bias is applied or are ON when negative bias is applied. This versatility is quite important since it gives a greater flexibility to the logic designer.

A device in which the current does not flow when the gate bias is zero, and flows only when a positive or negative gate bias is applied, is called an enhancement-mode device. Conversely, if the current flow occurs when the gate bias is zero and the device turns off when the gate bias is positive or negative, the device is said to operate in the depletion mode. The device we have discussed so far is an enhancement-mode device since, in our discussions, a positive gate bias was needed to cause inversion and channel formation.

To produce a depletion-mode device that is ON without any gate bias, the MOSFET fabrication is altered. As shown in figure 9.18, one starts with a p-type substrate and two n^+ contacts are placed. Additionally, in the depletion-mode device, one diffuses a thin layer of donors to produce a thin n-type channel between the n^+ contacts. The rest of the MOSFET is produced in the normal way by placing an oxide layer and a gate. The I-V characteristics of such a device are also shown in figure 9.18b.

The device discussed above can be fabricated in p-type or n-type substrates. In this device one has free carriers due to the doping and therefore the device is ON even if the gate bias is zero. The gate bias can now be used to turn the device OFF as shown.

The MOSFET can be used as a switching element in the same way as the bipolar devices or other FETs. Regardless of whether the FET is an enhancement or a depletion device, the FET carries current in one of the states of the switch. This causes power dissipation in the circuits. This is of great concern when the circuits are dense and power dissipation can cause serious heating problems. This can be avoided by using the NMOS and PMOS devices together, as will be discussed next.

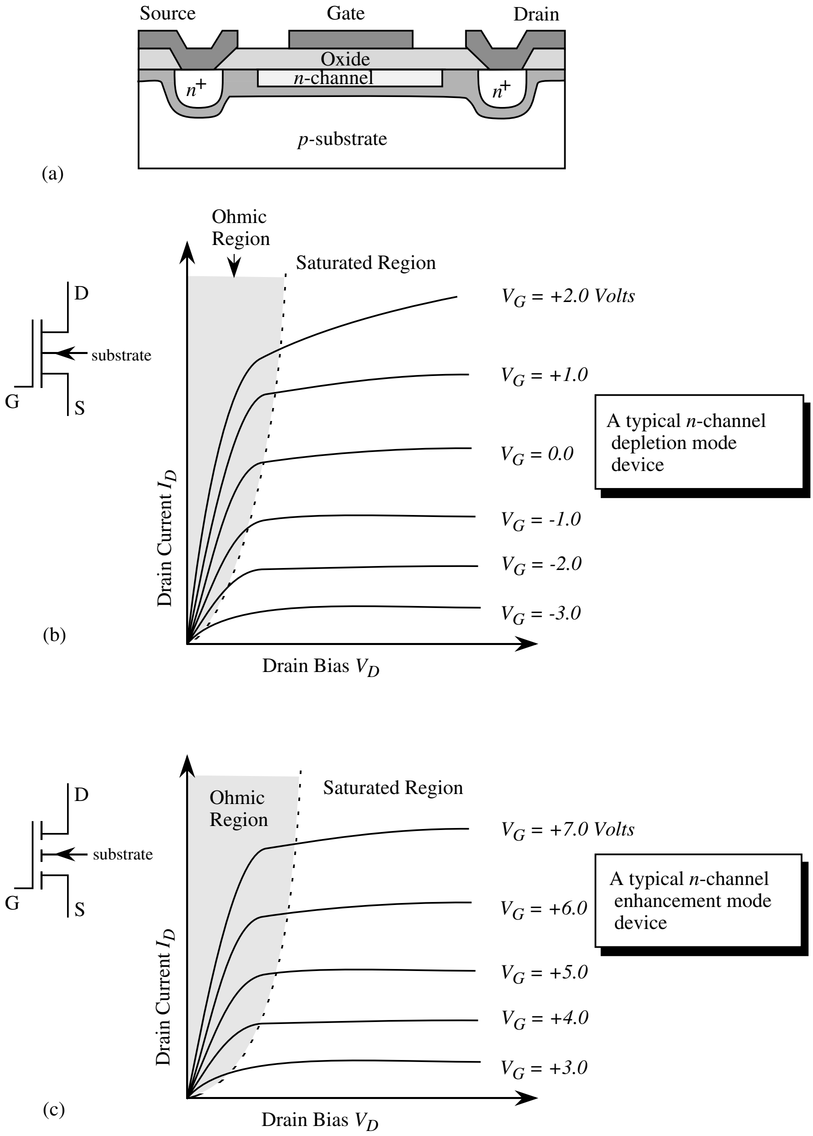


Figure 9.18: (a) A schematic of a depletion MOSFET fabricated in a p-type substrate, with an n-channel. (b) In the depletion mode, the device is ON at zero gate bias. To turn the device OFF, a negative gate bias is required as shown. The device symbol is also shown. (c) The I-V characteristics showing the device behavior in the enhancement mode. The device symbol is also shown.

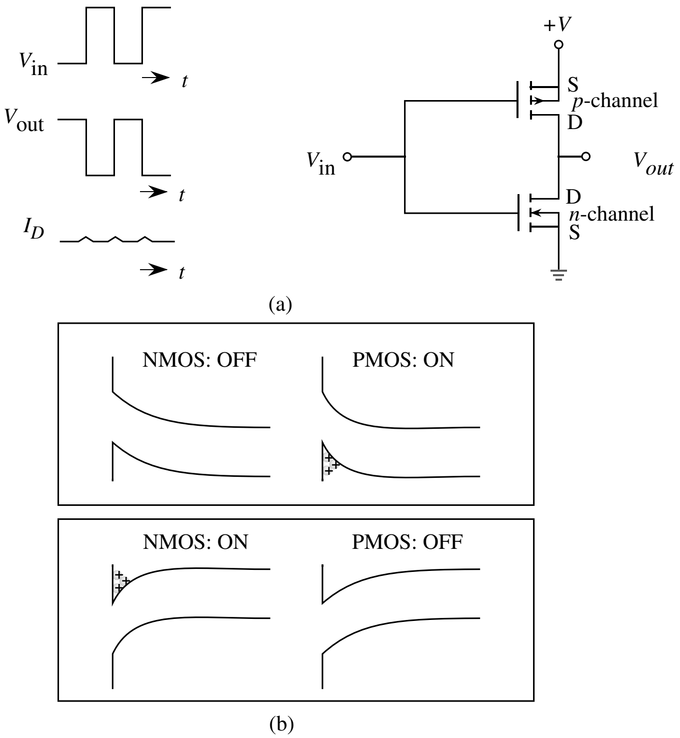


Figure 9.19: (a) A complementary MOS structure shown to function as an inverter. The circuit draws current only during the input voltage switching. (b) A schematic of the CMOS structure.

9.5.4 Complementary MOSFETs

It is possible to greatly reduce the power dissipation problem if an enhancement-mode n-channel device is connected to an enhancement-mode p-channel device in series. This is the complementary MOSFET or CMOS and is fabricated on the same chip, as shown in figure 9.19. In the CMOS inverter shown, the drains of the n- and p-MOSFET are connected and form the output. The input is presented to the gates of the device as shown. The p-channel device has a negative threshold voltage while the n-channel device has a positive threshold voltage. When a zero input voltage V_{in} is applied, the voltage between the source and gate of the n-channel device is zero, turning it OFF. However, the voltage between the gate and source of the p-channel device is $-V$ since the source of the p-channel device is at $+V$. This turns the p-channel device ON.

Thus the p-channel is ON and the n-channel is OFF so that the output voltage is V_{out} = V. No current flows in the devices since they are connected in series.

When a positive gate bias is applied, the n-channel device is ON, while the p-channel device is OFF. The output voltage V_{out} = 0. Once again no current flow occurs since the devices are in series and one of them is OFF. As can be seen, for input of 1 (High) or 0 (Low), one of the devices remains OFF. Thus the CMOS does not consume power when it is holding the information state. Only during switching is there a current flow. This low power consumption property of the CMOS makes it very attractive for high density applications, such as for semiconductor memories and processors. However, it must be noted that the device is much more complex to fabricate. Also, since the p-type transport is much poorer than the n-type transport, one has to take special care to design the two devices to have similar performances. In Chapter 10 we discuss applications of CMOS in digital and analog circuits.

Example 9.11 An n-channel MOSFET is formed in a p-type substrate with a substrate doping of $N_a = 10^{14}$ cm^{-3}. The oxide thickness is 500 Å and $\phi_{ms} = -0.83$ V. Calculate the threshold voltage and check whether the device is an enhancement- or depletion-mode device. If the device threshold voltage is to be changed by 0.5 V by ion implanting the channel by dopants, calculate the density of dopants needed. Assume that the dopant charge is all placed near the Si-SiO$_2$ interface within a thickness of 0.1 μm. Temperature is 300 K.

The position of the Fermi level is given by

$$\phi_F = 0.026\ \ln\left(\frac{10^{14}}{1.5\times 10^{10}}\right) = 0.228\text{ V}$$

The threshold voltage is

$$\begin{aligned}
V_T &= \phi_{ms} + 2\phi_F + \frac{\left[4e\epsilon_s N_a\left|\phi_F\right|\right]^{1/2}}{C_{ox}} \\
&= -0.83 + 0.456 \\
&\quad + \frac{\left[4\times(1.6\times10^{-19})(11.9\times8.85\times10^{-14})(10^{14})(0.228)\right]^{1/2}(5\times10^{-6})}{(3.9\times8.85\times10^{-14})} \\
&= -0.318\text{ V}
\end{aligned}$$

In this device there is an inversion layer formed even at zero gate bias and the device is in the depletion mode. To increase the threshold voltage by + 0.5 V, i.e., to convert the device into an enhancement-mode device, we need to place more negative charge in the channel. If we assume that the excess acceptors are placed close to the semiconductor-oxide region (i.e., within the distance W_{max}), the shift in threshold voltage is simply (N_a^{2D} is the areal density of the acceptors implanted)

$$\Delta V_T = \frac{eN_a^{2D}}{C_{ox}}$$

or

$$\begin{aligned} N_a^{2D} &= \frac{\Delta V_T C_{ox}}{e} = \frac{(0.5)(3.9 \times 8.85 \times 10^{14})}{(1.6 \times 10^{-19})(5 \times 10^{-6})} \\ &= 2.16 \times 10^{11} \text{ cm}^{-2} \end{aligned}$$

The dopants are distributed over a thickness of 0.1 μm. The dopant density is then

$$N_a = \frac{2.16 \times 10^{11}}{10^{-5}} = 2.16 \times 10^{16} \text{ cm}^{-3}$$

The use of controlled implantation can be very effective in shifting the threshold voltage.

9.6 IMPORTANT ISSUES AND FUTURE CHALLENGES IN REAL MOSFETS

In the discussions above, we have made a number of simplifying assumptions. These assumptions allowed us to obtain simple analytical expressions for the I-V relationships for the device. However, in real devices a number of important effects cause the device behavior to differ from our simple results. In this section we will briefly examine the important issues that control the performance of real MOSFETs and discuss future challenges. A summary of these challenges is shown in figure 9.32.

9.6.1 Subthreshold Conduction

As device dimensions are shrunk below 50 nm the behavior of the device below threshold or in the sub-threshold regime becomes critical. The analysis up to now has assumed that the device turns on abruptly at a gate voltage above threshold or

$$V_g - V_{th} \gtrsim 0^*$$

that no current flows at gate voltages below V_{th}. As shown in figure 9.20, this assumption does not account for current that flows through the channel in the region below strong inversion or in the weak inversion regime which is defined as the region where the surface band bending ψ_s is in the range,

$$\phi_F < \psi_s < 2\phi_F$$

Note that in strong inversion as we move from the source to the drain, the voltage across the oxide, V_{ox} decreases and the band bending in the semiconductor, ψ_S increases by a magnitude equal to $V_c(x)$. Now, let us compare this to the weak-inversion case: Figure 9.21 and illustrate the fundamental difference between current flow and the evolution of the band diagram between

*This is equivalent to a band bending of $\psi_s = 2\phi_F$ where ϕ_F is the bulk potential $= E_{iB} - E_{FP}$ and ψ_S is the band bending measured from the bulk

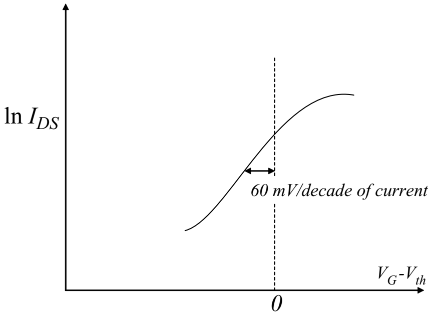


Figure 9.20: Observation: The device current does not abruptly turn-ff below threshold but decreases monotonically at a slope of 60 mV/decade of current.

the source and the drain for the case above threshold and below threshold. For $V_g > V_{th}$ and in the linear region, the full channel remains in strong inversion. The electron quasi-Fermi level in the channel follows the voltage variation and therefore drops by an amount equal to the channel potential, $V_c(x)$, and subsequently the band bending required to sustain inversion increases by this value. The current flow in this case is given by electron drift and

$$J_n = \sigma_{ch}(x)\mathcal{E}(x)$$

everywhere in the channel. This increased band bending increases V_{th} as a $f(x)$ by an amount V_c and hence the channel charge decreases monotonically, given by $n_{s\ inv}(x) = C_{ox}\left(V_g - V_{th}(x)\right)$. One critical element in the diagram is that the oxide voltage decreases and the band bending in the semiconductor increases as we go toward the drain. This can also be understood as the decrease in the channel inversion charge (negative) results in a reduced positive image charge on the gate and therefore a reduced band bending in the oxide.

The analysis of current transport in the subthreshold regime is less clear than the case above threshold. figure 9.22 shows the band diagram of the device operating in the subthreshold regime with zero bias on the source and drain regions. On applying a bias to the drain relative to the source, current could be carried either by diffusion, drift or a combination of both. figure 9.23 shows the band diagram assuming that the dominant current transport is by drift and figure 9.24 shows the case if the current transport is mainly by diffusion. The first case assumes that the applied bias drops uniformly along the channel and the second assumes that the bias drops primarily adjacent to the drain with very little drop along the channel. We will now show that the latter, diffusive transport dominates. In the weak inversion regime the maximum charge in the inversion layer is small (less than the bulk majority carrier concentration). If drift were to be true the band bending in the semiconductor has to increase continuously toward the drain which because of the constant gate voltage requires that the band bending in the oxide decreases by the same amount. The only manner that the oxide voltage can decrease is by having the inversion

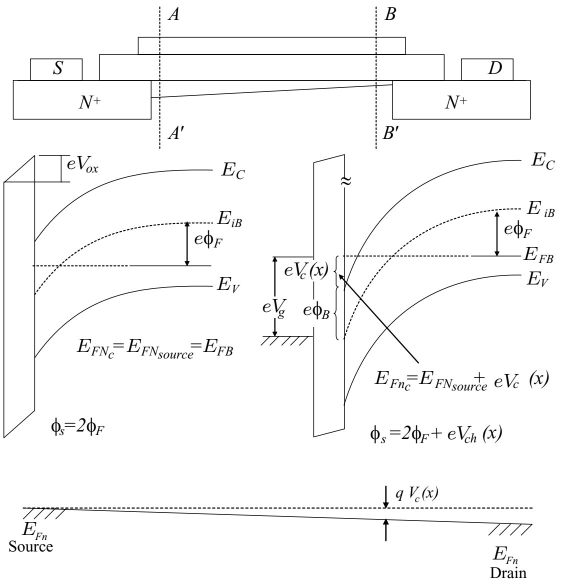


Figure 9.21: Band diagrams taken at the source side (AA') and the drain side (BB') of the gate

charge decrease faster than the increase in the depletion charge. This is not possible since the device is in weak inversion and the inversion charge is very small. Hence the change in band bending in the semiconductor and the oxide is minimal as one approaches the drain which is equivalent to saying that the lateral field in the channel is small. The very maximum voltage drop in the channel is ϕ_F to keep the channel in weak inversion throughout the channel but even this is not achievable when we consider the arguments based on the boundary condition placed by the gate as described above. The combination of very small voltage drop in the channel coupled with the small charge in the weak inversion layer makes drift currents minimal in the channel. Another way to physically understand the picture is to recognize that in the weak inversion regime the junction between the drain and the channel is closer to a reverse biased junction and hence absorbs most of the applied voltage as is shown in figure 9.24.

In this instance, the inversion charge in the channel in the absence of generation and recombination is obtained as the solution of the diffusion equation

$$n_{ch}(x) = n_{source}\left(1 - \frac{x}{L_{ch}}\right)$$

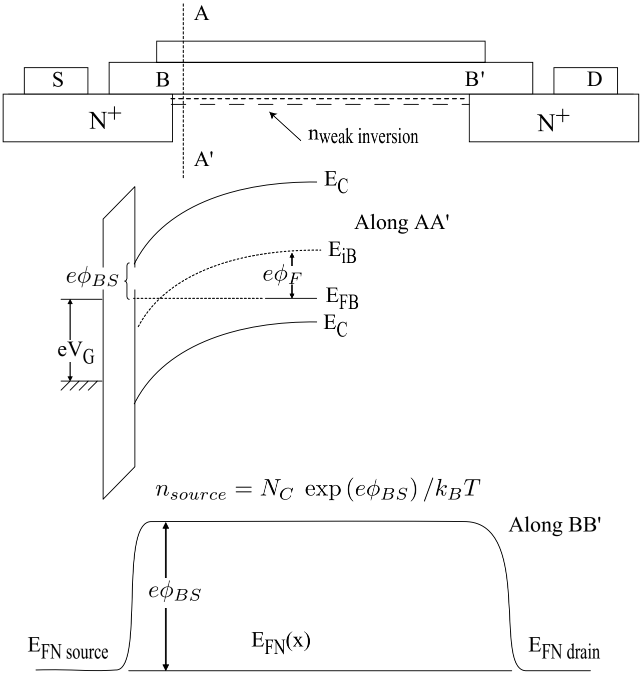


Figure 9.22: Band diagram as viewed along the gate-substrate axis, and along the source-drain axis with $E_{Fsource} = E_{Fchannel}$.

much like the electron profile in a narrow base bipolar transistor

$$\therefore I_{DS} = eD_n \frac{n_{source}}{L_{channel}}$$

the magnitude of n_{source} is limited by the barrier at the source end of the channel

$$n_{source} = N_C \ \exp\left(-e\phi_{BS}/k_BT\right)$$

or

$$I_{DS} = q\frac{D_n}{L_n}N_C \ \exp\left(-e\phi_{BS}/k_BT\right)$$

to get I vs. $(V_g - V_{th})$ we need to relate ϕ_{BS} to $(V_g - V_{th})$ which is readily done by analyzing the band diagram along AA'.

$$V_g - V_{fb} = \psi_s + \frac{1}{C_{ox}}\sqrt{2\epsilon_s e N_A \psi_s}$$

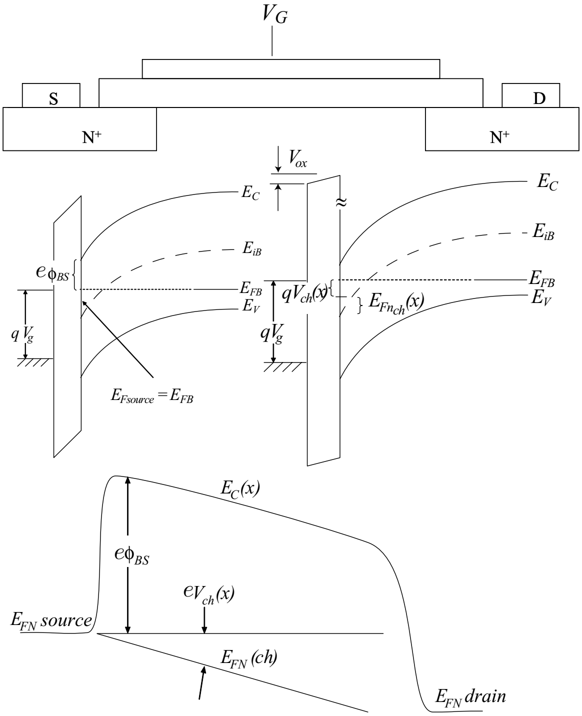


Figure 9.23: Band diagrams for a nMOSFET at the source, drain, and along the direction of transport for a device in the drift regime (for illustrative purposes only).

at threshold

$$V_{th} - V_{fb} = 2\phi_F + \frac{1}{C_{ox}}\sqrt{2\epsilon_s e N_A \left(2\phi_F\right)}$$

below threshold,

$$V_g - V_{fb} = \psi_S \quad (\phi_F < \psi_S < 2\phi_F) \quad + \frac{1}{C_{ox}}\sqrt{2\epsilon_s e N_A \psi_s}$$

$$\therefore (V_g - V_{th}) = \psi_s - 2\phi_F \simeq -\phi_{BS}$$

where we have neglected the terms in the square root

$$\therefore I_{DS} = \frac{eD_n N_C}{L_{Ch}} \exp\left(\frac{V_G - V_{th}}{k_B T}\right)$$

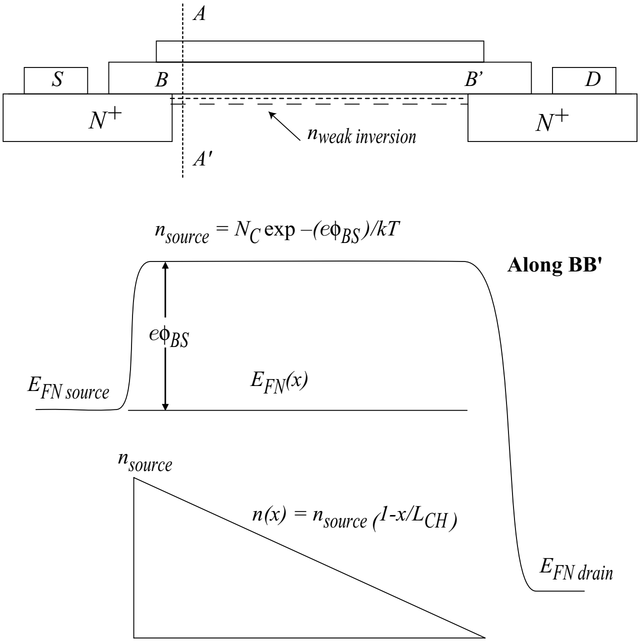


Figure 9.24: Band diagrams for a nMOSFET at the source, drain, and along the direction of transport for a device where diffusion current dominates

where $V_g < V_{th}$. This gives us the desired subthreshold slope in current of 60 mV/decade. Deviations from this can occur if

1. Charge sharing occurs i.e. the gate charge is not imaged in the semiconductor but on the electrodes as well (short channel effect, or traps in the system)

2. C_{ox} is small or the aspect ration is small

3. Voltage division occurs - for example due to poor contacts

4. Gate leakage occurs.

5. Leakage through the buffer occurs.

Figure 9.25 shows the impact that non-idealities have on subthreshold leakage.

$$\text{Band to Band tunneling} \propto \exp\left(\frac{V_{bias}}{E_G}\right)$$

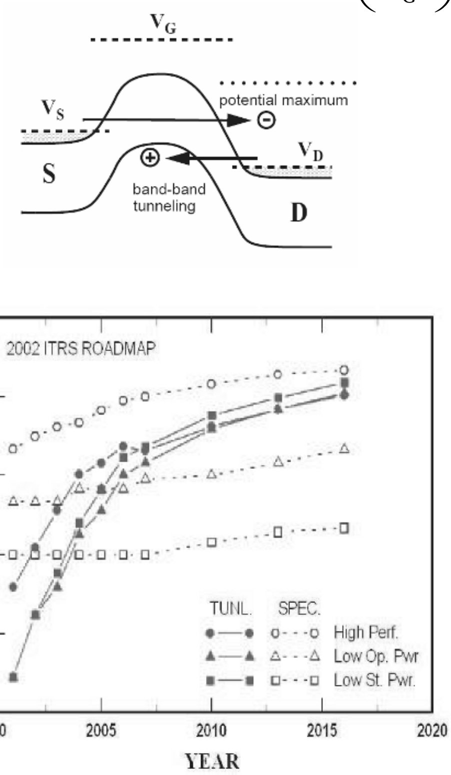


Figure 9.25: As gate lengths in MOSFETs are reduced, subthreshold leakage increases due to drain induced barrier lowering (DIBL) and band-to-band tunneling, as illustrated at the top of this figure. At the bottom, we show the ITRS roadmap for subthreshold leakage in future devices. Illustrations from Solomen et. al., IEDM 2003.

9.6.2 Mobility Variation with Gate Bias

In our simple model for carrier transport, we regarded the carrier mobility as having no dependence upon the gate bias. As the gate bias is changed the electron density in the channel changes. The electron density in turn is related to the surface field $\mathcal{E}_s$ normal to the channel by equation 9.3.5. Thus, if the sheet charge n_s increases the surface field also increases. The increased electric field forces electrons closer to the Si-SiO$_2$interface. As a result, the electrons suffer a greater degree of scattering from the interface roughness and oxide impurities, and the mobility degrades.

Mobility Variation with Channel Field

The mobility of electrons (holes) in silicon is not independent of longitudinal field as well, but is high at low field and becomes smaller at high fields where the velocity saturates. Velocity saturation typically occurs because at higher fields the rate of phonon emission increases and the rate of energy gained form the electric field equals the rate of energy loss to the crystal primarily through phonons. This phenomenon has a classical analog in the terminal velocity achieved by a person in a parachute or rain drops etc.. As a result, the current calculated by our simple model is much larger than the current observed in real devices. More realistic device modeling approaches use a more accurate description of the velocity-field relationship. A common expression used for the velocity-field relation is (see figure 9.26a)

$$v(\mathcal{E}) = \frac{\mu\mathcal{E}}{1 + \frac{\mu\mathcal{E}}{v_s}} \tag{9.6.1}$$

where v_s is the saturation velocity ($\sim 10^7$ cm/s) and $\mathcal{E}$ is the local longitudinal field in the channel. Use of this expression in calculating drain current causes a reduction in current by a factor of $\sim (1 + \mu V_{DS}/v_s L)$.

In figure 9.26 we show a comparison of the current-voltage relations calculated using the constant-mobility model and the more accurate saturation velocity model.

Channel Length Modulation in Saturation Region

In our simple model, once V_{DS} exceeds $V_D(sat)$ and the channel pinches off at the drain end, the current is assumed to remain independent of V_{DS}. The current in the channel is inversely proportional to the channel length. We have so far assumed that the channel length is the metallurgical channel length. However, the L that appears in the current-voltage relation represents the distance under the gate from the source side to the pinch-off point, as shown in figure 9.27a. As V_{DS} increases beyond $V_D(sat)$, the pinch-off point comes closer to the source side, thus effectively decreasing the channel length. This produces a change in the channel length ΔL (see figure 9.27b) and the current increases as

$$I_D = \frac{L}{L - \Delta L(V_{DS})} I_D(sat) \tag{9.6.2}$$

where $I_D(sat)$ is the current calculated assuming a fixed channel length. The effect results in an increase in the output conductance of the device. A similar effect occurs in MESFETs and JFETs. It is common to represent the increase in drain current arising from channel-length modulation by an expression

$$I_D = I_D(L = \text{fixed})(1 + \lambda V_{DS}) \tag{9.6.3}$$

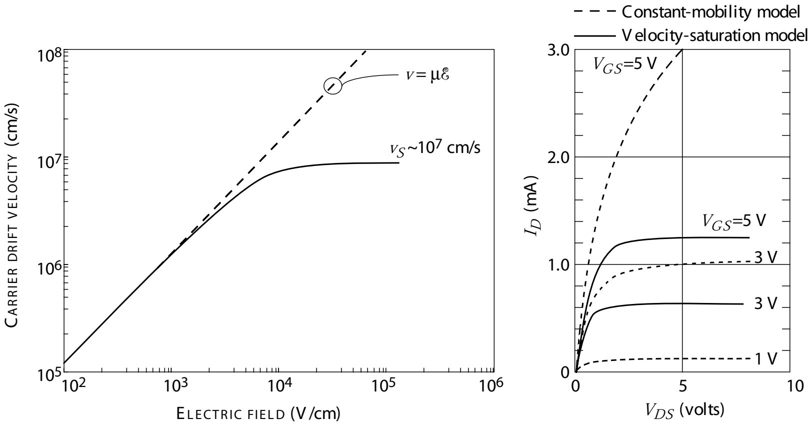


Figure 9.26: (a) Velocity-field relation for the constant-mobility model and saturation-velocity model. (b) $I_D - V_{DS}$ relations for a MOSFET using the constant-mobility model and the more accurate saturation-velocity model.

where $I_D(L = \text{fixed})$ is the current calculated assuming the channel length is fixed.

To a first approximation we can evaluate the change in effective channel length by assuming that the excess potential ΔV_{DS} falls across the region L. This gives

$$\Delta L = \sqrt{\frac{2\epsilon}{eN_a}}\left[\sqrt{\phi_{fb} + V_{DS}(sat) + \Delta V_{DS}} - \sqrt{\phi_{fb} + V_{DS}(sat)}\right] \qquad (9.6.4)$$

where

$$\Delta V_{DS} = V_{DS} - V_{DS}(sat)$$

This is also referred to as V_{dp} in the Grebene and Ghandhi analysis presented in chapter 8 on FETs and is only defined for $V_{ds} > V_{ds}(sat)$ Following the analysis of chapter 8 the drain resistance

$$r_d = \frac{\Delta V_{DS}}{\Delta I_{DS}} \text{ or } r_d = \frac{\Delta V_{dp}}{\Delta I_{DS}}$$

is given by

$$r_d = \frac{\pi V_{dp}}{2I_D}\left(\frac{L}{\tilde{d_{ox}}}\right)$$

where $V_{dp} = V_{DS} - V_{DS}(SAT)$, and $\tilde{d_{ox}}$ is $\epsilon/\epsilon_{ox} \cdot d_{ox}$ the equivalent oxide thickness. This emphasizes the need to reduce oxide thickness as we shrink the gate length, L; a high aspect ratio design.

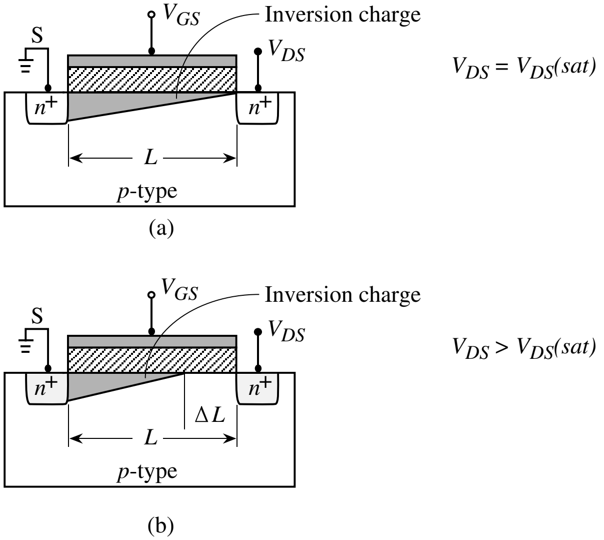


Figure 9.27: (a) A schematic of the MOSFET channel when $V_{DS} = V_{DS}(sat)$. (b) A schematic showing the decrease in the effective channel length for $V_{DS} > V_{DS}(sat)$.

9.6.3 Important Effects in Short-Channel MOSFETs

Advances in lithographic techniques are allowing MOSFET channel lengths to shrink to sizes below 1.0 μm. Experimental devices with channel lengths smaller than 0.1 μm have been fabricated. The force for miniaturization is coming from the need for dense circuits for high-density memory and logic applications as well as from the need for high-frequency microwave devices. In short-channel devices the simple models we developed for the current-voltage characteristics become quite invalid for quantitative description. In addition to the effects discussed in the previous subsection for long-channel devices, specific issues relating to short-channel devices also play an important role. Important issues that need to be considered are V_T lowering; surface scattering, velocity saturation and overshoot; hot carrier generation, impact ionization and drain induced barrier lowering and punch through. Some of these are now discussed.

Gate Leakage

As gate lengths are reduced, gate oxide thickness must also be reduced to maintain a constant aspect ratio. Currently, the SiO_2 gates in MOSFETs are only a few monolayers thick. Future devices will require high-K dielectric gates, as shown in figure 9.30, which allows for the physical thickness to be large while maintaining a small equivalent oxide thickness, $\tilde{d}_{ox}$

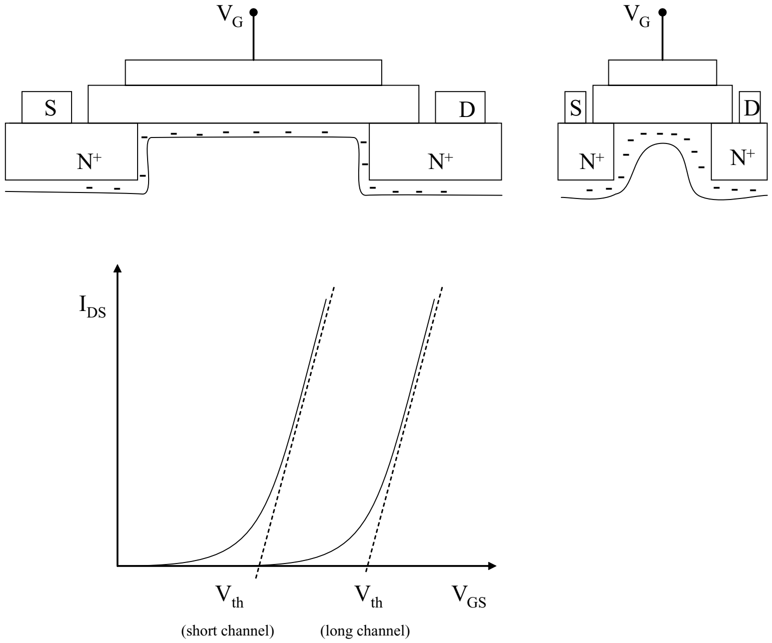


Figure 9.28: Schematic of long and short channel MOSFETs (above) with corresponding current-voltage characteristics below.

Three-Dimensional Transport

In our simple model for the MOSFET current, we assumed that the current flow was one-dimensional and we could use the gradual channel approximation. For a very short-channel device, the current flow is not just parallel to the gate, but one has to consider the current flow from the source and drain side, which is highly two-dimensional. Also, if the gate width Z is small, the transport becomes truly three-dimensional, requiring enormous computations to do a proper device simulation.

Charge Sharing and VT lowering (Drain Induced Barrier Lowering)

It is observed that the threshold voltage of a MOSFET becomes increasingly negative as the gate length of the device shrinks with all other parameters remaining the same. In conventional analysis there is no dependence of V_T on the gate length or channel length. This is because conventional analysis assumes that the band bending in the semiconductor and hence the onset

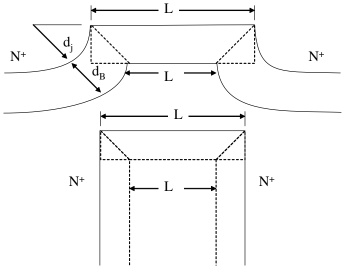


Figure 9.29: Schematic of the channel of a short-channel MOSFET showing the definitions for relevant length parameters.

of strong inversion (threshold) is determined by the 1-dimensional potential distribution from the gate to the substrate. This in effect neglects the effect of the source and drain contacts on the charge and hence band bending in the channel. Figure 9.28 shows the impact of the contacts on the band bending in the channel. The depletion due to the source and drain contacts encroaches substantially under the gate, increasing the band bending and hence decreasing the additional gate voltage required to create strong inversion compared to the long channel case. This is shown in figure 9.29 where the source and drain regions are assumed to be cylindrical with radius d_j and the depletion depth of extent d_B. At strong inversion the conduction band at the surface is close to the source and hence the surface band bending is similar to band bending at the $n^+ - p$ junctions giving a uniform value of d_B for small values of V_{DS}. The amount of charge that images on the gate electrode is assumed under a trapezoidal approximation to be:

$$Q_B' = -eN_A d_B \left(\frac{L + L'}{2} \right)$$

In the long channel case:

$$Q_B = -eN_A d_B L$$

or the charge in the shaded regions image on the gate and not on the contacts. Thus the reduced bulk charge is the source of the reduced threshold voltage from

$$V_T = 2\phi_F + V_{fb} - \frac{Q_B}{C_{ox}}$$

and

$$\hat{V}_T = 2\phi_F + V_{fb} - \frac{Q_B'}{C_{ox}}$$

or

$$\Delta V_T = \frac{Q_B - Q_B'}{C_{ox}} = \frac{Q_B}{C_{ox}} \left[1 - \frac{Q_B'}{Q_B} \right]$$

geometrical analysis gives

$$\frac{Q'_B}{Q_B} = 1 - \frac{d_j}{L}\left[\sqrt{1+\frac{2d_B}{d_j}} - 1\right]$$

As L increases, $Q'_B/Q_B \rightarrow 1$ approaching the long channel case as expected. As d_B/d_j becomes small (the case for large d_j) which is shown in figure 9.29 then

$$\frac{Q'_B}{Q_B} = 1 - \frac{d_B}{L}$$

In general

$$\frac{Q'_B}{Q_B} - 1 - \frac{\beta_1 d_B}{L}$$

this leads to

$$\Delta V_T = 2\beta_1 \frac{\epsilon_s}{\epsilon_{ox}} \frac{t_{ox}}{L} \left(2\phi_F + V_{BS}\right)$$

when V_{BS} is the substrate bias and β_1 is a parameter based on specific geometry.

Hot Electron Effects

As the channel lengths shrink, the electric fields in the channel increase if the supply voltages are kept fixed. The carriers become very "hot," i.e., they acquire higher kinetic energies than the thermal energy in such devices. These hot electrons can be injected into the oxide barrier causing a tunneling gate current. They can also cause deterioration of the device by breaking bonds in the semiconductor-oxide interface region or causing oxide charging. This damage is especially dangerous since over a period of time the device degrades and eventually the circuit based on the device loses its functionality. High fields also cause impact ionization near the drain end of the channel. To avoid hot electron devices, MOSFETs are being designed so that the electric field does not become very large in any region of the device.

9.6.4 Parasitic Bipolar Transistors and Latch-up in CMOS

CMOS circuits, while having the important benefit of low power consumption, have an important undesired property. This effect, known as latch-up, results from the presence of parasitic bipolar transistors present in integrated circuits. In figure 9.31 we show the origins of the parasitic bipolar transistors in a CMOS structure. We can see that in the CMOS structure there is an npn bipolar transistor and a pnp transistor in close proximity. As can be seen, the npn and pnp transistors form a positive feedback circuit. The resistances R_1, R_2, R_3, and R_4 are parasitic resistances associated with the n-substrate and p-well regions, as shown.

If we examine the two-terminal current between A and B as a function of bias, we find that up to a certain bias, V_L, the current is very low ($\sim \mu$A range). However, above this critical voltage V_L (related to the punch through of the transistor, typically $\sim$ 10 V), the two transistors start to conduct and the current rises abruptly to the level of milliamperes. The current is now controlled by the resistors R_3 and R_4. This phenomenon is called latch-up. Latch-up can occur whenever the voltages applied to input or output cause forward biasing of pn junctions in the devices and

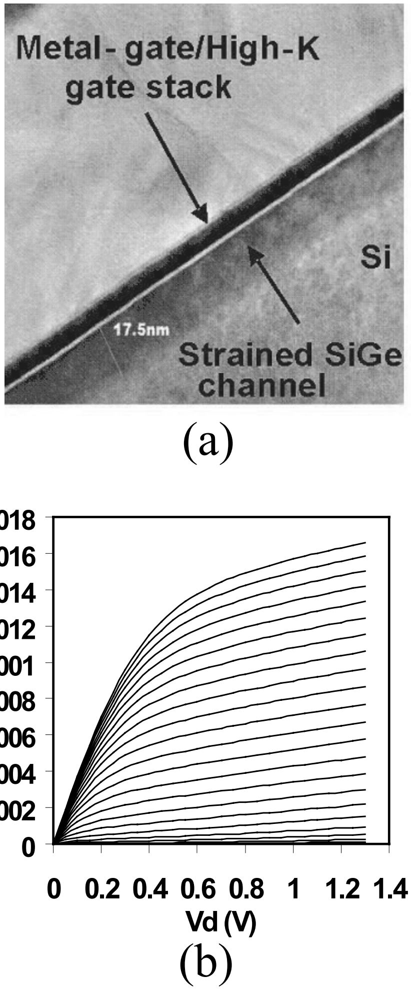


Figure 9.30: (a) TEM of MOSFET structure employing a high-K gate dielectric and a strained SiGe channel. (b) Device I-V characteristics. Figures courtesy of R. Chau, Intel corp.

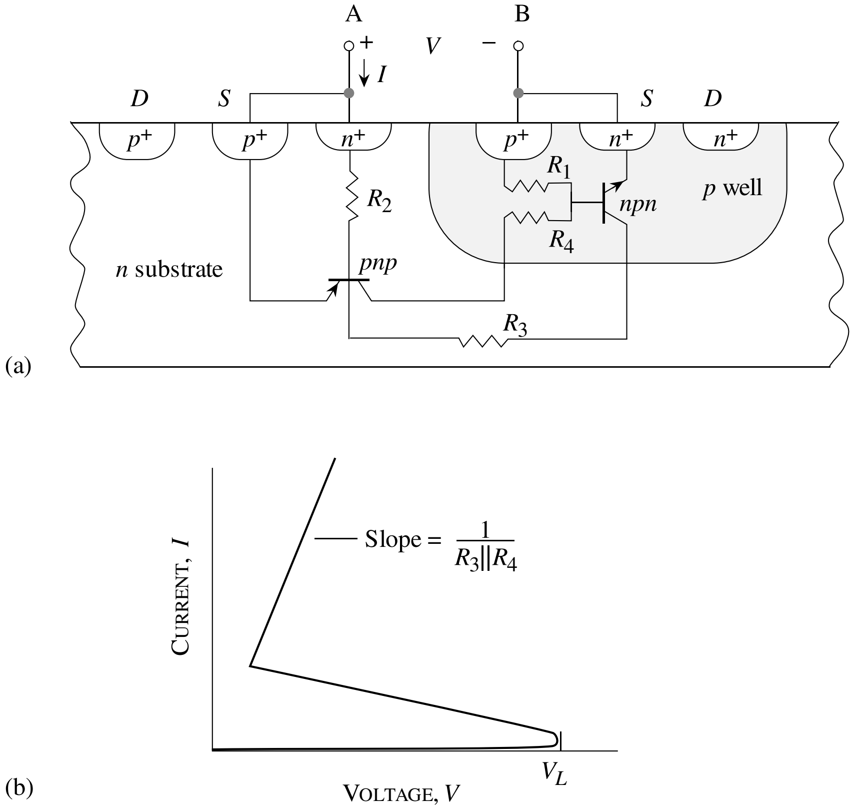


Figure 9.31: (a) a schematic of the parasitic effects that lead to CMOS latch-up problems. (b) current versus voltage effect. the onset of latch-up is represented by a sharp rise in the parasitic current.

can cause permanent damage to the chips. To avoid latch-up it is important that device design be such that the bipolar transistor gain is low.

9.7 SUMMARY

In this chapter we have discussed the basic operating principles of one of the most important devices in solid state electronics. The MOS capacitor and the MOSFET are key devices in almost

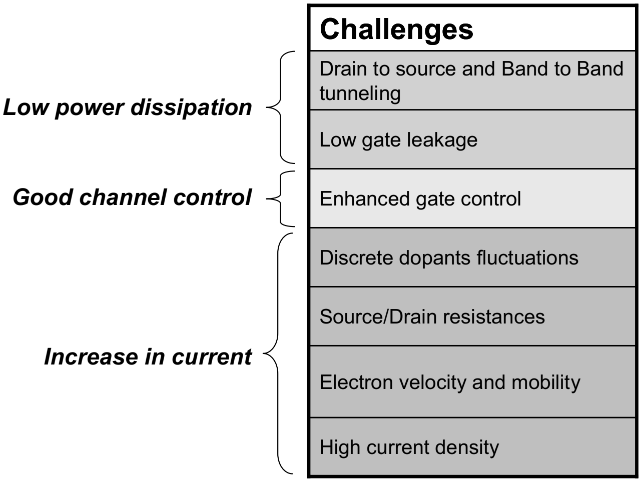


Figure 9.32: Challenges to the future of MOSFETs.

all electronic components. As devices continue to be scaled new challenges continue to be faced as is summarized in figure 9.32. The solutions will come in the form of high K dielectrics, structures with enhanced gate control such as the FINFET or Tri-gate structures, and probably new semiconductors such as GaAs, InGaAs, and InSb based MOSFETs. The future direction is truly unpredictable and therefore very exciting for research.

9.8 PROBLEMS

Assume a temperature of 300 K unless explicitly stated otherwise.

- **Section 9.3**

Problem 9.1 Calculate the maximum space charge width W_{max} in p-type silicon doped at $N_a = 10^{16}\ \mathrm{cm}^{-3}$and at $10^17\ \mathrm{cm}^{-3}$.

Problem 9.2 A p-type silicon has a uniform doping of $N_a = 10^{16}\ \mathrm{cm}^{-3}$. Calculate the surface potential needed to cause strong inversion.

Problem 9.3 A 50 Å oxide is grown on p-type silicon with $N_a = 5 \times 10^{15}\ \mathrm{cm}^{-3}$. Assume that the oxide charge is negligible and calculate the surface potential and gate voltage to create inversion at the surface. Calculate the value of W_{max} for the device. The flat band voltage is -1.0 V.

Problem 9.4 An Al-gate MOS capacitor has an oxide thickness of 100 Å and an oxide charge density of 3×10^{11} cm^{-2}. The charge is positive. Calculate (a) the flat band voltage, (b) the turn-on voltage. Also, draw the energy band diagram and electric field profile of the structure at the onset of inversion. N_a = $5\times10^{15}\mathrm{cm}^{-3}$.

Problem 9.5 An Al-gate transistor is fabricated on a p-type substrate with an oxide thickness of 600 Å. The measured threshold voltage is V_T = 1.0 V, and the $p-$type doping is $5\times10^{16}\mathrm{cm}^{-3}$. Calculate the fixed charge density in the oxide.

• Section 9.4

Problem 9.6 An n-channel MOS capacitor has a doping of $N_a = 10^{15}$ cm^{-3}. The gate oxide thickness is 500 Å. Calculate the capacitances C_{ox}, C_{fb}, and C_{min} for the capacitor.

Problem 9.7 Show that if $\rho(x)$ is the distribution of charge density in the $\mathrm{SiO_2}$2 region of thickness d_{ox}, the shift in the flat band voltage is given by

$$\Delta V_{fb} = -\frac{1}{C_{ox}}\int_0^{d_{ox}} \frac{x\rho(x)dx}{d_{ox}}$$

(Use Gauss' law for electric field due to a thin sheet of charge density. Then use the superposition principle.)

Problem 9.8 Calculate the shift in the flat band voltage using the result of problem 9.7 for the following oxide charge distributions: (a) $Q'_{ss} = 10^{11}$ cm^{-2} is at the Si-$\mathrm{SiO_2}$ interface; (b) the same charge is uniformly distributed in the oxide; (c) the charge is at the gate-$\mathrm{SiO_2}$ interface. The oxide thickness is 500 Å. Assume that the charge is positive.

Problem 9.9 The small signal capacitance of a $(Metal - SiO_2 - Si - Metal)$ MOS capacitor is equal to a series connection of two capacitors.
(a) One capacitor is formed by a plate in bulk Si and the other plate at the $SiO_2 - Si$ interface.
(b) The second capacitor, has its plates separated by the oxide layer.
Prove this using Gauss' law.

• Section 9.5

Problem 9.10 Consider an n-channel MOSFET with a Z/L ratio of 15, a threshold voltage of 0.5 volt, mobility, μ_n = 500 $\mathrm{cm^2/V\cdot s}$, and d_{ox} = 700 Å. Calculate the drain current and transconductance of the device (a) at V_{DS} = 0.2 V; (b) in the saturation region. The gate voltage is 1.5 V for both cases. Assume that the p-type doping is small.

Problem 9.11 Consider an ideal n-channel MOSFET with the following parameters:

Flat band voltage,	V_{fb}	=	-0.9 V
Channel width,	Z	=	25 μm
Channel mobility,	μ_n	=	450 $\mathrm{cm^2/\,V\cdot s}$
Channel length,	L	=	1.0 μm
Oxide thickness,	d_{ox}	=	500 Å
Channel doping,	N_a	=	5×10^{14} cm^{-3}

Calculate and plot I_D versus V_{DS} for $0 \leq V_{DS} \leq 5$ V and for V_{GS} values of 0, 1, 2, 3 volts. Also, draw the locus of the $V_D(sat)$ points for each curve.

Problem 9.12 Consider an ideal p-channel MOSFET with the following parameters:

Channel width,	Z	$=$	$25\ \mu\text{m}$
Channel mobility,	μ_p	$=$	$250\ \text{cm}^2/\ \text{V} \cdot \text{s}$
Channel length,	L	$=$	$1.0\ \mu\text{m}$
Oxide thickness,	d_{ox}	$=$	500 Å
Threshold voltage,	V_T	$=$	-0.8 V

Calculate and plot I_D vs. V_{DS} for $-0.5 \leq V_{DS} \leq 0$ V for a gate bias of $V_{GS} = 0, -1, -2, -3$ V. Assume that the background doping is very small.

Problem 9.13 In the text we used the criterion that inversion occurs when $V_s = 2\phi_F$. Calculate the channel conductivity near the Si-Si0$_2$ interface for two MOS devices with the following parameters:

$$\begin{array}{lll} N_a = 5 \times 10^{13}\ \text{cm}^{-3} & \qquad & N_a = 5 \times 10^{15}\ \text{cm}^{-3} \\ \mu_n = 600\ \text{cm}^2/\ \text{V} \cdot \text{s} & \qquad & \mu_n = 600\ \text{cm}^2/\ \text{V} \cdot \text{s} \end{array}$$

The problem shows the rather arbitrary way of defining the inversion condition.

Problem 9.14 An n-channel and a p-channel MOSFET have to be designed so that they both have a saturated current of 5 mA when the gate-to-source voltage is 5 V for the n-MOS and -5 V for the p-MOS. The other parameters of the devices are:

Oxide thickness,	d_{ox}	$=$	500 Å
Electron mobility,	μ_n	$=$	$500\ \text{cm}^2/\ \text{V} \cdot \text{s}$
Hole mobility,	μ_p	$=$	$300\ \text{cm}^2/\ \text{V} \cdot \text{s}$
V_T for the n-MOS,		$=$	$+0.7$ V
V_T for the p-MOS,		$=$	-0.7 V

What is the Z/L ratio for the n-MOSFET and the p-MOSFET?

Problem 9.15 An n-channel MOSFET has the following parameters:

Oxide thickness,	d_{ox}	$=$	500 Å
p-type doping,	N_a	$=$	$10^{16}\ \text{cm}^{-3}$
Flat band voltage,	V_{fb}	$=$	-0.5 V
Channel length,	L	$=$	$1.0\ \mu\text{m}$
Channel width,	Z	$=$	$15\ \mu\text{m}$
Channel mobility,	μ_n	$=$	$500\ \text{cm}^2/\ \text{V} \cdot \text{s}$

Plot $\sqrt{I_D(sat)}$ versus V_{GS} over the range $0 \leq I_D(sat) \leq 1\ mA$ for the source-to-body voltage of V_{SB} = 0, 1, 2 V.

Problem 9.16 Consider a p-channel MOSFET with oxide thickness of 500 Å, and $N_d = 10^{16}$ cm^{-3}. Calculate the body-to-source voltage needed to shift the threshold voltage from the V_{BS} = 0 results by -1.0 V.

Problem 9.17 A NMOS with V_T of 1.5 V is operated at V_{GS} = 5 V and I_DS = 100 μA. Determine if the device is in linear or saturation regime.

$$k = \frac{\mu Z C_{ox}}{L} = 20\ \mu A/V^2$$

Problem 9.18 In the text we considered a criterion for inversion $V_s = 2\phi_F$. Consider another criterion that asserts that inversion occurs when the channel conductivity near the interface is $0.1(\Omega\text{cm})^{-1}$. Calculate the surface potential bending needed to satisfy this criterion when the channel has a p-type doping of: (a) 10^{14} cm^{-3}; (b) 10^{15} cm^{-3}; (c) 10^{16} cm^{-3}. Compare the surface band bending arising from this new criterion to be a value of V_s given by the criterion used in the text (μ_n = 600 cm^2/V·s).

Problem 9.19 Threshold bias for an n-channel MOSFET: In the text we used a criterion that the inversion of the MOSFET channel occurs when $V_s = \psi_s = 2\phi_F$ where $e\phi_F = (E_{Fi} - E_F)$. Consider another criterion in which we say that inversion occurs when the electron density at the Si/SiO$_2$ interface becomes 10^{16} cm^{-3}. Calculate the gate threshold voltage needed for an MOS device with the following parameters for the two different criteria:

$$\begin{aligned} d_{ox} &= 500\ \text{Å} \\ \phi_{ms} &= 1.0\ \text{V} \\ N_a &= 10^{13}\ \text{cm}^{-3} \end{aligned}$$

Problem 9.20 A frequently needed quantity in experimental studies of MOS transistors is ψ_S, the surface potential.
(a) Show that when the gate voltage V$_G$ is changed in a MOS capacitor biased in the depletion region, it is possible to find the corresponding change in ψ_S, by using the measured capacitance of the MOS system. The change is calculated from the relation

$$\psi_S(V_{G2}) - \psi_S(V_{G1}) = \int_{V_{G1}}^{V_{G2}} (1 - \frac{C}{C_{ox}}) dV_G \qquad (9.8.1)$$

(b) If V$_{G1}$ is taken as V$_{FB}$ (flat-band voltage), sketch a low frequency MOS capacitance curve for p-type silicon bulk. Normalize it to C$_\infty$ and indicate by shading an area of the curve equal to $\Delta\phi_S$.

Problem 9.21 Consider the Si MOSFET structure in figure 9.33. Calculate the threshold voltage when the p-type region is doped at 10^{17}cm^{-3} uniformly as shown in Fig. 9.26b. Because of problems during processing, I lose Boron atoms from 50 nm of the Si and it gets magically incorporated uniformly in the oxide and provides unit negative charge per atom there. The resultant doping in the Si is shown below in Fig. 9.26c. Calculate the new threshold voltage V$_{TH}$ of the structure. Assume ϕ_{MS} = 0 eV.

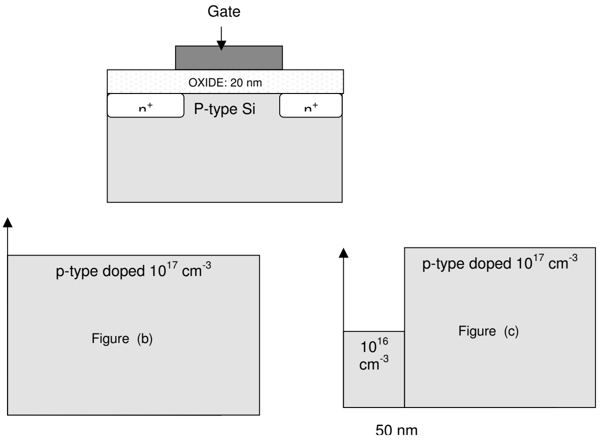


Figure 9.33: Figure for problem 9.21.

Problem 9.22 Consider that a MOS system on p-type silicon is biased to deep depletion by the sudden deposition of a total charge Q_G on the gate at time $t = 0$. Carrier generation in the space charge region at the silicon surface results in a charging current for the channel charge Q_n according to the net generation rate equation

$$J_G = \frac{qn_i x_i}{2\tau_0}$$

where τ_0 is the maximum recombination rate, and x_i is the width of the space charge region. This allows us to write

$$\frac{dQ_n}{dt} = -\frac{qn_i(x_d - x_{df})}{2\tau_0}$$

where x_d is the (time dependent) depletion region width at the surface. The quantity x_{df} is the space charge region width at thermal equilibrium; that is, when $x_d = x_{df}$, channel charging by generation is zero.

(a) Show that the time evolution of Q_n is governed by the differential equation

$$Q_n + \left(\frac{2\tau_0 N_A}{n_i}\right)\left(\frac{dQ_n}{dt}\right) = -(Q_G - qN_A x_{df})$$

(b) Solve this equation subject to the BC that $Q_n(t = 0) = 0$, and thus show that the characteristic time to form the surface inversion layer is of the order

$$T \sim \frac{2N_A\tau_0}{n_i}$$

9.9 DESIGN PROBLEMS

Problem 9.1 Consider an n-MOSFET made from Si-doped p-type at $N_a = 10^{16}$ cm^{-3} at 300 K. The source and drain contacts are ohmic (negligible resistance) and are made from

n^+-doped regions. The other parameters for the device are the following:

$$\begin{aligned}
V_{fb} &= -1.0\text{ V} \\
\mu_n &= 500\text{ cm}^2\text{ V}^{-1}\text{ s}^{-1} \\
\mu_p &= 100\text{ cm}^2\text{ V}^{-1}\text{ s}^{-1} \\
\text{Gate length} &= 2.0\ \mu\text{m} \\
\text{Gate width} &= 20.0\ \mu\text{m} \\
d_{ox} &= 500\text{ Å}
\end{aligned}$$

(a) Calculate the channel conductivity near the Si-SiO$_2$ interface under flat band condition and at inversion. Use the condition $V_s = 2\phi_F$ for inversion.
(b) Calculate the electron and hole densities at the Si-SiO$_2$ interface on the source and drain side of the gate when the gate bias is $V_T + 0.5$ V and $V_{DS} = 1.0$ V.
(c) Calculate the saturation current in the channel for the gate bias specified above.
(d) If the gate voltage is such that the Si bands are flat, $estimate$ the current density in in the channel for a drain bias of 1.0 V.

Problem 9.2 Consider an n-MOSFET made from Si-doped p-type at $N_a = 5 \times 10^{16}$ cm^{-3} at 300 K. The other parameters for the device are the following:

$$\begin{aligned}
V_{fb} &= -0.5\text{ V} \\
\mu_n &= 600\text{ cm}^2\text{ V}^{-1}\text{ s}^{-1} \\
\mu_p &= 100\text{ cm}^2\text{ V}^{-1}\text{ s}^{-1} \\
\text{Gate length} &= 1.5\ \mu\text{m} \\
\text{Gate width} &= 50.0\ \mu\text{m} \\
d_{ox} &= 500\text{ Å}
\end{aligned}$$

The inversion condition is $V_s = 2\phi_F$.
(a) Calculate the threshold voltage V_T.
(b) Calculate the channel current when the gate bias is $V_T + 1.5\ V$ and the drain bias is 1.0 V.
(c) Estimate the ratio of the electron velocities in the channel on the source side and the drain side of the gate for the biasing in part (ii).

Problem 9.3 Consider an n-MOSFET made from Si doped p-type at

$N_a = 5 \times 10^{16}$ cm^{-3} at 300 K. The other parameters for the device are the following:

$$\begin{aligned} V_{fb} &= -0.5 \text{ V} \\ \mu_n &= 600 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1} \\ \mu_p &= 200 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1} \\ \text{Gate length} &= 1.5\ \mu\text{m} \\ \text{Gate width} &= 50.0\ \mu\text{m} \\ d_{ox} &= 500 \text{ Å} \end{aligned}$$

The inversion condition is $V_s = 2\phi_F$. Assume that the electrons induced under inversion are in a region 200 Å wide near the Si/SiO$_2$ interface.
(a) Calculate the channel conductivity near the Si-SiO$_2$ interface under flat band condition and at inversion. Use the condition $V_s = 2\phi_F$ for inversion.
(b) Calculate the threshold voltage.

Problem 9.4 Consider an n-MOSFET at room temperature made from Si-doped p-type. To characterize the device C-V measurements are done for the MOS capacitor. It is found from the low-frequency measurements that the maximum and minimum capacitances per unit area are 1.72×10^{-7} F/cm^2 and 2.9×10^{-8} F/cm^2. The other parameters for the device are the following:

$$\begin{aligned} \mu_n &= 600 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1} \\ \text{Gate length} &= 1.5\ \mu\text{m} \\ \text{Gate width} &= 50.0\ \mu\text{m} \end{aligned}$$

(a) Calculate the oxide thickness.
(b) Estimate the p-doping level in the channel.
(c) Calculate the channel current at saturation when the gate bias is $V_T + 1.5\ V$.

9.10 FURTHER READING

- **General**

  - E. H. Nicollian and J. R. Brews, MOS Physics and Technology (Wiley, New York, 1982).
  - D. A. Neamen, Semiconductor Physics and Devices: Basic Principles (Irwin, Boston, MA, 1997).
  - R. F. Pierret, Field Effect Devices (Vol. 4 of the Modular Series on Solid State Devices, Addison-Wesley, Reading, MA, 1990).

- S. M. Sze, Physics of Semiconductor Devices (Wiley, New York, 1981).
- W. M. Werner, “The Work Function Difference of the MOS System with Aluminum Field Plates and Polycrystalline Silicon Field Plates” Solid State Electronics, **17**, 769-75 (1974).
- M. Zambuto, Semiconductor Devices (McGraw-Hill, New York, 1989).

Chapter 10

COHERENT TRANSPORT AND MESOSCOPIC DEVICES

10.1 INTRODUCTION

In quantum mechanics electrons are waves (or wavepackets) which have a discrete charge (1.6×10^{-19} C), amplitude and phase and have a spin ($1/2\hbar$). Yet none of the electronic devices we have considered explicitly use these features. Conventional electronic devices do not use the wave nature of electrons (e.g interference effects are not used), nor is the discrete nature of electron charge reflected in the current or conductance. The spin of electrons is also not directly used in diodes or transistors. There are several reasons for this. The devices are large so that scattering effects dominate and electron phase information is lost. Also the number of electrons is very large (say in billions or more) so that the discrete nature of electron charge is unimportant. Finally in traditional semiconductors there is no simple way to distinguish electron spin.

Charge transport in devices discussed so far is described within Born approximation or the Fermi golden rule. This involves free flight and scattering processes. While such an approach is quite relevant to modern microelectronic devices there are a number of important issues that are not described by this approach. These issues relate to the wave nature of the electrons, the discrete nature of charge in current flow and the spin of electrons. As semiconductor devices evolve and shrink, these issues are becoming increasingly important. In this chapter we will discuss some transport issues and devices that come into prominence as devices become smaller and smaller. In particular we will discuss devices that exploit electron phase, discrete electron charge and electron spin.

Let us recall how scattering is influenced by crystal quality and device dimensions. In figure 10.1 we show several types of structural properties of materials. In figure 10.1a we show a perfect crystal where there are no sources of scattering. Of course, in a real material we have phonon related fluctuations even in a perfect material. However, for short times or at very low temperatures it is possible to consider a material with no scattering. There are several types of transport that are of interest when there is no scattering: i) ballistic transport, where electrons

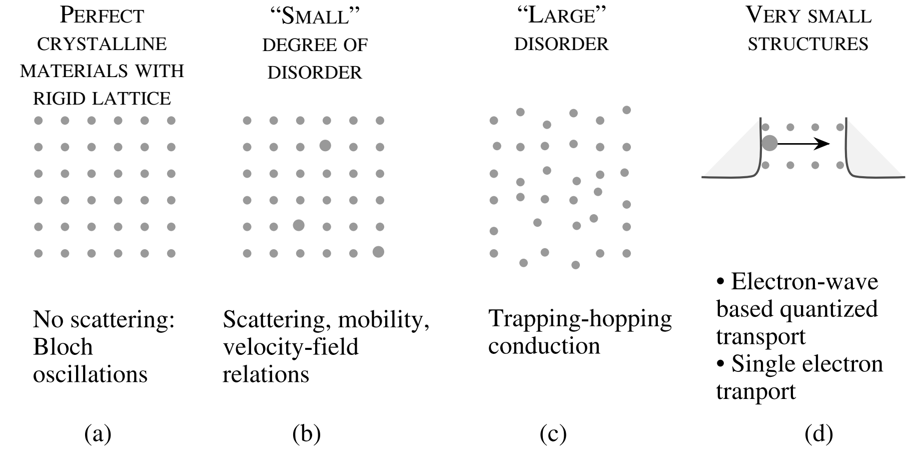

PERFECT CRYSTALLINE MATERIALS WITH RIGID LATTICE
"SMALL" DEGREE OF DISORDER
"LARGE" DISORDER
VERY SMALL STRUCTURES
No scattering: Bloch oscillations
Scattering, mobility, velocity-field relations
Trapping-hopping conduction
• Electron-wave based quantized transport
• Single electron tranport
(a)
(b)
(c)
(d)

Figure 10.1: A schematic of how levels of structural disorder and size impact electronic properties of a material.

move according to the modified Newton's equation; and ii) Bloch oscillations, where electrons oscillate in k-space as they reach the Brillouin zone edge, as will be discussed in section 10.2. In addition we can have tunneling type transport as well as quantum interference effects. These are discussed in Sections section 10.3 and section 10.4. The wave nature of electrons and the quantization of charge also leads to conductance quantization and Coulomb blockade effects. Finally if the spin of electrons can be manipulated novel devices can result.

In figure 10.1b we show the case where there is a small degree of disorder. This is the situation where Born approximation can be used and transport under these conditions has been discussed in the previous chapters. In figure 10.1c we show the case where the structural disorder is large. This happens in amorphous materials and leads to localized states (band tails) and transport that is described by "hopping" behavior. Transport in disordered semiconductors (or amorphous semiconductors) is relatively poor and used primarily for low cost applications such as thin film transistors for display. Such devices are not useful for high performance devices which are the primary focus of this text.

Finally in figure 10.1d we show the case for devices that are very small (several tens of atoms across). Such structures are called mesoscopic structures and are increasingly becoming important as fabrication technology improves. Mesoscopic structures have a number of very interesting and potentially important transport properties. Single electron effects as well as spin effects are manifested in such structures.

10.2 ZENER-BLOCH OSCILLATIONS

In a perfect crystal electrons see a periodic potential and according to Bloch theorem an electron wavefunction is described by a plane wave with a central cell periodic part. Of course the crystal has to be rigid since lattice vibrations even in a defect-free structure will cause scattering. There are many interesting effects that occur when electrons move without scattering in crystals. One such effect is Zener-Bloch oscillations. The equation of motion of electrons in an electric field is simply

$$\hbar \frac{d\boldsymbol{k}}{dt} = e\mathcal{E} \tag{10.2.1}$$

In the absence of any collisions the electron will simply start from the bottom of the band (figure 10.2) and go along the E vs k curve until it reaches the Brillouin zone edge.

It must be noted that just as the electron sees a periodic potential in real space in a crystal the bandstructure E vs $\boldsymbol{k}$ is also periodic in k-space. The electron at the zone edge is thus "reflected" as shown in figure 10.2 and now starts to lose energy in its motion in the field. The k-direction of the electron changes sign as the electron passes through the zone edge representing oscillations in k-space and consequently in the real space. These oscillations are called the Zener-Bloch oscillations.

If we have a spatial periodicity defined by distance a the bandstructure is periodic in the reciprocal vector $\Gamma = 2\pi/a$. As a result the frequency of Bloch-Zener oscillation is

$$\omega_b = \frac{e\mathcal{E}a}{\hbar} \tag{10.2.2}$$

The oscillation frequency is quite high and can easily be in the several terrahertz regime. Note that the oscillations depend upon field direction since the edges of the Brillouin zone (see chapter 3) are at different points along different directions. From a practical device point of view it has not been possible to exploit Bloch oscillations since the scattering mechanisms are usually strong enough to cause a electron to scatter before it can go through a complete oscillation, it has not been possible to observe these oscillations.

If τ_{SC} is the scattering time oscillations can occur if we have the condition

$$\omega_b \tau_{SC} \geq 1 \tag{10.2.3}$$

From the oscillation condition given above we see that if the periodic distance in real space is increased, it will take less time to reach the zone edge and one can expect Bloch oscillations to survive. The periodicity can be increased by using superlattices. In figure 10.3 we show a schematic of the effect of enlarging the periodic distance (by making superlattices) on an energy band. On the top we show the energy band schematic of a crystal with a unit cell periodicity represented by the distance a. The zone edge in k-space is at $2\pi/a$. Now if a superlattice with a period na is made as shown in the lower panel the zone edge occurs at $2\pi/na$. Assuming that the scattering time is not changed much due to superlattice formation, it can be expected that an electron will be able to reach the superlattice zone edge without scattering, thus Bloch-Zener oscillations could occur. Although these considerations seem promising, real devices have not been created.

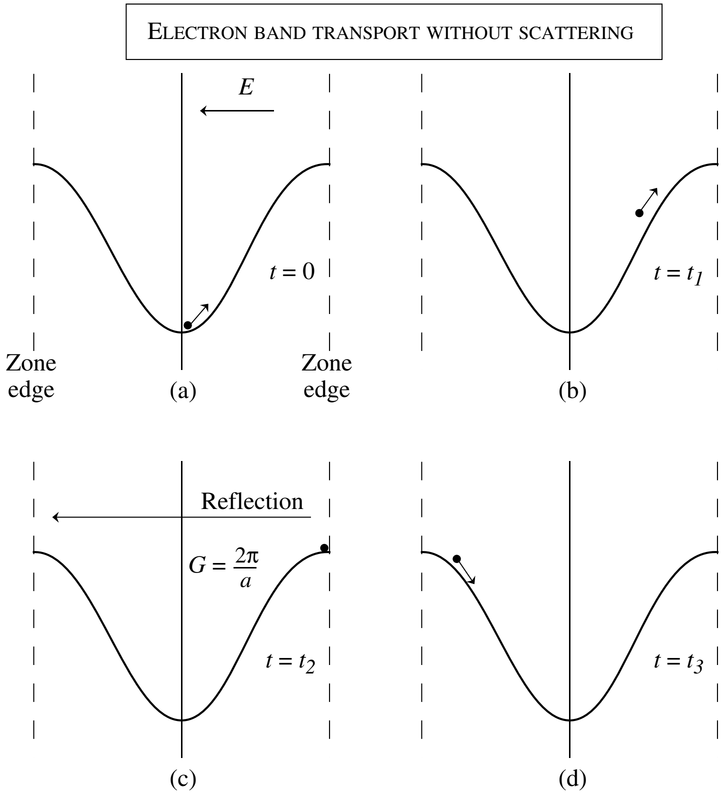


Figure 10.2: A schematic showing how an electron starting at $t = 0$ at the bottom of the conduction band (Γ-valley) travels up the E vs k diagram and gets reflected at the zone edge.

10.3 RESONANT TUNNELING

In absence of scattering the behavior of electron waves is similar to that of optical waves. Effects like filtering, interference and diffraction can occur. One class of devices that has been demonstrated and used for high performance applications is the one based on electron tunneling through heterostructures. Resonant tunneling is a very interesting phenomenon in which an electron passes through two or more classically forbidden regions sandwiching a classically allowed

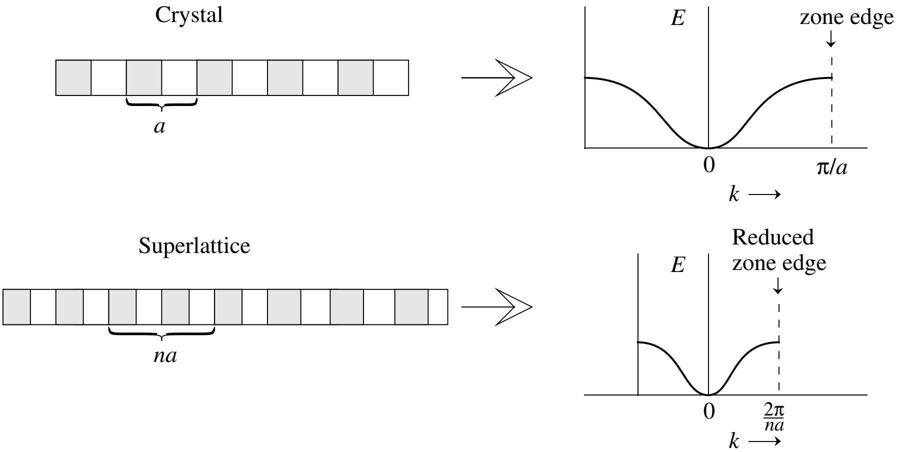


Figure 10.3: An increase in periodic spacing through the use of a superlattice can reduce the k-space an electron has to traverse before it reaches the zone edge. The reduced zone edge may allow the possibility of Bloch-Zener oscillations.

region. A particularly interesting outcome of resonant tunneling is "negative differential resistance." In Fig. figure 10.4a we show a typical potential profile for a resonant tunneling structure. As shown the double barrier structure of figure 10.4 has a quasi-bound ground state at energy E_0 as shown. The level E_0 is close to the level in the quantum well formed within the double barrier region but it is broadened due to the escape lifetime. The broadening comes from the Heisenberg energy-time uncertainty. If the electrons coming from the left have energies close to E_0 they are able to transmit through the structure. The operation of a resonant tunneling structure is understood conceptually by examining figure 10.4. At zero bias, point A, no current flows through the structure since the allowed level in the well is not aligned with the energy of electrons coming from the left.. At point B, when the Fermi energy lines up with the quasibound state, a maximum amount of current flows through the structure. Further increasing the bias results in the structure of point C, where the current through the structure has decreased with increasing bias (negative resistance). Applying a larger bias results in a strong thermionic emission current and thus the current increases substantially as shown at point D.

To understand the tunneling behavior, the potential profile (say, the conduction band lineup) is divided into regions of constant potential. The Schrödingerequation is solved in each region and the corresponding wavefunction in each region is matched at the boundaries with the wavefunctions in the adjacent regions as shown in figure 10.5.

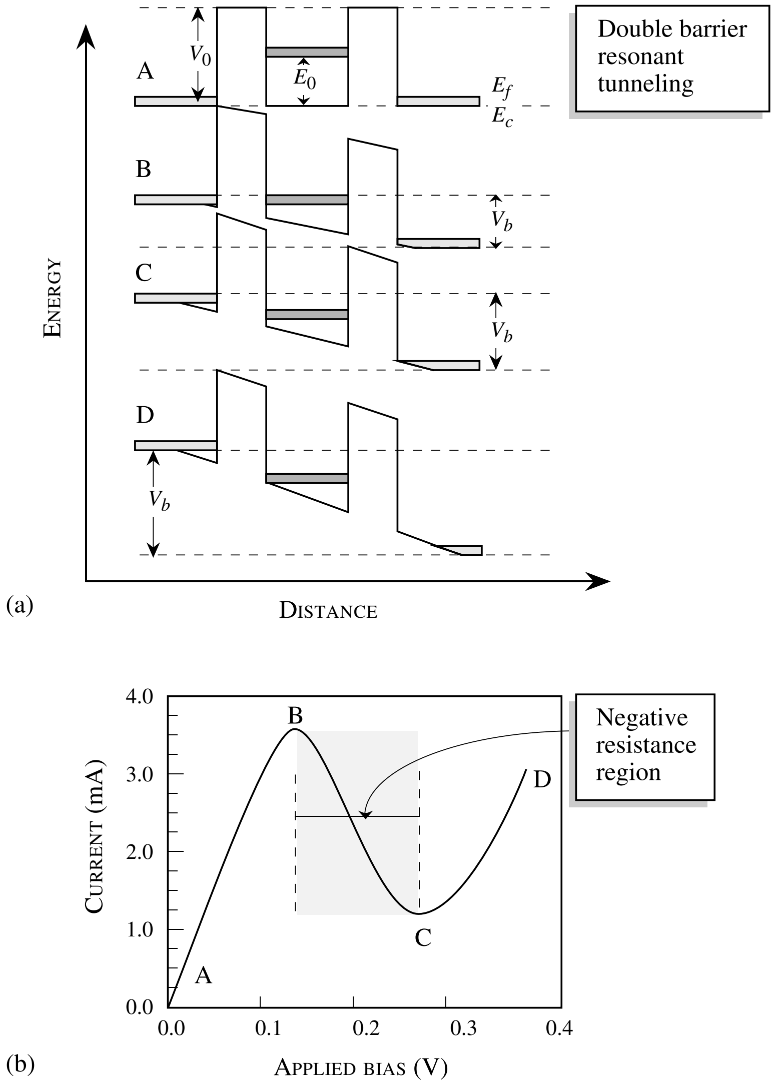


Figure 10.4: (a) A conceptual explanation of the operation of resonant tunneling devices showing the energy band diagram for different bias voltages. (b) Negative resistance region in the current–voltage characteristic for the resonant tunneling diode.

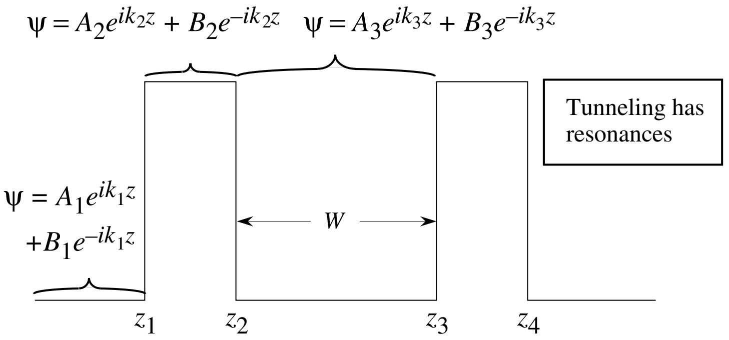

ψ = A₂e^{ik₂z} + B₂e^{−ik₂z}
ψ = A₃e^{ik₃z} + B₃e^{−ik₃z}
Tunneling has resonances
ψ = A₁e^{ik₁z} +B₁e^{−ik₁z}
W
z₁
z₂
z₃
z₄

Figure 10.5: Typical resonant tunneling structure with two barriers. The wavefunction in each region has a general form shown. By matching the wavefunctions and their derivatives at the boundaries one can obtain tunneling probabilities.

A simple application of this formalism is the tunneling of electrons through a single barrier of height V_0 and width a. The tunneling probability is given by

$$\begin{aligned} T_{1B}(E) &= \left|\frac{A_3}{A_1}\right|^2 \\ &= \frac{4E(V_0 - E)}{V_0^2 \sinh^2(\gamma a) + 4E(V_0 - E)} \end{aligned} \tag{10.3.1}$$

with

$$\gamma = \frac{1}{\hbar}\sqrt{2m(V_0 - E)} \tag{10.3.2}$$

If we have two barriers as shown in figure 10.4, the tunneling through the double barrier is given by

$$T_{2B} = \left[1 + \frac{4R_{1B}}{T_{1B}^2}\sin^2(k_1 W - \theta)\right]^{-1} \tag{10.3.3}$$

where R_{1B} is the reflection probability from a single barrier

$$R_{1B} = \frac{V_0^2 \sinh^2 \gamma a}{V_0^2 \sinh^2 \gamma a + 4E(V_0 - E)} \tag{10.3.4}$$

and θ is given by

$$\tan\theta = \frac{2k_1\gamma \cosh \gamma a}{(k_1^2 - \gamma^2)\sinh \gamma a} \tag{10.3.5}$$

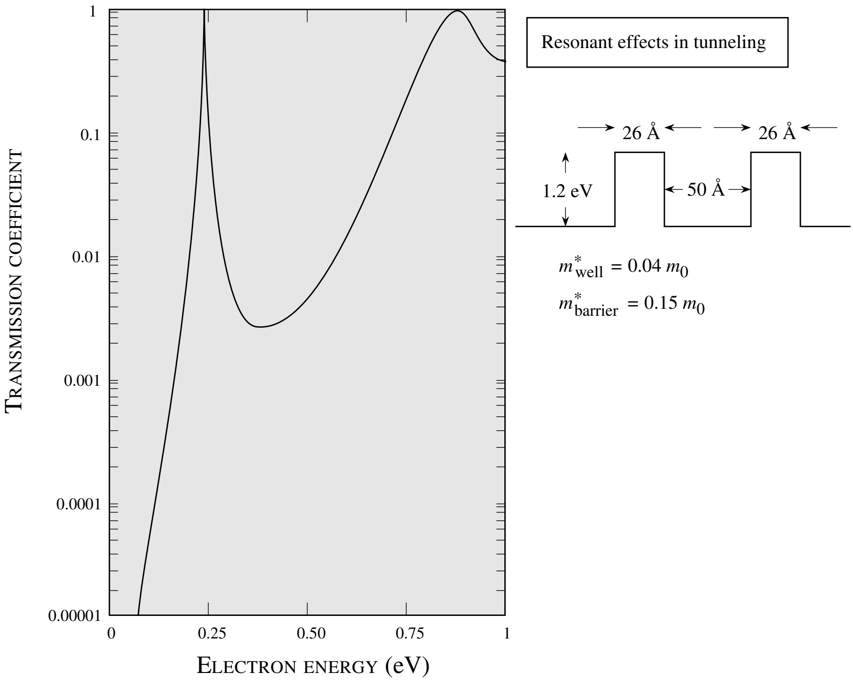


Figure 10.6: Transmission coefficient as a function of electron longitudinal energy for a double barrier structure

The wavevector k_1 is given by

$$\frac{\hbar^2 k_1^2}{2m^*} = E$$

While tunneling through a single barrier has no interesting feature, tunneling through a double barrier structure has interesting resonances as can be seen from the expression for T_{2B}. The calculated transmission probability as a function of longitudinal electron energy for a typical double barrier is shown in figure 10.6. The sharp peaks in the transmission probability correspond to resonant tunneling through the quasi-bound states in the quantum well formed between the two barriers. The tunneling probability reaches unity at energies corresponding to the quasi-bound states in the quantum well. To calculate the current density in the system we note that

$$\begin{aligned} \boldsymbol{J} &= ne\boldsymbol{v} \\ &= \frac{e}{4\pi^3\hbar}\int_0^\infty dk_\ell \int_0^\infty d^2k_t \left[f(E) - f(E^{'})\right] T(E_\ell)\frac{\partial E}{\partial \boldsymbol{k}_\ell} \end{aligned} \qquad (10.3.6)$$

where the longitudinal velocity is

$$\boldsymbol{v} = \frac{1}{\hbar}\frac{\partial E}{\partial \boldsymbol{k}_\ell}$$

and the net current is due to the electrons going from the left-hand side with energy E and from the right-hand side with energy $E^{'} = E + e\,|\boldsymbol{\mathcal{E}}|\,l = E + eV$ where $|\boldsymbol{\mathcal{E}}|$ is the electric field and l is the distance between the contacts on the two sides.

$$\begin{aligned} J &= \frac{e}{4\pi^3\hbar}\int d\boldsymbol{k}_\ell T(E_\ell)\frac{\partial E}{\partial \boldsymbol{k}_\ell}\int d^2\boldsymbol{k}_t\left[\frac{1}{\exp\left[(E_t+E_\ell-E_F)/k_BT\right]+1}\right. \\ &\quad \left. - \frac{1}{\exp\left[(E_t+E_\ell+eV-E_F)/k_BT\right]+1}\right] \end{aligned}$$

The transverse momentum integral can be simplified by noting that

$$\begin{aligned} d^2k_t &= k_t\ dk_t\ d\phi \\ &= \frac{m^*\ dE_t\ d\phi}{\hbar^2} \end{aligned}$$

This gives

$$\begin{aligned} J &= \frac{em^*}{2\pi^2\hbar^3}\int dE_\ell T(E_\ell)\int_0^\infty dE_t\left[\frac{1}{\exp\left[(E_t+E_\ell-E_F)/k_BT\right]+1}\right. \\ &\quad \left. - \frac{1}{\exp\left[(E_t+E_\ell+eV-E_F)/k_BT\right]+1}\right] \\ &= \frac{em^*}{2\pi^2\hbar^3}\int_0^\infty T(E_\ell)\ \ln\left[\frac{1+\exp\left[(E_F-E_\ell)/k_BT\right]}{1+\exp\left[(E_F-E_\ell-eV)/k_BT\right]}\right] dE_\ell \qquad (10.3.7) \end{aligned}$$

In figure 10.7 we show typical current-voltage characteristics measured in resonant double barrier structures. The results show are for a InGaAs/AlAs structure with parameters shown. As can be seen a large peak to valley current ration can be obtained at room temperature. There is a region of negative resistance as expected from simple arguments. The negative resistance can be exploited for microwave devices or for digital applications.

10.4 QUANTUM INTERFERENCE EFFECTS

In a perfectly periodic potential the electron wavefunction has the form

$$\psi_k(r) = u_k(r)e^{ik\cdot r}$$

and the electron maintains its phase coherence as it propagates in the structure. However, in a real material electrons scatter from a variety of sources. In high-quality semiconductors (the material of choice for most information-processing devices) the mean free path is $\sim$ 100 Å at room temperature and $\sim$ 1000 Å at liquid helium. For sub-micron devices it is possible to see quantum interference effects at very low temperatures in semiconductor devices. These effects

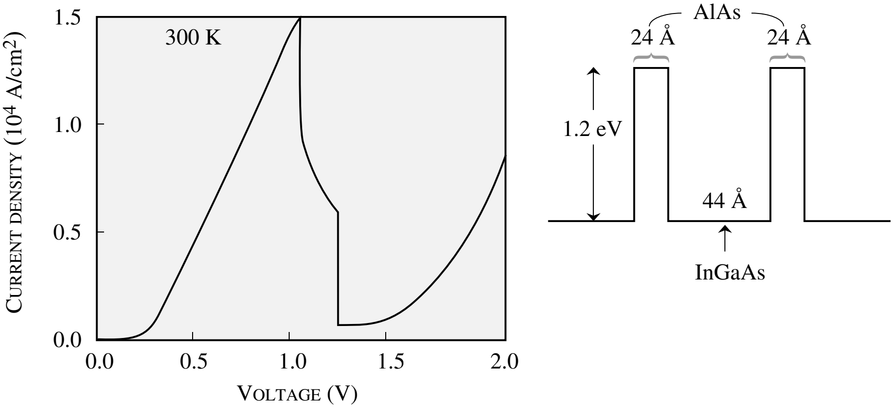


Figure 10.7: Room temperature current–voltage characteristics of an InGaAs/AlAs resonant tunneling diode.

can be exploited to design digital devices and switches operating at very low power levels. The general principle of operation is shown in figure 10.9. Electron waves travel from a source to a drain via two paths. At the output the intensity of the electron wave is (addition is coherent)

$$I(d) =\mid \psi_1(d) + \psi_2(d)^2 \tag{10.4.1}$$

If the waves are described by

$$\begin{aligned} \psi_1(x) &= Ae^{ik_1x} \\ \psi_2(x) &= Ae^{ik_2x} \end{aligned} \tag{10.4.2}$$

where k_1 and k_2 are the wavevectors of the electrons in the two paths. We have

$$I(d) = 2A^2[1 - \cos(k_1 - k_2)d] \tag{10.4.3}$$

If we can now alter the wavevectors of the electron (i.e., the value of $(k_1 - k_2)$) we can modulate the signal at the drain. This modulation can be done by using an electric bias to alter the kinetic energy of the electrons in one arm. In figure 10.8b we show a schematic of a split-gate device in which electrons propagate from the source to the drain either under one gate or the other. The ungated region is such that it provides a potential barrier for electron transport as shown by the band profile. Interference effects are then caused by altering the gate bias.

In quantum interference transistors, a gate bias is alters the potential energy seen by the electrons. The electron k-vector at the Fermi energy is given by (E_c is the bandedge i.e the subband

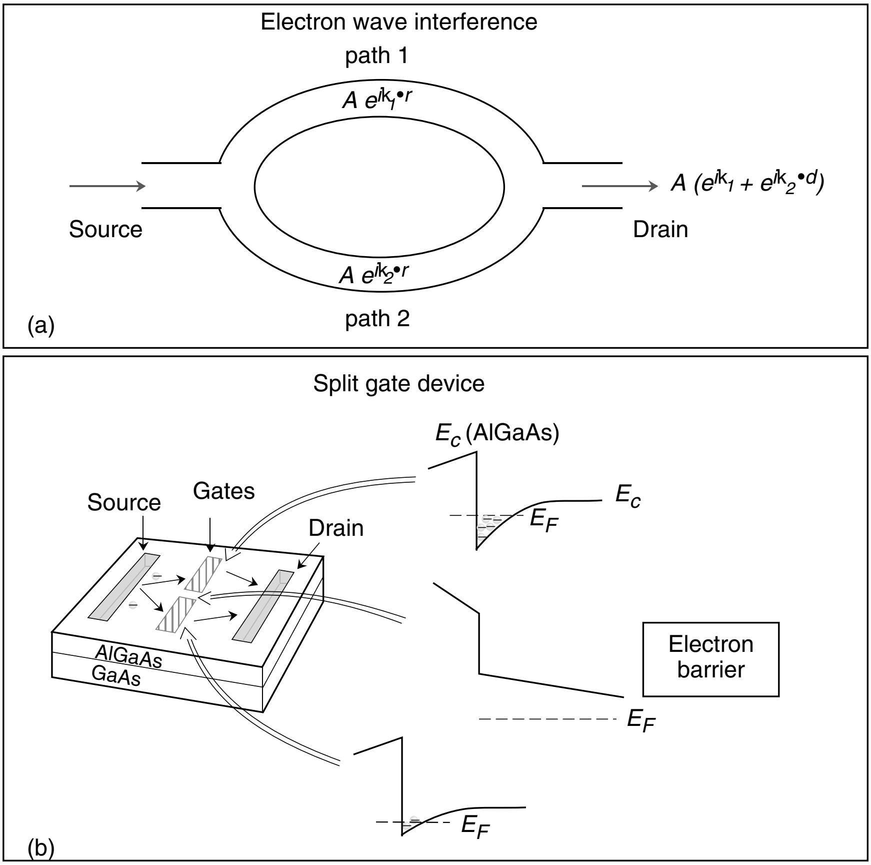


Figure 10.8: (a) A schematic of a coherent electron beam interference structure. (b) A schematic of a split-gate transistor to exploit quantum interference effects. Electrons propagate from the source to the drain under the two independently controlled gates in the 2-dimensional channel of AlGaAs/GaAs as shown.

energy in the quantum well)

$$E_F = E_c + \frac{\hbar^2 k^2}{2m^*} \tag{10.4.4}$$

By changing the position of E_F, one can alter the k-value. Thus one can develop quantum interference transistors.

10.5 MESOSCOPIC STRUCTURES

In mesoscopic structures single electron effects arising from phase coherence or from charge quantization become important. The structures are so small that density of states is not a continuous function but has discreteness to it. As a result mesoscopic structures show a number of interesting transport effects.

10.5.1 Conductance Fluctuations and Coherent Transport

In very small structures electron waves can flow from one contact to another maintaining phase coherence. Additionally the structures are so small that the change in electron number by unity creates observable effects. In structures that are $\sim$ 100–500 Å this occurs at low temperatures, since at high temperatures the random scattering due to phonons removes the coherence in the transport process. A dramatic manifestation of the phase coherence is the fluctuation seen in conductivity of mesoscopic structures as a function of magnetic field, electron concentration, etc.

The origin of the fluctuations can be understood on the basis of Landauer formalism which allows one to study transport in terms of the scattering processes directly. For simplicity consider a one-dimensional system with scattering centers. Each of these scatterers is characterized in terms of a transfer matrix which describes what fraction of the incident electron is "reflected" after scattering and what fraction is transmitted. The scatterer is described by the reflection and transmission coefficients shown in figure 10.9. The reflection and transmission coefficients are R and T for an incident wave from the left or right. We will provide a simple formulation to understand the origin of conductance fluctuations in mesoscopic structures. We assume a one dimensional flow of charge from one contact to another. This allows us to use 1-dimensional density of states to describe carrier density changes.

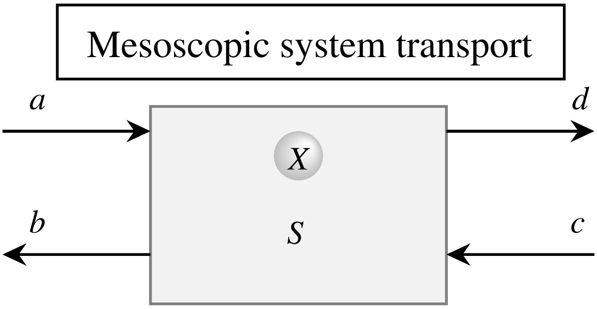


Figure 10.9: Transport in a mesoscopic structure. A schematic showing the effect of the scattering center S on electron waves a and c incident from the left and right respectively. The waves b and d emerge as a result of reflection and transmission.

In small structures where phase information is retained we will see that conductance has quantized behavior. The conductance is

$$G = \frac{\delta I}{\delta V} \tag{10.5.1}$$

Also

$$\delta I = \delta n.e.v_k \tag{10.5.2}$$

where the carrier velocity is

$$v_k = \frac{1}{\hbar}\frac{\delta E}{\delta k} \tag{10.5.3}$$

If δV is the potential change we have

$$\delta n = \frac{dn}{dE}(e\ \delta V) \tag{10.5.4}$$

Using these equations we get for a small change in current

$$\delta I = \frac{e^2}{\hbar}\frac{\delta n}{\delta}\delta V \tag{10.5.5}$$

The conductance is then

$$G = \frac{e^2}{\hbar}\frac{\delta n}{\delta k} \tag{10.5.6}$$

Now for a 1-dimensional case the number of electron states available per k-state is

$$\frac{dn}{dk} = \frac{1}{\pi} \tag{10.5.7}$$

so that (including the spin degeneracy factor of 2)

$$G = \frac{2e^2}{h} \tag{10.5.8}$$

The expression shows that the fundamental unit of conductance is $2e^2/h$. By using Landauer formalism where electrons are treated as incident, transmitted and reflected waves it can be shown that there is a remarkable universality in the magnitude of the fluctuations independent of the sample size, dimensionality and extent of disorder, provided the disorder is weak and the temperature is low (a few Kelvin). Such universal conductance fluctuations have been measured in a vast range of experiments involving magnetic field and Fermi level position (voltage).

In figure 10.10 we show experimental results of Wees, et al., carried out on a GaAs/AlGaAs MODFET at low temperatures. As shown, a pair of contacts are used to create a short channel of the high mobility region, and conductance is measured. The gates form a 1-dimensional channel in which the Fermi level and thus the electron wavefunctions can be altered. As can be seen from the figure 10.10, there are quantized steps in the conductance.

Transistors based on the mesoscopic effects described here are called single electron transistors. They promise low power operation although they require low temperatures. In figure 10.11 we show an SEM image of a single electron transistor where the conductance fluctuations discussed here have been observed.

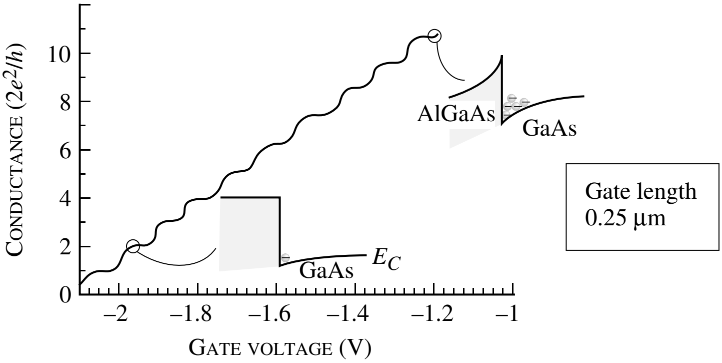


Figure 10.10: Experimental studies on conductance fluctuations arising in a GaAs/AlGaAs channel constricted by the structure shown. The results are for the channel conductance in units of $e^2/\pi\hbar\,(= 2e^2/h)$. (From the paper by B. J. van Wees, et al., Phys. Rev. Lett., **60**, 848 (1988).)

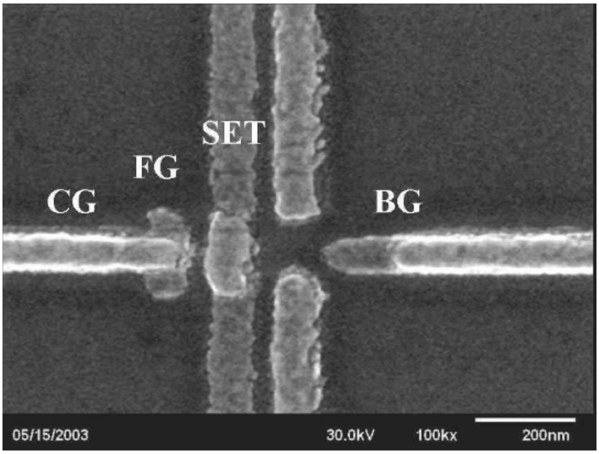


Figure 10.11: SEM image of a single electron transistor (SET) structure. Figure courtesy of Greg Snyder, University of Notre Dame.

10.5.2 Coulomb Blockade Effects

So far in our discussion we have not paid attention to the Coulombic repulsion between electrons. The reason is that in large systems the repulsion is negligible. However, in very small systems, electron charging energy effects arising from Coulomb interactions between electrons can become significant. This phenomenon is called the Coulomb blockade effect. We are familiar with the parallel plate capacitor with capacitance C and the relation between a charge increment ΔQ and the potential variation ΔV

$$C = \frac{\Delta Q}{\Delta V} \text{ or } \Delta V = \frac{\Delta Q}{C} \tag{10.5.9}$$

The capacitance is given by the spacing of the plates (d) and the area (A)

$$C = \frac{\epsilon A}{d} \tag{10.5.10}$$

Now consider a case where the capacitance decreases until a single electron on the capacitor causes a significant change in the voltage. The charging energy to place a single electron on a capacitor is

$$\Delta E = \frac{e^2}{2C} \tag{10.5.11}$$

and the voltage needed is

$$\frac{e}{2C} = \frac{80\, mV}{C(aF)} \tag{10.5.12}$$

where the capacitance is in units of 10^{-18} F(aF). If we write the charging energy as a thermal energy, k_BT_0, the temperature associated with the charging energy is

$$T_0 = \frac{e^2}{2k_BC} = \frac{928.5\, K}{C(aF)} \tag{10.5.13}$$

Coulomb blockade effects will manifest themselves if the sample temperature T is smaller than this effective charging temperature $T_0$and we expect the following to occur:

• When the capacitance reaches values approaching $\sim 10^{-18}$ F, each electron causes a shift in voltage of several 10s of millivolts.

• The charging energy of the capacitor, i.e., the energy needed to place a single extra electron becomes comparable to or larger than k_BT with T reaching 10 K or even 100 K if the capacitance becomes comparable to 10^{-18} F.

To get the small capacitors needed to generate Coulomb blockade effects at reasonable temperatures one has to use areal dimensions of $\lesssim$ 1000 Å× 1000 Å with spacing between the contacts reaching $\sim$ 50–100 Å. With such dimensions (using a relative dielectric constant of $\sim$ 10) we get capacitors with capacitances of the order of $\sim 10^{-16}$ F. The charging voltages are then $\sim$1 mV and T_0 $\sim$10 K. If the area of the capacitor is reduced further these values increase. It is possible to fabricate small capacitors with capacitance approaching 10^{-18} F.

In figure 10.12 we show the band profile of a typical tunnel junction capacitor which consists of two metal contacts separated by a thin tunneling barrier. In the absence of any Coulomb blockade we observe a monotonic increase in current with applied bias as shown in figure 10.12a.

In case the Coulomb blockade is significant we get a very different device behavior. In figure 10.12b we show the behavior for a structures where the charging energy is large enough to

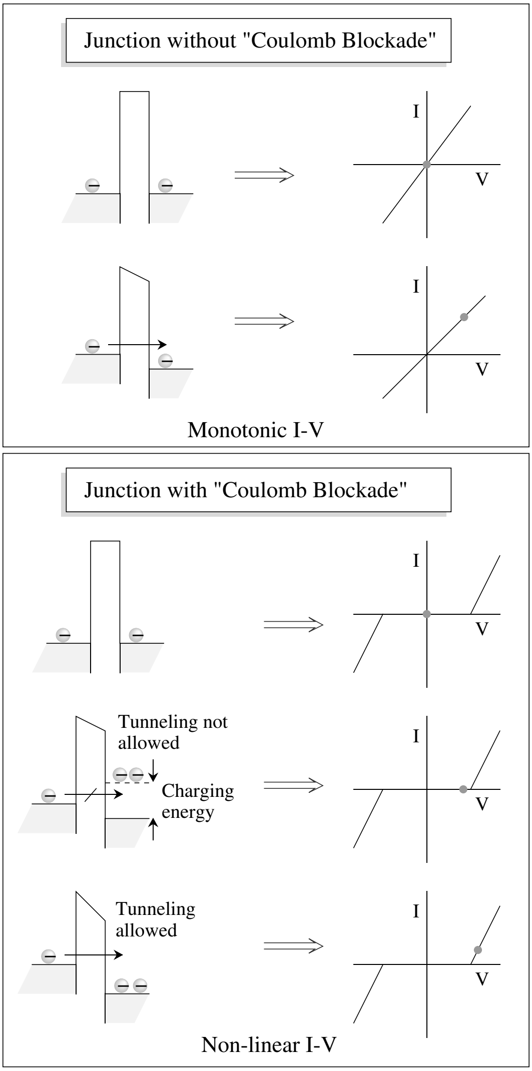


Figure 10.12: (a) A tunnel junction with large capacitance shows ohmic I–V characteristics. (b) In very small capacitance tunnel junctions the presence of a Coulomb blockade ensures no current flows until the voltage reaches a threshold value determined by the charging energy.

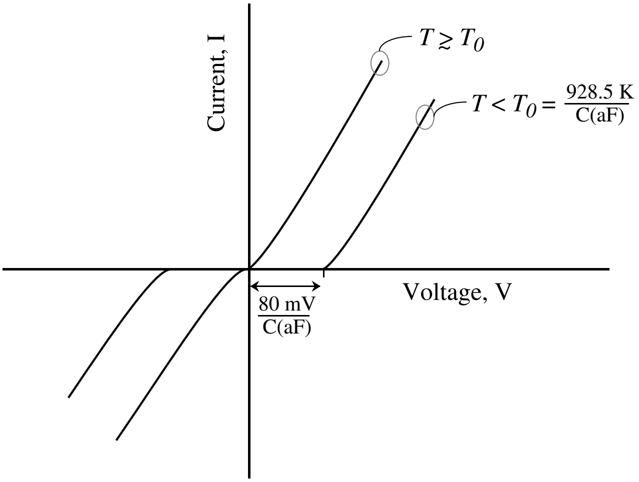


Figure 10.13: A schematic of how current voltage relations change as temperature is raised. Above T_0, defined in the figure, normal ohmic conduction occurs.

have measurable effects. At zero bias there is no net flow of electrons as usual. However at small biases smaller than the charging energy, an electron cannot move from the left to the right because that would raise the energy of the right side by $e^2/2C$ as shown. Once the voltage level (times electron charge) exceeds the charging energy, electrons can flow across the junction and we have ohmic behavior. The current–voltage relation shows a highly non-linear behavior as shown in figure 10.12b.

The effects sketched in figure 10.12b have a strong temperature dependence. As the temperature rises, the distribution of carriers in the contact is smeared by $\sim k_B T$. AS a result the Coulomb blockade effect survives only up to the temperature, T_0 defined above. In figure 10.13 we show how the current-voltage relations change when temperature is raised.

10.6 MAGNETIC SEMICONDUCTORS AND SPINTRONICS

In most semiconductors the asymmetry between spin up and spin down electrons is negligible even in presence of a magnetic field. As a result in existing electronic devices the spin of the electron is not relevant to current flow. The density of spin up and spin down electrons is the same unless a strong magnetic field is applied to select a particular state. The contacts used to inject electrons also usually have no spin selectivity. If spin selectivity can be created it should be possible to develop electronic devices that are dependent on the spin of the electrons much like optical devices are dependent on polarization of light. In optics the use of a polarizer, analyzer and modulator allow one to make switches. The same can be possible in electronics if electrons can be injected and extracted with spin selection.

In magnetic semiconductors it is possible to use ferromagnetic contacts to inject electrons with spin selectivity. Notable examples of magnetic semiconductors are InGaAsMn, CdMnTe, ZnMnSe, and HgMnTe. These semiconductors, known as diluted magnetic semiconductors, and their heterostructures with other semiconductors can now be fabricated and they offer a unique opportunity for the combined studies of semiconductor physics and magnetism. The magnetic semiconductors are fabricated by the usual epitaxial techniques like MBE or MOCVD and Mn is introduced as an extra ingredient. The Mn composition is usually $\leq 20\%$.

In recent years there has been a growing interest in a field known as "spintronics" (after spin and electronics). In conventional electronics, electron density is modulated to create devices for digital and analog applications. In spintronics the expectation is that one modulates the spin of electrons. As in quantum interference devices discussed in section 10.4, such a possibility promises very low power, high density devices. An important point to note in spin dependent devices is that usual scattering mechanisms that impact transport cause only very weak spin scattering. Thus an electron can maintain its spin value for several microns (or even 100 microns at low temperature). However, this does not mean that spin based transistors can function at high temperatures or for long channel lengths. Non-spin altering scattering processes are still important in spintronic devices.

In conventional electronic devices we ignore the electron spin. As noted above the main reason we have not worried about electron spin is that usually density of spin-up and spin-down electrons is the same and the spin splitting in the presence of a magnetic field is small. However, it is possible to prepare a semiconductor sample in a state where electrons in the conduction band have a much higher density of spin-down electrons. This can be done by using optical injection or electronic injection. Electrons (or other charged particles) interact with a magnetic field via a magnetic moment which is written as

$$\mu_s = -g\mu_B \mathrm{S} = \gamma\hbar \mathrm{S} \tag{10.6.1}$$

where S is the spin of the particle; g is known as the g-factor and characterizes the particle. The constant μ_B is the Bohr magneton and has a value

$$\mu_B = \frac{e\hbar}{2m} \tag{10.6.2}$$

The constant γ is called the gyromagnetic or magnetogyric ratio. The magnetic interacction associated with the spin is

$$H_{\text{spin}} = -\mu_s \cdot \boldsymbol{B} \tag{10.6.3}$$

Spin Injection and Spin Transistor

In ferromagnetic materials, once the material is magnetized, there is a strong selection of spin orientation (below the Curie temperature). If a ferromagnetic contact is used in a semiconductor

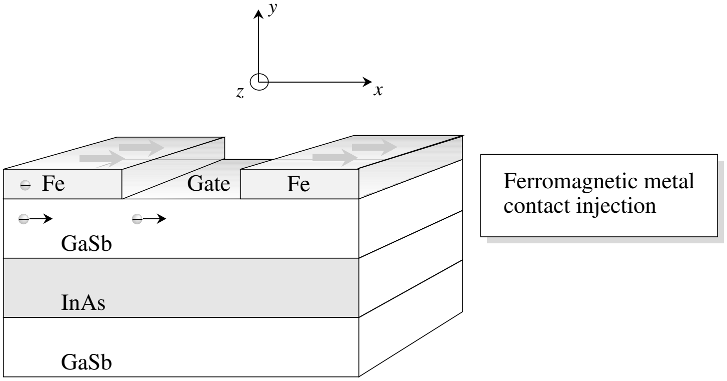


Figure 10.14: A schematic of a spin transistor in which electrons with a selected spin are injected into a 2-dimensional channel.

device it is possible to inject electrons or holes in a spin selected state using ferromagnetic contacts. In figure 10.14a we show a spin-transistor in which spin selected electrons are injected from an Fe contact acting as a source. The magnetized contact injects electrons with spin selected by the magnetization field and maintain this spin state as they travel throughout the device. The spin transistor exploits quantum interference effects with two nuances: i) spin select electrons can be injected into a transistor channel; ii) spin splitting of spin-up and spin-down states causes the two spin state electrons to have a different k-vector which can be controlled by a gate bias to create interference effects.

Using the geometry shown in figure 10.14a, electrons are injected into the 2-dimensional channel with a spin polarized along the $+x$ direction. These electrons may be written in terms of the spin-up (positive z-polarized) and spin-down (negative z-polarized) states

$$\langle x| \to \frac{1}{\sqrt{2}} \left(\langle\uparrow| + \langle\downarrow|\right) \tag{10.6.4}$$

Now consider the possibility where the energy of the spin-up and spin-down electrons is different as shown in figure 10.15. The splitting in the spin-up and spin-down states can occur due to external magnetic fields or internal spin-orbit effects combined with lack of inversion symmetry. These effects are strongest in narrow bandgap semiconductors where the conduction band states are influenced by the p-type valence band states.

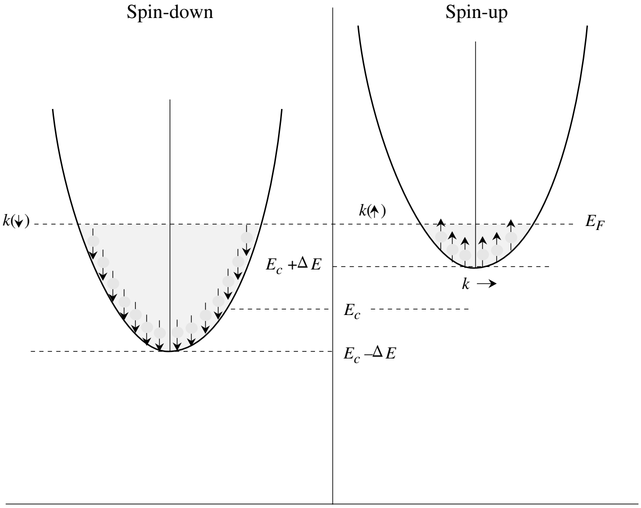


Figure 10.15: A schematic of the band profile of spin-up and spin-down electrons. At the Fermi energy the k-vector for spin-up and spin-down electrons are different.

The position of the Fermi level is the same for the spin-up and spin-down states as shown in figure 10.15. We have

$$\begin{aligned} E_F &= E_c - \Delta E + \frac{\hbar^2 k^2(\downarrow)}{2m^*} \\ &= E_c + \Delta E + \frac{\hbar^2 k^2(\uparrow)}{2m^*} \end{aligned} \tag{10.6.5}$$

As the electrons move down the channel the phase difference between spin-up and spin-down electrons changes according to the usual wave propagation equation

$$\Delta\theta = [k(\uparrow) - k(\downarrow)]\, L \tag{10.6.6}$$

where L is the channel length. The drain contact acts as a spin filter and only accepts electron states with spin in the x-direction. Thus the current flows if $\Delta\theta = 2n\pi$. Otherwise the current value is lower. Thus the spin transistor essentially behaves as an electrooptic modulator where the phase is controlled by the gate voltage which controls E_F.

10.7 PROBLEMS

- **Section 10.2**

Problem 10.1 Consider a GaAs sample in which fields of 10 kV/cm and 100 kV/cm is applied. Discuss the restrictions on scattering times under which Bloch oscillations can occur. Also calculate the frequency of oscillations.

Problem 10.2 Design a GaAs/AlAs superlattice structure in which Bloch oscillations could occur when the scattering rate is 10^{13} s^{-1} and the applied field is 100 kV/cm. Discuss possible effects that could prevent the observation of the oscillations.

Problem 10.3 Consider a Si crystal in which a field of 10^5 V/cm is applied. Calculate the Bloch oscillation period if the field is applied along the i) [100]; ii) [110], and iii) [111] directions. Discuss if these oscillations are feasible .

- **Section 10.3**

Problem 10.4 In the resonant tunnel structure the transmission probability vs. energy plot has resonances with a line width ΔE_n. Show that if E_n is the energy of the n^{th} resonance,

$$\Delta E_n \sim \frac{E_n T_{1B}}{\pi n}$$

where T_{1B} is the transmission through a single barrier.

Problem 10.5 Estimate the time an electron will take to tunnel through a resonant tunnel double barrier structure. You can use the Heisenberg relation $\Delta t \Delta E \sim \hbar$, where ΔE is the energy line width of the transmission resonance.

Problem 10.6 Consider a resonant tunneling structure with the following parameters:

$$\begin{aligned} \text{Barrier height, } V_0 &= 0.3 \text{ eV} \\ \text{Well size, } W &= 60 \text{ Å} \\ \text{Barrier width, } a &= 25 \text{ Å} \\ \text{Effective mass , } m^* &= 0.07\, m_0 \end{aligned}$$

Calculate and plot the tunneling probability of electrons as a function of energy for $0 < E < V_0$.

Problem 10.7 Consider a 0.1 μm AlGaAs/GaAs device in which a 2-dimensional gas is formed with a density of $n_{2D} = 10^{12}$ cm^{-2}. A split gate device is made from the structure. Estimate the minimum gate voltage needed to switch a quantum interference transistor. How does this compare to the voltage needed to switch regular FET?

Problem 10.8 In normal transistors the ON and OFF states of the device are produced by injecting and removing electrons in the device. Consider a Si device with an area of

2.0 μm×0.1 μm in which a 1 V gate bias changes the electron density in the channel from 10^{12} cm^{-2} to 10^{8} cm^{-2}, thus switching the device from ON to OFF. What is the switching energy?

Estimate the switching energy if quantum interference effects were used in the same device.

Problem 10.9 Consider a 2-dimensional electron channel in a AlGaAs/GaAs device. The gate length is 0.1 μm and gate width is 2.0 μm. The device is biased so that the electron density in the channel is 10^{12} cm^{-2}. How much will the electron number in the channel change if $\Delta\sigma = 2e^2/h$? Use a semi-classical model with mobility 10^5 cm^2/V·s.

Problem 10.10 Consider a metal-oxide-silicon capacitor. At what areal dimensions will it display Coulomb blockade effects at 300 K? The relative dielectric constant of SiO$_2$ is 3.9 and the oxide thickness is 25 Å.

Problem 10.11 Consider a single electron transistor based on a MOSFET in which the gate capacitance is 10^{-18} F. The gate capacitor state is altered by a single electron (at very low temperatures). Calculate the change in the device channel current if the device transconductance is

$$g_m = \frac{\delta I}{\delta V_G} = 1.0 \text{ S}$$

10.8 Further Reading

- **Mesoscopic Structures**
  - Articles in *Nanostructure Physics and Fabrication* (edited by M. A. Reed and W. P. Kirk, Academic Press, New York, 1989).
  - Datta, Supriyo, *Electronic Transport in Mesoscopic Systems* (Cambridge University Press, 1995).
  - Ferry, D. K., *Semiconductors* (Macmillan, New York, 1991).
  - Gradert, Hermann and H. Michel, Editors, *Single Charge Tunneling: Coulomb Blockade Phenomenon in Nanostructures* (NATO ASI Series, **B**, Physics, vol. 294), Plenum Publishing Corporation, 1992.
  - Janssen, Martin, *Fluctuations and Localization in Mesoscopic Electron Systems* (World Scientific Publishing Company, 1991).
  - Landauer, R., Philos. Mag., **21**, 863 (1970).
  - Murayama, Yoshinasa, *Mesoscopic Systems*, (John Wiley and Sons, 2001).
  - Physics Today, (Dec. 1988). Covers the important aspects of physics in mesoscopic structures.
  - Van Wees, B. J., H. Van Houten, C. W. J. Beenakker, J. L. Williamson, L. P. Kauwenhoven, D. van der Marel, and C. T. Foxon, Phys. Rev. Lett., **60**, 848 (1988).

Appendix A

LIST OF SYMBOLS

a	lattice constant (edge of the cube for the semiconductor fcc lattice)
B	base transport factor in a bipolar transistor
c	velocity of light
C_{ox}	oxide capacitance per unit area
C_{mos}	capacitance (per area) of an MOS capacitor
$C_{mos(min)}$	minimum capacitance (per area) of an MOS capacitor
$C_{mos(fb)}$	capacitance (per area) of an MOS capacitor under flatband conditions
C_{GS}, C_{GD}	gate to source and gate to drain capacitance in a FET
C_{DS}	drain to substrate capacitance in an FET
C_j, C_d	junction, diffusion capacitance in a p-n diode
D_n	electron diffusion coefficient
D_p	hole diffusion coefficient
D_b	diffusion coefficient in the base of a bipolar transistor
D_e	diffusion coefficient in the emitter of a bipolar transistor
D_c	diffusion coefficient in the collector of a bipolar transistor
e	magnitude of the electron charge
E	energy of a particle
E_F	Fermi level
E_{Fi}	intrinsic Fermi level

E_{Fn}	electron quasi-Fermi level
E_{Fp}	hole quasi-Fermi level
$E^e(E^h)$	energy of an electron (hole) in an optical absorption or emission measured from the bandedges
$E_c(E_v)$	conduction (valence) bandedge
$f(E)$	occupation probability of an electron state with energy E at equilibrium. This is the Fermi-Dirac function
$f^e(E)$	occupation function for an electron in non-equilibrium state. This is the quasi-Fermi function
$f^h(E)$	occupation function for a hole = $1 - f^e(E)$
f_τ	cutoff frequency for unit current gain
$f_{\max}$	available power gain is unity at this frequency
$\mathcal{E}$	electric field
F_{ext}	external force such as an electric or magnetic force
$g_{\rm m}$	transconductance of a transistor
g_D	output conductance of a transistor
G_L	electron-hole generation rate due to a light beam
$\hbar$	Planck's constant divided by 2π
h	channel thickness of a JFET or an MESFET
$h(x)$	depletion region thickness in an FET at position x along the source to drain channel
H	magnetic field
I_{ph}	photon particle current
I_E, I_B, I_C	emitter, base, and collector current in a BJT
I_{En}, I_{Ep}	electron, hole part of the emitter current in an npn BJT
I_D	drain current in an FET
I_o	reverse bias saturation current in a p-n diode
I_s	reverse bias saturation current in a Schottky diode
I_{GR}	generation recombination current in a diode
I_{GR}^o	prefactor for the generation recombination current

J	current density
J_L	photocurrent density
J_{ph}	photon particle current density

ℓ	mean free path between successive collisions

L_n	diffusion length for electron
L_p	diffusion length for holes

m_o	free electron mass
m_e^*	electron mass
m_h^*	hole mass
m_{dos}^*	density of states mass
m_σ^*	conductivity mass
m_{hh}^*	mass of the heavy hole
$m_{\ell h}^*$	mass of the light hole
m_r^*	reduced mass of the electron-hole system

M, M_e, M_h	multiplication factor, multiplication factor for electrons, mulitplication factor for holes

n	electron concentration in the conduction band
n_i	intrinsic electron concentration in the conduction band
n_d	electrons bound to the donors
$n_p(n_p)$	equilibrium electron density in the p-side (n-side) of a p-n junction

N_{cv}	joint density of states for electrons and holes
$N_e(E)$	density of states of electrons in the conduction band
$N_h(E)$	density of states of holes in the valence band
$N_c(E)$	effective density of states in the conduction band
$N_v(E)$	effective density of states in the valence band
N_d	donor density
N_a	acceptor density
$N_{2D}(E)$	2-dimensional density of states
N_t	density of impurity states (trap states)
N_{ab}	acceptor concentration in the base of an npn BJT
N_{de}	donor concentration in the emitter of an npn BJT
N_{dc}	donor concentration in the collector of npn BJT

p	momentum of a particle
p	hole concentration in the valence band
p_i	intrinsic hole concentration in the valence band
p_a	holes bound to acceptors
p_{cv}	momentum matrix element for an optical transition between the valence and conduction band
$p_n(p_p)$	equilibrium hole density in the n-side (p-side) of a p-n junction
P_{op}	optical power density (energy flow/sec/area)
Q_s	total charge (per area) in an MOS channel
Q_n	total mobile charge (per area) in an MOS channel
Q_{ss}	surface charge density in an MOS capacitor
Q_{dep}	depletion charge (per area) in an MOS channel
R_{spon}	total rate at which an electron-hole system recombines to emit photons by spontaneous recombination
R_s, R_G, R_D	parasitic resistances associated with the source, gate and drain of a transistor respectively
R_L	load resistance
R^*	Richardson constant in a Schottky barrier
t_{tr}	transit time of a carrier through a channel
T	tunneling probability
$U(r)$	position dependent potential energy
v	velocity of the electron
v_s	saturation velocity of the carrier (electron, hole)
V_{bi}	built-in voltage
V_G	gate bias (referred to the source)
V_D	drain bias
V_p	pinch-off voltage to deplete the channel of an FET
V_T	threshold gate bias for pinch-off
V_{fb}	flat band voltage. Voltage needed to make the semiconductor bands flat in an MOS capacitor

V_{SB}	source to body (substrate) potential
$V_r(V_f)$	reverse (forward) bias voltage in a diode
V_{BE}, V_{BC}	base to emitter, base to collector bias in a bipolar transistor
V_{pt}	punchthrough voltage

$W_n(W_p)$	depletion region edge on the n-side (p-side) of a p-n junction
W	depletion region width
W_b, W_{bn}	base width, neutral base width of a bipolar transistor

α	optical absorption coefficient
α	current transfer ratio in a bipolar transistor
α_R	reflection loss coefficient in an optical cavity
α_{imp}	impact ionization coefficient for electrons
β	base to collector current amplification factor in a BJT
β_{imp}	impact ionization coefficient for holes
γ_e	emitter efficienty of a bipolar transistor
γ_{inj}	injection efficiency of a p-n diode for electron (hole) current
ΔE_g	bandgap difference between two materials
$\Delta E_c, \Delta E_v$	band discontinuity in the conduction, valence band in a heterostructure
ϵ_o	free space permittivity
ϵ	product of the relative dielectric constant and ϵ_o
ψ	electron wavefunction
$\sigma_n(\sigma_p)$	electron (hole) capture cross-section for an impurity
σ	conductivity of a material
μ	mobility of a material
$\mu_n(\mu_p)$	electron (hole) mobility
τ_{sc}	scattering time between successive collisions. Also called relaxation time
ω	frequency
τ_o	rate at which an electron recombines radiatively with a hole at the same momentum value
τ_r	radiative recombination time for e-h pair
τ_{nr}	non-radiative recombination time for a e-h pair
τ_n	lifetime of an electron to recombine with a hole
τ_p	lifetime of a hole to recombine with an electron
τ_{sd}	storage delay time in a diode
δn	excess electron density in a region. This is the density above the equilibrium density
δp	excess hole density in a region
ϕ_m	metal work function

χ_s	electron affinity of a semiconductor
ϕ_s	work function of a semiconductor
ϕ_{ms}	difference between a metal and semiconductor work function
ϕ_b	barrier height seen by electrons coming from a metal towards a semiconductor

Appendix B

BOLTZMANN TRANSPORT THEORY

Transport of electrons in solids is the basis of many modern technologies. The Boltzmann transport theory allows us to develop a microscopic model for macroscopic quantities such as mobility, diffusion coefficient, and conductivity. This theory has been used in Chapter 8 to study transport of electrons and holes in materials. In this appendix we will present a derivation of this theory.

B.1 BOLTZMANN TRANSPORT EQUATION

In order to describe the transport properties of an electron gas, we need to know the distribution function of the electron gas. The distribution would tell us how electrons are distributed in momentum space or k-space (and energy-space) and from this information all of the transport properties can be evaluated. We know that at equilibrium the distribution function is simply the Fermi-Dirac function

$$f(E) = \frac{1}{\exp\left(\frac{E-E_F}{k_BT}\right)+1} \tag{B.1}$$

This distribution function describes the equilibrium electron gas and is $\underline{\text{independent}}$ of any collisions that may be present. While the collisions will continuously remove electrons from one $\boldsymbol{k}$-state to another, the net distribution of electrons is always given by the Fermi-Dirac function as long as there are no external influences to disturb the equilibrium.

To describe the distribution function in the presence of external forces, we develop the Boltzmann transport equation. Let us denote by $f_{\mathbf{k}}(\boldsymbol{r})$ the local concentration of the electrons in state $\boldsymbol{k}$ in the neighborhood of $\boldsymbol{r}$. The Boltzmann approach begins with an attempt to determine how $f_{\mathbf{k}}(\boldsymbol{r})$ changes with time. Three possible reasons account for the change in the electron distribution in k-space and r-space:

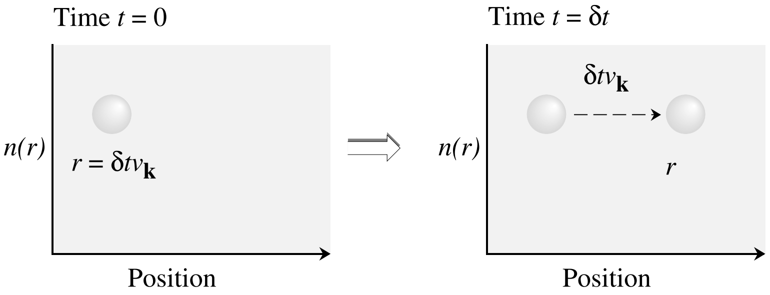


Figure B.1: At time $t = 0$ particles at position $\boldsymbol{r} - \delta t \boldsymbol{v}_{\mathbf{k}}$ reach the position $\boldsymbol{r}$ at a later time δt. This simple concept is important in establishing the Boltzmann transport equation.

1. Due to the motion of the electrons (diffusion), carriers will be moving into and out of any volume element around $\boldsymbol{r}$.

2. Due to the influence of external forces, electrons will be changing their momentum (or $\boldsymbol{k}$-value) according to $\hbar \, d\boldsymbol{k}/dt = \boldsymbol{F}_{ext}$.

3. Due to scattering processes, electrons will move from one $\boldsymbol{k}$-state to another.

We will now calculate these three individual changes by evaluating the partial time derivative of the function $f_{\mathbf{k}}(\boldsymbol{r})$ due to each source.

B.1.1 Diffusion-Induced Evolution of $f_{\mathbf{k}}(\boldsymbol{r})$

If $\boldsymbol{v}_{\mathbf{k}}$ is the velocity of a carrier in the state $\boldsymbol{k}$, in a time interval t, the electron moves a distance $t \; \boldsymbol{v}_{\mathbf{k}}$. Thus the number of electrons in the neighborhood of $\boldsymbol{r}$ at time δt is equal to the number of carriers in the neighborhood of $\boldsymbol{r} - \delta t \; \boldsymbol{v}_{\mathbf{k}}$ at time 0, as shown in figure B.1

We can thus define the following equality due to the diffusion

$$f_{\mathbf{k}}(\boldsymbol{r}, \delta t) = f_{\mathbf{k}}(\boldsymbol{r} - \delta t \; \boldsymbol{v}_{\mathbf{k}}, 0) \tag{B.2}$$

or

$$\begin{aligned} f_{\mathbf{k}}(\boldsymbol{r}, 0) + \frac{\partial f_{\mathbf{k}}}{\partial t} \cdot \delta t &= f_{\mathbf{k}}(\boldsymbol{r}, 0) - \frac{\partial f_{\mathbf{k}}}{\partial \boldsymbol{r}} \cdot \delta t \; \boldsymbol{v}_{\mathbf{k}} \\ \left. \frac{\partial f_{\mathbf{k}}}{\partial t} \right|_{\text{diff}} &= -\frac{\partial f_{\mathbf{k}}}{\partial \boldsymbol{r}} \cdot \boldsymbol{v}_{\mathbf{k}} \end{aligned} \tag{B.3}$$

B.1.2 External Field-Induced Evolution of $f_{\mathbf{k}}(\boldsymbol{r})$

The crystal momentum $\boldsymbol{k}$ of the electron evolves under the action of external forces according to Newton's equation of motion. For an electric and magnetic field ($\mathcal{E}$ and $\boldsymbol{B}$), the rate of change of $\boldsymbol{k}$ is given by

$$\dot{\boldsymbol{k}} = \frac{e}{\hbar}\left[\mathcal{E} + \boldsymbol{v}_{\mathbf{k}} \times \boldsymbol{B}\right] \tag{B.4}$$

In analogy to the diffusion-induced changes, we can argue that particles at time $t = 0$ with momentum $\boldsymbol{k} - \dot{\boldsymbol{k}}\ \delta t$ will have momentum $\boldsymbol{k}$ at time δt and

$$f_{\mathbf{k}}(\boldsymbol{r}, \delta t) = f_{\mathbf{k}-\dot{\mathbf{k}}\delta t}(\boldsymbol{r}, 0) \tag{B.5}$$

which leads to the equation

$$\begin{aligned} \left.\frac{\partial f_{\mathbf{k}}}{\partial t}\right|_{\text{ext. forces}} &= -\dot{\boldsymbol{k}}\frac{\partial f_{\mathbf{k}}}{\partial \boldsymbol{k}} \\ &= \frac{-e}{\hbar}\left[\mathcal{E} + \frac{\boldsymbol{v} \times \boldsymbol{B}}{c}\right] \cdot \frac{\partial f_{\mathbf{k}}}{\partial \boldsymbol{k}} \end{aligned} \tag{B.6}$$

B.1.3 Scattering-Induced Evolution of $f_{\mathbf{k}}(\boldsymbol{r})$

We will assume that the scattering processes are <u>local</u> and <u>instantaneous</u> and change the state of the electron from $\boldsymbol{k}$ to $\boldsymbol{k}^{'}$. Let $W(\boldsymbol{k}, \boldsymbol{k}^{'})$ define the rate of scattering from the state $\boldsymbol{k}$ to $\boldsymbol{k}^{'}$ if the state $\boldsymbol{k}$ is occupied and $\boldsymbol{k}^{'}$ is empty. The rate of change of the distribution function $f_{\mathbf{k}}(\boldsymbol{r})$ due to scattering is

$$\left.\frac{\partial f_{\mathbf{k}}}{\partial t}\right)_{\text{scattering}} = \int \left[f_{\mathbf{k}'}\ (1 - f_{\mathbf{k}})\ W(\boldsymbol{k}^{'}, \boldsymbol{k}) - f_{\mathbf{k}}\ (1 - f_{\mathbf{k}'})\ W(\boldsymbol{k}, \boldsymbol{k}^{'})\right] \frac{d^3k^{'}}{(2\pi)^3} \tag{B.7}$$

The $(2\pi)^3$ in the denominator comes from the number of states allowed in a k-space volume $d^3k^{'}$. The first term in the integral represents the rate at which electrons are coming from an occupied $\boldsymbol{k}^{'}$ state (hence the factor $f_{\mathbf{k}'}$) to an unoccupied $\boldsymbol{k}$- state (hence the factor $(1 - f_{\mathbf{k}})$). The second term represents the loss term.

Under steady-state conditions, there will be no net change in the distribution function and the total sum of the partial derivative terms calculated above will be zero.

$$\left.\frac{\partial f_{\mathbf{k}}}{\partial t}\right)_{\text{scattering}} + \left.\frac{\partial f_{\mathbf{k}}}{\partial t}\right)_{\text{fields}} + \left.\frac{\partial f_{\mathbf{k}}}{\partial t}\right)_{\text{diffusion}} = 0 \tag{B.8}$$

Let us define

$$g_{\mathbf{k}} = f_{\mathbf{k}} - f_{\mathbf{k}}^0 \tag{B.9}$$

where $f_{\mathbf{k}}^0$ is the equilibrium distribution.

We will attempt to calculate $g_{\mathbf{k}}$, which represents the deviation of the distribution function from the equilibrium case.

Substituting for the partial time derivatives due to diffusion and external fields we get

$$-\boldsymbol{v}_{\mathbf{k}} \cdot \nabla_r f_{\mathbf{k}} - \frac{e}{\hbar}\left(\mathcal{E} + \frac{\boldsymbol{v}_{\mathbf{k}} \times \boldsymbol{B}}{c}\right) \cdot \nabla_k f_{\mathbf{k}} = \left.\frac{-\partial f_{\mathbf{k}}}{\partial t}\right)_{\text{scattering}} \tag{B.10}$$

Substituting $f_{\mathbf{k}} = f_{\mathbf{k}}^0 + g_{\mathbf{k}}$

$$\begin{aligned} & -\boldsymbol{v}_{\mathbf{k}} \cdot \nabla_r f_{\mathbf{k}}^0 - \tfrac{e}{\hbar}\left(\mathcal{E} + \boldsymbol{v}_{\mathbf{k}} \times \boldsymbol{B}\right) \nabla_k f_{\mathbf{k}}^0 \\ & = -\left.\tfrac{\partial f_{\mathbf{k}}}{\partial t}\right)_{\text{scattering}} + \boldsymbol{v}_{\mathbf{k}} \cdot \nabla_r g_{\mathbf{k}} + \tfrac{e}{\hbar}\left(\mathcal{E} + \boldsymbol{v}_{\mathbf{k}} \times \boldsymbol{B}\right) \cdot \nabla_k g_{\mathbf{k}} \end{aligned} \tag{B.11}$$

We note that the magnetic force term on the left-hand side of equation B.11 is proportional to

$$\boldsymbol{v}_{\mathbf{k}} \cdot \frac{e}{\hbar}\left(\boldsymbol{v}_{\mathbf{k}} \times \boldsymbol{B}\right)$$

and is thus zero. We remind ourselves that (the reader should be careful not to confuse $E_{\mathbf{k}}$, the particle energy and $\mathcal{E}$, the electric field)

$$\boldsymbol{v}_{\mathbf{k}} = \frac{1}{\hbar}\frac{\partial E_k}{\partial \boldsymbol{k}}$$

and (in semiconductor physics, we often denote μ by E_F)

$$f_{\mathbf{k}}^0 = \frac{1}{\exp\left[\frac{E_{\mathbf{k}}-\mu}{k_BT}\right] + 1}$$

Thus

$$\begin{aligned} \nabla_r f^0 &= \frac{-\left[\exp\left(\frac{E_{\mathbf{k}}-\mu}{k_BT}\right)\right]}{\left[\exp\left(\frac{E_{\mathbf{k}}-\mu}{k_BT}\right)+1\right]^2} \nabla_r \left(\frac{E_{\mathbf{k}} - \mu(\boldsymbol{r})}{k_BT(r)}\right) \\ &= k_BT \cdot \frac{\partial f^0}{\partial E_{\mathbf{k}}}\left[-\frac{\nabla\mu}{k_BT} - \frac{(E_{\mathbf{k}} - \mu)}{k_BT^2}\nabla T\right] \end{aligned}$$

$$\nabla_r f^0 = \frac{\partial f^0}{\partial E_{\mathbf{k}}}\left[-\nabla\mu - \frac{(E_{\mathbf{k}} - \mu)}{T}\nabla T\right] \tag{B.12}$$

Also

$$\begin{aligned} \nabla_k f^0 &= \frac{\partial f^0}{\partial E_{\mathbf{k}}} \cdot \nabla_k E_k \\ &= \hbar\boldsymbol{v}_{\mathbf{k}}\frac{\partial f^0}{\partial E_{\mathbf{k}}} \end{aligned} \tag{B.13}$$

Substituting these terms and retaining terms only to second-order in electric field (i.e., ignoring terms involving products $g_{\mathbf{k}} \cdot \mathcal{E}$), we get, from equation B.11,

$$\begin{aligned} & -\tfrac{\partial f^0}{\partial E_{\mathbf{k}}} \cdot \boldsymbol{v}_{\mathbf{k}} \cdot \left[-\tfrac{(E_{\mathbf{k}}-\mu)}{T} \nabla T + e\mathcal{E} - \nabla \mu \right] \\ & = -\tfrac{\partial f}{\partial t} \Big)_{\text{scattering}} + \boldsymbol{v}_{\mathbf{k}} \cdot \nabla_r g_{\mathbf{k}} + \tfrac{e}{\hbar} \left(\boldsymbol{v}_{\mathbf{k}} \times \boldsymbol{B} \right) \cdot \nabla_k g_{\mathbf{k}}. \end{aligned} \tag{B.14}$$

The equation derived above is the Boltzmann transport equation.

We will now apply the Boltzmann equation to derive some simple expressions for conductivity, mobility, etc., in semiconductors. We will attempt to relate the microscopic scattering events to the measurable macroscopic transport properties. Let us consider the case where we have a uniform electric field $\mathcal{E}$ in an infinite system maintained at a uniform temperature.

The Boltzmann equation becomes

$$-\frac{\partial f^0}{\partial E_{\mathbf{k}}} \boldsymbol{v}_{\mathbf{k}} \cdot e\mathcal{E} = -\frac{\partial f_{\mathbf{k}}}{\partial t} \Big)_{\text{scattering}} \tag{B.15}$$

Note that only the deviation $g_{\mathbf{k}}$ from the equilibrium distribution function above contributes to the scattering integral.

As mentioned earlier, this equation, although it looks simple, is a very complex equation which can only be solved analytically under fairly simplifying assumptions. We make an assumption that the scattering induced change in the distribution function is given by

$$-\frac{\partial f_{\mathbf{k}}}{\partial t} \Big)_{\text{scattering}} = \frac{g_{\mathbf{k}}}{\tau} \tag{B.16}$$

We have introduced a time constant τ whose physical interpretation can be understood when we consider what happens when the external forces have been removed. In this case the perturbation in the distribution function will decay according to the equation

$$\frac{-\partial g_{\mathbf{k}}}{\partial t} = \frac{g_{\mathbf{k}}}{\tau}$$

or

$$g_{\mathbf{k}}(t) = g_{\mathbf{k}}(0) e^{-t/\tau} \tag{B.17}$$

The time τ thus represents the time constant for relaxation of the perturbation as shown schematically in figure B.2 The approximation which allows us to write such a simple relation is called the relaxation time approximation (RTA).

According to this approximation

$$\begin{aligned} g_{\mathbf{k}} &= -\frac{\partial f_{\mathbf{k}}}{\partial t} \Big)_{\text{scattering}} \cdot \tau \\ &= \frac{-\partial f^0}{\partial E_{\mathbf{k}}} \, \tau \boldsymbol{v}_{\mathbf{k}} \cdot e\mathcal{E} \end{aligned} \tag{B.18}$$

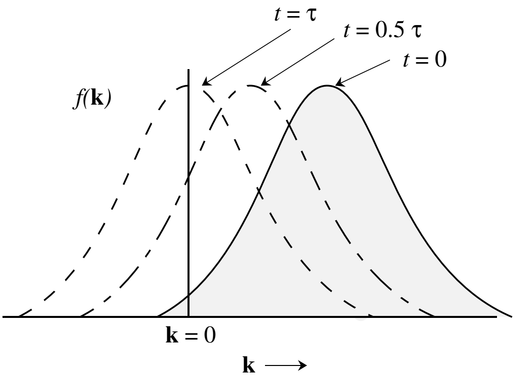


Figure B.2: This figure shows that at time $t = 0$, the distribution function is distorted by some external means. If the external force is removed, the electrons recover to the equilibrium distribution by collisions.

Note that we have not defined how τ is to be calculated. We have merely introduced a simpler unknown that still needs to be determined. The k-space distribution function may be written as

$$\begin{aligned} f_{\mathbf{k}} &= f_{\mathbf{k}}^0 - \left(\frac{\partial f_{\mathbf{k}}^0}{\partial E_{\mathbf{k}}}\right) e\tau \boldsymbol{v}_{\mathbf{k}} \cdot \mathcal{E} \qquad \text{(B.19)} \\ &= f_{\mathbf{k}}^0 - \left(\nabla_k f_{\mathbf{k}}^0\right) \cdot \frac{\partial \boldsymbol{k}}{\partial E_{\mathbf{k}}} \cdot e\tau \boldsymbol{v}_{\mathbf{k}} \cdot \mathcal{E} \qquad \text{(B.20)} \end{aligned}$$

Using the relation

$$\hbar \frac{\partial \boldsymbol{k}}{\partial E_{\mathbf{k}}} \cdot \boldsymbol{v}_{\mathbf{k}} = 1$$

We have

$$\begin{aligned} f_{\mathbf{k}} &= f_{\mathbf{k}}^0 - \left(\nabla_{\mathbf{k}} f_{\mathbf{k}}^0\right) \cdot \frac{e\tau\mathcal{E}}{\hbar} \\ &= f_{\mathbf{k}}^0 \left(\boldsymbol{k} - \frac{e\tau\mathcal{E}}{\hbar}\right) \qquad \text{(B.21)} \end{aligned}$$

This is a very useful result which allows us to calculate the non-equilibrium function $f_{\mathbf{k}}$ in terms of the equilibrium function f^0. The recipe is very simple—shift the original distribution function for $\boldsymbol{k}$ values parallel to the electric field by $e\tau\mathcal{E}/\hbar$. If the field is along the z-direction, only the distribution for k_z will shift. This is shown schematically in figure B.3. Note that for the equilibrium distribution function, there is an exact cancellation between positive velocities

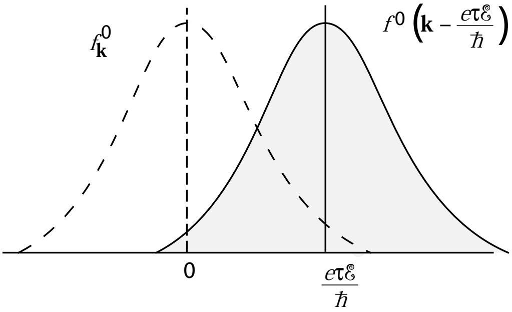


Figure B.3: The displaced distribution function shows the effect of an applied electric field.

and negative velocities. When the field is applied, there is a net shift in the electron momenta and velocities given by

$$\begin{aligned}\delta\boldsymbol{p} &= \hbar\delta\boldsymbol{k} = -e\tau\mathcal{E} \\ \delta\boldsymbol{v} &= -\frac{e\tau\mathcal{E}}{m^*}\end{aligned} \tag{B.22}$$

This gives, for the mobility,

$$\mu = \frac{e\tau}{m^*} \tag{B.23}$$

If the electron concentration is n, the current density is

$$\begin{aligned}\boldsymbol{J} &= ne\delta\boldsymbol{v} \\ &= \frac{ne^2\tau\mathcal{E}}{m^*}\end{aligned}$$

or the conductivity of the system is

$$\sigma = \frac{ne^2\tau}{m^*} \tag{B.24}$$

This equation relates a microscopic quantity τ to a macroscopic quantity σ.

So far we have introduced the relaxation time τ, but not described how it is to be calculated. We will now relate it to the scattering rate $W(\boldsymbol{k}, \boldsymbol{k}^{'})$, which can be calculated by using the Fermi golden rule. We have, for the scattering integral,

$$\left.\frac{\partial f}{\partial t}\right)_{\text{scattering}} = \int \left[f(\boldsymbol{k}^{'})(1 - f(\boldsymbol{k}))W(\boldsymbol{k}^{'}, \boldsymbol{k}) - f(\boldsymbol{k})(1 - f(\boldsymbol{k}^{'}))W(\boldsymbol{k}, \boldsymbol{k}^{'})\right] \frac{d^3\boldsymbol{k}^{'}}{(2\pi)^3}$$

Let us examine some simple cases where the integral on the right-hand side becomes simplified.

Elastic Collisions

Elastic collisions represent scattering events in which the energy of the electrons remains unchanged after the collision. Impurity scattering and alloy scattering discussed in Chapter 8 fall into this category. In the case of elastic scattering the principle of microscopic reversibility ensures that

$$W(\boldsymbol{k},\boldsymbol{k}^{'}) = W(\boldsymbol{k}^{'},\boldsymbol{k}) \tag{B.25}$$

i.e., the scattering rate from an initial state $\boldsymbol{k}$ to a final state $\boldsymbol{k}^{'}$ is the same as that for the reverse process. The collision integral is now simplified as

$$\begin{aligned}\left.\frac{\partial f}{\partial t}\right)_{\text{scattering}} &= \int \left[f(\boldsymbol{k}^{'}) - f(\boldsymbol{k})\right] W(\boldsymbol{k},\boldsymbol{k}^{'}) \frac{d^3\boldsymbol{k}^{'}}{(2\pi)^3} \\ &= \int \left[g(\boldsymbol{k}^{'}) - g(\boldsymbol{k})\right] W(\boldsymbol{k},\boldsymbol{k}^{'}) \frac{d^3\boldsymbol{k}^{'}}{(2\pi)^3}\end{aligned} \tag{B.26}$$

The simple form of the Boltzmann equation is (from equation B.17)

$$\begin{aligned}\frac{-\partial f^0}{\partial E_{\mathbf{k}}}\boldsymbol{v}_{\mathbf{k}}\cdot e\mathcal{E} &= \int (g_{\mathbf{k}} - g_{\mathbf{k}'})\ W(\boldsymbol{k},\boldsymbol{k}^{'}) d^3k^{'} \\ &= \left.\frac{-\partial f}{\partial t}\right)_{\text{scattering}}\end{aligned} \tag{B.27}$$

The relaxation time was defined through

$$\begin{aligned}g_{\mathbf{k}} &= \left(\frac{-\partial f^0}{\partial E}\right) e\mathcal{E}\cdot\boldsymbol{v}_{\mathbf{k}}\cdot\tau \\ &= \left.\frac{-\partial f}{\partial t}\right)_{\text{scattering}} \cdot \tau\end{aligned} \tag{B.28}$$

Substituting this value in the integral on the right-hand side, we get

$$\frac{-\partial f^0}{\partial E_{\mathbf{k}}}\boldsymbol{v}_{\mathbf{k}}\cdot e\mathcal{E} = \frac{-\partial f^0}{\partial E_{\mathbf{k}}} e\tau\mathcal{E}\cdot \int (\boldsymbol{v}_{\mathbf{k}} - \boldsymbol{v}_{\mathbf{k}'})\ W(\boldsymbol{k},\boldsymbol{k}^{'})\ d^3k^{'} \tag{B.29}$$

or

$$\boldsymbol{v}_{\mathbf{k}}\cdot\mathcal{E} = \tau \int (\boldsymbol{v}_{\mathbf{k}} - \boldsymbol{v}_{\mathbf{k}'})\ W(\boldsymbol{k},\boldsymbol{k}^{'})\ d^3k^{'}\cdot\mathcal{E} \tag{B.30}$$

and

$$\frac{1}{\tau} = \int W(\boldsymbol{k},\boldsymbol{k}^{'}) \left[1 - \frac{\boldsymbol{v}_{\mathbf{k}'}\cdot\mathcal{E}}{\boldsymbol{v}_{\mathbf{k}}\cdot\mathcal{E}}\right] d^3k^{'} \tag{B.31}$$

In general, this is a rather complex integral to solve. However, it becomes considerably simplified for certain simple cases. Consider, for example, the case of isotropic parabolic bands and elastic scattering. In figure B.4 we show a geometry for the scattering process. We choose a coordinate axis where the initial momentum is along the z-axis and the applied electric field is

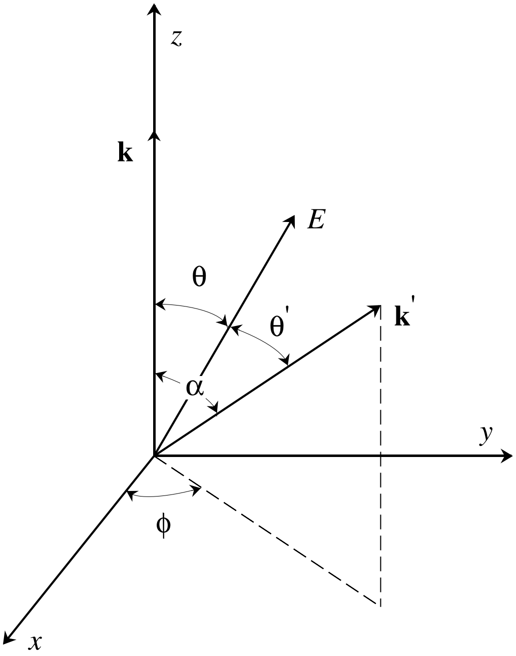


Figure B.4: Coordinate system illustrating a scattering event.

in the y-z plane. The wavevector after scattering is given by $\boldsymbol{k}^{'}$ represented by the angles α and ϕ. Assuming that the energy bands of the material is isotropic, $|\boldsymbol{v}_{\mathbf{k}}| = |\boldsymbol{v}_{\mathbf{k}'}|$. We thus get

$$\frac{\boldsymbol{v}_{\mathbf{k}'} \cdot \mathcal{E}}{\boldsymbol{v}_{\mathbf{k}} \cdot \mathcal{E}} = \frac{\cos\theta^{'}}{\cos\theta} \tag{B.32}$$

We can easily see from figure B.4 that

$$\cos\theta^{'} = \sin\theta \sin\alpha \sin\phi + \cos\theta\cos\alpha$$

or

$$\frac{\cos\theta^{'}}{\cos\theta} = \tan\theta \sin\alpha \sin\phi + \cos\alpha$$

When this term is integrated over ϕ to evaluate τ, the term involving $\sin\phi$ will integrate to zero for isotropic bands since $W(\boldsymbol{k}, \boldsymbol{k}^{'})$ does not have a ϕ dependence, only an α dependence. Thus

$$\frac{1}{\tau} = \int W(\boldsymbol{k}, \boldsymbol{k}^{'})\ (1 - \cos\alpha)\ d^3k^{'} \tag{B.33}$$

This weighting factor $(1 - \cos\alpha)$ confirms the intuitively apparent fact that large-angle scatterings are much more important in determining transport properties than small-angle scatterings. Forward-angle scatterings ($\alpha = 0$), in particular, have no detrimental effect on σ or μ for the case of elastic scattering.

Inelastic Collisions

In the case of inelastic scattering processes, we cannot assume that $W(\boldsymbol{k}, \boldsymbol{k}^{'}) = W(\boldsymbol{k}^{'}, \boldsymbol{k})$. As a result, the collision integral cannot be simplified to give an analytic result for the relaxation time. If, however, the system is non-degenerate, i.e., $f(E)$ is small, we can ignore second-order terms in f and we have

$$\left.\frac{\partial f}{\partial t}\right|_{\text{scattering}} = \int \left[g_{\mathbf{k}'} W(\boldsymbol{k}^{'}, \boldsymbol{k}) - g_{\mathbf{k}} W(\boldsymbol{k}, \boldsymbol{k}^{'})\right] \frac{d^3\boldsymbol{k}^{'}}{(2\pi)^3} \tag{B.34}$$

Under equilibrium we have

$$f^0_{\mathbf{k}'} W(\boldsymbol{k}^{'}, \boldsymbol{k}) = f^0_{\mathbf{k}} W(\boldsymbol{k}, \boldsymbol{k}^{'}) \tag{B.35}$$

or

$$W(\boldsymbol{k}^{'}, \boldsymbol{k}) = \frac{f^0_{\mathbf{k}}}{f^0_{\mathbf{k}'}} W(\boldsymbol{k}, \boldsymbol{k}^{'}) \tag{B.36}$$

Assuming that this relation holds for scattering rates in the presence of the applied field, we have

$$\left.\frac{\partial f}{\partial t}\right|_{\text{scattering}} = \int W(\boldsymbol{k}, \boldsymbol{k}^{'}) \left[g_{\mathbf{k}'} \frac{f^0_{\mathbf{k}}}{f^0_{\mathbf{k}'}} - g_{\mathbf{k}}\right] \frac{d^3\boldsymbol{k}^{'}}{(2\pi)^3} \tag{B.37}$$

The relaxation time then becomes

$$\frac{1}{\tau} = \int W(\boldsymbol{k}, \boldsymbol{k}^{'}) \left[1 - \frac{g_{\mathbf{k}'}}{g_{\mathbf{k}}} \frac{f^0_{\mathbf{k}}}{f^0_{\mathbf{k}'}}\right] \frac{d^3\boldsymbol{k}^{'}}{(2\pi)^3} \tag{B.38}$$

The Boltzmann is usually solved iteratively using numerical techniques.

B.2 AVERAGING PROCEDURES

We have so far assumed that the incident electron is on a well-defined state. In a realistic system the electron gas will have an energy distribution and τ, in general, will depend upon the energy of the electron. Thus it is important to address the appropriate averaging procedure for τ. We will now do so under the assumptions that the drift velocity due to the electric field is much smaller than the average thermal speeds so that the energy of the electron gas is still given by $3k_BT/2$.

Let us evaluate the average current in the system.

$$\boldsymbol{J} = \int e\, \boldsymbol{v}_{\mathbf{k}}\, g_{\mathbf{k}}\, \frac{d^3k}{(2\pi)^3} \tag{B.39}$$

The perturbation in the distribution function is

$$\begin{aligned} g_{\mathbf{k}} &= \frac{-\partial f^0}{\partial E_{\mathbf{k}}} \tau \boldsymbol{v}_{\mathbf{k}} \cdot e\mathcal{E} \\ &\approx \frac{f^0}{k_B T}\, \boldsymbol{v}_{\mathbf{k}} \cdot e\mathcal{E} \end{aligned} \tag{B.40}$$

If we consider a field in the x-direction, the average current in the x-direction is from equation B.39 and B.40

$$\langle J_x \rangle = \frac{e^2}{k_B T} \int \tau\, v_x^2\, f^0\, \frac{d^3 k}{(2\pi)^3}\ \ |\boldsymbol{\mathcal{E}}|_x \tag{B.41}$$

The assumption made on the drift velocity ensures that $v_x^2 = v^2/3$, where $\boldsymbol{v}$ is the total velocity of the electron. Thus we get

$$\langle J_x \rangle = \frac{e^2}{3 k_B T} \int \tau\, v^2\, f^0(\boldsymbol{k})\, \frac{d^3 k}{(2\pi)^3}\ \ |\boldsymbol{\mathcal{E}}|_x \tag{B.42}$$

Now we note that

$$\begin{aligned} \frac{1}{2} m^* \langle v^2 \rangle &= \frac{3}{2} k_B T \\ \Rightarrow k_B T &= m^* \langle v^2 \rangle / 3 \end{aligned}$$

also

$$\begin{aligned} \langle v^2\ \tau \rangle &= \frac{\int v^2\ \tau\ f^0(\boldsymbol{k})\ d^3k/(2\pi)^3}{\int f^0(\boldsymbol{k})\ d^3k/(2\pi)^3} \\ &= \frac{\int v^2\ \tau\ f^0(\boldsymbol{k})\ d^3k/(2\pi)^3}{n} \end{aligned} \tag{B.43}$$

Substituting in the right-hand side of equation B.42, we get (using $3k_B T = m \left\langle v^2 \right\rangle$)

$$\begin{aligned} \langle J_x \rangle &= \frac{n e^2}{m^*} \frac{\langle v^2 \tau \rangle}{\langle v^2 \rangle} |\boldsymbol{\mathcal{E}}|_x \\ &= \frac{n e^2}{m^*} \frac{\langle E \tau \rangle}{\langle E \rangle} |\boldsymbol{\mathcal{E}}|_x \end{aligned} \tag{B.44}$$

Thus, for the purpose of transport, the proper averaging for the relaxation time is

$$\langle\langle \tau \rangle\rangle = \frac{\langle E\tau \rangle}{\langle E \rangle} \tag{B.45}$$

Here the double brackets represent an averaging with respect to the perturbed distribution function while the single brackets represent averaging with the equilibrium distribution function.

For calculations of low-field transport where the condition $v_x^2 = v^2/3$ is valid, one has to use the averaging procedure given by equation B.45 to calculate mobility or conductivity of the semiconductors. For most scattering processes, one finds that it is possible to express the energy dependence of the relaxation time in the form

$$\tau(E) = \tau_0 (E/k_B T)^s \tag{B.46}$$

where τ_0 is a constant and s is an exponent which is characteristic of the scattering process. We will be calculating this energy dependence for various scattering processes in the next two chapters. When this form is used in the averaging of equation B.45, we get, using a Boltzmann distribution for $f^0(\boldsymbol{k})$

$$\langle\langle\tau\rangle\rangle = \tau_0 \frac{\int_0^\infty [p^2/(2m^* k_B T)]^s \ \exp[-p^2/(2m^* k_B T)] \ p^4 \ dp}{\int_0^\infty \exp[-p^2/(2m^* k_B T)] \ p^4 \ dp} \tag{B.47}$$

where $\boldsymbol{p} = \hbar\boldsymbol{k}$ is the momentum of the electron.

Substituting $y = p^2/(2m^* k_B T)$, we get

$$\langle\langle\tau\rangle\rangle = \tau_0 \frac{\int_0^\infty y^{s+(3/2)} e^{-y} dy}{\int_0^\infty y^{3/2} e^{-y} dy} \tag{B.48}$$

To evaluate this integral, we use Γ-functions which have the properties

$$\begin{aligned} \Gamma(n) &= (n-1)! \\ \Gamma(1/2) &= \sqrt{\pi} \\ \Gamma(n+1) &= n\,\Gamma(n) \end{aligned} \tag{B.49}$$

and have the integral value

$$\Gamma(a) = \int_0^\infty y^{a-1} e^{-y} dy \tag{B.50}$$

In terms of the Γ-functions we can then write

$$\langle\langle\tau\rangle\rangle = \tau_0 \frac{\Gamma(s+5/2)}{\Gamma(5/2)} \tag{B.51}$$

If a number of different scattering processes are participating in transport, the following approximate rule (Mathiesen's rule) may be used to calculate mobility:

$$\frac{1}{\tau_{tot}} = \sum_i \frac{1}{\tau_i} \tag{B.52}$$

$$\frac{1}{\mu_{tot}} = \sum_i \frac{1}{\mu_i} \tag{B.53}$$

where the sum is over all different scattering processes.

Appendix C

DENSITY OF STATES

In semiconductor devices we use the effective mass approximation to describe the properties of electrons in a crystal. Using the effective mass picture the Schrödinger equation for electrons can be written as a “free’ electron problem with a background potential V_0,

$$\frac{-\hbar^2}{2m^*}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)\psi(r) = (E - V_0)\psi(r)$$

A general solution of this equation is

$$\psi(r) = \frac{1}{\sqrt{V}}\ \exp\left(\pm ik \cdot r\right)$$

and the corresponding energy is

$$E = \frac{\hbar^2 k^2}{2m} + V_0$$

where the factor $1/\sqrt{V}$ in the wavefunction occurs because we wish to have one particle per volume V or

$$\int_V d^3r \mid \psi(r) \mid^2 = 1$$

We assume that the volume V is a cube of side L.

An important aspect of electronic bands is the density of states which tells us how many allowed energy levels there are between two energies. To obtain macroscopic properties independent of the chosen volume V, two kinds of boundary conditions are imposed on the wavefunction. In the first one the wavefunction is considered to go to zero at the boundaries of the volume, as shown in figure C.1a. In this case, the wave solutions are standing waves of the form sin($k_x x$) or cos($k_x x$), etc., and k-values are restricted to positive values:

$$k_x = \frac{\pi}{L}, \frac{2\pi}{L}, \frac{3\pi}{L} \cdots$$

Periodic boundary conditions are shown in figure C.2b. Even though we focus our attention on a finite volume V, the wave can be considered to spread in all space as we regard the entire space as made up of identical cubes of sides L. Then

$$\begin{array}{rcl} \psi(x,y,z+L) & = & \psi(x,y,z) \\ \psi(x,y+L,z) & = & \psi(x,y,z) \\ \psi(x+L,y,z) & = & \psi(x,y,z) \end{array}$$

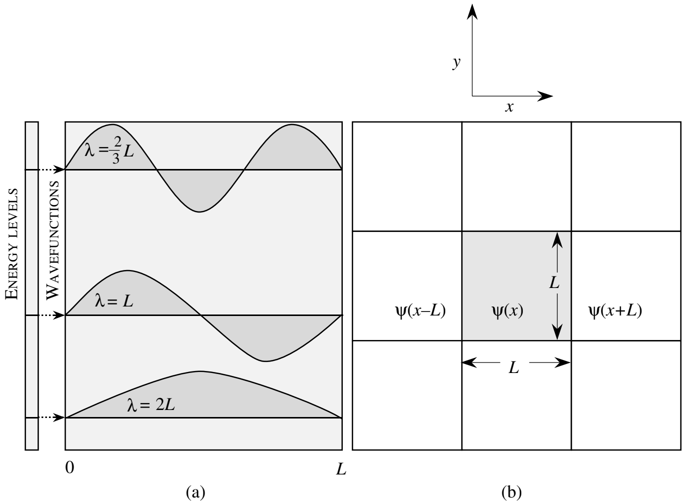


Figure C.1: Two types of boundary conditions. A schematic showing (a) the stationary boundary conditions; (b) the periodic boundary conditions.

In this case the allowed values of k are (n are integers—positive and negative)

$$k_x = \frac{2\pi n_x}{L}; \;\; k_y = \frac{2\pi n_y}{L}; \;\; k_z = \frac{2\pi n_z}{L}$$

If L is large, the spacing between the allowed k-values is very small. Also it is important to note that the results one obtains for properties of the particles in a large volume are independent of whether we use the stationary or periodic boundary conditions. It is useful to discuss the volume in k-space that each electronic state occupies. As can be seen from figure C.2, this volume is (in three dimensions)

$$\left(\frac{2\pi}{L}\right)^3 = \frac{8\pi^3}{V} \tag{C.1}$$

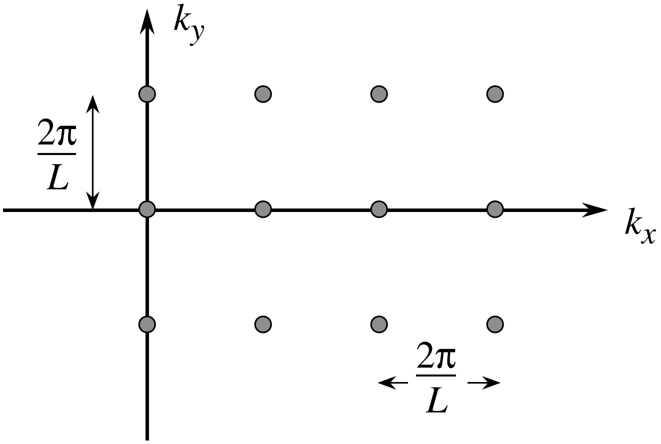


Figure C.2: k-Space volume of each electronic state. The separation between the various allowed components of the k-vector is $\frac{2\pi}{L}$.

If Ω is a volume of k-space, the number of electronic states in this volume is

$$\frac{\Omega V}{8\pi^3}$$

It is easy to verify that stationary and periodic boundary conditions lead to the same density of states value as long as the volume is large.

Density of States for a Three-Dimensional System

Important physical properties in materials such as optical absorption, transport, etc., are intimately dependent upon how many allowed states there are. Density of states is the number of available electronic states per unit volume per unit energy around an energy E. If we denote the density of states by $N(E)$, the number of states in a unit volume in an energy interval dE around an energy E is $N(E)dE$. To calculate the density of states, we need to know the dimensionality of the system and the energy versus k relation that the particles obey. We will choose the particle of interest to be the electron, since in most applied problems we are dealing with electrons. Of course, the results derived can be applied to other particles as well. For the free electron case we have the parabolic relation

$$E = \frac{\hbar^2 k^2}{2m^*} + V_0$$

The energies E and $E + dE$ are represented by surfaces of spheres with radii k and $k + dk$, as shown in figure C.3. In a three-dimensional system, the k-space volume between vector k and $k + dk$ is (see figure C.3a) $4\pi k^2 dk$. We have shown in equation C.1 that the k-space volume per electron state is $(\frac{2\pi}{L})^3$. Therefore, the number of electron states in the region between k and $k + dk$ is

$$\frac{4\pi k^2 dk}{8\pi^3} V = \frac{k^2 dk}{2\pi^2} V$$

Denoting the energy and energy interval corresponding to k and dk as E and dE, we see that the number of electron states between E and E + dE per unit volume is

$$N(E)\, dE = \frac{k^2 dk}{2\pi^2}$$

Using the E versus k relation for the free electron, we have

$$k^2 dk = \frac{\sqrt{2} m^{*3/2} (E - V_0)^{1/2} dE}{\hbar^3}$$

and

$$N(E)\, dE = \frac{m^{*3/2} (E - V_0)^{1/2} dE}{\sqrt{2}\pi^2 \hbar^3}$$

The electron can have a spin state $\hbar/2$ or $-\hbar/2$. Accounting for spin, the density of states obtained is simply multiplied by 2

$$N(E) = \frac{\sqrt{2} m^{*3/2} (E - V_0)^{1/2}}{\pi^2 \hbar^3}$$

Density of States in Sub-Three-Dimensional Systems

In quantum wells electrons are free to move in a 2-dimensional space. The two-dimensional density of states is defined as the number of available electronic states per unit area per unit energy around an energy E. Similar arguments as used in the derivation show that the density of states for a parabolic band (for energies greater than V_0) is (see figure C.3b)

$$N(E) = \frac{m^*}{\pi \hbar^2}$$

The factor of 2 resulting from spin has been included in this expression. Finally, we can consider a one-dimensional system often called a “quantum wire.” The one-dimensional density of states is defined as the number of available electronic states per unit length per unit energy around an energy E. In a 1D system or a “quantum wire” the density of states is (including spin) (see figure C.3c)

$$N(E) = \frac{\sqrt{2} m^{*1/2}}{\pi \hbar} (E - V_0)^{-1/2}$$

Notice that as the dimensionality of the system changes, the energy dependence of the density of states also changes. As seen in figure C.4, for a three-dimensional system we have $(E - V_0)^{1/2}$ dependence, for a two-dimensional system we have no energy dependence, and for a one-dimensional system we have $(E - V_0)^{-1/2}$ dependence. The changes in density of states with dimensions are exploited in electronic and optoelectronic devices.

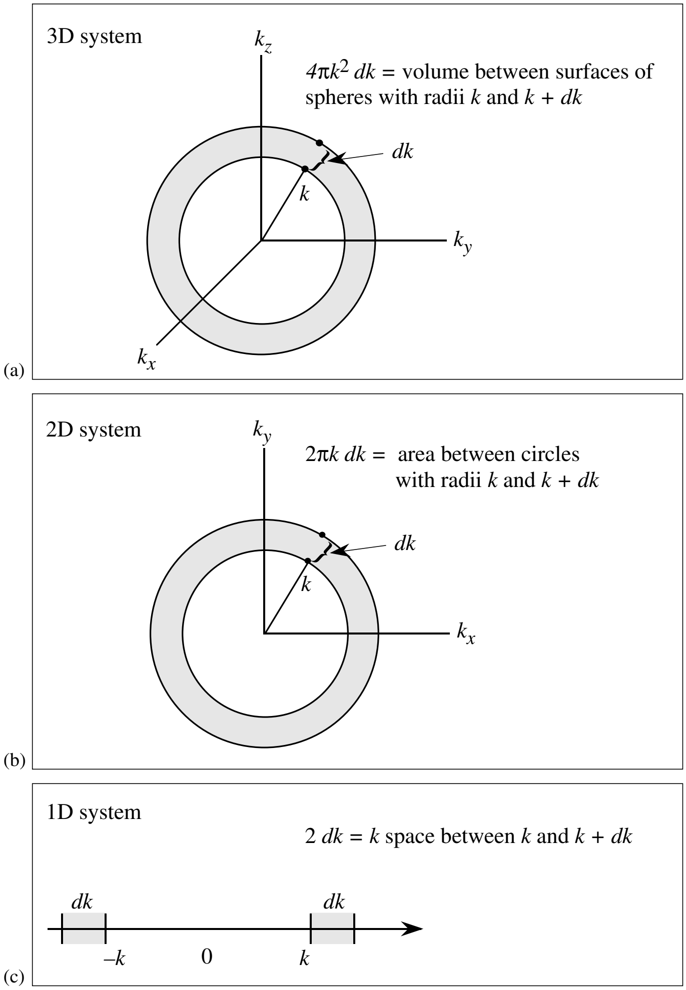

3D system
k_z
$4\pi k^2\,dk$ = volume between surfaces of spheres with radii k and $k + dk$
dk
k
k_y
k_x
(a)
2D system
k_y
$2\pi k\,dk$ = area between circles with radii k and $k + dk$
dk
k
k_x
(b)
1D system
$2\,dk$ = k space between k and $k + dk$
dk
dk
$-k$
0
k
(c)

Figure C.3: Geometry used to calculate density of states in three, two, and one dimensions. By finding the k-space volume in an energy interval between E and $E+dE$, one can find the number of allowed states.

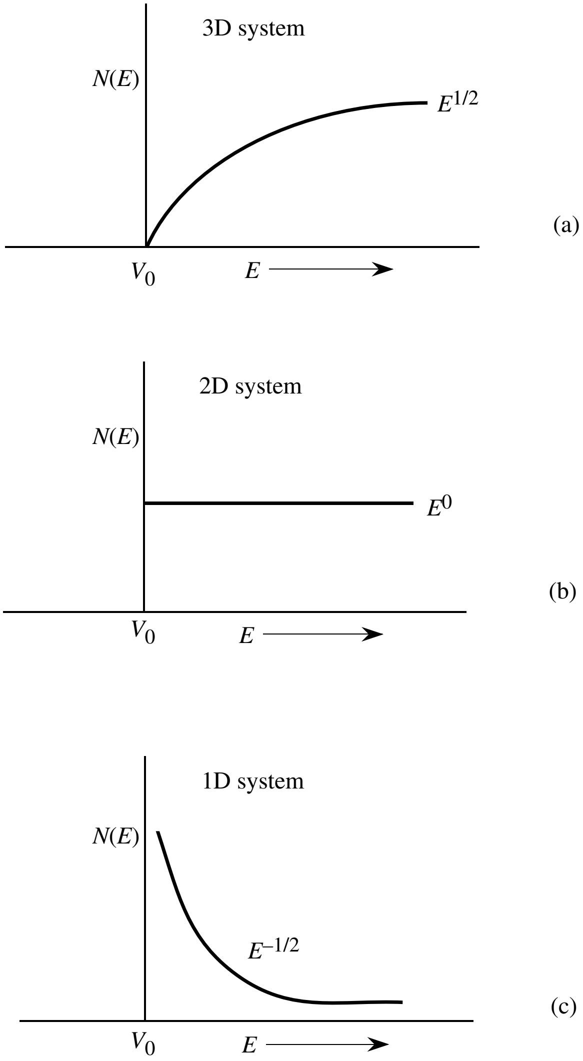


Figure C.4: Energy dependence of the density of states in (a) three-dimensional, (b) two-dimensional, and (c) one-dimensional systems. The energy dependence of the density of states is determined by the dimensionality of the system.

Appendix D

IMPORTANT PROPERTIES OF SEMICONDUCTORS

The data and plots shown in this Appendix are extracted from a number of sources. A list of useful sources is given below.

- S. Adachi, J. Appl. Phys., 58, R1 (1985).
- H.C. Casey, Jr. and M.B. Panish, Heterostructure Lasers, Part A, "Fundamental Principles;" Part B, "Materials and Operating Characteristics," Academic Press, N.Y. (1978).
- Landolt-Bornstein, Numerical Data and Functional Relationship in Science and Technology, Vol. 22, Eds. O. Madelung, M. Schulz, and H. Weiss, Springer-Verlog, N.Y. (1987). Other volumes in this series are also very useful.
- S.M. Sze, Physics of Semiconductor Devices, Wiley, N.Y. (1981). This is an excellent source of a variety of useful information on semiconductors.
- "World Wide Web;" A huge collection of data can be found on the Web. Several professors and industrial scientists have placed very useful information on their websites.

Material	Structure	Lattice Constant (Å)	Density (gm/cm^3)
C	Diamond	3.5668	3.5153
Si	Diamond	5.431	2.329
Ge	Diamond	5.658	5.323
GaAs	Zinc Blende	5.653	5.318
AlAs	Zinc Blende	5.660	3.760
InAs	Zinc Blende	6.058	5.667
GaN	Wurtzite	$a = 3.175$; $c = 5.158$	6.095
AlN	Wurtzite	$a = 3.111$; $c = 4.981$	3.255
SiC	Zinc Blende	4.360	3.166
Cd	hcp	$a = 2.98$; $c = 5.620$	8.65
Cr	bcc	2.88	7.19
Co	hcp	$a = 2.51$; $c = 4.07$	8.9
Au	fcc	4.08	19.3
Fe	bcc	2.87	7.86
Ag	fcc	4.09	10.5
Al	fcc	4.05	2.7
Cu	fcc	3.61	8.96

Table D.1: Lattice constants and density of some semiconductors.

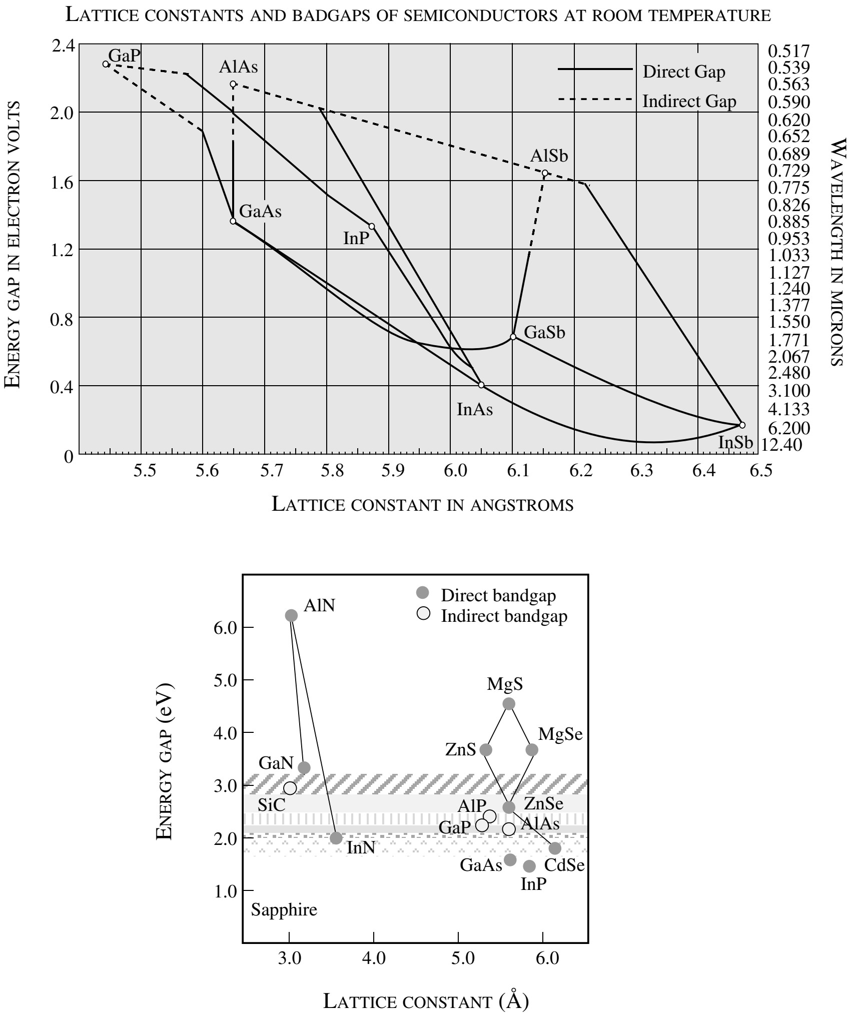


Figure D.1: Lattice constants and bandgaps of semiconductors at room temperature.

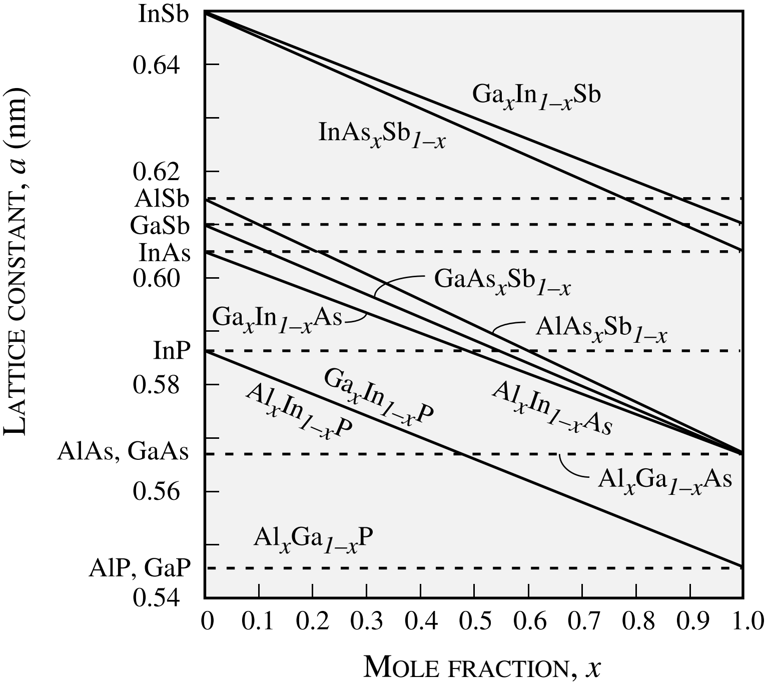


Figure D.2: Lattice constants of several alloy systems.

		Experimental Energy Gap E_g (eV)		
Semi-conductor	**Type of Energy Gap**	**0 K**	**300 K**	**Temperature Dependence of Energy Gap (eV)**
AlAs	Indirect	2.239	2.163	2.239–6.0 x $10^{-4}T^2$/(T + 408)
GaP	Indirect	2.338	2.261	2.338–5.771 x $10^{-4}T^2$/(T + 372)
GaAs	Direct	1.519	1.424	1.519–5.405 x $10^{-4}T^2$/(T + 204)
GaSb	Direct	0.810	0.726	0.810-–3.78 x $10^{-4}T^2$/(T + 94)
InP	Direct	1.421	1.351	1.421–3.63 x $10^{-4}T^2$/(T + 162)
InAs	Direct	0.420	0.360	0.420–2.50 x $10^{-4}T^2$/(T + 75)
InSb	Direct	0.236	0.172	0.236–2.99 x $10^{-4}T^2$/(T + 140)
Si	Indirect	1.17	1.11	1.17–4.37 x $10^{-4}T^2$/(T + 636)
Ge	Indirect	0.66	0.74	0.74–4.77 x $10^{-4}T^2$/(T + 235)

Table D.2: Energy gaps of some semiconductors along with their temperature dependence.

Material	Electron Mass (m_0)	Hole Mass (m_0)
AlAs	0.1	
AlSb	0.12	$m_{\rm dos} = 0.98$
GaN	0.19	$m_{\rm dos} = 0.60$
GaP	0.82	$m_{\rm dos} = 0.60$
GaAs	0.067	$m_{\rm lh} = 0.082$ $m_{\rm hh} = 0.45$
GaSb	0.042	$m_{\rm dos} = 0.40$
Ge	$m_l = 1.64$ $m_{\rm t} = 0.082$	$m_{\rm lh} = 0.044$ $m_{\rm hh} = 0.28$
InP	0.073	$m_{\rm dos} = 0.64$
InAs	0.027	$m_{\rm dos} = 0.4$
InSb	0.13	$m_{\rm dos} = 0.4$
Si	$m_l = 0.98$ $m_{\rm t} = 0.19$	$m_{\rm lh} = 0.16$ $m_{\rm hh} = 0.49$

Table D.3: Electron and hole masses for several semiconductors. Some uncertainty remains in the value of hole masses for many semiconductors.

Compound	Direct Energy Gap E_g (eV)
Al_xIn_{1-x} P	$1.351 + 2.23x$
Al_xGa_{1-x} As	$1.424 + 1.247x$
Al_xIn_{1-x} As	$0.360 + 2.012x + 0.698x^2$
Al_xGa_{1-x} Sb	$0.726 + 1.129x + 0.368x^2$
Al_xIn_{1-x} Sb	$0.172 + 1.621x + 0.43x^2$
Ga_xIn_{1-x} P	$1.351 + 0.643x + 0.786x^2$
Ga_xIn_{1-x} As	$0.36 + 1.064x$
Ga_xIn_{1-x} Sb	$0.172 + 0.139x + 0.415x^2$
GaP_xAs_{1-x}	$1.424 + 1.150x + 0.176x^2$
$GaAs_xSb_{1-x}$	$0.726 + 0.502x + 1.2x^2$
InP_xAs_{1-x}	$0.360 + 0.891x + 0.101x^2$
$InAs_xSb_{1-x}$	$0.18 + 0.41x + 0.58x^2$

Table D.4: Compositional dependence of the energy gaps of the binary III-V ternary alloys at 300 K. (After Casey and Panish (1978).)

	Bandgap (eV)		Mobility at 300 K (cm^2/V-s)	
Semiconductor	300 K	0 K	Elec.	Holes
C	5.47	5.48	1800	1200
GaN	3.4	3.5	1400	350
Ge	0.66	0.74	3900	1900
Si	1.12	1.17	1500	450
α-SiC	3.00	3.30	400	50
GaSb	0.72	0.81	5000	850
GaAs	1.42	1.52	8500	400
GaP	2.26	2.34	110	75
InSb	0.17	0.23	80000	1250
InAs	0.36	0.42	33000	460
InP	1.35	1.42	4600	150
CdTe	1.48	1.61	1050	100
PbTe	0.31	0.19	6000	4000
$In_{0.53}Ga_{0.47}As$	0.8	0.88	11000	400

Table D.5: Bandgaps along with electron and hole mobilities in several semiconductors. Properties of large bandgap materials (C, GaN, SiC) are continuously changing (mobility is improving), due to progress in crystal growth. Zero temperature bandgap is extrapolated.

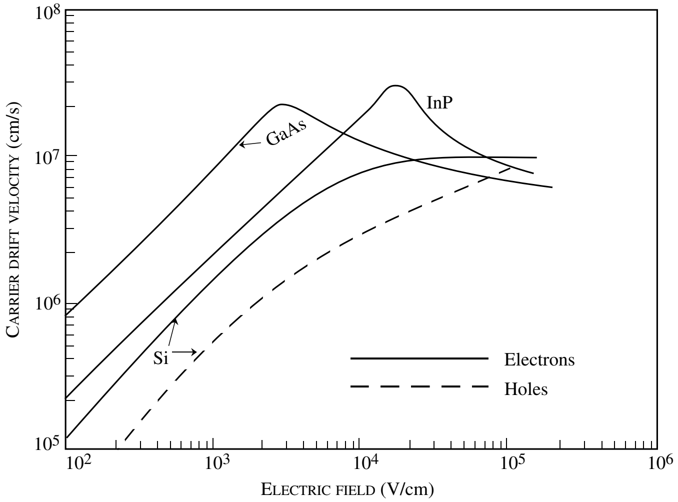


Figure D.3: Velocity-Field relations for several semiconductors at 300 K.

Material	Bandgap (eV)	Breakdown electric field (V/cm)
GaAs	1.43	4 x 10^5
Ge	0.664	10^5
InP	1.34	
Si	1.1	3 x 10^5
$In_{0.53}Ga_{0.47}As$	0.8	2 x 10^5
C	5.5	10^7
SiC	2.9	2-3 x 10^6
SiO_2	9	–10^7
Si_3N_4	5	–10^7
GaN	3.4	2 x 10^6

Table D.6: Breakdown electric fields in some semiconductors.

Appendix E

BEYOND THE DEPLETION APPROXIMATION

In the depletion approximation the contribution of mobile charges to the electrostatics of the depletion region was neglected. This allowed one to accurately define depletion region edges beyond which the material was neutral. A schematic of this structure is shown below in figure E.1.

However, this picture is not physical because the mobile charges cannot abruptly go to zero but will decrease in a manner predicted by the law of the junction where

$$n = n_{n0} e^{\frac{-q\Psi}{k_B T}} \tag{E.1}$$

where Ψ is the band bending measured from the bulk. This is shown schematically in figure E.2.

As the mobile charge concentration decreases exponentially with band bending the net charge in the regions close to the depletion region edge is no longer given by the depletion charge, but as is always the case in general, the sum of all mobile and fixed charges. Studying the p-side of the junction

$$\rho = \frac{-eN_A^- + ep_p(\Psi)}{\epsilon}$$

and $\frac{\partial \mathcal{E}}{\partial x}$ derivates from the linear relationship when the charge is constant. This leads to "skirts" in the $\mathcal{E}$ vs. x relationship. It also raises the question, "what is the depletion region edge?" The depletion region edge is defined by extrapolating the linear region of the curve (where the mobile charges are negligible) to zero. We recognize that the area under the $\mathcal{E}$ vs. x is the built-in voltage of the junction, V_{bi}. This is obviously larger than the area of the triangle, specifically by the area of the "skirts" shown shaded in figure E.2. We will show shortly that each of the areas is of the order $\frac{k_B T}{e}$, the thermal voltage. Hence the area under the triangular $\mathcal{E}$ vs. x curved bounded by $-W_p$ and $+W_n$ is

$$V_{bi}^{'} = V_{bi} - \frac{k_B T}{e} - \frac{k_B T}{e} = V_{bi} - \frac{2k_B T}{e} \tag{E.2}$$

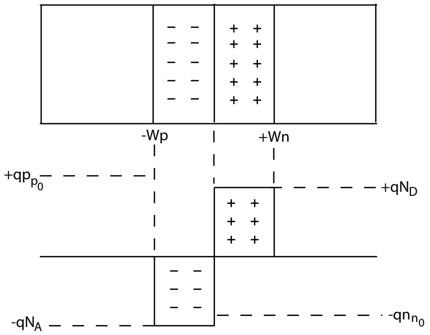


Figure E.1: Schematic of a $p - n$ junction within the depletion approximation.

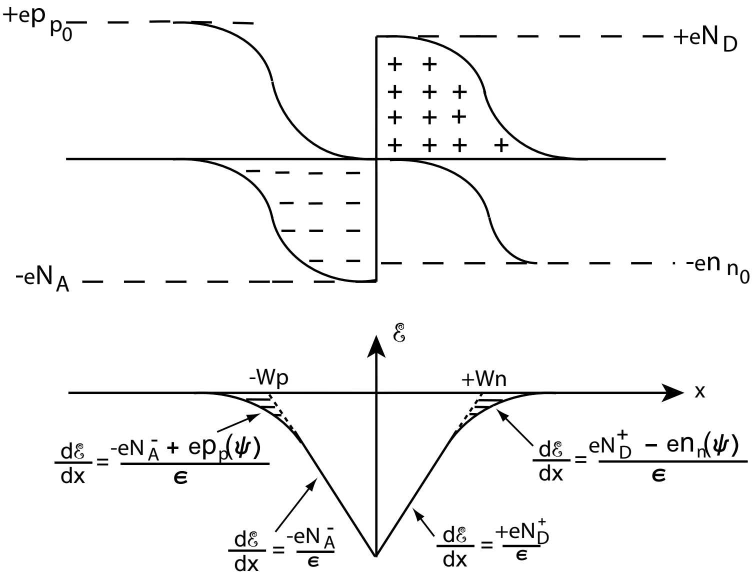

+ep$_{p_0}$
+eN$_D$
-eN$_A$
-en$_{n_0}$
ℰ
-Wp
+Wn
x
$\frac{d\mathcal{E}}{dx} = \frac{-eN_A^- + ep_p(\psi)}{\epsilon}$
$\frac{d\mathcal{E}}{dx} = \frac{eN_D^+ - en_n(\psi)}{\epsilon}$
$\frac{d\mathcal{E}}{dx} = \frac{-eN_A^-}{\epsilon}$
$\frac{d\mathcal{E}}{dx} = \frac{+eN_D^+}{\epsilon}$

Figure E.2: Schematic of a p-n junction within the Gummel correction to the depletion approximation

This is called the Gummel correction to the built-in voltage. To apply the depletion approximation and calculate parameters related to electrostatics such as depletion region width, depletion capacitance etc., it is necessary to substitute $V_{bi}^{'}$ for V_{bi} in previous formulae. Hence,

$$W = \sqrt{\frac{2\epsilon_s}{e}\left(\frac{1}{N_A} + \frac{1}{N_D}\right)(V_{bi}^{'})} \tag{E.3}$$

note that in a Schottky barrier the correction due to the thermal broadening of carriers (which occurs over a Debye Length, L_D) occurs in only the semiconductor and hence

$$V_{bi}^{'} = V_{bi} - \frac{k_B T}{e}$$

for a Schottky barrier.

The Gummel correction is arrived at by solving Poisson's equation in the depletion region including the contribution of mobile charges. Consider the band diagram of a p-n junction in figure E.3. For the purpose of our analysis we will only consider the p-type semiconductor. The analysis is equivalent for the n-side. The governing equations are

$$\frac{d^2\Psi}{dx^2} = -\frac{\rho(x)}{\epsilon} \quad (\,Poisson's\ \ equation\,) \tag{E.4}$$

and

$$\rho(x) = q(N_D^+ - N_A^- + p_p - n_p) \tag{E.5}$$

where N_D^+ and N_A^- are the ionized donors and acceptors respectively with the latter dominant in the p-region. In the bulk of the semiconductor charge neutrality requires $\rho(x) = 0$ or from equation E.5

$$N_D^+ - N_A^- = n_{p0} - p_{p0} \tag{E.6}$$

Applying equation E.6 to equation E.5 and equation E.4 we get the resultant Poisson's equation

$$\frac{d^2\Psi}{dx^2} = -\frac{e}{\epsilon}\left[(p_p - p_{p0}) - (n_p - n_{p0})\right] \tag{E.7}$$

From Boltzmann statistics and figure E.3 we know $p_p = p_{p0}e^{-\frac{e\Psi}{k_BT}}$ and $n_p = n_{p0}e^{+\frac{e\Psi}{k_BT}}$ or

$$\frac{d^2\Psi}{dx^2} = -\frac{e}{\epsilon}\left[p_{p0}(e^{-\frac{e\Psi}{k_BT}} - 1) - n_{p0}(e^{+\frac{e\Psi}{k_BT}} - 1)\right] \tag{E.8}$$

Recognizing that $\left(\frac{\partial\Psi}{\partial x}\right) d\left(\frac{\partial\Psi}{\partial x}\right) = \frac{\partial^2\Psi}{\partial x^2}d\Psi$ we can integrate equation E.8 from the bulk towards the junction

$$\int_0^{\frac{\partial\Psi}{\partial x}} \left(\frac{\partial\Psi}{\partial x}\right) d\left(\frac{\partial\Psi}{\partial x}\right) = -\frac{e}{\epsilon}\int_0^{\Psi}\left[p_{p0}(e^{-\frac{e\Psi}{k_BT}} - 1) - n_{p0}(e^{+\frac{e\Psi}{k_BT}} - 1)\right] d\Psi \tag{E.9}$$

Using $\mathcal{E} = \frac{-\partial\Psi}{\partial x}$ we get

$$\mathcal{E} = \left(\frac{2k_BT}{\epsilon}p_{p0}\left[\left(e^{-\frac{e\Psi}{k_BT}} + \frac{e\Psi}{k_BT} - 1\right) + \frac{n_{p0}}{p_{p0}}\left(e^{-\frac{e\Psi}{k_BT}} - \frac{e\Psi}{k_BT} - 1\right)\right]\right)^{\frac{1}{2}} \tag{E.10}$$

For our purposes of understanding the origin of the Gummel correction we will evaluate equation E.10 for a condition of mild depletion, where Ψ is small and positive, of such magnitude that

$$\mathcal{E} = \left(\frac{2k_BT}{\epsilon}p_{p0}\right)^{\frac{1}{2}}\left(\frac{e\Psi}{k_BT} - 1\right)^{\frac{1}{2}} \tag{E.11}$$

where the $2^{\underline{nd}}$ term in parentheses in equation E.10 is neglected because of the $\frac{n_{p0}}{p_{p0}}$ pre-factor and the $e^{-\frac{e\Psi}{k_BT}}$ is neglected because Ψ is positive. Thus

$$\mathcal{E} = \sqrt{\frac{2p_{p0}}{\epsilon e}\left(\Psi - \frac{k_BT}{e}\right)}$$

This is identical to the depletion approximation except for Ψ being replaced by $\Psi - \frac{k_BT}{e}$. This reflects the reduced electric field because of the effect of mobile charges (in our case holes) at the depletion region edge.

Therefore, the depletion region edge is defined by using the depletion approximation while reducing the built-in potential by $\frac{k_BT}{q}$ at each depletion region edge as shown in figure E.2 and stated in equation E.3.

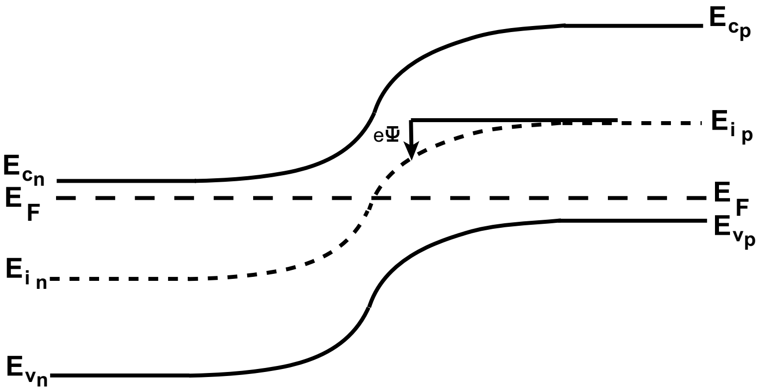


Figure E.3: Band diagram of a p-n junction showing the references used to describe $e\Psi$.

Appendix F

DESIGN OF GRADED HETEROJUNCTIONS FOR BIPOLAR TRANSISTORS

This appendix discusses the design of graded heterojunctions for bipolar transistors using an example from the text (Example 5.3).

Consider four different n-p$^+$ Al$_{0.3}$Ga$_{0.7}$As/GaAs heterojunctions with N$_D = 10^{17}$ and $N_A = 5 \times 10^{18}$. The AlGaAs in these junctions is graded from $x = 0$ to $x = 0.3$ over $X_{Grade} = 0$ (abrupt), $X_{Grade} = 100\ \mathring{A}$, $X_{Grade} = 300\ \mathring{A}$, and $X_{Grade} = 1\ \mu m.$ Calculate and plot the energy band diagrams for the above four cases. Assume the dielectric constant of AlGaAs to be the same as that of GaAs.

Solution: E$_g$ = 1.8 eV for Al$_{0.3}$Ga$_{0.7}$As, and E$_g$ = 1.42 eV for GaAs. ΔE$_g$ = 0.374 eV, ΔE$_C$ = 0.237 eV, and ΔE$_V$ = 0.137 eV. On the AlGaAs emitter side, the conduction band energy relative to the Fermi level far away from the junction is given by

$$\phi_n = \frac{E_C - E_F}{e} = \frac{kT}{e}\, ln(\frac{N_C}{n}) = 0.0323\ V. \tag{F.1}$$

Since the p-GaAs is degenerately doped, Joyce-Dixon statistics must be applied:

$$\phi_p = E_V - E_F = \frac{kT}{e}\, (ln(\frac{p}{N_V}) + A_1(\frac{p}{N_V}) + A_2(\frac{p}{N_V})^2) = 0.011\ V. \tag{F.2}$$

The built-in potential in the conduction band, ϕ_{bi} is given by

$$\phi_{bi} = \frac{1}{e}(E_g(GaAs) + \Delta E_c) - (\phi_p + \phi_n) = 1.62\ V. \tag{F.3}$$

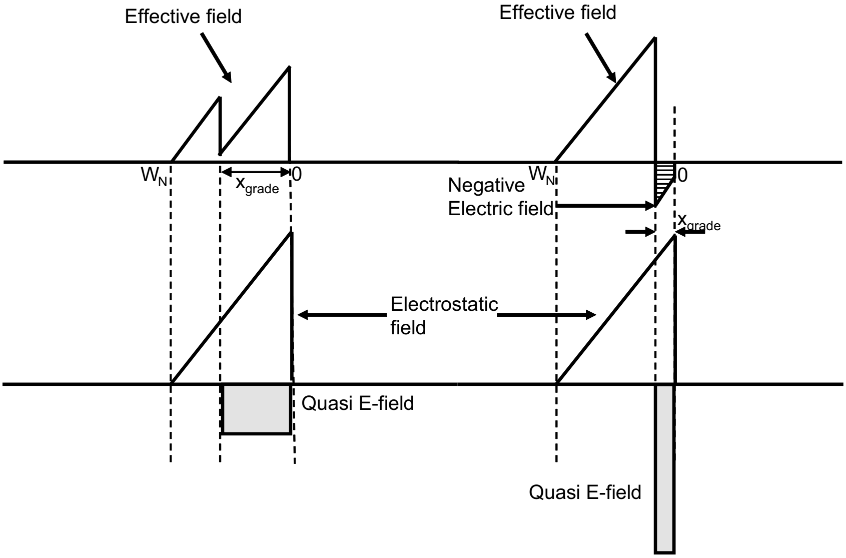


Figure F.1: Electric field in a graded heterojunction. If the grading distance is too short (Right), the quasi-electric field can cause the effective electric field to reverse direction, leading to a barrier in the conduction band. When designed correctly, the quasi-electric field magnitude is lower than the electrostatic field (Left).

Assuming x$_{D1}$ and x$_{D2}$ are the depletion thicknesses in the n and p regions, and solving,

$$N_D x_{D1} = N_A x_{D2} \tag{F.4}$$

$$\frac{e}{2\epsilon}(N_D W_n^2 + N_A W_p^2) = \phi_{bi} \tag{F.5}$$

W$_n$ = 1.5 × 10^{-5}, and W$_p$ = 3.0 × 10^{-7}.

Since W$_n$ and W$_p$ are known, the electrostatic potential can now be calculated. The band profiles are found by superimposing the electrostatic and quasi-electric fields. The quasi electric field is given by $-\frac{\Delta E_C}{e\ x_{grade}}$ for the conduction band, and $\frac{\Delta E_V}{e\ x_{grade}}$ for the valence band.

In the 100 $\AA$ and 300 $\AA$ cases, we can assume that the depletion width is much larger than the grading distance. The electric field in the conduction band is given by the following equations:

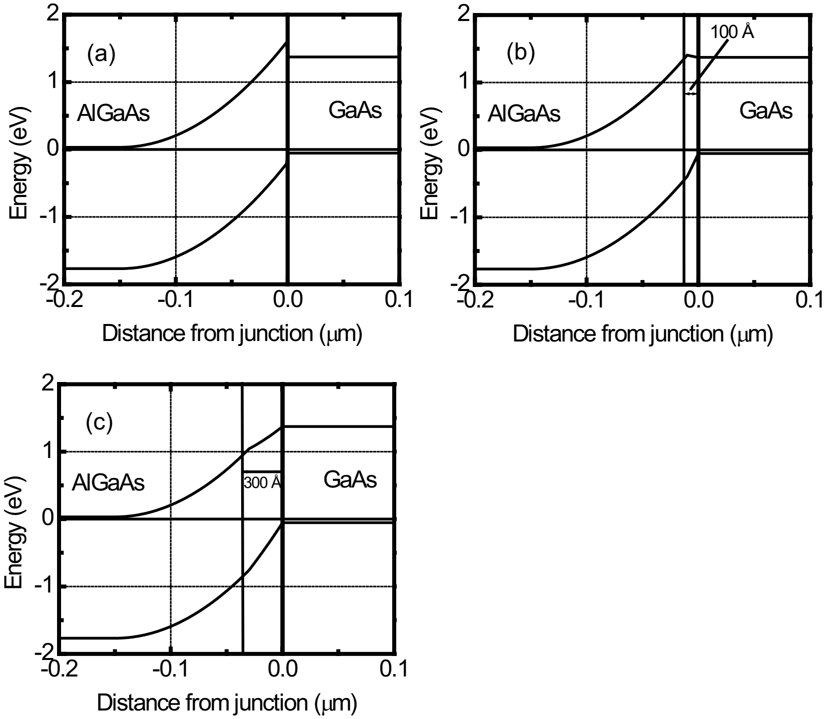


Figure F.2: Calculated band profiles for the graded heterojunctions with (a) abrupt, (b)100$\mathring{A}$, and (c) 300$\mathring{A}$ grade. There is no bump in the 300$\mathring{A}$ case.

$$0 < x < x_{grade} \qquad \mathcal{E} = -\frac{eN_D}{\epsilon}(1 - \frac{x}{W_n}) + \frac{\Delta E_C}{e\, x_{grade}} \tag{F.6}$$

$$x_{grade} < x < W_n : \qquad \mathcal{E} = -\frac{eN_D}{\epsilon}(1 - \frac{x}{W_n}) \tag{F.7}$$

$$x > W_n : \qquad \mathcal{E} = 0 \tag{F.8}$$

The equations describing the valence band potential are similar except that ΔE_C is replaced by $-\Delta E_V$.

The final potential is found by integrating the piecewise electric field function above. The conduction band at the junction ($x = 0$) is given by

$$\phi(0) = \phi_n + \frac{eN_D W_n^2}{2\epsilon}. \tag{F.9}$$

The conduction band profile is given by the following equations.

$$0 < x < x_{grade} \qquad E_C = \phi(0) - \frac{eN_D}{\epsilon}(x - \frac{x^2}{2W_n}) + \frac{\Delta E_C x}{e\, x_{grade}} \tag{F.10}$$

$$x_{grade} < x < W_n : \qquad E_C = \phi(0) - \frac{eN_D}{\epsilon}(x - \frac{x^2}{2W_n}) + \frac{\Delta E_C}{e} \tag{F.11}$$

$$x > W_n : \qquad E_C = \phi_n \tag{F.12}$$

In the case where the AlGaAs is graded over 1 μm, the quasi-electric field is very small compared to the electrostatic field. The electrostatic depletion depth is therefore much smaller than the grading distance. The junction behaves almost like a n-GaAs/p-GaAs homojunction, and very little performance advantage is gained from using a heterojunction.

The band profiles for the three different grading conditions, (a) abrupt grade, (b) 100 $\AA$ grade, and (c) 300 $\AA$ grade are shown in Figure 2.

The quasi-electric field can create an undesirable bump in the conduction band if not designed correctly, as seen in Figure 2 for the abrupt and the 100 $\AA$ case. The 300 $\AA$ grade is best suited for the HBT since it does not lead to a barrier to electron flow.

INDEX