SOLID STATE ELECTRONIC DEVICES

SECOND EDITION

D.K. Bhattacharya
Scientist
Solid State Physics Laboratory
New Delhi

Rajnish Sharma
Dean (Academics),
Chitkara University,
Himachal Pradesh

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries.

Published in India by
Oxford University Press
YMCA Library Building, 1 Jai Singh Road, New Delhi 110001, India

First Edition published in 2007
Second Edition published in 2013

ISBN-13: 978-0-19-808457-0
ISBN-10: 0-19-808457-9

Typeset in Times New Roman
by Time Digitech Private Limited, Noida
Printed in India by Yash Printographics, Noida 201301

About the Authors

D.K. Bhattacharya A PhD from the University of Delhi, he has over two decades of experience as a practising semiconductor scientist including a long association with the MEMS Division, Solid State Physics Laboratory, New Delhi. He has published more than 50 research papers in national and international journals and conference proceedings. He is also the author of *Engineering Physics* (OUP, 2010).

Rajnish Sharma A PhD from Kurukshetra University and National Physical Laboratory, New Delhi, he has served BITS, Pilani as a faculty for almost six years. He has successfully handled a research project funded by the Ministry of Science and Technology, Government of India and has presented his work at the MRS Symposium held at San Francisco and at Risoe National Laboratory, Denmark. His current research interests are in the area of device technologies for next-generation ICs and circuit design at radio frequencies.

Preface to the Second Edition

Solid state electronic devices are devices that are built entirely from solid materials and in which electrons, or other charge carriers, are confined entirely within the solid material. Prior to the use of solid state devices, most electronic devices used vacuum tubes, in which electricity passed through various elements inside a heated vacuum tube. In many applications, solid state devices have replaced vacuum tubes since they last longer and are smaller, cheaper, and more efficient and reliable.

While solid state devices can be built from crystalline, polycrystalline, and amorphous solids, the building material is most often a crystalline semiconductor. Common solid state devices include transistors, microprocessor chips, and dynamic random access memory (DRAM) chips. The first solid state device was the 'cat's whisker' detector which was first used in 1930s as radio receivers. The **transistor**, invented in 1947 by Bell Labs and named an IEEE Milestone in 2009, was the first solid state device to come into commercial use in the 1960s. More recently, the integrated circuit (IC), the light-emitting diode (LED), and the liquid crystal display (LCD) have evolved as further examples of solid state devices.

The semiconductor industry is seeing steady change, based on continued technological innovations and the requirement of reducing cost. The next big development in semiconductor manufacturing is the transition from 300 mm wafers to 450 mm wafers, which can lead to reduction in die cost and subsequently production cost. Other exciting recent developments in semiconductor industry are the extreme ultraviolet (EUV) lithography and nano-chip technology. A nano-chip can carry billions of transistors, and its applications include high-performance servers and supercomputers, virtual reality and advanced electronic games, and ultra-fast telecommunications devices. It is these types of applications that could lead to a new revolution in how electronics goods are designed and manufactured.

Recognizing the importance of this industry in our everyday lives, this second edition of *Solid State Electronic Devices* offers an improved coverage of the fundamental concepts of solid state electronics. This edition is an attempt to incorporate most of the feedback received from educators in terms of improvement of content presented in the book.

About the Book

This book contains 14 chapters which provide a thorough coverage of solid state electronic devices. Starting with the fundamentals of solid state physics such as electron dynamics, growth and crystal properties of semiconductors, energy bands, and charge carriers, the book goes on to the study of p-n junctions, metal–semiconductor contacts, BJTs, and FETs (including JFET, MESFET, MOSFET, and HEMT technologies). An analysis of special devices, such as opto-electronic devices, microwave devices, and power devices, follows. Finally, the book covers ICs, MEMS, rectifiers, and power supplies.

Each chapter has been divided into small sections that are independent in themselves. A unique feature of the book is the additional information in shaded boxes at the end of relevant sections. These inputs offer a few extra facts outside the limits of the curriculum to draw the interest and attention of students.

A large number of numerical problems along with answers and hints have been provided at the end of each chapter. This is followed by a recapitulation of the salient features of the chapter. Numerous review exercises and numerical problems along with answers and hints have been provided as well. A list of references is presented at the end of the book for those interested in further reading. A special attempt has been made to include topics that are a part of courses offered by a large cross-section of educational institutions.

New to the Second Edition

- New sections such as reciprocal lattice, band structure modification, electrons and holes in quantum wells, Early effect, short channel MOSFET I-V characteristics, and photoluminescence and electroluminescence
- Detailed explanation and coverage
- New illustrations

Extended Chapter Material

Chapter 1 Applications of cathode ray tubes (CRTs) have been added.

Chapter 2 Reciprocal lattice, diffraction due to crystal planes, floatzone (FZ) method of crystal growth, and four-probe method for the measurement of conductivity of semiconductors are explained.

Chapter 3 The concept of phonon has been introduced and band structure modification in semiconductors has been described.

Chapter 4 Deep impurity levels, Auger recombination process, and gradient in quasi-Fermi levels have been described.

Chapter 7 Formation of practical ohmic contacts has been explained and a new section on quantum confinement of carriers has been introduced.

Chapter 8 Input and output characteristics of BJTs, Early effect in BJTs, thermal runaway and thermal stability in BJTs, Kirk effect, and Webster effect have been described.

Chapter 9 Significantly expanded by introducing sections on short channel effect, control of threshold voltage, substrate bias effects, sub-threshold characteristics, equivalent circuits for MOSFET, MOSFET scaling and hot electron effects, drain-induced barrier lowering, short channel and narrow width effect, gate-induced drain leakage, and comparison of BJTs with MOSFETs.

Chapter 10 The phenomenon of photoluminescence has been explained.

Chapter 11 Two power semiconductor devices have been introduced – Gate Turn-off Thyristor (GTO) and Insulated-Gate Bipolar Transistor (IGBT), and also the formula to calculate intrinsic stand-off ratio has been included.

Chapter 12 The process of photolithography, different etching techniques such as wet etching and dry etching, and Moore's law have been described.

Coverage

Chapter-wise details of content coverage are as follows:

Chapter 1 presents important aspects of electron dynamics, including the analysis of motion of charged particles under the influence of electric and magnetic fields. The last section of this chapter discusses the cathode ray tube in detail.

Chapter 2 presents the salient properties of semiconductor materials with a special emphasis on the crystalline forms of these materials. Important concepts used for classifying and defining crystal lattices have been outlined in this chapter. Bulk and epitaxial growth techniques have also been included.

Chapter 3 introduces semiconductor physics, including energy bands and charge carriers. *E*-**k** diagrams have been discussed in detail as they form the basis of understanding the operation of many solid state electronic devices. It also presents a comprehensive treatment of the two most important physical processes, namely drift and diffusion of charge carriers. To make the treatment complete, a section has been added to include graded impurity profiles.

Chapter 4 offers details about the nature and behaviour of excess carriers created by external stimuli. A combination of drift and diffusion also governs the lives of excess carriers. It also presents the very important concept of the continuity equation that forms the core of any useful device physics model.

Chapter 5 initiates the study of solid state devices with the *p*-*n* junction. Beginning with a brief discussion about the different methods used for fabricating *p*-*n* junctions, the chapter also presents forward- and reverse-biased junctions.

Chapter 6 discusses the small-signal model of a *p*-*n* junction. An important aspect of *p*-*n* junctions is the way it behaves when the polarity of the applied bias is suddenly changed. A thorough treatment of transients that are critical to many applications is provided in this chapter.

Chapter 7 discusses metal–semiconductor contacts or junctions, including ohmic and Schottky contacts.

Chapter 8 discusses the bipolar junction transistor, without which the stupendous growth in the field of solid state electronics would not have been possible. The treatment includes different models of the device along with an explanation of their relative merits and demerits. Ordinary transistors cannot be operated at very high frequencies. Methodologies involved in the design of high-frequency transistors are presented in this chapter.

Chapter 9 focuses on the field effect transistor which is a key ingredient for the development of IC technology. Since the MOS structure forms the basis of this device, the chapter also includes important features of this structure.

Chapter 10 describes opto-electronic devices, such as photovoltaic cells, photo-detectors, LEDs, and laser diodes, which find a variety of applications in the field

of light wave communication. It also includes the basic principle of operation and important semiconductor structures involved.

Chapter 11 develops a focused understanding of the special measures required to enable solid state devices to handle high power, through the discussion of some important aspects pertaining to power devices. Some special devices, such as semiconductor-controlled rectifiers (SCR) and unijunction transistors (UJT) are also presented.

Chapter 12 discusses how the individual devices discussed up to Chapter 11 are put together using standard processes to build widely used ICs. The three most important technologies— namely MOSFET, MESFET, and bipolar technologies—are discussed along with their salient features. This chapter also includes a discussion on micro-electromechanical systems.

Chapter 13 focuses on some important microwave devices such as Gunn, IMPATT, TRAPATT, and BARITT diodes. This chapter attempts to present the basic physics of these devices so that students can understand their operation at microwave frequencies.

Chapter 14 discusses some basic circuits such as rectifiers, filters, and regulators that combine to form power supplies, which are one of the most important application areas of solid state devices. This is because all modern electronic systems need power supplies to operate. The last section of this chapter deals with switched mode power supply (SMPS), which is being increasingly used in various electronic systems.

Authors would be grateful for further suggestions and feedback with regard to this edition.

D.K. Bhattacharya

Rajnish Sharma

Symbols

F	Force
a	Acceleration
m	Mass
W	Work done
V	Voltage
m_0	Rest mass
B	Magnetic field
r	Radius
e	Electronic charge
p_1	Pitch
D	Deflection
S_e	Deflection sensitivity for electric deflection
S_m	Deflection sensitivity for magnetic deflection
p	Pressure
k_s	Segregation coefficient
ψ	Wave function
m_{eff}	Effective mass
n_0	Equilibrium electron concentration
p_0	Equilibrium hole concentration
n_i	Intrinsic carrier concentration
$f(F)$	Fermi–dirac distribution function
N_c	Effective density of states in conduction band
m_n*	Density of states effective mass
N_V	Effective density of states in valence band
N_d	Donor concentration
N_a	Acceptor concentration
μ	Mobility
V_H	Hall voltage
R_H	Hall coefficient
D_n	Diffusion coefficient of electrons
D_p	Diffusion coefficient of holes
G	Generation rate
R	Recombination rate
E_g	Band gap
δ_n	Excess electron concentration
δ_p	Excess hole concentration
E_t	Trap energy
$\tau_n 0$	Minority carrier electron lifetime
$\tau_p 0$	Minority carrier hole lifetime
s	Surface recombination velocity
V_{bi}	Built-in voltage
C_{dep}	Depletion region capacitance
p_n	Hole concentration in *n*-type semiconductor
n_p	Electron concentration in *p*-type semiconductor
L_n	Diffusion length of electrons
L_p	Diffusion length of holes
σ	Conductivity
g	Conductance
c_d	Diffusion capacitance
ρ_C	Specific contact resistance
I_C	Collector current
I_B	Base current
I_E	Emitter current
β	Common-emitter current gain
A_V	Voltage gain
R_e	Diffusion resistance
C_{je}	Emitter junction capacitance
C_p	Parasitic capacitance between base and emitter
v_S	Electron saturation velocity
R_C	Collection-region series resistance
C_S	Collection-to-substrate capacitance
C_{BC}	*B-c* junction capacitance
f_T	Cut-off frequency
HBT	Heterojunction bipolar transistor
ϕ_n	Potential barrier for electron injection
ϕ_p	Potential barrier for hole injection
V_{CC}	Collector supply voltage
V_{BB}	Base supply voltage
t_D	Delay time of transistor
t_R	Rise time of transistor
t_S	Storage time of transistor
L	Channel length of JFET
W	Channel width of JFET
$2d$	Channel depth of JFET
V_{GS}	Gate voltage with respect to source for JFET

V_{DS}	Drain voltage with respect to source for JFET
R	Resistance
ρ	Resistivity
A	Area
N_d	Donor concentration
W_d	Width of depletion region
I_D	Drain current
$V_{DS(\text{sat})}$	Drain–source voltage at saturation
ε_S	Dielectric constant of semiconductor
V	Voltage
I_P	Pinch-off current
V_P	Pinch-off voltage
g_D	Drain conductance
V_T	Threshold voltage
HEMT	High electron mobility transistor
JFET	Junction field-effect transistor
MODFET	Modulation doped field-effect transistor
MOS	Metal–oxide–semiconductor
E_C	Conduction band edge
E_{Fi}	Intrinsic fermi energy level
E_F	Fermi energy level
E_v	Valence band edge
ϕ	Potential
ϕ_s	Surface potential
N_a	Acceptor concentration
ϕ_{pF}	Difference between E_{Fi} and E_F for a p-type semiconductor
W_{dT}	Maximum space-charge width
ϕ_{nF}	Difference between E_{Fi} and E_F for an n-type semiconductor
ϕ'_m	Modified work function of metal
ϕ_m	Work function of metal
χ	Electron affinity of semiconductor
χ'	Modified electron affinity of semiconductor
$V_{\text{oxide}(0)}$	Potential drop across the oxide for zero gate voltage
ϕ_{s0}	Surface potential for zero applied gate voltage
ϕ_{ms}	Work function difference
V_{ox}	Voltage across oxide
Q_S	Charge per unit area in the semiconductor
ρ_S	Charge density in semiconductor
Q_M	Charge per unit area in metal
E_{ox}	Electric field across oxide
t_{ox}	Oxide thickness
ε_{ox}	Dielectric constant of oxide
C	Total capacitance
C_{ox}	Oxide capacitance
C_J	Depletion-layer capacitance
C_{min}	Minimum capacitance of MOS under strong inversion
V_{FB}	Flat band voltage
Q_o	Charge per unit area within oxide
E_0	Electric field within oxide
Q_{it}	Interface-trapped charge
Q_f	Fixed oxide charge
Q_{ot}	Oxide-trapped charge
Q_m	Mobile ionic charge
Z	Channel width
g'_D	Channel conductance
MOSFET	Metal–oxide–semiconductor field-effect transistor
E	Energy
ν	Frequency
h	Planck's constant
C	Velocity of light
E_g	Energy band gap.
I_ν	Photon flux with frequency ν
α	Absorption coefficient
g'	Generation rate of EHP due to photons
I_L	Photocurrent
I_{SC}	Short-circuit current
V_{OC}	Open-circuit current
P	Power

V_m	Voltage across solar cell under maximum power condition
I_m	Current through solar cell under maximum power condition
P_m	Maximum power
P_{in}	Incident optical power
η	Conversion efficiency of solar cell
FF	Fill factor
AM	Air mass
R_S	Series resistance
ΔE_c	Conduction band step
ΔE_v	Valence band step
σ	Conductivity
G_L	Photo generation rate of excess carriers
v_d	Drift velocity
t_n	Electron transit time
G_{ph}	Photoconductor gain
τ_t	Transit time
f_m	Modulating frequency
θ_C	Critical angle
$\overline{n}$	Refractive index
P_T	Maximum rated power
R_{ON}	On-state resistance
R_{CH}	Channel contribution to resistance
R_D	Drain contact resistance
R_{th}	Thermal resistance
$R_{\text{th}(d\text{-}p)}$	Thermal resistance between device and package
$R_{\text{th}(p\text{-}s)}$	Thermal resistance between package and heat sink
$R_{\text{th}(s\text{-}a)}$	Thermal resistance between heat sink and ambient
$P_{D(\text{max})}$	Maximum power dissipation
T_{dev}	Device temperature
$T_{j(\text{max})}$	Maximum junction temperature
V_{BB}	Base-to-base voltage
V_P	Peak-point voltage
V_v	Valley-point voltage
G	Conductance
$R_{\square}$	Sheet resistance
MEMS	Micro-electromechanical system
DRIE	Deep reactive ion etching
LIGA	LI (Röntgen lIthographie), G (Galvanik), A (Abformung).
SMPS	Switched mode power supply
V_m	Peak voltage
R_L	Load resistor
v_L	Load voltage
I_{rms}	Rms current
P_{ac}	Ac power
η	Conversion efficiency
PRV	Peak reverse voltage
γ	Ripple factor
T	Time period
ω	Angular frequency
I_Z	Zener diode current
V_i	Input voltage
V_o	Output voltage
P_{max}	Maximum power
t_{ON}	On time
RFI	Radio frequency interference
IMPATT	Impact-avalanche transit-time
TRAPATT	Trapped-plasma avalanche-triggered-transit
BARITT	Barrier-injected transit-time
μ_Γ	Mobility in central valley
v_Z	Avalanche-zone velocity
MSM	Metal–semiconductor–metal

Brief Contents

Detailed Contents

1 Electron Dynamics

- Conduction of electricity through gases
- Motion of charged particles in electric field
- Motion of charged particles in magnetic field
- Motion of charged particles in combined electric and magnetic field
- Cathode ray tube

Learning Objectives

After going through this chapter the student will be able to

- understand how gases conduct electricity
- understand the different regions of glow discharge
- derive expressions for the position and velocity of a charged particle moving in an electric field
- derive expressions for energy acquired and transit time of a charged particle moving in an electric field
- derive expression for the radius, time period, and pitch of a charged particle moving in a magnetic field
- understand important characteristics of the motion of charged particles in combined electric and magnetic field
- understand important parts of a cathode ray tube
- understand the focussing mechanism using electric and magnetic fields
- derive expressions for the deflection sensitivity of electric and magnetic deflection systems
- compare electric and magnetic deflection systems
- understand important characteristics of fluorescent screens
- solve numericals based on the motion of charged particles in electric and magnetic fields
- solve numericals based on electric and magnetic deflection systems

Introduction

We will start our study of electronics first by understanding the phenomenon of conduction of electricity through gases. Though solid-state electronics has robbed a lot of glory of this very interesting field, one cannot afford to forget that the foundations of modern electronics were established on this phenomenon. Vacuum tubes, invented in 1904 by J. A. Fleming, used the phenomenon of conduction of electricity through gases to a great extent. Many modern devices such as neon signs, night lamps, and tubelights also use this phenomenon. This chapter begins with an attempt at understanding the issues involved in considering the flow of electricity through gases. A great many applications in electronics and related instrumentation are based on the ability to modify the dynamics of charged particles in an electric field, a magnetic field, or a combination of the two. We will study why electrons modify their trajectory in the presence of these fields. Although we will not discuss these issues here, it is pertinent to remember that some natural phenomena such as lightning and Northern Lights are also due to interaction between charges and gases. We will then discuss a very important instrument that combines all that we will learn in this chapter, namely the cathode ray tube or simply the CRT. This instrument is at the heart of modern electronic devices such as oscilloscopes, television, and so on.

1.1 Conduction of Electricity through Gases

Under normal temperature and pressure conditions, gas is an excellent insulator. At extremely high temperatures or very low pressures, gas starts conducting electricity. Suppose we have an arrangement in which a given volume of gas can be enclosed with the facility to apply different magnitudes of potential difference (pd) across two electrodes at the two ends of the tube containing the gas. Let us also assume that the tube can be evacuated to varying degrees with the help of a suitable vacuum pump, as shown schematically in Fig. 1.1.

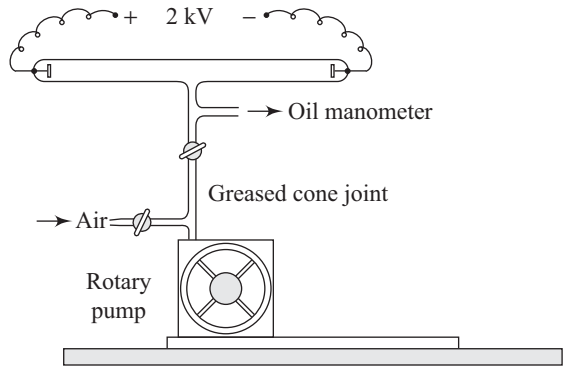


Fig. 1.1 Set-up for evacuating the tube to varying degrees

At atmospheric pressure, as the pd applied across the pair of electrodes is increased, a *dark discharge* occurs first. A small flow of current can be recorded, but there is no accompanied visible evidence of discharge. As the pd is increased further, a visible *brush discharge* makes its appearance. A typical example is the bluish streamers of the brush discharge into the air. If the pd is increased further, the streamers extend from one electrode to the other, the discharge being described as a *spark discharge*. If large current densities are possible then an *arc discharge* can be obtained. At pressures below 1 atm, a *glow discharge* occurs instead of a spark.

1.1.1 Glow Discharge

A steady supply electro-magnetic force (emf) of at least 2 kV across the electrodes placed within an enclosed gas is needed to observe the phenomenon of glow discharge. The glow discharge characteristics observed as the pressure is reduced to various measured values are shown in Fig. 1.2. At pressures of the order of 1 cmHg and above, thin streamers of a colour that depends on the nature of the gas are observed. The most striking effects however start at a pressure of around 1 mmHg. A well-defined region can then be observed as shown in Fig. 1.2.

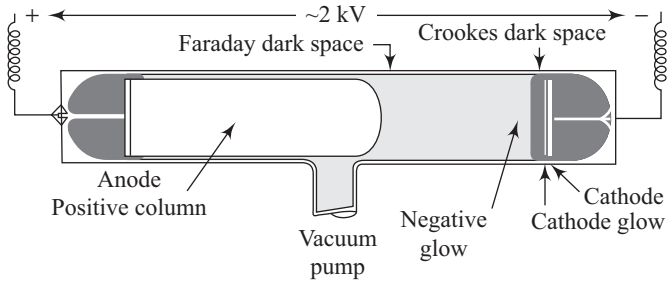


Fig. 1.2 Glow discharge characteristics

As we proceed from the cathode to the anode, the following regions can be identified:

1. A thin luminous layer is spread over a part or the whole of the cathode surface. This luminosity is called the *cathode glow*.
2. The *Crookes dark space* occurs next to the cathode glow. The length of this dark space is gas pressure dependent, increasing with decreasing pressure. An empirical relation between the dark space length, d, and the gas pressure, p, is given by

$$d = \frac{A}{p} + B \tag{1.1}$$

   where A and B are constants.
3. Beyond the Crookes dark space there appears a luminous region called the *negative glow*.

4. On proceeding further towards the anode, we come across an almost dark region called the *Faraday dark space*. Considerable variation is observed in the width of this region even when the pressure is maintained constant.
5. Finally, on reaching the surface of the anode, we get the *positive column*. This is the most brilliant region of the glow discharge. The region may be continuous or striated, depending upon the pressure and the current density. These striations are most prominent at pressures below 1 mmHg. On reducing the pressure further, the dark regions between the bright spaces in the striations increase in width.

If the length of the discharge tube is increased, the positive column length also increases. The other regions of the glow discharge remain nearly constant in length.

If the pressure inside the discharge tube is reduced, the Crookes dark space increases in width till the pressure reaches between 10^{-2} and 10^{-4} mmHg depending upon the tube dimensions and the current density. At this pressure, the dark space pervades the entire tube, i.e., the glow discharge disappears. This state is called the *black discharge*. This expansion of the dark space is accompanied with fluorescence of the glass walls of the tube.

1.2 Motion of a Charged Particle in Electric Field

We will discuss the motion of a charge particle kept in a uniform electric field. The treatment to follow makes the following assumptions:

(i) The charge density is low enough for us to consider the mutual repulsive force to be negligible.
(ii) The charged particle moves in high vacuum, therefore no collisions with gas atoms take place.
(iii) The mass of the charged particle is so small that gravitational forces can be neglected in comparison to electrostatic forces.

The force **F** experienced by a charged particle carrying a charge q when kept in an electric field **E** is given by

$$\mathbf{F} = q\mathbf{E} \tag{1.2}$$

As shown in Fig. 1.3, the force **F** and electric field **E** are vector quantities directed from the positive to the negative electrode.

For the case of an electron carrying a charge e we get

$$\mathbf{F}_e = -e\mathbf{E} \tag{1.3}$$

Using Newton's second law of motion, we can write

$$\mathbf{F}_e = -e\mathbf{E} = m\mathbf{a} \tag{1.4}$$

yielding

$$\mathbf{a} = \frac{d\mathbf{u}}{dt} = \frac{-e\mathbf{E}}{m} \tag{1.5}$$

where **u** represents the velocity of the electron and m its mass.

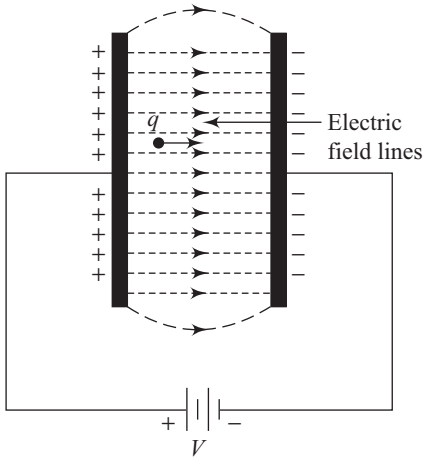


Fig. 1.3 Force experienced by a charge kept in a uniform electric field

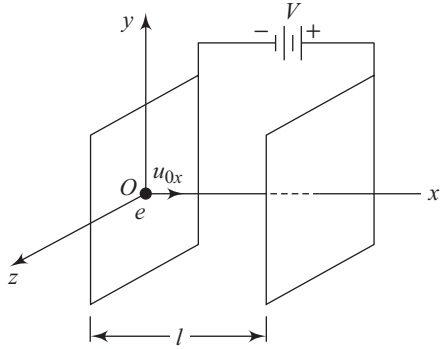


Fig. 1.4 Motion of an electron between infinite parallel electrodes

Let us at this stage make a simplifying assumption. We will assume that the electron is released at the origin of the chosen coordinate system with an initial velocity u_{0x} in the x-direction. Let us also assume that a voltage V has been applied between the two electrodes placed a distance l apart as shown in Fig. 1.4.
We can then write

$$E_x = \frac{-V}{l} \tag{1.6}$$

Obviously for the chosen conditions, we have

$$E_y = E_z = 0 \tag{1.7}$$

Using Eqn (1.5), we can write the equations of motion as

$$a_x = \frac{du_x}{dt} = \frac{-eE_x}{m} \tag{1.8}$$

$$a_y = \frac{du_y}{dt} = 0 \tag{1.9}$$

$$a_z = \frac{du_z}{dt} = 0 \tag{1.10}$$

By integrating Eqn (1.8) we get

$$u_x = -\int \frac{eE_x}{m} dt + C_u \tag{1.11}$$

yielding

$$u_x = \frac{-eE_x t}{m} + C_u \tag{1.12}$$

We have assumed

$$u_x = u_{0x} \quad \text{at} \quad t = 0 \tag{1.13}$$

Using Eqn (1.3) in Eqn (1.12) we get

$$C_u = u_{0x}$$

resulting in

$$u_x = \frac{-eE_x t}{m} + u_{0x} \tag{1.14}$$

Using Eqn (1.6) in Eqn (1.14) implies

$$u_x = \frac{eV}{ml} t + u_{0x} \tag{1.15}$$

Suppose the electron has initial velocities u_{0y} and u_{0z} at $t = 0$ in the y and z directions respectively. Since the field is directed in the x direction, the velocity components u_{0y} and u_{0z} remain unaltered with time. The total velocity u at any given time t is then given by

$$u = \sqrt{u_x^2 + u_{0y}^2 + u_{0z}^2} \tag{1.16}$$

The position of the electron at any time t can be obtained by integrating Eqn (1.15). This leads to

$$x = \int u_x dt = \frac{eV}{2ml} t^2 + u_{0x} t + c_x \tag{1.17}$$

where the constant c_x can be evaluated using the boundary condition $x = 0$ at $t = 0$ in Eqn (1.17), which yields

$$c_x = 0 \tag{1.18}$$

Equation (1.17) thus gets simplified to

$$x = \frac{eV}{2ml} t^2 + u_{0x} t \tag{1.19}$$

1.2.1 Energy Acquired by an Electron

The external electric field does work on the electron. This work increases the kinetic energy of the electron.

The work done, W, is given by

$$W = \int_A^B \mathbf{F}_e \cdot d\mathbf{l} \tag{1.20}$$

where A and B represent the initial and final positions, respectively.

For the situation presented in Fig. 1.4, we have

$$W = -e \int_A^B E_x dx = \frac{m(u_x^2 - u_{0x}^2)}{2} \tag{1.21}$$

The integral of the electric field intensity is equal to the negative of the applied potential difference, *V,* between the initial and final points. Thus, Eqn (1.21) results in

$$W = Ve = \frac{m(u_x^2 - u_{0x}^2)}{2} \tag{1.22}$$

The energy acquired by the electron is not dependent upon the specific path taken to go from A to B due to the conservative nature of electric field.

Equation (1.22) can be rewritten as

$$u_x^2 - u_{0x}^2 = \frac{2Ve}{m} \tag{1.23}$$

If the initial velocity $u_{0x} = 0$, then Eqn (1.23) results in

$$u = \sqrt{\frac{2Ve}{m}} \tag{1.24}$$

Substituting the value of e/m for an electron,
Eqn (1.24) yields

$$u = 5.94 \times 10^5 \sqrt{V} \tag{1.25}$$

Equation (1.25) is valid under non relativistic ($u \ll c$) conditions.

The energy acquired by the electron, W, is given by

$$W = \frac{1}{2} mu^2 = eV = 1.602 \times 10^{-19}\ \text{V} \tag{1.26}$$

1.2.2 Electron Transit Time

Referring to Fig. 1.3, the velocity u_p with which the electron strikes the plate p can be obtained using the principle of conservation of energy, we can write

$$\frac{1}{2} mu_p^2 = eV + \frac{1}{2} mu_{0x}^2 \tag{1.27}$$

so that

$$u_p = \sqrt{\frac{2eV}{m} + u_{0x}^2} \tag{1.28}$$

The transit time t_{kp} taken to travel between the plates is given by

$$t_{kp} = \frac{l}{\left(\frac{u_{0x} + u_p}{2}\right)} \tag{1.29}$$

If the electron leaves the first plate with zero initial velocity, then we can write

$$t_{kp} = \frac{2l}{u_p} \tag{1.30}$$

- Electric field lines originate from positive charges and terminate at negative charges. Positive charges are therefore sometimes referred to as *sources* and negative charges as *sinks*. This nomenclature is analogous to the description of fluid flow.
- Poisson equation relates the potential, V, to charge density through the form,

$$\nabla^2 V = -\frac{\rho}{\varepsilon_0}$$

1.3 Motion of a Charged Particle in Magnetic Field

In this section we will discuss the motion of a charged particle in a magnetic field. The force $\mathbf{F}_B$ experienced by a charged particle q is given by

$$\mathbf{F}_B = q\mathbf{u} \times \mathbf{B} \tag{1.31}$$

where u is the velocity of the charged particle, $\times$ represents the cross product, and $\mathbf{B}$ is the magnetic flux density in Wb/m^2. A schematic representation for the electron is shown in Fig. 1.5.

In Fig. 1.5 the electron is assumed to be injected into the region of uniform magnetic field of flux density, $\boldsymbol{B}$, with its velocity making an angle ϕ with the positive direction of $\boldsymbol{B}$. The force, $\boldsymbol{F_B}$, on the electron can be written down using Eqn (1.31) in the form

$$F_B = -Beu \sin\phi \tag{1.32}$$

From Eqn (1.32) we can conclude that an electron injected in the direction of magnetic field would not experience any force since $\phi = 0$.

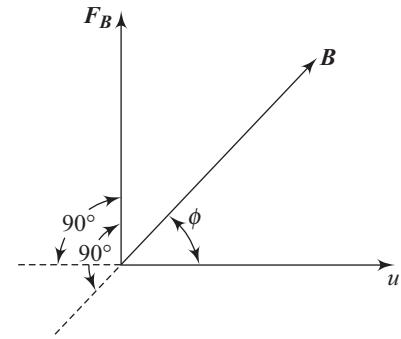


Fig. 1.5 Force experienced by an electron moving in a magnetic field

Thus an electron (any charged particle would do) would experience a force in a magnetic field only if the following two conditions are met:

(i) The charged particle is in motion.
(ii) The direction of motion is not along the magnetic field.

From Eqn (1.31) we can also infer that the force experienced by a charged particle moving in a magnetic field is always perpendicular to its velocity. *Thus a magnetic field does not do any work on a charged particle moving in its influence*. The kinetic energy of a charged particle moving in a magnetic field does not change; the magnetic field can only result in a change of the direction of motion of the charged particle.

Figure 1.6 shows an electron's trajectory when it enters a region of uniform perpendicular magnetic field. The velocity of the charged particle has been

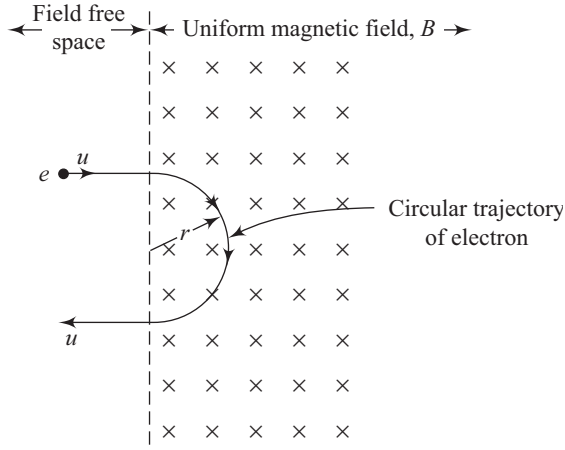


Fig. 1.6 Motion of an electron in a uniform perpendicular magnetic field

assumed to be u and the flux density of magnetic field has been assumed to be B.

The magnitude of the force experienced by the electron is given by

$$F_B = Beu \tag{1.33}$$

Thus a constant force would be felt by the electron. This constant force will in turn produce a constant acceleration at right angles to the direction of motion of the electron. *Such a situation implies that the electron trajectory is circular*. The force due to the magnetic field provides the necessary centripetal force. Thus

$$Beu = \frac{mu^2}{r} \tag{1.34}$$

which results in

$$r = \frac{mu}{eB} \tag{1.35}$$

Angular velocity, ω, is then given by

$$\omega = \frac{u}{r} = \frac{eB}{m} \tag{1.36}$$

The time period, T, required by the electron to complete one revolution is

$$T = \frac{2\pi}{\omega} = \frac{2\pi m}{eB} \tag{1.37}$$

The time period, T, of the revolution is thus independent of the electron velocity. A higher velocity results in a higher radius of the circular path to ensure a constant time period.

An interesting situation is one in which the electron acquires velocity u by being accelerated through a potential V. In such a situation

$$u = \sqrt{\frac{2\,eV}{m}} \tag{1.38}$$

Putting the expression for u in Eqn (1.35) yields

$$r = \frac{m}{eB}\sqrt{\frac{2\,eV}{m}} = \frac{3.37 \times 10^{-6}\sqrt{V}}{B} \tag{1.39}$$

In case the electron has a constant speed but its direction of motion when it enters the region of magnetic field is not perpendicular to the field, the trajectory will be a helical path. This is because the component of velocity perpendicular to the field will yield a circular motion, which will be superimposed on a constant velocity in the perpendicular direction. A schematic diagram of this situation is shown in Fig. 1.7.

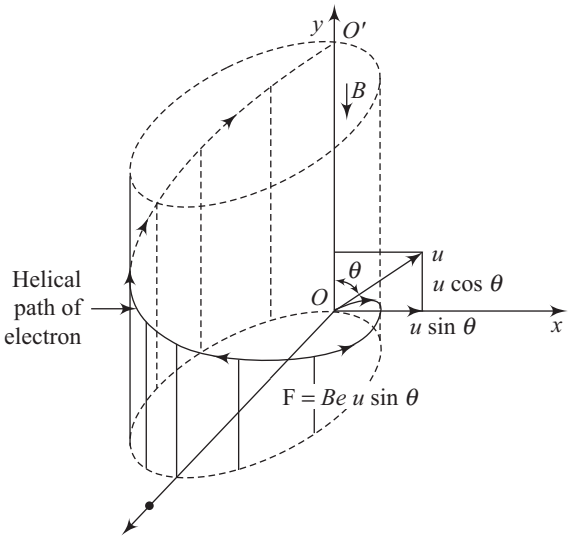


Fig. 1.7 Helical path of an electron moving in a uniform magnetic field

The electron in Fig. 1.7 has an initial velocity u in the x-y plane, making an angle θ with the y-axis. The electron starts from the origin. The uniform magnetic field has a flux density B along the negative y-axis. The component of velocity $u \cos \theta$ will not be affected by the magnetic field, ensuring that the electron keeps moving in the y-direction with constant velocity $u \cos \theta$. The perpendicular component $u \sin \theta$ will give rise to a force F_B, which also provides the centripetal force. The projection of the helical path of the electron on the x-z plane is a circle with its radius r given by

$$r = \frac{mu \sin\theta}{Be} \tag{1.40}$$

The time T required to complete one revolution is given by

$$T = \frac{2\pi r}{u \sin\theta} = \frac{2\pi m}{Be} \tag{1.41}$$

The distance travelled by the electron along the *y*-axis during a time interval *T* is called the *pitch* of the helix. Using Eqn (1.41) we can see that the pitch is given by

$$p_1 = u \cos\theta \, T = \frac{2\pi mu \cos\theta}{Be} \tag{1.42}$$

Thus for small θ values the pitch is independent of the angle θ. This property is made use of for focussing electron beams produced by electron guns in cathode ray tubes and high-voltage X-ray tubes. The beams consisting of divergent electrons are forced to follow helical trajectories of varying radii but same pitch by applying a magnetic field parallel to the axis of the tube.

1.4 Motion of a Charged Particle in Combined Electric and Magnetic Field

When a charged particle q moves in a region having electric and magnetic field, the net force $\mathbf{F}_T$, experienced by it is given by

$$\mathbf{F}_T = q\mathbf{E} + q\mathbf{u} \times \mathbf{B} \tag{1.43}$$

The trajectory of a charged particle under the influence of a combined electric and magnetic field is in general quite complex. The following simple cases however deserve special mention.

Case I Suppose the charged particle starts from rest and enters a region having electric and magnetic fields that are parallel to each other. The magnetic field then has no effect on the motion of the charged particle, and it moves in the direction of the electric field. The trajectory of the charged particle is thus a straight line.

Case II A charged particle is given an initial velocity and then injected into a region having the electric and magnetic fields perpendicular to each other. Furthermore, the electric and magnetic field directions are perpendicular to the direction of the initial velocity of the charged particle. In such circumstances, for a certain ratio of the magnitudes of electric and magnetic fields, the path of the charged particle is not deflected. This occurs because the force on the charged particle due to the electric field is completely balanced out by the force due to the magnetic field. Under these conditions we have

$$qE = quB \tag{1.44}$$

leading to

$$\frac{E}{B} = 5.94 \times 10^5 \sqrt{V} = u \tag{1.45}$$

Thus the charged particle continues to move in a straight line as long as the condition indicated in Eqn (1.45) holds true.

Case III A charged particle starts from rest and enters a region having electric and magnetic fields that are perpendicular to each other. The charged particle first experiences a force due to the electric field and starts moving in its direction. As long as the velocity of the particle is low, the magnetic field has no influence on its motion. When the velocity picks up, a sideways deflection takes place due to the magnetic field. The charged particle ultimately turns around and comes to rest at a point that corresponds to its initial position slightly displaced to one side. This action then repeats itself. *The path traversed by the charged particle is a cycloid*.

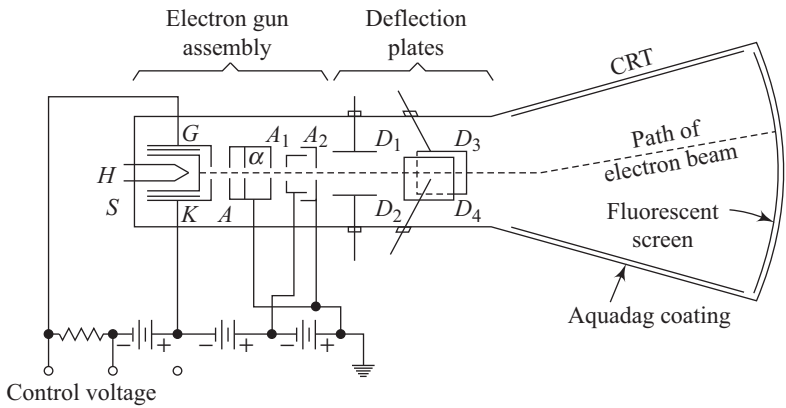
1.5 Cathode Ray Tube

The cathode ray tube (CRT) works on the principles of electron ballistics (the study of the trajectory of electrons in electric and magnetic fields). It has diverse applications such as cathode ray oscilloscope, television, and radar. The cathode ray tube can produce visual display of electrical effects with extremely high speed due to the high velocity of electrons striking its screen.

Figure 1.8 is a schematic diagram of a cathode ray tube with electric focussing and deflection. Cathode ray tubes using magnetic deflection have current-carrying coils mounted on the side of the tube instead of deflection plates.

The electron gun consists of a heated cathode K which emits electrons. A control grid G helps in varying the electron current density. A is the accelerating electrode and anodes A_1 and A_2 serve as a focussing arrangement.

Cathode K is generally in the form of a nickel cylinder coated with barium oxide. A heater filament wire, H, indirectly heats cathode K. A nickel cylinder



H—Heater
K—Cathode
S—Heat shield
G—Control grid
A—Accelerating electrode
a—Aperture in *A*
A_1—First anode of focussing lens
A_2—Second anode of focussing lens
D_1D_2—Vertical deflection plates
D_3D_4—Horizontal deflection plates

Fig. 1.8 Schematic diagram of the cathode ray tube with electric focussing and deflection

surrounds the cathode at a clearance of a few micrometres. This second cylinder serves as a heat shield and helps to concentrate the electrons to form a fine beam. As shown in Fig. 1.8, the control grid G consists of a cylinder surrounding the cathode and heat shield. It has an aperture in the centre of the end surface. The electrode is maintained at a low, negative potential with respect to the cathode. There is a provision for applying additional variable control voltage on this grid. This allows a control on the number of electrons in the beam and hence ultimately sets the brightness of the spot on the screen. The accelerating electrode A is maintained at a high, positive dc voltage with respect to the cathode. Due to this applied potential difference the electrons emitted by the cathode are accelerated to a very high speed. The divergent electrons in the beam are removed as the electron beam passes through one or two apertures in the accelerating electrode cylinder. During their transit the electrons in the beam experience a mutual *coulombic repulsive force*, making them spread away from the main beam. Some focussing arrangement is therefore required to bring the beam to a sharp focus at the screen. Two methods of focussing are generally practised: (i) focussing with electric fields and (ii) focussing with magnetic fields. The electric- and magnetic-field-based focussing arrangements used to focus electrons are somewhat analogous to the way in which optical lenses are used to control and focus light. These arrangements are therefore sometimes aptly referred to as electric and magnetic lenses for electrons. The final beam consists of a fine pencil of electrons having extremely high velocities directed along the major axis of the tube. The situation reminds one of bullets being shot out of the barrel of a gun, thereby justifying the name *electron gun*.

1.5.1 Focussing with Electric Fields

The anodes A_1 and A_2 shown in Fig. 1.8 constitute an electron lens system that focusses the beam into a fine spot on a fluorescent screen. The inside surface of the glass tube between the final anode (A_2) and the screen is coated with a conducting layer of graphite particles, termed as *aquadag*. This conducting layer is ultimately electrically connected to the final anode. The two anodes, A_1 and A_2, are kept at a positive potential with respect to the cathode, the potential of the first anode being lower than that of the second anode. The second anode is either kept at the same potential or at a potential slightly higher than the accelerating electrode. The second anode has an aperture that provides a well defined electron beam emerging from it. The ultimate velocity with which the electron strikes the screen depends upon the potential difference between the cathode and the second anode.

The electron gun sketched in Fig. 1.8 contains a three-lens system. The first lens consists of the cathode surface, the control grid, and the accelerating electrode. A schematic diagram of the first lens is shown in Fig. 1.9 along with the equipotential contours.

The electrons emitted from common points such as P_1 and P_2 on the surface of the cathode in different directions converge at P_1' and P_2', respectively in the

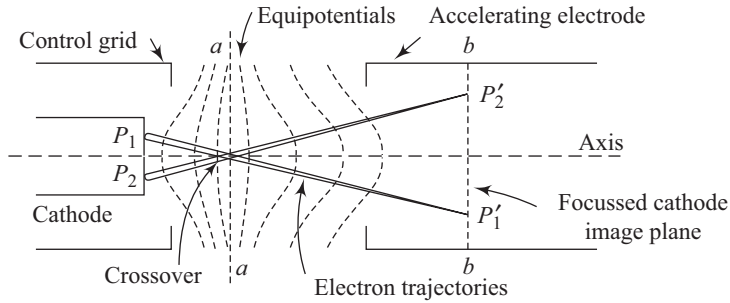


Fig. 1.9 Focussing action of electric field

plane *bb* under the influence of the lens system. All the electrons emitted from different points on the cathode surface cross the axis in the plane *aa*. The electron beam has the minimum diameter at the plane *aa*, which is termed as *crossover*. The crossover has a diameter much smaller on the screen by the subsequent lens action that follows.

The accelerating electrode and the first anode constitute the second lens. The third lens consists of the first anode and the second anode. The equipotential contours for the corresponding electric field between the first and the second anode are sketched in Fig. 1.10.

The electrons move normal to these equipotential contours. The shape of these contours decides the converging action of the lenses, which are in turn dependent upon the geometry or the ratio of the diameters of the cylindrical electrodes and the ratio of the potentials of the electrodes with respect to the cathode surface. The ratio of the potential of the first anode to that of the second anode is generally kept in the region of 1.5. The potentials of the accelerating electrode and the second anode are usually kept fixed at a few kilovolts. The potential of the first anode is kept variable to realize optimum focussing of the crossover on the screen.

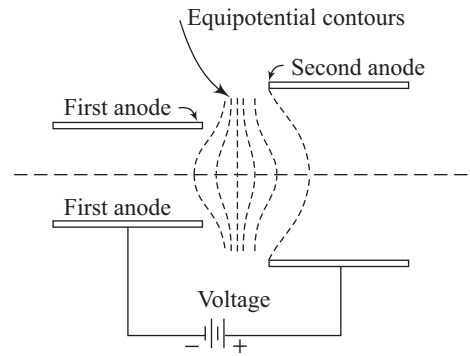


Fig. 1.10 Equipotentials between the first and second anodes

The ultimate size of the focussed spot on the fluorescent screen is dependent upon the magnification of the lens system and the size of the crossover. The aim is to have the smallest possible size of the spot on the screen. The magnification of the lens system can be made appreciably less than unity. Also, the aperture in the electron gun is made very small, thereby eliminating more divergent electrons. The minimized spot size achieved is however limited by the mutual repulsion between electrons and unequal electron velocities at the emission point.

The brightness at the centre of the spot is the maximum and falls gradually away from it. This is due to the fact that although the highest number of electrons are focussed towards the centre of the spot, not all of them can be focussed to a single point on the screen. As can be seen from Fig. 1.8, the fluorescent screen is given a slight curvature to ensure that the focussing of the beam remains undisturbed as it is deflected in different directions.

1.5.2 Focussing with Magnetic Field

Some cathode ray tubes use magnetic field to focus the electron beam. The requisite magnetic field can be produced either by a permanent magnet or by using current-carrying coils with an axis coincident with the beam. Cathode ray tubes using magnetic focussing use a single anode. Cathode ray tubes used in picture tubes of television receivers and radar indicators use a concentrated magnetic field caused by a coil. The *image orthicon* camera tube in television uses extended coils to produce a uniformly distributed magnetic field along the entire path of the electrons.

A schematic diagram of a typical magnetic focussing system is shown in Fig. 1.11.

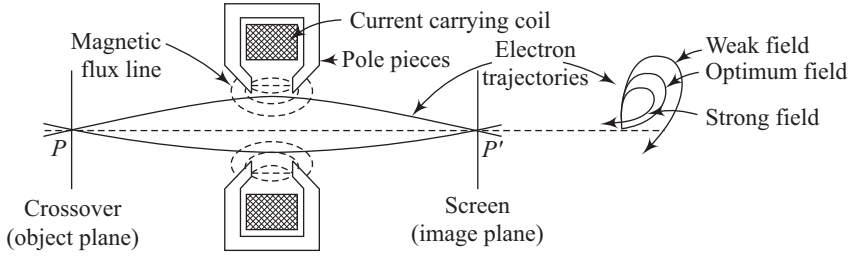


Fig. 1.11 Magnetic focussing system

To keep the figure simple, a localized concentrated magnetic field has been assumed. We have already learnt that an electron that has a component of velocity at right angles to the magnetic flux direction experiences a force that makes it rotate about the axis. The combined effect of this rotation and the axial velocity bends the path of the electrons through an arc. It is possible to adjust the strength of the magnetic field such that all the electrons leaving point P on the object plane at different angles with the axis of the system are deflected so that they return to a common point P' on the image plane or screen. The crossover is thus made to focus on to a point on the screen using the magnetic field. The magnetic field strength can be varied by varying the dc current through the coil. This adjustment is critical for effective focussing. A weak magnetic field would not bend the electron paths sufficiently to ensure their meeting at a single point on the screen. A very strong field on the other hand would lead to too much bending as shown in Fig. 1.11.

1.5.3 Deflection Systems

The deflection of the electron beam emanating from the electron gun can be achieved using either electric or magnetic field. In a typical electric deflection system, two pairs of deflection plates at right angles to each other are used (see Fig. 1.8). The parallel plane plates D_1 and D_2 are horizontally oriented whereas the plates D_3 and D_4 are oriented vertically. On application of an applied voltage between plates D_1 and D_2, the electrons get attracted towards the positive plate. This gives the electrons a vertical velocity component, which gets added to their axial velocity. The spot on the screen thus gets vertically displaced with reference to the original undisplaced position of the spot. Plates D_1 and D_2 are therefore called *vertical deflection plates* or *y* plates. Similarly, a voltage applied between plates D_3 and D_4 deflects the electron beam horizontally towards the more positive plate. Plates D_3 and D_4 are therefore called *horizontal deflection plates* or *x* plates.

A magnetic deflection system uses a set of current-carrying coils mounted on the sides of the cathode ray tube, as shown schematically in Fig. 1.12.

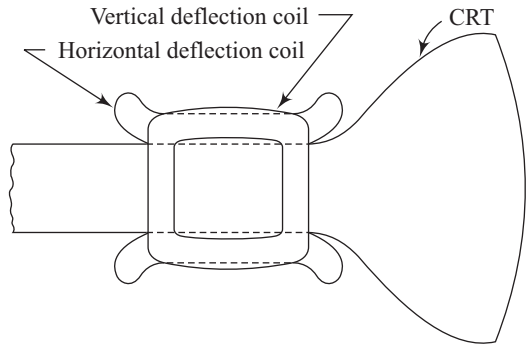


Fig. 1.12 Position of the current-carrying coils for producing magnetic deflection of the beam

When a current is made to flow through the vertical deflection coil, the resulting magnetic field has flux lines that are horizontal and perpendicular to the direction of motion of electrons. This results in a force on the electrons in the vertical direction. Current flow through the horizontal deflection coil gives rise to a force on the electron beam in the horizontal direction.

Electric deflection

In this section we will study electric deflection in detail. Figure 1.13 shows a pair of vertical deflection plates of length *l* and separated by distance *d*.

Suppose, a voltage V is applied between the plates. Further assume that an electron that has been accelerated through voltage V_1 enters the region within the plates with an initial velocity u_{0x}.

The acceleration a_y experienced by the electron as it passes through the electric field between the plates is given by

$$a_y = \frac{du_y}{dt} = \frac{-eE_y}{m} \tag{1.46}$$

where

$$E_y = \frac{-V}{d} \tag{1.47}$$

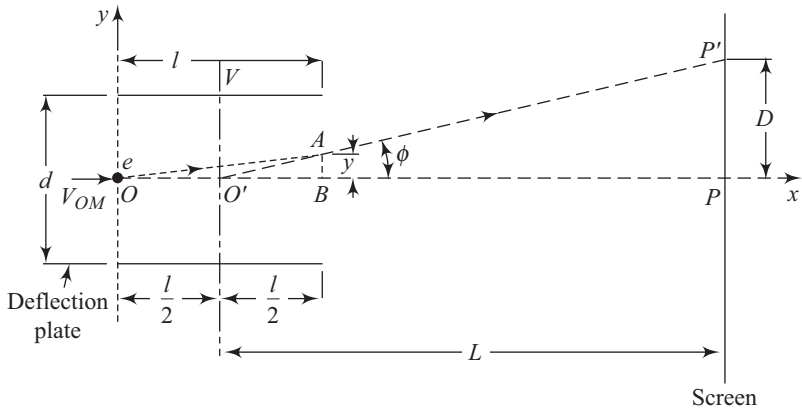


Fig. 1.13 Deflection produced by electric field

Putting E_y from Eqn (1.47) into Eqn (1.46) leads to

$$a_y = \frac{du_y}{dt} = \frac{eV}{md} \tag{1.48}$$

Assuming a constant V, integration of Eqn (1.48) results in

$$u_y = \frac{eV}{md}t \tag{1.49}$$

On further integration of Eqn (1.49) we get

$$y = \frac{eV}{md}\frac{t^2}{2} \tag{1.50}$$

The velocity component u_{0x} of the electron in the x direction remains constant during its travel within the plates. Thus

$$x = u_{0x}t \tag{1.51}$$

Using Eqs (1.51) and (1.50) we can write

$$y = \frac{eV}{2mdu_{0x}^2}x^2 \tag{1.52}$$

Thus the trajectory of the electron between points O and A is parabolic. The path AP' of the electron is a straight line. The direction of AP' is along the tangent to the parabolic path at point A. The slope of the parabola at $x = l$ can be evaluated by differentiating Eqn (1.52), which yields

$$\frac{dy}{dx} = \frac{eV}{mdu_{0x}^2}l = \tan\phi \tag{1.53}$$

For triangle $O'AB$, we can write

$$O'B = \frac{y}{\tan\phi} = \frac{eVl^2}{2mdu_{0x}^2\tan\phi} = \frac{l}{2} \tag{1.54}$$

Thus one can conclude that point O' lies at the centre of the deflecting plates.

Suppose L represents the distance between the fluorescent screen and the centre of the deflection plates. The deflection D of the spot on the screen of the cathode ray tube is then

$$D = L \tan \phi = \frac{LeVl}{mdu_{0x}^2} \tag{1.55}$$

From the conservation of energy principle we can write

$$\frac{1}{2} mu_{0x}^2 = eV_1$$

leading to

$$u_{0x}^2 = \frac{2eV_1}{m} \tag{1.56}$$

Putting the expression for u_{0x}^2 from Eqn (1.56) into Eqn (1.55) yields

$$D = \frac{LeVl(m)}{md(2eV_1)} = \frac{LVl}{2dV_1} \tag{1.57}$$

The deflection sensitivity, S_e, of the electric deflection system is defined as the magnitude of deflection, D, in metres per volt of deflecting potential, V. Thus

$$S_e = \frac{D}{V} = \frac{Ll}{2dV_1} \tag{1.58}$$

We can, therefore, draw the following conclusions about deflection sensitivity from Eqn (1.58).

(i) Deflection sensitivity is directly proportional to the distance L between the centre of deflection plates and the screen.
(ii) It increases with increase in the length of the deflection plates, l.
(iii) It is inversely proportional to the spacing d between the deflection plates.
(iv) It is inversely proportional to the anode voltage V_1.

There are, however, other considerations to be taken into account while attempting to increase the deflection sensitivity. If the length l of the deflection plates is made too large, the electron beam would strike the plates for large deflections. The same consequence is linked to reducing the inter-plate spacing to very low values. To circumvent these problems, some cathode ray tube designs use larger deflection plates that are kept inclined at an angle with respect to the axis of the tube to ensure a wider spacing at the outer edges. Any increase in L leads to an increase in the mechanical size of the cathode ray tube. The accelerating voltage V_1 needs to be low to ensure a high deflection sensitivity, but it needs to be high enough to ensure maximum spot brightness and good frequency response. A good cathode ray tube design has therefore to make suitable compromises on the above factors for optimization. A typical deflection sensitivity is of the order of 0.2–0.8 mm/V. The simple analysis presented here has neglected the effect of fringing field at the edges of the deflection plates. Taking this fringing field into account leads to an effective length of the deflection plates that exceeds the physical length by approximately the spacing between the plates.

The electrons in the beam lose their energy on striking the screen. A part of this energy is utilized in the emission of light and the rest is dissipated as heat. It is therefore advisable to avoid a stationary spot at any one point of the screen for an extended period of time. This safeguards the screen material from overheating.

Magnetic deflection

Electrons can also be deflected using magnetic fields. Figure 1.14 shows an electron with an initial velocity u_{0x} entering a region having a perpendicular magnetic field of flux density B.

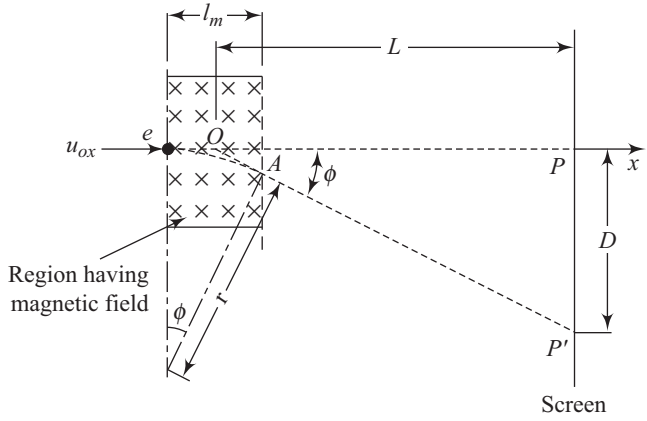


Fig. 1.14 Deflection due to magnetic field

In the schematic diagram the magnetic field has been assumed to act over an axial distance l_m. The electron emerges out of the magnetic field at point A and then travels in a straight line along the tangent to the circular path at point A. Point O represents the point of intersection between the axis of the tube and the tangential line AP'. The screen is at a distance of L metres from point O. Unlike electric deflection, point O will not, in general, lie at the centre of l_m. However, for small angles of deflection, wherein $\cos\phi \cong 1$, point O can be assumed to approximately lie at the centre of the magnetic field region l_m. From Fig. 1.14 we can write

$$\sin\phi = \frac{l_m}{r} \tag{1.59}$$

where r is the radius of the circular motion of the electron within the region l_m. Using Eqn (1.35), we can write

$$r = \frac{mu_{0x}}{eB} \tag{1.60}$$

Using Eqn (1.60) in Eqn (1.59) leads to

$$\sin\phi = \frac{eBl_m}{mu_{0x}} \tag{1.61}$$

Also, from Eqn (1.41)

$$\tan \phi = \frac{D}{L} \tag{1.62}$$

In actual practice r has large values and B has small values and l_m is also reasonably small. Under these conditions the deflection angle ϕ is small, and we can write

$$\tan \phi \cong \sin \phi \tag{1.63}$$

Using Eqs (1.61) and (1.62) in Eqn (1.63) leads to

$$\frac{D}{L} = \frac{eBl_m}{mu_{0x}}$$

implying that

$$D = \frac{eBl_m L}{mu_{0x}} = \frac{Bl_m L\sqrt{e/m}}{\sqrt{2V}} \tag{1.64}$$

where u_{0x} has been substituted using Eqn (1.38).

Deflection sensitivity, S_m, *in a magnetic deflection system is defined as the amount of deflection, D, per unit magnetic flux density*, B. Thus

$$S_m = \frac{D}{B} = \frac{l_m L\sqrt{e/m}}{\sqrt{2V}} \tag{1.65}$$

From Eqn (1.65) it is clear that S_m is

(a) directly proportional to length l_m in which the magnetic field exists,
(b) directly proportional to the distance L between the fluorescent screen and the centre of the field, and
(c) inversely proportional to the square root of the anode voltage, V.

Magnetic deflection sensitivity can also be defined as the magnitude of deflection produced when a current of 1 mA is made to flow through the deflection coils. When defined in this manner, it has units of mm/mA.

Comparison of electric deflection and magnetic deflection

The two deflection systems we have discussed so far have special fields of applications because of their special properties. Magnetic field coils require large currents (due to large inherent losses) implying large power requirements. On the other hand, electric deflection needs little or no power. Deflection produced in a magnetic deflection system based on air-cored coil is a direct function of the coil current. The coil current on the other hand is a function of the time integral of the applied voltage. Thus magnetic deflection is unsuitable for direct display of applied voltage. Furthermore, electric deflection is usable at much higher frequencies as compared to magnetic deflection. This makes electric deflection very suitable for use in instruments that operate at high frequencies.

Cathode ray tubes emit a small number of negative ions in addition to electrons. Some stray, negatively charged gas atoms are also present in the tube. The negative ions if allowed to bombard a particular part of the fluorescent screen for a long time can lead to a brownish appearance at the centre, also called *ion spot* or *burning*. These negative ions have the same charge as electrons but are much heavier. In electric deflection these negative ions are deflected through the same angle as are electrons. In magnetic deflection, however, the deflection produced is much smaller for ions in comparison to electrons. As a consequence, the ions continue to bombard a very small region around the centre of the undeflected position of the beam. This long-term bombardment damages the screen, rendering it insensitive to electron bombardment. In some applications electric deflection is preferred to magnetic deflection for this reason. In some applications, such as television, magnetic deflection is the preferred method of electron deflection. Here, some innovative techniques have to be incorporated to get over the ion spot or ion blemish problem. In one such technique a very thin aluminium coating is used on the inner surface of the fluorescent screen. The coating is thin enough for high-velocity electrons to pass through and excite the fluorescent screen but stop the larger ions. The coating also serves as a reflector of light to the front of the tube. In the absence of this reflector this light is lost into the interior of the tube.

An *ion trap* is another technique that can be used to avoid ion spotting. A schematic diagram of a typical ion trap is shown in Fig. 1.15.

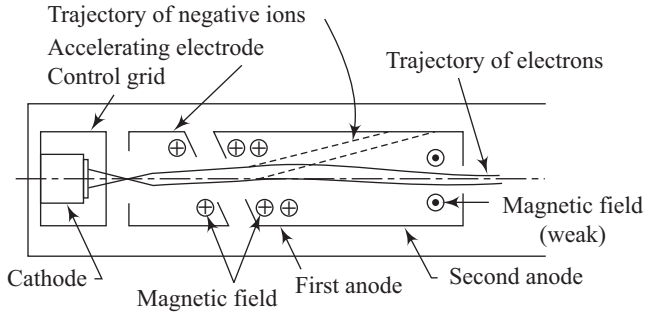


Fig. 1.15 Ion trap for avoiding ion spotting

In this technique, the electric field between the accelerating electrode and the first anode is given a transverse component by using a slot between these electrodes inclined at an angle with the major axis of the electron gun. The transverse component of electric field produces equal transverse deflection for electrons and negative ions. A magnetic field of suitable strength is also imposed in the region of transverse electric field, nullifying the deflection of the electrons. Negative ions, being much heavier, are not much affected by the magnetic field. As a consequence of this arrangement the electrons start moving along the axis of the electron gun, whereas the negative ions continue to be deflected (as shown in Fig. 1.15). Ultimately the negative ions strike the sides of the second anode and get trapped.

Fluorescent screens

We will begin this section by first understanding the meaning of *luminescence*, *fluorescence*, and *phosphorescence*. The term *luminescence* refers to the visible and non-visible radiation given off by a certain class of materials both during and after excitation. *Fluorescence* is the luminescence during excitation. *Phosphorescence* is the luminescence occurring after excitation. In cathode ray tubes, this is the radiation given off by the fluorescent material after excitation through the electron beam has ceased. Since both fluorescence and phosphorescence are involved in the operation of cathode ray tubes, it is more apt to refer to the screen action as *cathode luminescence*. Materials that show luminescence are called *phosphors*. The electron bombardment on this material gives off visible light. Since the screen is well insulated from other electrodes by the glass, the constant electron bombardment tends to make the screen acquire a negative potential, leading to repulsion of further electrons. This negative charge build-up needs to be removed. In practice, the electrons striking the screen produce not only light but also electrons through a process called *secondary emission*. These secondary electrons (the number of electrons emitted may be more than one per incident electron) are attracted to the graphite coating, popularly known as *aquadag*, on the inside of the walls of the glass bulb, which is connected to the second anode of the electron gun. Table 1.1 lists some important characteristics of a few common phosphors.

Table 1.1 Characteristics of common fluorescent screens

Phosphor No.	Fluorescent colour	Persistence (milliseconds)	Application
P-1	Green	Medium short (30–50)	General purpose oscilloscopes
P-2	Blue-green	Long	Transient visualization
P-3	Yellow-green	Medium (50)	Oscilloscopes
P-4	White	Short (5)	Television
P-5	Blue	Very short (0.005)	Fast photographic oscillography
P-7	Blue-white; yellow	Very long	Radar screens
P-11	Blue	Very short (0.01)	Photographic oscillography
P-12	Orange	Long	Radar screens
P-16	Bluish purple	Very short (0.01–0.1)	Television pick-up
P-19	Orange	Long	Radar indicators

Applications of cathode ray tube

Cathode ray tubes with large screen and incorporating sweep circuits are used for generation of images in television receivers. These CRT based television receivers are being replaced with modern devices such as TFT monitors, LCD,

LED, plasma monitors, and so on. Cathode ray tubes are also used in radar to give visual indication of position (azimuth angle, elevation angle, and distance) of a target. These targets could be aeroplanes, ships, etc. By far, the most important application of CRT is in cathode ray oscilloscopes that are used as a test-equipment for several types of analysis such as study of waveform, measurement of voltages, measurement of currents, and measurement of frequency. To study the waveform of a given alternating voltage, it is applied to the *y*-input terminals of the oscilloscope. An internally generated saw-tooth shaped time base voltage is applied to the *x*-deflection plates of the oscilloscope. The oscilloscope screen then displays the input waveform for analysis. The whole process is shown schematically in Fig. 1.16. In this figure, (a) is the input waveform, (b) is the time base voltage, and (c) is the oscilloscope display.

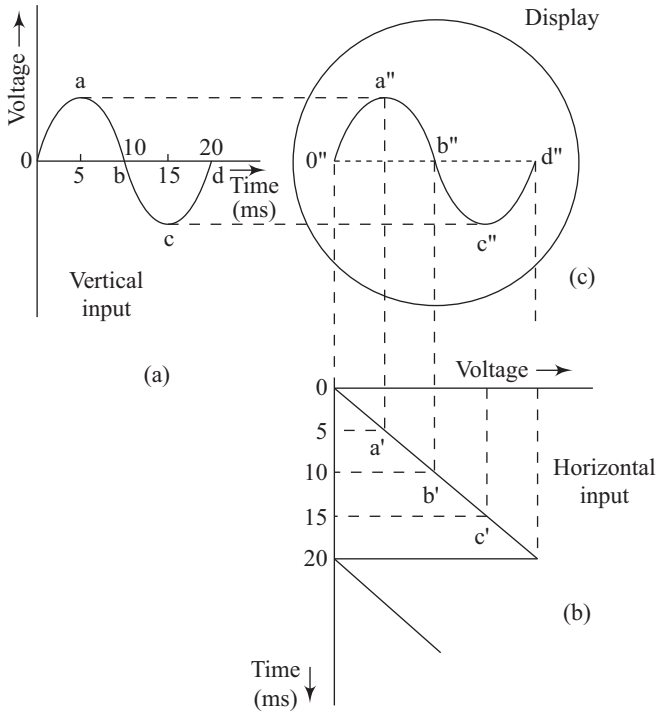


Fig. 1.16 Display process in a cathode ray oscilloscope

The size of the pattern obtained on the screen can be varied by adjusting the gain control. The pattern obtained on the screen becomes steady only when the time period of the input alternating voltage is either equal to or is an exact submultiple of the time period of time base voltage. The time period of the time base voltage can be varied using the coarse and fine frequency controls. The pattern displayed can then be used to carryout measurements of voltages, currents, frequency, phase difference, etc.

- The study of motion of electrons in electric and magnetic fields is called electron ballistics.
- The wave-form display on the face of a CRT can be measured visually against a set of horizontal and vertical scale marks, called a graticule.
- A storage CRT can retain the display much longer, up to several hours, after the image was first written on the phosphor.

Solved Problems

1.1 An electron is accelerated through a potential of 20,000 V. Calculate the percentage change in the mass of the electron at this speed.

Solution

The speed u after acceleration through a potential V is given by

$$u = \sqrt{\frac{2Ve}{m}} = 5.94 \times 10^5 \sqrt{V} \tag{1.1.1}$$

Putting the given value of V in Eqn (1.1.1), we get

$$u = 5.94 \times 10^5 \sqrt{20,000} = 8.40 \times 10^7 \text{ m/s} \tag{1.1.2}$$

The mass of the electron moving with velocity m_u is related to the rest mass m_0 of the electron through the expression

$$m_u = \frac{m_0}{\sqrt{1 - \left(\frac{u}{c}\right)^2}} \tag{1.1.3}$$

where c is the velocity of light and has a magnitude of 3×10^8 m/s. Putting the value of u from Eqn (1.1.2) into Eqn (1.1.3) yields

$$m_u = \frac{m_0}{\sqrt{1 - \left(\frac{8.40 \times 10^7}{3 \times 10^8}\right)^2}} = 1.042 m_0$$

The change in mass is given by

$$m_u - m_0 = (1.042 - 1) m_0 = 0.042 m_0$$

Thus the percentage change in the mass of the electron is 4.2%.

1.2 Two parallel plane plates A and B are separated by a distance of 3 mm. A potential difference of 400 V is applied between the two plates. An electron enters the region between the plates through a small hole in plate A. Calculate (i) the velocity with which the electron strikes the plate B; (ii) the kinetic energy acquired by the electron in joules and eV; (iii) the transit time of the electron as it travels from plate A to plate B. Assume that the initial velocity of the electron is zero.

Solution

(i) Velocity u_B with which the electron strikes plate B is given by

$$u_B = \sqrt{\frac{2Ve}{m}} = 5.94 \times 10^5 \times \sqrt{V} \tag{1.2.1}$$

Putting the given value of V in Eqn (1.2.1) yields

$$u_B = 11.8 \times 10^6 \text{ m/s}$$

(ii) Kinetic energy acquired by the electron in joules, KE, is

$$\text{KE} = \text{eV} = 1.602 \times 10^{-19} \times 400 = 6.408 \times 10^{-17} \text{ J}$$

KE in eV is

$$\text{KE} = 400 \text{ eV}$$

(iii) Transit time, t_{AB}, is given by

$$t_{AB} = \frac{2l}{u_B} = \frac{2 \times 3 \times 10^{-3}}{11.88 \times 10^6} = 0.505 \times 10^{-9} \text{ s} = 0.505 \text{ ns}$$

1.3 An electron enters a region having a perpendicular magnetic field of flux density 0.02 Wb/m^2. The initial speed of the electron is 5×10^7 m/s. Calculate the radius of the circular path followed by the electron. Assume $e = 1.6 \times 10^{-19}$C and $m_e = 9.1 \times 10^{-31}$ kg.

Solution

The radius r of the circular path followed by the electron is given by

$$r = \frac{mu}{eB} \tag{1.3.1}$$

Putting the given values in Eqn (1.3.1) yields

$$r = \frac{9.1 \times 10^{-31} \times 5 \times 10^7}{1.6 \times 10^{-19} \times 0.02}$$

resulting in

$$r = 1.42 \times 10^{-2} \text{ m}$$

1.4 A cathode ray tube electric deflection system consists of deflection plates 3 cm long, with a uniform spacing of 4 mm between them. The fluorescent screen is 30 cm away from the centre of the deflection plates. Determine the deflection sensitivity if the final anode has a potential of 2.5 kV.

Solution

The deflection sensitivity S_e is given by

$$S_e = \frac{Ll}{2dV_1} \tag{1.4.1}$$

where the symbols have the same meaning as in Eqn (1.58).

Putting the given values in Eqn (1.4.1), we get

$$S_e = \frac{30 \times 10^{-2} \times 3 \times 10^{-2}}{2 \times 4 \times 10^{-3} \times 2500}$$

leading to

$$S_e = 4.5 \times 10^{-4} \text{ m/V}$$

1.5 An electron is placed in a cathode ray tube. An electric field of 3×10^4 N/C is generated by a nearly uniform distribution of charges on the deflecting plates of the tube. The direction of the electric field is shown in Fig. 1.5.1. Find the acceleration of the electron and comment on the result obtained.

Solution

Using the given data we can find the electric force being exerted on the electron by the electric field in the vertically upward direction.

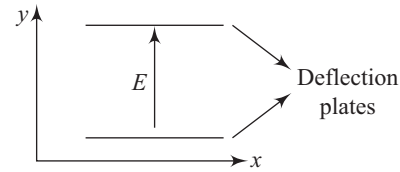


Fig. 1.5.1 Direction of the electric field

$$\mathbf{F} = q\mathbf{E}$$
$$= (-e) \times E$$

If the y-axis is vertically upward

$$E_y = 3 \times 10^4 \text{ N/C and}$$

$$E_x = 0 \text{ [i.e., the force has only } y\text{-component]}$$

$$F_y = (-e) \times E_y$$
$$= (-1.6 \times 10^{-19} \text{ C}) \times (3 \times 10^4 \text{ N/C})$$
$$= -4.8 \times 10^{-15} \text{ N}$$

Now, as per Newton's second law, we can say that this force gives the electron an acceleration

$$a_y = \frac{F_y}{m_e} = \frac{-4.8 \times 10^{-15} \text{ N}}{9.1 \times 10^{-31} \text{ kg}}$$
$$= -5.3 \times 10^{15} \text{ m/s}^2$$

The negative sign tells us that the direction of this acceleration is downward, i.e., opposite to the direction of the electric field.

Comment: In this calculation, we have neglected the acceleration due to gravity, since this acceleration is much smaller than that due to the electric field.

1.6 The beam of electrons which hits the phosphor screen in a cathode ray oscilloscope has to undergo complex deflection mechanisms. An experiment on the motion of electron beam in an oscilloscope was conducted. A potential of 2000 V was applied to the vertical deflection plates. Estimate the velocity with which the electron beam will travel.

Solution

Assuming initial velocity

$$u_{0x} = 0$$

$$u \text{ (i.e., velocity of electron beam)} = \sqrt{\frac{2Ve}{m}}$$

$$= \sqrt{\frac{2 \times 1.6 \times 10^{-19} \times 2000}{9.1 \times 10^{-31}}}$$

$$= 26.5 \times 10^6 \text{ m/s}$$

1.7 For the situation described in Problem 1.6, assume that the electron leaves the first plate with zero initial velocity. How much time will it take to cover a distance of 5 cm?

Solution

Length to be covered $(l) = 5$ cm

u_p (as calculated in Problem 1.6)

$$= 26.5 \times 10^6 \text{ m/s}$$

$$\therefore \quad t = \frac{2l}{u_p}$$

$$= \frac{2 \times 5 \text{ cm}}{26.5 \times 10^6 \times 10^2 \text{ cm/s}}$$

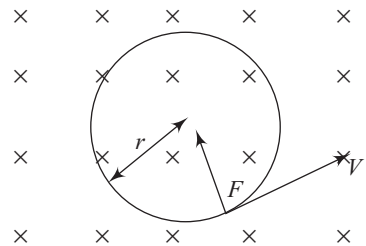
$$= \frac{10}{26.5 \times 10^8} = 3.7 \text{ ns}$$

1.8 If a particle moves in a uniform magnetic field, find an expression for the radius of the circular orbit along which this particle will move.

Solution

The above figure shows a region with a uniform magnetic field, directed perpendicularly into the plane of the page.

Suppose that the given positively charged particle has an initial velocity in the plane of the page. This initial velocity is perpendicular to the magnetic field. Magnetic force is then in the plane of the page, perpendicular to both the velocity and the magnetic field.



Now, the acceleration caused by this force has magnitude

$$a = \frac{F}{m} = \frac{qvB}{m}$$

and its direction is perpendicular to the velocity. Such an acceleration is characteristic of uniform circular motion. Thus, the particle will move in a circle of radius r with an acceleration given by

$$a = \frac{qvB}{m}$$

This acceleration will play the role of the centripetal acceleration $\frac{v^2}{r}$, that is,

$$\frac{qvB}{m} = \frac{v^2}{r}$$

$$\therefore \quad r = \frac{mv}{qB}$$

Recapitulation

- A steady supply of at least 2 kV and pressure below 1 atm. leads to glow discharge.
- Crookes dark space length, d, and the gas pressure, p, are related through the empirical relation

$$d = \frac{A}{p} + B$$

- The velocity, u, acquired by an electron starting from rest and accelerating through potential V is given by

$$u = \sqrt{\frac{2Ve}{m}}$$

- The time period, T, of revolution of an electron in a perpendicular magnetic field is given by

$$T = \frac{2\pi m}{eB}$$

- In a non-perpendicular magnetic field the path of the electron is a helix with pitch p_1 given by

$$p_1 = \frac{2\pi mu\cos\theta}{Be}$$

  where θ is the initial angle of launch with respect to the magnetic field.
- The deflection sensitivity, S_e, in an electric deflection system is given by

$$S_e = \frac{Ll}{2dV_1}$$

- The deflection sensitivity, S_m, in a magnetic deflection system is given by

$$S_m = \frac{l_m L\sqrt{e/m}}{\sqrt{2V}}$$

Exercises

Review Questions

1.1 Explain the following terms:
(a) dark discharge (b) brush discharge
(c) spark discharge (d) arc discharge

1.2 Describe the different regions of a glow discharge using a suitable schematic diagram.

1.3 What is black discharge?

1.4 Derive expressions for the velocity and position of an electron moving in a region with applied uniform electric field $E_y \hat{j}$. Assume the initial velocity of the electron to be zero.

1.5 Derive expressions for increase in the kinetic energy of an electron moving in a region of uniform electric field. Assume the electron to enter the region with zero initial velocity.

1.6 Derive an expression for the time period of revolution of an electron in a perpendicular magnetic field.

1.7 An electron enters a region of magnetic field with an initial velocity in a direction that is not perpendicular to the magnetic field. Draw a schematic diagram of the trajectory of the electron.

1.8 A charged particle is moving in a region having electric and magnetic fields. The electric field, magnetic field, and initial velocity are perpendicular to one another. Under what condition would the charged particle pass through this region undeflected?

1.9 Draw a schematic diagram showing the important parts of a cathode ray tube.

1.10 Explain with the help of a suitable schematic diagram the process of focussing of electron beams using electric field.

1.11 How is magnetic field used to focus a beam of electrons?

1.12 Derive an expression for the deflection sensitivity of the electric deflection system.

1.13 Derive an expression for the deflection sensitivity of a magnetic deflection system.

1.14 What is an ion spot? How is it avoided using suitable designs?

1.15 Differentiate between fluorescence, luminescence, and phosphorescence.

1.16 Explain the important regions of a glow discharge.

1.17 How does displacement of a charged particle moving in a uniform electric field depend upon the transit time?

1.18 Derive an expression for the transit time of a charged particle between two parallel plates maintained at different potentials.

1.19 Derive an expression for the radius of the circular path of an electron moving in a perpendicular magnetic field.

1.20 Derive an expression for the pitch of an electron moving in a non-perpendicular magnetic field.

1.21 Explain three important components of a CRT.

1.22 Differentiate between the focussing effect of electric field and magnetic field with suitable diagrams.
1.23 On what factors does deflection sensitivity for the electric and magnetic deflection system depend?
1.24 Derive the functioning of an ion trap.
1.25 Why does cathode luminescence more aptly describe the screen action in a CRT?
1.26 What type of devices are replacing CRT based television receivers?
1.27 How are CRTs used in radar?
1.28 Give a schematic representation of the display process in cathode ray oscilloscope?
1.29 What are phosphors?
1.30 What is *aquadag*?

Problems

1.1 A potential difference of 2.5 kV is used to accelerate an electron. Calculate the final speed of the electron and the percentage change in its mass at the final velocity.

$$\left[\textit{Hint: } u = \sqrt{\frac{2Ve}{m}} \text{ and } m = \frac{m_0}{\sqrt{1 - \left(\frac{u}{c}\right)^2}} \right]$$

Ans. $u = 2.97 \times 10^7$ m/s; 0.5%

1.2 Consider the situation presented in solved numerical 1.1. Assume that the initial velocity of the electron is 2×10^6 m/s directed towards plate *B*. Calculate (i) velocity with which the electron strikes plate *B*; (ii) kinetic energy acquired by the electron in joules and eV; (iii) transit time of the electron as it travels from plate *A* to plate *B*.

$$\left[\textit{Hint: } u_B = \sqrt{\frac{2eV}{m} + u_A^2} \right]$$

Ans. (i) 12.04×10^6 m/s; (ii) 652.33×10^{-19} J, 407.71 eV; (iii) 0.43 ms

1.3 A parallel plate diode consists of a cathode and an anode spaced 4 mm apart. The anode is kept at a dc potential of 250 V with respect to the cathode. Determine the velocity and the distance travelled by an electron after 0.6 ns if the initial velocity of the electron is zero.

$$\left[\textit{Hint: } u_x = \frac{eV}{ml} t \text{ and } x = \frac{eV}{2ml} t^2 \right]$$

Ans. 6.596×10^6 m/s; 1.98×10^{-3} m

1.4 An electron enters a region of uniform perpendicular magnetic field of flux density 0.04 Wb/m^2. The initial speed of the electron is 6×10^7 m/s. Calculate the radius of the circular path of the electron. Assume $e = 1.6 \times 10^{-19}$ C and $m_e = 9.1 \times 10^{-31}$ kg.

$$\left[Hint: r = \frac{m_e u}{eB} \right]$$

Ans. 0.853 cm

1.5 An electron with a speed of 5×10^7 m/s enters a region having a uniform magnetic field at an angle of 60° with the electron's velocity. Determine the magnetic flux density required to make the electron go through a helical path of 0.8 m diameter. Also calculate the time taken by the electron to complete one revolution.

$$\left[Hint: r = \frac{mu \sin\theta}{Be} \text{ and } T = \frac{2\pi m}{Be} \right]$$

Ans. 0.308×10^{-3} Wb/m^2; 115 ns

1.6 An ionized hydrogen atom and an electron, both having an initial velocity corresponding to 250 V, are projected perpendicular to a uniform magnetic field of flux density 5×10^{-3} wb/m^2. Calculate the ratio of radii of the circular paths followed by the particles.
[*Hint:* $m_h = 1837\, m_e$]

Ans. $\frac{r_h}{r_e} = \sqrt{1837} = 42.9$

1.7 The electric deflection system of a cathode ray tube consists of deflection plates 4 cm long with a uniform spacing of 3 mm between them. The fluorescent screen is located 30 cm from the centre of the deflection plates. Determine the deflection sensitivity in mm/V if the anode is maintained at a potential of 3 kV.

$$\left[Hint: S_e = \frac{Ll}{2dV_1} \right]$$

Ans. 0.375 mm/V

1.8 The deflection plates of a cathode ray tube are 4 cm long and have a spacing of 5 mm between them. The screen of the tube is at a distance of 25 cm from the centre of the deflection plates. The final anode voltage is 2000 V. A voltage of 50 V is applied between the deflection plates. Determine the displacement of the spot produced on the screen.

$$\left[Hint: D = \frac{LVl}{2dV_1} \right]$$

Ans. 2.5 cm

1.9 A cathode ray tube uses a magnetic deflection system. The system consists of a screen located 25 cm from the centre of the deflection coils. The length of

the region having a uniform magnetic field along the tube's axis is 3 cm. A deflection of 2 cm is required on the screen. Calculate the amount of magnetic flux density required if the final anode voltage is 1000 V. Also calculate the deflection sensitivity.

$$\left[Hint: B = \frac{D\sqrt{2V}}{l_m L\sqrt{e/m}} \right]$$

Ans. 2.846×10^{-4} Wb/m^2; 70.27 m/(Wb/m^2)

1.10 An electron moving through an electric field is observed to have an acceleration of 10^{18} cm/s^2 in the x-direction. What must be the magnitude and direction of the electric field that produces this acceleration?

Ans. -5.7×10^4 N/C

1.11 Our planet earth possesses both a magnetic field as well as an atmospheric electric field. On a typical clear day, electric field strength is about 100 N/C and it points vertically downward. Considering this value of electric field and also taking gravity into account, what should be the acceleration of a dust particle of mass 1×10^{-18} kg carrying a single electron charge?

Ans. 6.2 m/s^2

1.12 In an X-ray tube, electrons are exposed to an electric field of 8×10^5 N/C. What is the force on an electron? What is its acceleration?

Ans. 1.3×10^{-13} N, 1.4×10^{17} m/s^2

1.13 An electron of speed 4.0×10^5 m/s is observed to move in a circular orbit of radius 0.4 m in a magnetic field. What is the strength of the magnetic field that will lead to such circular motion?

Ans. 1.0×10^{-2} T

1.14 An electron enters a region having a pd of V volts. The final velocity of the electron as it exits the region is 9×10^7 m/s. Calculate the accelerating potential if the initial velocity is zero.

Ans. 22,958 V

1.15 An electron enters a region having a perpendicular magnetic field. Initial speed of the electron is 10^7 m/s. Radius of the circular path followed by the electron is 1.5×10^{-2} m. Determine the magnitude of the magnetic field. Assume $e = 1.6 \times 10^{-19}$ C and $m_e = 9.1 \times 10^{-31}$ kg.

Ans. 3.8×10^{-3} wb/m^2.

2 Growth and Crystal Properties of Semiconductors

- Types of solids
- Crystal lattices
- Atomic bonding
- Imperfections and impurities in solids
- Bulk crystal growth
- Epitaxial growth

Learning Objectives

After going through this chapter the student will be able to

- understand different types of solids and define amorphous, polycrystalline, and single crystal solids
- understand the concepts of crystal lattice including unit cell and primitive cell
- understand different types of atomic bonding
- understand the nature of imperfections and impurities in solids
- define Miller indices, Van der Waals bond, ionic bond, covalent bond, and metallic bond
- understand point and line defects and substitutional and interstitial impurities
- understand the salient features of bulk and epitaxial crystal growth including pseudomorphic layers, reaction chamber, and critical thickness
- solve numericals based on hard sphere packing, Miller indices, density calculation from known values of lattice constants, surface density of atoms in different planes, and the Czochralski growth technique

Introduction

The present age is being called the age of electronics. To be more precise, it can be called the age of semiconductor electronics. From individual discreet devices to very large scale integrated circuits, the progress in semiconductor technology has been stupendous. Why do semiconductors behave the way they do is the topic of discussion in this chapter. To answer this we need to know the different types of solids and the arrangement of individual atoms in them. Also some typical techniques of epitaxial growth would be covered in this chapter.

2.1 Semiconductor Materials

Semiconductors comprise a group of materials that have their electrical conductivities lying between metals and insulators. Furthermore, their conductivities can be modified by introducing controlled amounts of desirable impurities. Optical excitation and temperature also affect their electrical conductivity. Semiconductors are of two types, namely *elemental semiconductors* and *compound semiconductors*. Silicon and germanium are two common elemental semiconductors. Elemental semiconductors are found in group IV of the periodic table and are composed of single species of atoms. Compound semiconductors on the other hand are made up of a combination of more than one single species of atoms. Some common compound semiconductors are made up of either a combination of group III and group V atoms (GaAs is the most common example) or a combination of group II and group VI atoms (ZnS is a prominent example). Two group IV elements can also combine to form a compound semiconductor. Table 2.1 lists some elemental and compound semiconductors.

More complicated *ternary* (Al_x $Ga_{1-x}As$) and *quarternary* (InGaAsP) compound semiconductors are also finding increasingly sophisticated applications.

2.2 Types of Solids

There are three general types of solids depending upon the arrangement of the constituent atoms. The three types are crystalline/single, polycrystalline, and amorphous crystal solids. Atoms are arranged in a regular manner in a crystalline solid. A three-dimensional repetition of a certain basic pattern constitutes the solid. In a single crystal the periodicity of the basic pattern extends throughout the material. The periodicity of the structure gets interrupted at certain boundaries in polycrystalline material. These so called boundaries are called grain boundaries. Specific grain sizes may vary from several Angstroms to macroscopic dimensions. An interesting situation is one in which the size of the grains within the grain boundaries becomes comparable to the size of the basic pattern.

Table 2.1 (a) The region of the periodic table where semiconductors occur; (b) Elemental and compound semiconductors

(a)	II	III	IV	V	VI
		B	C	N	
		Al	Si	P	S
	Zn	Ga	Ge	As	Se
	Cd	In		Sb	Te

(b)	Elemental	IV compounds	Binary III–V compounds	Binary II–VI compounds
	Si	SiC	AlP	ZnS
	Ge	SiGe	AlAs	ZnSe
			AlSb	ZnTe
			GaN	CdS
			GaP	CdSe
			GaAs	CdTe
			GaSb	
			InP	
			InAs	
			InSb	

The periodicity is completely absent at this stage and the material is now called amorphous. Figure 2.1 is a schematic two-dimensional representation of the three types of solids.

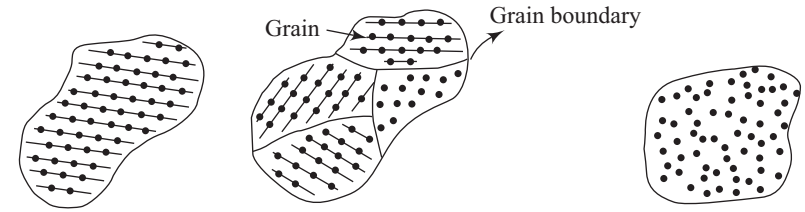


Fig. 2.1 Three types of solids

2.3 Crystal Lattices

Crystalline materials consist of a periodic arrangement of the constituent atoms. In this section we will discuss some important characteristics of crystalline materials.

2.3.1 Unit Cell

We will here focus our attention on single crystals. As outlined above, in a single crystal a basic pattern consisting of a single atom or a group of atoms is repeated

at regular intervals in all the three dimensions. *This periodic arrangement of atoms in a crystal is called a lattice*. Atoms can be arranged in different forms in a lattice. The distance between atoms and the relative orientation of the atoms can have different forms. Every lattice consists of a regular repetition of a fundamental unit called a *unit cell*. The entire crystal can be constructed from this unit cell. Figure 2.2 shows an infinite two-dimensional array of lattice atoms.

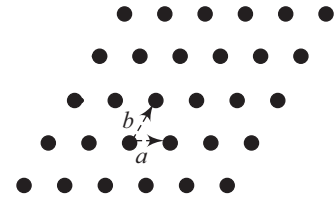


Fig. 2.2 Two-dimensional single crystal lattice

Each dot represents a particular atom and is referred to as a lattice point. Any lattice point can be translated through a distance a in one direction and a distance b in a second noncolinear direction to create the two-dimensional lattice. There is no need for the two translation directions to be perpendicular to each other. A three-dimensional lattice can be created using translation in a third noncolinear direction through a distance c. A general three-dimensional lattice can be obtained by a periodic repetition of a fundamental unit called a *unit cell*. A unit cell is thus a small volume of the crystal that reproduces the entire crystal through periodic repetition. There is nothing unique about a unit cell as shown schematically for a two-dimensional lattice in Fig. 2.3.

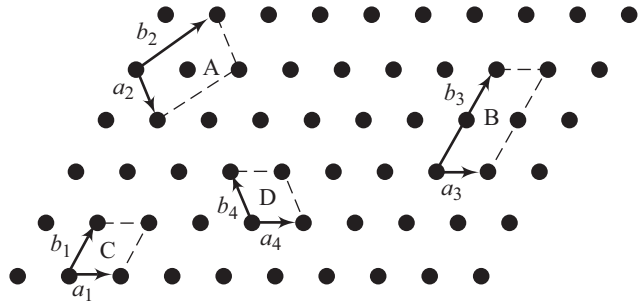


Fig. 2.3 Two-dimensional single crystal lattice showing various possible unit cells

The smallest unit cell that can be repeated periodically to form a crystal is called a *primitive cell*. The primitive cell is however not always a good unit cell to represent a lattice. Very often a different choice of a unit cell leads to a better representation. A unit cell involving orthogonal directions can, for example, lead to some simplifications. In general a unit cell can be characterized by a set of three vectors—**a**, **b**, and **c**—which may or may not be orthogonal or equal in lengths (Fig. 2.4). A lattice point is indistinguishable from another lattice point if the displacement vector between the two can be represented by

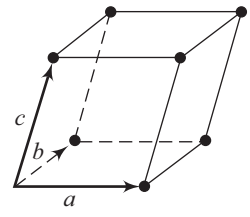


Fig. 2.4 A generalized unit cell

$$\mathbf{r} = p\mathbf{a} + q\mathbf{b} + s\mathbf{c} \tag{2.1}$$

where p, q, and s are integers and vectors **a**, **b**, and **c** are called *basis vectors*. These basis vectors are also referred to as *direct basis vectors*. For a given set of direct basis vectors, we can define three reciprocal lattice basis vectors **a***, **b***, **c*** as,

$$\mathrm{a}^* = 2\pi \frac{\mathbf{b}\times\mathbf{c}}{\mathbf{a}\cdot\mathbf{b}\times\mathbf{c}},\ \mathrm{b}^* = 2\pi \cdot \frac{\mathbf{c}\times\mathbf{a}}{\mathbf{a}\cdot\mathbf{b}\times\mathbf{c}},\ \mathrm{c}^* = 2\pi \cdot \frac{\mathbf{a}\times\mathbf{b}}{\mathbf{a}\cdot\mathbf{b}\times\mathbf{c}} \tag{2.2}$$

From Eqn (2.2) we can see that $\mathbf{a}\cdot\mathbf{a}^* = 2\pi,\ \mathbf{a}\cdot\mathbf{b}^* = 0$, etc. The general reciprocal lattice vector **g** is given by,

$$\mathbf{g} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^* \tag{2.3}$$

where h, k, and l are integers.

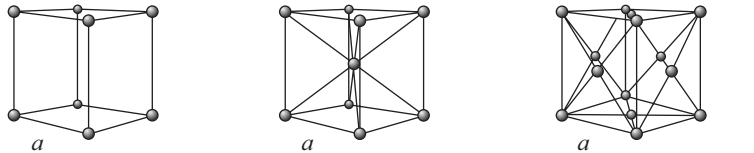
Using Eqs (2.1) and (2.2) we can conclude that $\mathbf{g}\cdot\mathbf{r} = 2\pi \times$ integer; thus each vector of the reciprocal lattice is normal to a set of planes in the direct lattice. Furthermore, the volume v_c^* of a unit cell of the reciprocal lattice happens to be inversely proportional to the volume v_c of unit cell involving the direct lattice. The exact expression is

$$v_c^* = \frac{(2\pi)^3}{v_c} \tag{2.4}$$

where, $v_c = \mathbf{a}\cdot\mathbf{b}\times\mathbf{c}$ (2.5)

2.3.2 Cubic Lattices

The simplest form of a lattice is called a cubic lattice. It is called so because the unit cell for such a lattice is a cubic volume. For a cubic lattice, the vectors **a**, **b**, and **c** are equal in length and perpendicular to each other. The three most common cubic lattices are simple cubic (sc), body-centered cubic (bcc), and face-centered cubic (fcc). These three lattice types are shown in Fig. 2.5.



(a) Simple cubic (b) Body-centered cubic (c) Face-centered cubic

Fig. 2.5 Unit cells for cubic lattices

The simple cubic structure consists of an atom located at each corner. The body-centered cubic structure has an additional atom at the centre of the cube. The face-centered cubic structure has additional atoms on each face of the cube.

2.3.3 Crystal Planes and Directions

Semiconductor devices are generally fabricated on surfaces of semiconductor wafers (circular pieces of semiconductor crystals of varying sizes). Many properties of semiconductors are also direction dependent. A standard method of

indicating planes and directions in semiconductors is, therefore, very useful. A set of three integers h, k, and l shown within brackets [i.e., (hkl)] is generally used to indicate a particular plane. These integers can be found using the following simple procedure:

1. Determine the intercepts of the plane with the crystal axes and express them as integral multiples of basis vectors. A translation of the plane with respect to the origin is permitted as long as the direction of the plane is maintained.
2. Find the reciprocals of the three integers obtained in step (1) and reduce them to the smallest set of integers maintaining their relationship. Designate these integers h, k, and l.
3. Finally label the plane as (hkl).

Some common lattice planes for a cubic crystal are shown in Fig. 2.6.

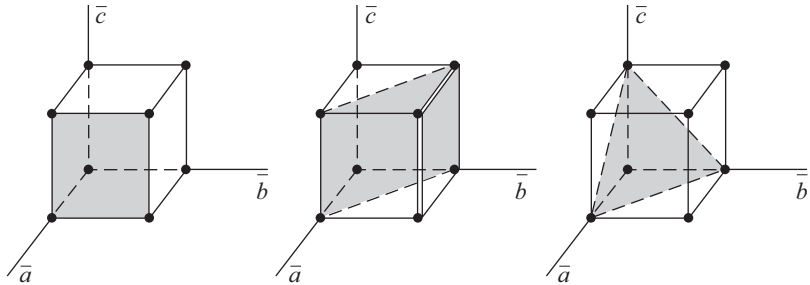


Fig. 2.6 Three lattice planes for a cubic lattice (a) (100) plane, (b) (110) plane, (c) (111) plane

The set h, k, l is called the Miller indices. These numbers define a set of parallel planes inside the lattice. Many planes in a lattice are equivalent. Any given plane with a set of Miller indices can be shifted within the lattice by choosing a suitable position and orientation of the unit cell. For example, the cube faces of a cubic lattice are crystallography equivalent since the unit cell can be rotated in various directions without affecting its form and appearance. Equivalent planes are shown enclosed in curly brackets { }. The six equivalent faces of a cubic lattice are designated as {100} and are shown schematically in Fig. 2.7.

Directions in a lattice are determined using the following procedure:

1. Choose the basis vectors with a suitable origin.
2. Express the vector components for the particular direction in multiples of the basis vectors.
3. Reduce the three integers to their smallest values while maintaining the relationship between them.
4. Express the specific direction within square brackets, e.g., [abc].

The body diagonal of a cubic lattice has a [111] direction as shown in Fig. 2.8.

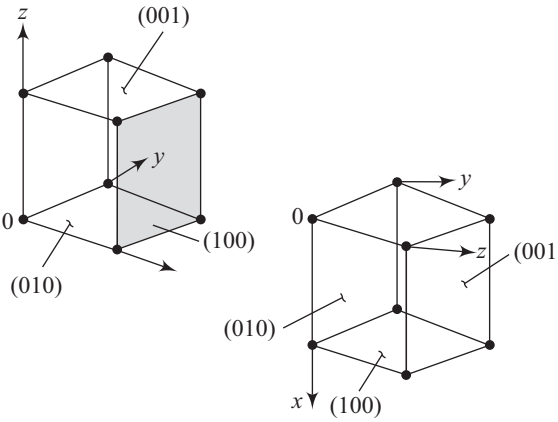


Fig. 2.7 Equivalent {100} planes

Many directions in a lattice are equivalent. This depends upon the choice of the orientation of axes. Equivalent directions are expressed within angular brackets < >. Some equivalent <100> directions are shown in Fig. 2.9.

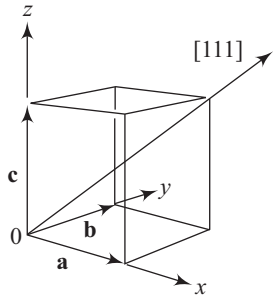


Fig. 2.8 Schematic diagram of the [111] direction

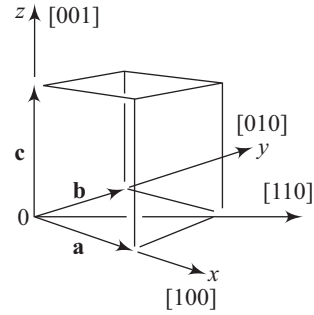


Fig. 2.9 ⟨100⟩ equivalent directions

For cubic lattices, the [*hkl*] direction is perpendicular to the *hkl* plane as shown in Fig. 2.10.

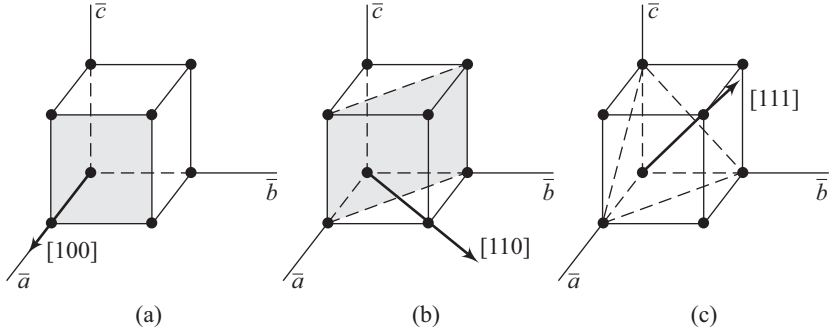


Fig. 2.10 Orthogonality of three lattice directions and planes: (a) (100) plane and [100] direction, (b) (110) plane and [110] direction, (c) (111) plane and [111] direction

The crystal structure of a crystalline material can be studied using *X*-ray diffraction (XRD) technique. In this technique, a beam of *X*-rays having a single wavelength that is of the same order of magnitude as the atomic spacing of the material is made to strike the material. These *X*-rays are then scattered in all directions. Though most of the radiation scattered from one atom cancels out the scattered radiation from other atoms, *X*-rays striking certain crystallographic planes at specific angles get reinforced instead of cancelling out. This phenomenon is called *diffraction*. The condition that needs to get fulfilled for this reinforcement to take place is given by,

$$2d_{hkl} \sin \theta = \lambda$$

where d_{hkl} represents the interplanar spacing between the planes and λ is the wavelength of the incident *X*-rays.

2.3.4 Diamond Lattice

The modern semiconductor industry uses a variety of semiconductor materials such as Si, Ge, GaAs for different applications. Many semiconductors have a *diamond lattice* as the basic lattice structure. Elemental semiconductors such as Si and Ge have a pure diamond lattice. Some compound semiconductors such as GaAs have a *zincblende lattice*, which is also closely related to a diamond lattice with two different types of atoms in the lattice. We will now discuss some salient features of the diamond lattice. The unit cell of a diamond lattice is shown in Fig. 2.11.

The diamond lattice can be obtained by inserting one fcc lattice into another fcc lattice displaced along the space diagonal by one-fourth of its length. Figure 2.12 shows a tetrahedron shown by a dotted line.

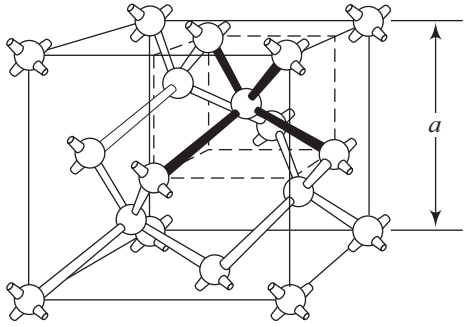


Fig. 2.11 The diamond unit cell

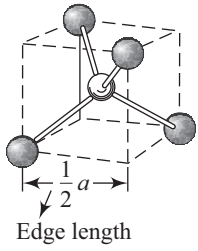


Fig. 2.12 Schematic diagram of a tetrahedron within the diamond unit cell

In the diamond lattice every atom is surrounded by four nearest neighbours located at the apexes of the tetrahedron with an edge $a/2$.

All atoms shown in Fig. 2.11 are similar. If alternate atoms differ, then what results is a zincblende structure. Thus in a zincblende structure the two interpenetrating sublattices are composed of different atoms. For example, in a GaAs lattice one of the sublattice is of Ga and the other of As (see Fig. 2.13). Most compound semiconductors have the zincblende structure though some II–VI compound semiconductors display a slightly different structure, called *wurtzite lattice*.

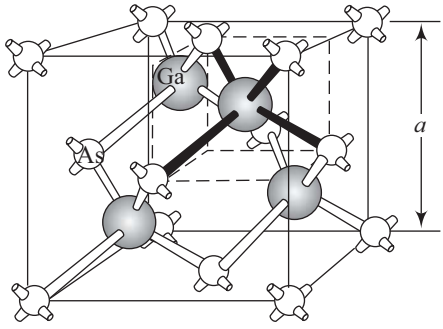


Fig. 2.13 The zincblende (sphalerite) lattice of GaAs

- There are lattices of two types: the translational Bravais lattice and the lattice with a basis.
- The diamond lattice is an example of a three-dimensional lattice with a basis.
- Some solids have two or more crystal structures, each of which is stable for a particular range of temperatures and pressures.

2.4 Atomic Bonding

A solid exists in that state due to the forces of interaction between its constituent particles. These forces of interaction result in bonds between the atoms of the solid. We will now learn the salient features of some of these bonds.

2.4.1 Van der Waals Bond

The most general bond that exists between any two atoms or molecules is due to Van der Waals forces. You would know that these forces are introduced in order to explain the equation of state of real gases. Van der Waals equation is of the form

$$\left(p + \frac{a}{V^2}\right)(V - b) = RT \tag{2.6}$$

where the correction terms a/V^2 and b account for the forces of attraction and repulsion acting between the real gas molecules, respectively. Van der Waals forces manifest in the interaction between molecules with saturated chemical bonds (O_2, H_2, N_2, CH_2, etc.) and also between the atoms of inert gases. This makes it possible for them to exist in both liquid and solid states.

2.4.2 Ionic Bond

Atoms occupying places in the periodic table next to inert gases have a tendency to acquire the electronic configuration of inert gases by either accepting or giving away electrons. The valence electron of alkali metals, which immediately follow the inert gases, moves outside the closed shell and thus is only weakly connected to the nucleus. The halides immediately preceding the inert gases lack an electron required to complete a stable shell, which is a characteristic of an inert gas. Therefore, the halides exhibit a high affinity towards an excess electron.

Atoms such as metals and halides are bonded in the following way. First, a recharging of the atoms takes place. The electron from the metal moves over to the halide, enabling it to become a positively charged ion. The haloid atom in turn becomes negatively charged. The interaction between these ions is governed by Coulomb's law as applicable to two opposite charges. Such a bond is known as an ionic bond. Some common ionic crystals are sodium chloride (NaCl), potassium iodide (KI), rubidium bromide (RbBr), etc.

2.4.3 Covalent Bond

In a covalent bond neighbouring atoms share their valence electrons. A hydrogen molecule is an example of *covalent bonding*. Each hydrogen atom has one electron and needs one more electron to complete the lowest shell. Covalent bonding leads to the sharing of electrons between atoms so that the valence energy shell of each atom is full.

Group IV atoms such as silicon (Si) and germanium (Ge) also form covalent bonds. These elemental semiconductors have four valence electrons. They need four more electrons to complete the valence energy shell. Let us understand this using the example of silicon. Each silicon atom has four valence electrons. It also has four nearest neighbours. Each valence electron is shared with the valence electron of a neighbouring atom in a covalent bond as shown in Fig. 2.14.

A hydrogen atom in a hydrogen molecule does not have additional electrons after forming a covalent bond with another hydrogen atom. However, the silicon atom has additional electrons with which it can form covalent bonds with electrons

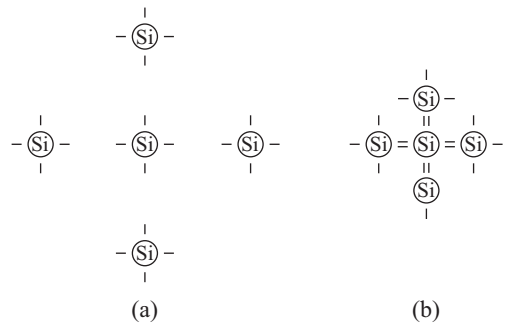


Fig. 2.14 (a) Silicon valence electrons and (b) covalent bonding between neighbouring electrons

of neighbouring atoms. The process can go on, and silicon arrays can lead to a large crystal. In this crystal, each silicon atom has four nearest neighbours and eight shared electrons. As discussed earlier, these nearest neighbours along with the host atom have a tetrahedral structure and have a diamond lattice structure. We can thus see that the atomic bonding and crystalline structure are correlated.

2.4.4 Metallic Bond

Metals are a special group of substances occupying places at the beginning of every period of the periodic table. Metallic bond cannot be explained by the presence of ionic or covalent bond. This is because this type of the bond exists between identical atoms having identical affinity for electrons. Metallic atoms do not have enough valence electrons to form covalent bonds with their nearest neighbours. A copper atom, for example, has one valence electron and thus can form a bond only with a single atom. However, it is known that in the copper lattice every atom is surrounded by 12 neighbours, which is impossible without the presence of suitable lines of force. In metallic atoms the external valence electrons are rather weakly coupled to the nucleus. The atoms come so close together in liquid and solid states that the valence electrons are able to leave their respective parent atoms and wander throughout the lattice. This results in a homogeneous distribution of negative charge in the lattice. The metallic bond is due to the interaction of positive ions with the electron gas. All the atoms of the crystal take part in the so called collectivization (no special attachment to respective atoms) process.

2.5 Imperfections and Impurities in Solids

In our discussion on solids so far we have limited ourselves to ideal perfect systems. Though technology has reached high levels of perfection, the ideal single-crystal structure is still nearly impossible to reach. Two types of irregularities usually manifest themselves in a real crystalline structure. One of them is imperfections in the basic arrangement of atoms constituting the crystal and the other is the introduction of impurities during the various growth and processing steps. These imperfections and impurities can play a key role in the electrical characteristics of the material. We would now give a brief outline of this important topic.

2.5.1 Imperfections

There are several imperfections in a crystal lattice. Some of the more important ones are the following.

Thermal vibrations An ideal single crystal consists of atoms at well-defined lattice sites separated by a constant distance between themselves. At any finite

temperature the crystal has a temperature-dependent thermal energy. This thermal energy in turn leads to random vibration of the atoms about an equilibrium lattice point. The distance between atoms is no more a constant but fluctuates randomly, disrupting the ideal geometric arrangement of atoms. This imperfection is called *lattice vibration*. Some electrical parameters are affected due to this imperfection.

Point, line, and plane defects When the absence of an atom or dislocation in a crystal is limited to individual lattice sites, the defect is called a point defect. If an atom is missing from a particular lattice site, the defect is called a *vacancy*. The presence of an atom at locations between lattice sites is called an *interstitial*. Figure 2.15 is a schematic representation of a vacancy and an interstitial.

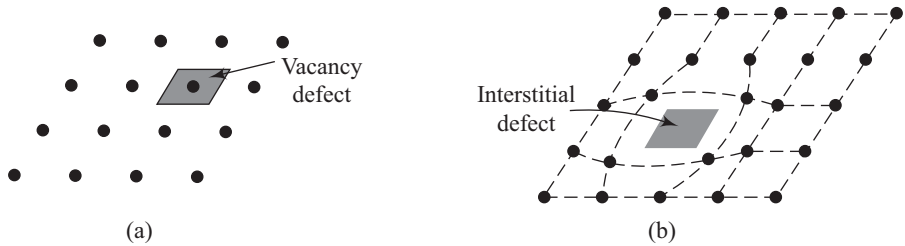


Fig. 2.15 Two-dimensional single-crystal lattice showing (a) vacancy defect and (b) interstitial defect

Vacancies and interstitials change the electrical properties of a material. This is often due to the deviations produced in the nature of chemical bonding between atoms. Sometimes a vacancy and an interstitial may occur in close proximity. Thus atoms may move from their natural sites to interstitials, thereby creating a vacancy. This vacancy–interstitial defect is called a *Frenkel defect*. *Frenkel defects* produce effects that are characteristically different from the ones produced by simple vacancies or interstitials.

More complex defects can also occur in crystals. One such defect is *Schottky defect*. In this type of defect the vacancies are not accompanied by a simultaneous transition of atoms to interstitials. If an entire row of atoms is missing from the normal lattice site, the defect is called a *line defect*. A line defect is also referred to as a *line dislocation*. A two-dimensional schematic diagram of a line dislocation is shown in Fig. 2.16.

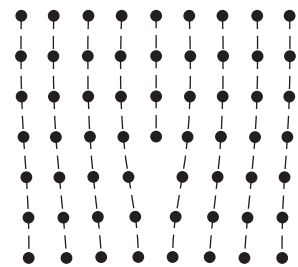


Fig. 2.16 A line dislocation in a two-dimensional lattice

Line dislocations also disturb the ideal atomic periodicity and affect the ideal atomic bonds in the crystal. Due to these reasons, a line dislocation can alter the electrical properties of a semiconductor material.

2.5.2 Impurities

Contemporary refining techniques cannot guarantee absolute purity of crystals. A very pure material may contain up to 10^{-9} per cent of impurities, which corresponds to an impurity atom concentration of about 10^{17} m^{-3} of the material. This level of impurity concentration corresponds to a grain of rye contained in about 10 tonnes of wheat. Such levels of impurities can, however, have appreciable effects in altering chemical, optical, electrical, and mechanical properties of the material. When impurity atoms are located at the normal lattice sites, they are called *substitutional impurities*. On the other hand, if the impurity atoms are located between normal lattice sites, they are called *interstitial impurities*. A schematic representation of these two types of impurities is shown in Fig. 2.17.

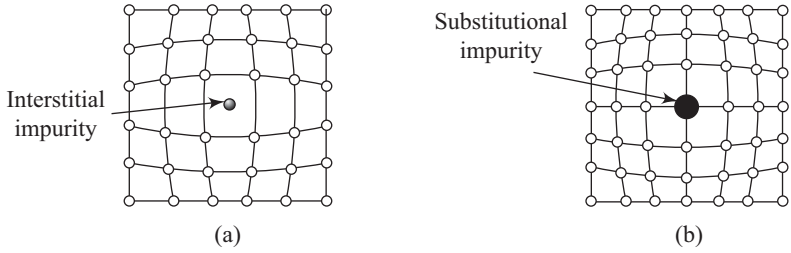


Fig. 2.17 (a) Interstitial and (b) substitutional impurities

The controlled addition of desirable impurities to modify the conductivity of a semiconductor material is called *doping*. Impurities such as B and Al are called acceptor impurities in silicon because they can accept electrons, whereas impurities such as P and As are called donor impurities in silicon because they can contribute additional electrons. These acceptor and donor impurities are substitutional impurities in silicon. Some impurities such as Mn and Ni are interstitial impurities in silicon. Doping of an impurity is carried out using two main techniques, namely *impurity diffusion* and *ion implantation*.

Impurity diffusion For carrying out impurity diffusion, the semiconductor crystal is kept at a high temperature (around 1000°C) gaseous atmosphere of the desired impurity atom. At such high temperatures the crystal atoms can move randomly around their lattice sites. Vacancies can be created due to this random motion and impurity atoms can move through the crystal by hopping from vacancy to vacancy. In the impurity diffusion process, the impurities move from a region of high concentration to a region of lower concentration within the crystal. When the temperature is reduced, the impurity atoms get permanently frozen in the substitutional lattice sites. Diffusion of desirable impurities is a key process step towards the development of complex semiconductor devices and circuits.

Ion implantation This is another process by which a known amount of dopant atoms can be introduced into a semiconductor. In this process, a beam of high-energy (50–100 keV) dopant ions is directed towards the semiconductor surface.

The ions penetrate the crystal and come to rest at some average depth. Thus ion implantation allows the device engineer to introduce impurity atoms into specific regions of the semiconductor crystal. This is a relatively lower temperature process than diffusion. All this is however achieved at a price. The incident impurity ions collide with the crystal atoms, leading to lattice-displacement damage. This displacement damage can be removed by a process called *thermal annealling*, in which the temperature of the crystal is raised for a short duration. Thus ion implantation is always followed by a thermal annealling step.

2.6 Bulk Crystal Growth

The tremendous growth in the semiconductor industry is completely dependent upon the supply of high-purity single-crystal semiconductor material. Using specialized techniques it is possible to obtain silicon with impurities less than one part in 10 billion. Such a high level of purity requires careful handling and treatment at all the stages of the fabrication process.

2.6.1 Starting Material

The starting raw material for a Si crystal is silicon dioxide (SiO_2). Silicon dioxide reacts with carbon in the form of coke using an arc furnace. Temperatures used are in the region of 1800°C. This reduces SiO_2 according to the reaction

$$SiO_2 + 2C \rightarrow Si + 2CO \tag{2.7}$$

The silicon obtained is called metallurgical grade silicon Si (MGS) and has impurities such as Fe, Al, and heavy metals at several hundred to several thousand parts per million (ppm). (You would recollect that 1 ppm of Si corresponds to an impurity level of 5×10^{16} cm^{-3}). The metallurgical grade Si is not single crystal and is not pure enough for electronic applications. It needs to be further refined to electronic grade Si (EGS), which has impurities in the region of parts per billion. MGS is reacted with dry HCl to form trichlorosilane ($SiHCl_3$) according to the reaction

$$Si + 3HC1 \rightarrow SiHC1_3 + H_2 \tag{2.8}$$

Chlorides of some impurities present ($FeCl_3$, for example) are also formed during this process. $SiHCl_3$ is a liquid having a boiling point of 32°C, whereas the chlorides of impurities have different boiling points. Fractional distillation is then used to separate pure $SiHCl_3$ from the impurities. $SiHCl_3$ is then reacted with H_2 to yield EGS, according to the reaction

$$2SiHC1_3 + 2H_2 \rightarrow 2Si + 6HC1 \tag{2.9}$$

2.6.2 Single-crystal Ingots

The EGS obtained in the above process is polycrystalline in nature. The most common technique to convert the polycrystalline EGS to single-crystal ingots is

the *Czochralski method*. EGS is melted in a quartz-lined graphite crucible by heating it to the melting point of Si (1412°C). Resistive heating is used to achieve this. A seed crystal is then lowered into the molten EGS and then raised slowly. The seed crystal serves as a template for growth. The crystal is also rotated slowly as it grows to ensure temperature uniformity and to provide a slight stirring effect. A schematic diagram of the Czochralski growth process is shown in Fig. 2.18.

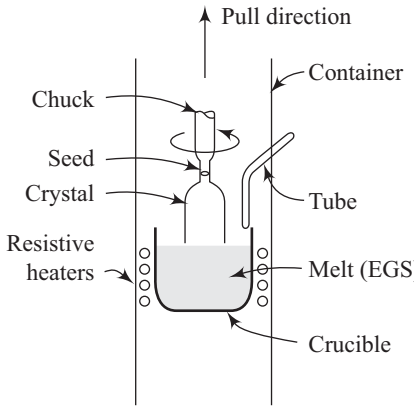


Fig. 2.18 Schematic diagram of Czochralski growth technique

The shape of ingots is determined by two competing processes. On one hand, the crystal structure gives the growing ingot a polygonal shape. On the other hand, the surface tension of the melt supports a circular cross-section just like any other liquid that is allowed to acquire the equilibrium shape. Large ingots are nearly circular in cross-section. The ingots so grown contain some undesirable impurities. These are removed by using a technique called *zone refining*. A high-temperature coil or RF induction coil is made to slowly pass along the length of the boule. A thin layer of liquid is formed. A distribution of impurities takes place between the two phases. This distribution is decided by the segregation coefficient, which is the ratio of the concentration of impurities in the solid to the corresponding concentration in the liquid. A segregation coefficient of 0.2 implies that the concentration of impurities in the liquid is a factor of 5 greater than that in the solid. As the liquid zone moves through the material, the impurities get driven along with the liquid. When the RF coil is made to repeat the process several times, most impurities are present at the end of the bar. This end is then cut off.

The fully grown boule is then mechanically trimmed to a proper diameter. A flat perpendicular to the [110] direction [or the (110) plane] is then ground over the entire length of the boule. The boule is then sliced into wafers. Most silicon ingots are grown along the ⟨100⟩ direction. For ⟨100⟩ Si wafers, the {110} cleavage planes are orthogonal to each other. The notch is then used to orient the individual integrated circuit chips along the {110} planes so that sawings of individual chips can be carried out with high yield.

The individual wafers so produced then undergo chemical–mechanical polishing. The process uses a slurry of very fine particles of SiO_2 in a basic NaOH solution. The end result is a surface with a mirror-like finish. Several applications require only one of the two surfaces to be polished to this degree. Some special processing sequences however require mirror finish on both sides of the wafer.

Intentional doping is also carried out to obtain Si wafers with specific electrical characteristics.

Another crystal growth technique is called the floatzone (FZ) technique. A schematic of the typical floatzone technique is shown in Fig. 2.19.

In this process, the refined polycrystalline material is first made into a rod. A small portion of the rod is then melted locally using a radio-frequency (RF) coil, while the rod is slowly rotated. The melting starts from the lowermost portion of the rod which also has the seed crystal. The seed crystal thus determines the orientation of the growing crystal. The RF coil is slowly moved upwards. This moves the molten zone upwards towards the top of the rod. The lower portion of the rod then solidifies and forms single crystal silicon. Unike the Czochralski no crucible is used in this technique, thus eliminating contamination from the crucible. This results in good doping uniformity in the upper crystal layers by optimizing the rod rotation and RF coil movement parameters.

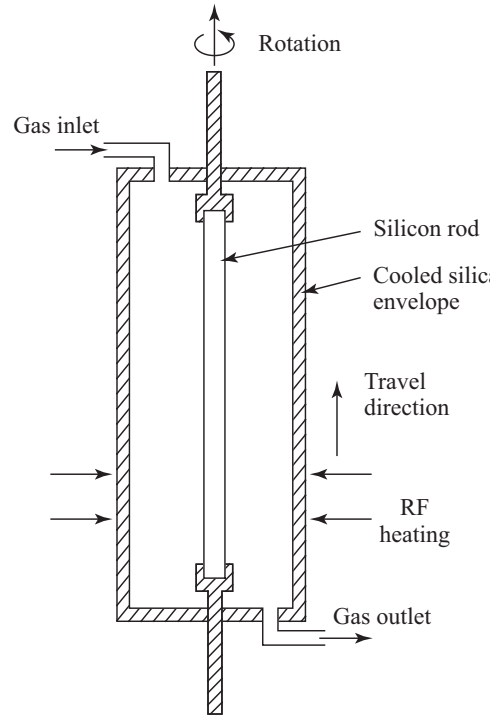


Fig. 2.19 Schematic of floatzone growth technique

Conductivity of semiconductors is generally measured using the four-probe measurement method. A schematic of a typical set-up is shown in Fig. 2.20.

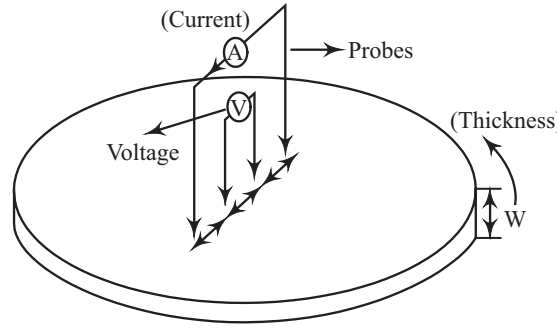


Fig. 2.20 Schematic of four-probe measurement technique

The set-up consists of four equally spaced probes that can be placed on the surface of the semiconductor. In Fig. 2.20 the inter-probe distance is indicated by s. The two outer probes are used to drive a small current through the semiconductor. The two inner probes are used to measure voltage using a very high impedance meter. The high impedance meter ensures negligible current flow through the contacts, thereby eliminating the effect of contact resistance. For sample thickness $w \ll s$, the conductivity of the sample is given by

$$\sigma = \frac{I \ln 2}{v \pi w}$$

The equation for σ is valid only if the thickness of the wafer is such that the current is uniformly distributed throughout the thickness and the distance between the probes is much smaller than the diameter of the wafer.

If $w \gg s$, conductivity σ, is given by

$$\sigma = \frac{I}{2\pi s V}$$

The four contacts can also be realized on the corners of a square as shown in Fig. 2.21 This structure is called van der Pauw measurement set-up. The expression for σ remains the same if the current source is connected between two adjacent corners and voltage is measured between the other two.

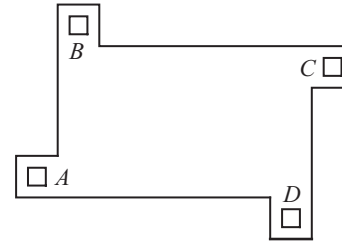


Fig. 2.21 Van der Pauw configuration

2.7 Epitaxial Growth

The process of growing a single-crystal layer with a specific orientation on a substrate wafer is called epitaxial growth or simply epitaxy. Epitaxial growth plays a very important role in the development of semiconductor devices. When the epitaxial layer grown is the same as the substrate material, epitaxy is referred to as *homoepitaxy*. If, on the other hand, the growing layer is of a different material from the substrate, the process is termed *heteroepitaxy*. The substrate serves as a seed for the growth of the epitaxial layer, but the temperatures involved are much lower than those required to melt the material. The growing crystalline layer maintains the structure and orientation of the substrate. Thus in heteroepitaxy the substrate and the epitaxial layer material must have the same lattice constant and lattice structure. Thus GaAs can be grown on Ge substrates. Heteroepitaxy also opens up some very interesting possibilities. GaAs and AlAs both have the zincblende structure with a lattice constant of 5.65 Å. Therefore, the composition x of the ternary compound $Al_x Ga_{1-x} As$ can be chosen to have any value without affecting its lattice-matching property to a GaAs wafer. The particular device requirement determines the choice of a specific value of composition x.

Figure 2.22 shows the energy band gap E_g for several III–V ternary compounds as a function of the lattice constant for a variation over a composition range. Different ternary compound systems are shown in the figure. For ternary compound InGaAs, the band gap changes from 0.36 to 1.43 eV and the lattice constant changes from 6.06 Å to 5.65 Å as one goes from InAs to GaAs. Any particular binary substrate is not suitable for the growth of the entire range from InAs to GaAs. A particular composition of $In_{0.53}Ga_{0.47}As$ is, however, possible to be grown lattice matched on an InP substrate as shown in the figure.

Quarternary alloys such as InGaAsP can permit an even broader range of semiconductor characteristics with lattice-matched substrates.

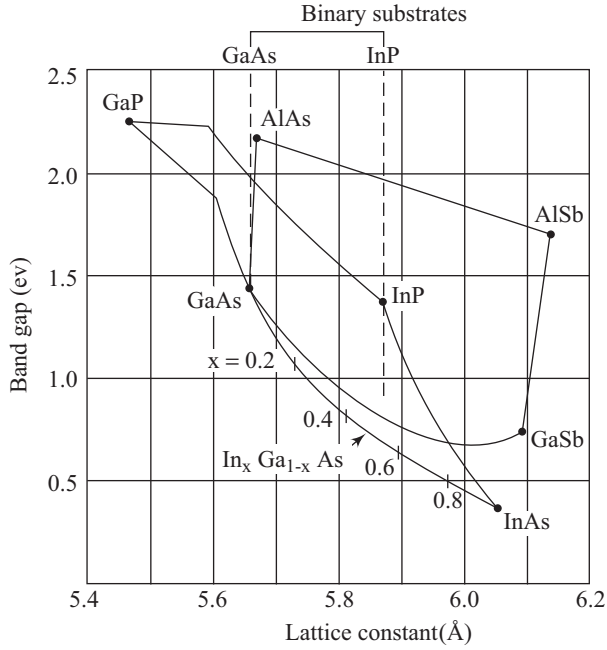


Fig. 2.22 Lattice constants of III–V compounds

The epitaxial layers have a varied set of applications. For applications involving some materials, it may not always be possible to grow these layers on perfectly matched single-crystal substrates. For some applications, very thin (~ 100 Å) epitaxial layers on lattice-mismatched crystals are needed. For thin layers with mismatch within a few per cent, the epitaxial layer grows with a lattice constant in compliance with the substrate crystal. As the lattice constant of the epitaxial layer adapts to the substrate, the layer is in compression or tension along the surface plane. Such layers are strained and are called *pseudomorphic layers*. This situation persists up to a particular thickness, called the *critical layer thickness*. The magnitude of the critical layer thickness depends upon the magnitude of lattice mismatch. Beyond the critical thickness, the strain energy leads to the formation of defects called *misfit dislocations*. *Strained-layer superlattice* (SLS) structures are grown using thin alternating layers of slightly mismatched crystal layers. In an SLS, alternate layers are in tension and compression. *The effective lattice constant of an SLS is the average of the two bulk constituent materials*. Figure 2.23 shows a schematic diagram of SiGe–Si heteroepitaxial system with two different conditions.

2.7.1 Vapour-phase Epitaxy

The growth of epitaxial layers using crystallization from the vapour phase is called vapour-phase epitaxy (VPE).

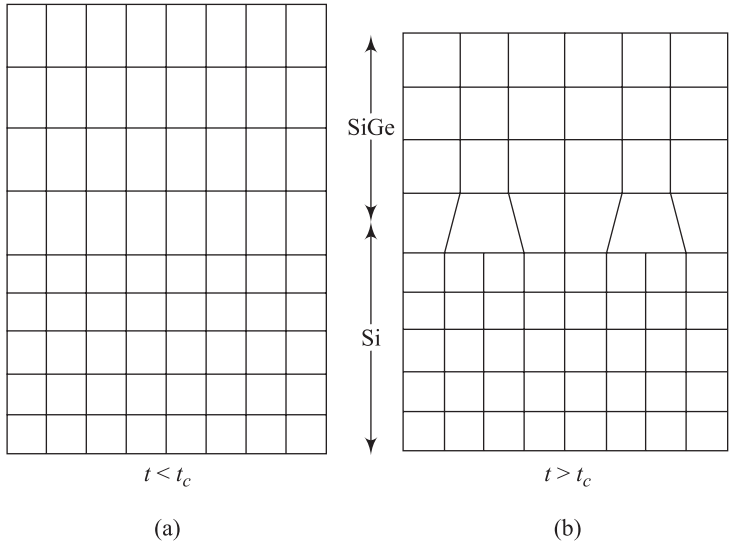


Fig. 2.23 Heteroepitaxy of a SiGe layer on Si, (a) For layer thicknesses less than the critical layer thickness, t_c, pseudomorphic growth occurs (b) Above t_c, misfit dislocations form at the interface

In one method, silicon tetrachloride gas reacts with hydrogen gas to give Si and anhydrous HCl according to the following reaction, at ~1150–1250°C:

$$SiCl_4 + 2H_2 \rightleftharpoons Si + 4HCl \tag{2.10}$$

The growth of an epitaxial silicon layer is possible on heated substrates. The HCl remains gaseous at growth temperatures and is swept out of the reactor. Sharp profiles of the dopant concentration are possible to be achieved using this technique. The arrangement consists of a chamber, into which the gases can be introduced and a provision exists for heating the substrate. The chamber is called a *reaction chamber* or simply a reactor. A schematic diagram of a barrel-type Si VPE reactor is shown in Fig. 2.24.

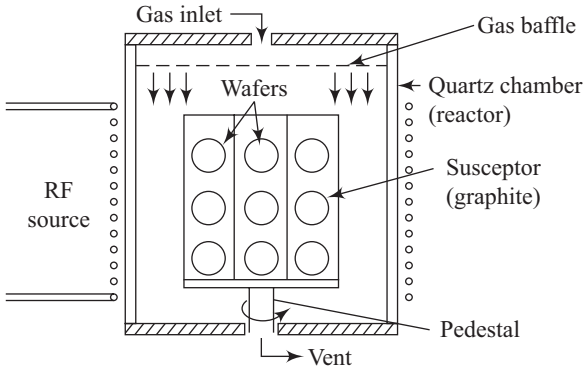


Fig. 2.24 A barrel-type reactor for Si VPE

The two gases along with other dopant gases are introduced into the reactor through the gas inlet. The Si wafers are kept on a graphite susceptor, which is heated with an RF heating coil. Tungsten halogen lamps can also be used for heating the silicon wafers.

Other reactions like pyrolysis of silane (SiH_4) are also used for obtaining epitaxial layers of silicon. Pyrolysis of silane takes place at 1000°C and involves the following reaction:

$$SiH_4 \rightarrow Si + 2H_2 \tag{2.11}$$

The vapour-phase epitaxial growth technique is also used for obtaining epitaxial layers of III–V compounds such as GaAs, GaP, and ternary alloy GaAsP. Compound semiconductors can also be obtained as epitaxial layers using a technique called metal-organic vapour-phase epitaxy (MOVPE). GaAs epitaxial layers can be obtained by reacting arsine with the organometallic compound trimethylgallium to form GaAs according to the reaction

$$(CH_3)_3\,Ga + AsH_3 \rightarrow GaAs + 3CH_4\,(\text{methane}) \tag{2.12}$$

This reaction takes place at around 700°C. Other organometallic compounds such as trimethyl-aluminium can be added to the gas mixture mentioned above to obtain epitaxial layers of AlGaAs.

2.7.2 Liquid-phase Epitaxy

Liquid-phase epitaxy technique makes use of the fact that a compound of a semiconductor with another element may have a lower melting point than the semiconductor itself. The semiconductor seed substrate is held in the liquid compound. The temperature profiles are such that the substrate does not melt. A single-crystal semiconductor epitaxial layer then grows on the seed substrate as the solution is slowly cooled. In another version of this technique, a GaAs substrate is dipped in molten gallium saturated with GaAs. The melt is supercooled just below its solidification point and an epitaxial layer is then obtained on the crystalline GaAs substrate. A schematic diagram of a liquid-phase epitaxy (LPE) system is shown in Fig. 2.25. A slider is used to move the GaAs substrate over the surface of the molten material.

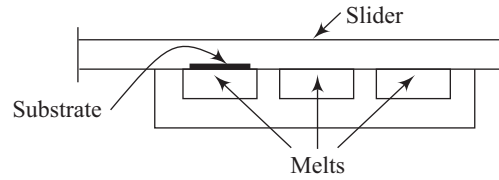


Fig. 2.25 Schematic diagram of a liquid-phase epitaxy system

2.7.3 Molecular Beam Epitaxy

Molecular beam epitaxy (MBE) is a sophisticated evaporation technique performed in ultra high vacuum. The substrate is held in a high-vacuum environment and elemental species are evaporated from ovens (also referred to as effusion or K-cells) and impinged upon the heated substrate where they get

deposited into single-crystal epitaxial layers. GaAs epitaxial layer can be obtained using Ga, As, Al, Si effusion cells. Silicon and aluminium serve to provide dopant atoms. Extremely sharp profiles can be obtained with resolutions of virtually one atomic layer at substrate temperatures in the region of 500°C to 600°C for GaAs. A schematic diagram of a typical GaAs MBE system is shown in Fig. 2.26.

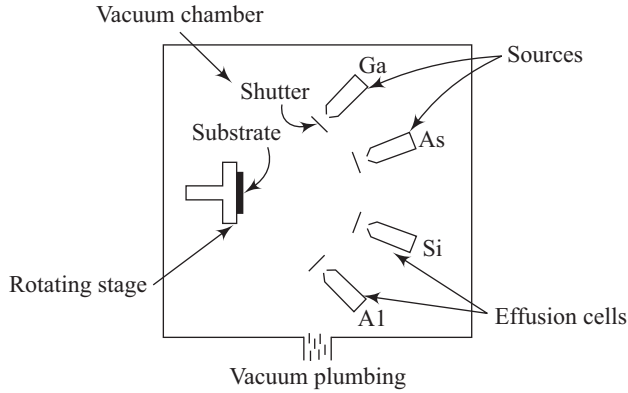


Fig. 2.26 A typical molecular beam epitaxy system

- Silicon MBE systems use an electron beam to create an atomic beam of silicon.
- GaAs crystals are grown using the liquid encapsulated Czochralski (LEC) technique.
- LPE is very popular in university environment.

Solved Problems

2.1 Hard spheres are packed in a bcc lattice in such a manner that the atom at the centre just touches the atoms present at the corners of the cube. Calculate the fraction of the bcc unit cell volume filled with hard spheres.

Solution

Each corner sphere of the bcc unit cell is shared with eight neighbouring cells. Thus each cell contains one eighth of a sphere at all the eight corners. Each unit cell also contains one central sphere.

$$\text{Spheres per unit cell} = 8 \times \frac{1}{8}(\text{corner}) + 1(\text{centre}) = 2$$

Nearest neighbour distance (along diagonal AE in Fig. 2.1.1) $= \dfrac{a\sqrt{3}}{2}$

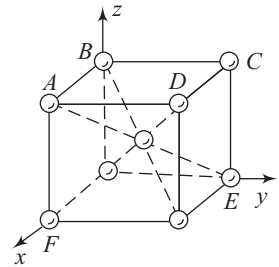


Fig. 2.1.1

$$\text{Radius of each sphere} = \frac{a\sqrt{3}}{4}$$

$$\text{Volume } V = \frac{4\pi}{3}\left[a\left(\frac{\sqrt{3}}{4}\right)\right]^3 = \frac{\pi a^3\sqrt{3}}{16}$$

The maximum fraction f of the unit cell filled is given by

$$f = \frac{\text{No. of spheres} \times \text{volume of one sphere}}{\text{volume of unit cell}}$$

leading to

$$f = \frac{2 \times \pi a^3\sqrt{3}/16}{a^3} = \frac{\pi\sqrt{3}}{8} = 0.68$$

Thus 68% of the bcc unit cell volume is filled with hard spheres.

2.2 Describe the plane shown in Fig. 2.2.1 in terms of the corresponding Miller indices.

Solution

The given plane has intercepts a, $2b$, and $3c$ along the three crystal axes.

Lattice points in the three-dimensional lattice are given by the expression,

$$\mathbf{r} = p\mathbf{a} + q\mathbf{b} + s\mathbf{c} \qquad (2.2.1)$$

From Fig. 2.2.1 and Eqn (2.2.1), we can conclude

$$p = 1,\ q = 2, \text{ and } s = 3$$

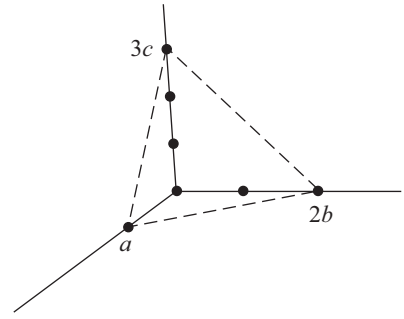


Fig. 2.2.1 Plane in a crystal

Taking reciprocals we get

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

Multiplying all the numbers by the lowest common denominator 6, we have

(6, 3, 2)

Thus the plane we have depicted in Fig. 2.2.1 is denoted as (6, 3, 2).

2.3 Lattice constants of Si and GaAs are 5.43×10^{-8} cm and 5.65×10^{-8} cm, respectively. Calculate the densities of Si and GaAs. The atomic weights of Si, Ga, and As are 28.1, 69.7, and 74.9, respectively.

Solution

(a) Si

Lattice constant $a = 5.43 \times 10^{-8}$ cm

Also,

n = No. of atoms/cells = 8

Atomic concentration $N = \frac{8}{a^3} = \frac{8}{(5.43 \times 10^{-8})} = 5 \times 10^{22}$ atoms/cc

$$\text{Density} = \frac{5 \times 10^{22} \times 28.1}{6.02 \times 10^{23}} = 2.33 \text{ g/cm}^3$$

(b) GaAs

In the GaAs crystal each cell has 4 Ga and 4 As atoms. Atomic concentration of Ga and As $= \frac{4}{a^3} = \frac{4}{(5.65 \times 10^{-8})^3} = 2.22 \times 10^{22}$ atoms/cc.

Thus,

$$\text{Density} = \frac{2.22 \times 10^{22} (69.7 + 74.9)}{6.02 \times 10^{23}} = 5.33 \text{ g/cm}^3$$

2.4 Calculate the surface density of atoms in the (111) plane of a body-centered cubic structure. Assume the lattice constant $a = 5$Å. Also assume the atoms to be hard spheres, with closest atoms touching each other.

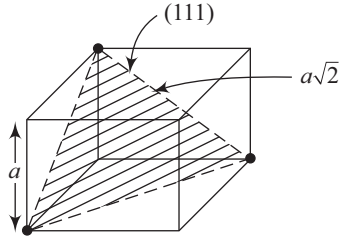


Fig. 2.4.1

Solution

To calculate planar concentration n_{hkl} on a given (hkl) plane, only atoms whose centres lie on bound area A are to be considered. Fig. 2.4.1 shows a (111) plane for a bcc crystal.

The shaded area is an equilateral triangle defined by face diagonals of length $a\sqrt{2}$.

$$\text{Height of the triangle} = a\sqrt{\frac{3}{2}}$$

Thus, the area of the triangular portion (shaded) is

$$\frac{1}{2} a\sqrt{2} \times a\sqrt{\frac{3}{2}} = \frac{a^2\sqrt{3}}{2}$$

Each atom at the corner contributes 1/6 to this area. Thus the planar concentration $n_{(111)}$ is given by

$$n_{(111)} = \frac{\frac{3}{6}}{a^2 \frac{\sqrt{3}}{2}} = \frac{1}{a^2\sqrt{3}}$$

Putting the given value of a we get

$$n_{(111)} = \frac{1}{(5 \times 10^{-10})^2 \sqrt{3}} = 2.3 \times 10^{18} \text{ atoms/m}^2$$

2.5 Si crystals are being grown using the Czochralski technique. The requirement is to have 5×10^{16} P atom/cm^3 in the ingot.

(a) Calculate the concentration of P atoms that should be present in the melt to result in the desired impurity concentration in the growing crystal in the initial phase of growth. Assume that the segregation coefficient for P in Si is $k_s = 0.35$.

(b) The initial load of Si in the crucible is 4 kg. How many grams of P should be added to the melt? The atomic weight of P is 31.

Solution

(a) The segregation coefficient k_S is given by

$$k_S = \frac{C_S}{C_L} \tag{2.5.1}$$

where C_S and C_L represent impurity concentrations in the solid and liquid, respectively.

Thus,

$$C_L = \frac{C_S}{k_S}$$

Putting the given values in expression (2.5.1), we get

$$C_L = \frac{5 \times 10^{16}}{0.35} = 1.43 \times 10^{17} \text{ cm}^{-3}$$

(b) At such a low P concentration, the density of Si is not expected to change much. Assuming the density of Si to be 2.33 g/cm^3, the volume V of the melt is

$$V = \frac{4000}{2.33} = 1717 \text{ cm}^3 \text{ of Si}$$

$$\text{No. of P atoms} = 1.43 \times 10^{17} \times 1717 = 2.46 \times 10^{20} \text{ atoms}$$

$$\text{Wt of P} = \frac{2.46 \times 10^{20} \times 31}{6.02 \times 10^{23}} = 12.67 \times 10^{-3} \text{ g}$$

The concentration of P atoms in the growing crystal is 1/0.35 times that in the melt. Silicon is thus used up more rapidly than P and the melt becomes richer in P as the growth proceeds. To obtain a uniformly doped ingot, the factor k_S is modified by suitably adjusting the pull rate. Czochralski growth systems have computerized controls for adjusting pull rate, temperature, etc.

2.6 List some of the most popular semiconductors which you have studied. Which one of these is the most popular and why?

Solution

Some of the most common semiconductors are silicon (Si), germanium (Ge), and gallium arsenide (GaAs).

Si is the most popular among all these due to the following fundamental advantages.

(i) Si possesses a higher value of band gap $= 1.1$ eV at room temperature in comparison to its counterpart germanium (Ge), which has an energy

band gap = 0.7 eV. Due to this reason Si poses less of a problem with regard to leakage current in electronic devices.

(ii) Si is abundant in nature. Si originally comes from silica and earth's crust contains around 97% silica.

(iii) Although GaAs possesses a higher value of band gap = 1.43 eV as compared to Si and Ge, it is particularly used for faster switching applications. GaAs boasts this feature due to very high mobility of carriers, especially electrons, as compared to that in case of silicon and germanium.

Other compound semiconductors are used for dedicated applications such as light emitting diodes (LEDs), lasers, and microwave devices.

2.7 Find the Miller indices of a plane that makes intercepts equal to 3, 4, and 5 times the basis vectors along the three crystal axes.

Solution

We can find out the intercepts in terms of basis vectors as

$$3\mathbf{a}, 4\mathbf{b}, \text{ and } 5\mathbf{c}$$

To calculate the Miller indices, reciprocals are first obtained for these intercepts. These reciprocals are

$$\frac{1}{3}, \frac{1}{4}, \text{and } \frac{1}{5}$$

Now we must reduce these fractions to the smallest triad of integers having the same ratio; so, multiplying these numbers by 60 we get Miller indices of plane as

$$\langle 20, 15, 12 \rangle$$

2.8 What is the basic lattice structure for:

(a) Silicon and germanium

(b) Gallium arsenide

How are these different?

Solution

Both silicon and germanium possess the 'diamond' lattice structure. In GaAs as well, atoms are arranged in the basic diamond structure, but are different on alternating sites. This is called the zincblende lattice and is typical of the III–V compounds.

2.9 (a) Considering a typical diamond lattice structure, find out the number of atoms in each cell. (b) If the value of lattice constant, L, for silicon has been provided as 5.43 Å, then estimate the density of atoms in silicon.

Solution

For a typical diamond structure eight atoms are shared by eight cells, six atoms are shared by two cells, and four atoms are internal to the cell.

(a) So, we can find the number of atoms in each cell $= \frac{8}{8} + \frac{6}{2} + 4 = 8$

(b) Volume of the cell is given by the cube of the lattice constant, L

$$\therefore \text{ Volume of the cell} = (5.43 \times 10^{-8})^3 \text{ cm}^3$$

$$\therefore \text{ Density} = \frac{8}{\text{Volume}} = 5 \times 10^{22} \text{ atoms cm}^{-3}$$

2.10 Determine the packing density [in terms of (mass per unit volume)] for silicon. Given Avogadro's number = 6.023×10^{23} atoms/mole.

Solution

Atomic weight of silicon = 28.09g/mole
Density of silicon atoms in diamond structure = 5×10^{22} cm^{-3} (from Problem 2.9)

$$\therefore \text{ Density (mass per unit volume)} = \frac{5 \times 10^{22} \text{ atoms/cm}^3 \times 28.09 \text{ g/mole}}{6.023 \times 10^{23} \text{ atoms/mole}}$$
$$= 2.33 \text{ g cm}^{-3}$$

Recapitulation

- Van der Waals bonding is due to Van der Waals forces between atoms and molecules.
- Metals and halides are bonded with ionic bonds.
- Sodium chloride, potassium iodide, and rubidium bromide are important ionic crystals.
- Covalent bonding leads to the sharing of electrons between atoms.
- Group IV atoms such as silicon and germanium form covalent bonds.
- In metallic atoms the external valence electrons are weakly coupled to the nucleus and get a collectivization with respect to the crystal.
- Imperfections and impurities play a key role in deciding the electrical characteristics of a material.
- Thermal vibrations lead to non-constant distances between atoms.
- Vacancies and interstitials are two important point defects.
- A vacancy and an interstitial occurring in close proximity lead to a Frenkel defect.
- In Schottky defect a vacancy is not accompanied by a simultaneous transition of atoms to interstitials.
- An entire row of atoms is missing in a line defect.
- Impurities can be interstitial or substitutional.
- In impurity diffusion process, the impurities move from a region of high concentration to a region of low concentration.
- In the ion implanation process, a beam of high-energy dopant ions is directed towards the substrate semiconductor.
- The displacement damage resulting in ion implantation can be removed with thermal annealling.

- Silicon dioxide is the starting raw material for obtaining Si crystals.
- The silicon obtained by reducing SiO_2 is called metallurgical grade silicon or MGS.
- $SiHCl_3$ is reacted with H_2 to yield electronic grade Si (EGS).
- Single-crystal Si ingots can be grown using the Czochralski method.
- The silicon boule is given a flat perpendicular to the [110] direction.
- In homoepitaxy, the epitaxial layer is the same as the substrate.
- In heteroepitaxy, the epitaxial layer is different from the substrate.
- Thin lattice mismatched layers are called pseudomorphic layers.
- The growth of epitaxial layers using crystallization from the vapour phase is called vapour-phase epitaxy.
- Liquid-phase epitaxy uses a compound of the semiconductor in the liquid form.
- Molecular beam epitaxy is a sophisticated evaporation technique carried out under ultra-high vacuum conditions.
- Semiconductors constitute a group of materials with electrical conductivities between metals and insulators.
- Conductivity of semiconductors can be modified by introducing controlled amounts of desirable impurities.
- Semiconductors are of two types, namely elemental and compound semiconductors.
- Elemental semiconductors are found in column IV of the periodic table.
- Some common compound semiconductors are made up of either a combination of column III and column V atoms or a combination of column II and column VI atoms.
- Ternary and quarternary compound semiconductors are being increasingly used for special applications.
- Amorphous, polycrystalline, and crystalline are the three general types of solids.
- In a single crystal a basic pattern unit is repeated periodically throughout the material.
- In polycrystalline materials the periodicity of the basic pattern gets interrupted at certain boundaries called grain boundaries.
- In amorphous materials the grain boundaries become comparable to the size of the basic pattern.
- The periodic arrangement of atoms in a crystal is called a lattice.
- The fundamental building block of a three-dimensional lattice is called a unit cell.
- The smallest unit cell is called a primitive cell.
- The displacement vector between two lattice points is given by a vector **r** expressible as

$$\mathbf{r} = p\mathbf{a} + q\mathbf{b} + s\mathbf{c}$$

  where p, q, and s are integers and **a**, **b**, and **c** are called basis vectors.
- Simple cubic, body-centered cubic, and face-centered cubic are three common cubic lattices.

- Miller indices h, k, l within braces, i.e. (hkl) are used to indicate specific planes in a crystal.
- Equivalent planes are shown enclosed in curly brackets { }.
- Direction in a crystal is expressed within square brackets, i.e. [abc].
- Equivalent directions are expressed within angular brackets ⟨⟩.
- Elemental semiconductors have a pure diamond lattice, which can be obtained by inserting one fcc lattice into another fcc lattice.
- In a diamond lattice every atom is surrounded by four nearest neighbours.
- Alternate atoms are different in a zincblende structure, which is otherwise identical to a diamond structure.
- $V_C^* = \dfrac{(2\pi)^3}{V_C}$
- $2d_{hkl} \sin\theta = \lambda$
- For the four-probe measurement system, $\sigma = \dfrac{I \ln^2}{V \pi W}$

Exercises

Review Questions

2.1 What are semiconductors?
2.2 Name some elemental and some compound semiconductors.
2.3 What are compound semiconductors?
2.4 What are quarternary compound semiconductors? Give one example.
2.5 Differentiate between crystalline, polycrystalline, and amorphous semiconductors with suitable diagrams.
2.6 What is a lattice?
2.7 Explain the concept of unit cell. What is a primitive cell?
2.8 Explain the concept of basis vectors.
2.9 Show simple cubic, body-centered cubic, and face-centered-cubic lattice with suitable sketches.
2.10 What are Miller indices? How are they determined?
2.11 How are directions within a crystal depicted?
2.12 Differentiate between diamond structure and zincblende structure.
2.13 What is a Van der Waals bond? Give one example.
2.14 How are ionic bonds formed?
2.15 Sketch the covalent bonds existing in silicon.
2.16 Explain the process of collectivization in the formation of metallic bond.
2.17 Explain the terms: (i) point defect (ii) line defect, and (iii) plane defect.
2.18 What are thermal vibrations?
2.19 Define vacancy and interstitial defects.
2.20 Explain substitutional and interstitial defects using suitable sketches. Give one example of each.
2.21 Differentiate between impurity diffusion and impurity ion implantation.
2.22 What is the role of annealling in the process of ion implantation?

2.23 Give the reduction reaction of SiO_2 resulting in Si.
2.24 How is MGS converted into EGS?
2.25 Draw a schematic set-up for the Czochralski growth process.
2.26 Differentiate between homoepitaxy and heteroepitaxy.
2.27 What are ternary and quarternary compounds?
2.28 What are pseudomorphic layers?
2.29 What is a reactor in a VPE system?
2.30 Briefly explain the VPE process.
2.31 Give a schematic diagram of the LPE process.
2.32 Explain the molecular beam epitaxy process with a suitable sketch.
2.33 Explain the different types of solids with suitable sketches.
2.34 Explain the terms, (i) lattice; (ii) unit cell; (iii) primitive cell; (iv) basis vectors.
2.35 What are Miller indices? Explain with suitable examples.
2.36 Indicate the salient features of a diamond lattice.
2.37 Give a brief account of the different types of atomic bonding. Give suitable examples.
2.38 How are point defects created in a crystal?
2.39 Explain the Czochralski technique of crystal growth.
2.40 What are the factors to be taken into account for carrying out successful heteroepitaxy?
2.41 Describe the vapour-phase epitaxy process with the help of a suitable sketch.
2.42 What is molecular beam epitaxy? Give its salient features.
2.43 What are reciprocal lattice vectors?
2.44 Write an expression for the volume of unit cell of the reciprocal lattice.
2.45 Give a schematic representation of FZ technique.
2.46 Describe the process of chemical–machanical polishing.
2.47 How is the XRD technique used for determining the interplanar spacing?
2.48 Describe the four-probe conductivity measurement method.
2.49 Give a schematic of the Van der Pauw configuration.

Problems

2.1 Hard spheres are packed in an fcc lattice with maximum packing density. Calculate the fraction of the cell filled.
[*Hint:* Atoms per unit cell = 1(corners) + 3 (from faces) = 4
Radius of each sphere = $\frac{1}{4}(a\sqrt{2})$]

Ans. 74%

2.2 Assume each atom to be a hard sphere with the surface of each atom in contact with the surface of its nearest neighbour. Calculate the percentage of the total unit cell volume that is occupied in a simple cubic lattice.
[*Hint:* A cubic unit cell contains only one sphere and the edge length is exactly equal to the diameter of the sphere]

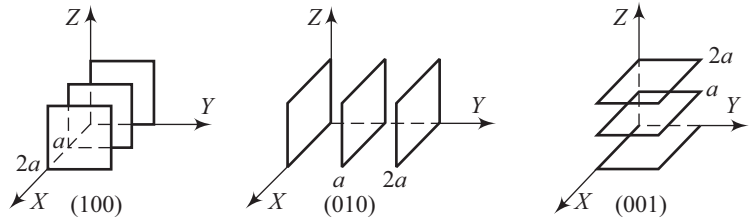
Ans. 52.4%

2.3 Find the percentage of the total unit cell volume that is occupied in a diamond lattice.
[*Hint:* There are 8 atoms per unit cell.]

Ans. 34%

2.4 Sketch the following principal planes in a cubic crystal:
(a) (100) (b) (010) (c) (001)

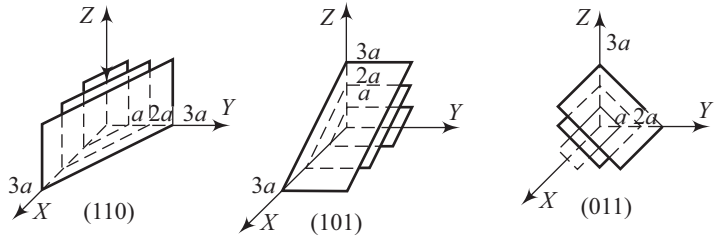
Ans.



(a) (b) (c)

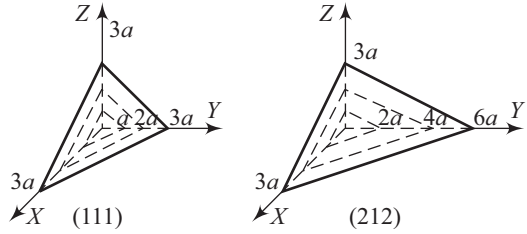
2.5 Sketch the planes (110), (101), and (011) in a cubic lattice.

Ans.

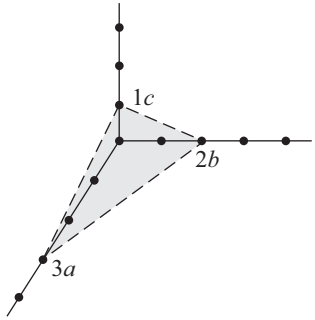


2.6 Show the planes (111) and (212) in a cubic crystal.

Ans.



2.7 Give the Miller indices for the plane shown in Fig. 2.P7.1.



Ans. (236)

2.8 Calculate the Miller indices for a plane that cuts the intercepts $a = \frac{1}{2}$, $b = 2$, and $c = \frac{1}{3}$ along the x, y, and z axes, respectively.

[*Hint:* Common denominator = 2]

Ans. (416)

2.9 A single-crystal material has a body-centered cubic structure with lattice constant $a = 5.5$ Å. Calculate the volume density of atoms in a unit cell.

[*Hint:* Each unit cell has $8 \times \frac{1}{8}$ (corner atom) + 1 (body centre atom) = 2 atoms]

Ans. 1.2×10^{22} atoms/cm^3

2.10 Calculate the surface density of atoms on the (110) plane of a silicon crystal.

[*Hint:* (110) plane has $4 \times \frac{1}{4} + 2 \times \frac{1}{2} + 2 = 4$ atoms/cell. Surface density $= \frac{4}{a(a\sqrt{2})}$ with $a = 5.43$ Å]

Ans. 9.59×10^{14} atoms/cm^2

2.11 Silicon single-crystal ingot is being grown using the Czochralski method. The desired level of P in the ingot is 10^{16} P atoms/cm^3.

(a) Calculate the concentration of P atoms in the melt assuming the segregation coefficient k_S to be 0.35.

(b) If the initial load of Si in the crucible is 6 kg. How many grams of P needs to be added? Assume the atomic weight of P to be 31.

[*Hint:* See solved numerical 2.5]

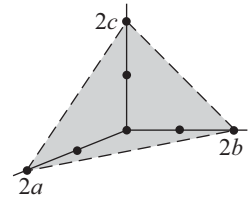
Ans. (a) 2.86×10^{16} cm^{-3}, (b) 3.79×10^{-3} g

2.12 B atoms at a concentration level of 5×10^{16} atoms/cm^3 are added to Si as a substitutional impurity. Calculate the percentage of the Si atoms that have been displaced in silicon single crystal.

[*Hint:* Volume density of Si atoms = 5×10^{22} cm^{-3}]

Ans. 1×10^{-4}%

2.13 Describe the plane shown in Fig. 2.P13.1 in terms of Miller indices.



Ans. (1,1,1)

2.14 The surface density of atoms in the (111) plane of a body-centered cubic structure is 4×10^{18} atom/m^2. Determine the lattice constant. Assume the atoms to be hard spheres, with the closest atoms touching each other.

Ans. 3.8 Å

3 Energy Bands and Charge Carriers in Semiconductors

- Formation of energy bands
- *E*-**k** diagrams
- Effective mass
- EHP formation
- Intrinsic and extrinsic semiconductors
- Fermi–Dirac distribution function
- Temperature dependence of carrier concentration
- Compensation
- Drift of charge carriers
- Lattice and impurity scattering
- High electric field effect
- Hall effect
- Diffusion of carriers
- Graded impurity concentration and induced field
- Einstein relation

Learning Objectives

After going through this chapter the student will be able to

- understand the concept of band formation and conduction and valence bands
- derive the *E*-**k** relation for a free electron and understand *E*-**k** diagrams for common semiconductors
- understand the concepts of electron, hole, and EHP formations
- understand the Fermi–Dirac distribution function and define Fermi level
- derive the expressions of equilibrium electron and hole concentrations and conductivity of a semiconductor
- define mobility and conductivity effective mass
- differentiate between lattice-scattering-based mobility and impurity-scattering-based mobility
- derive the expressions for Hall voltage and diffusion current density
- derive the Einstein relation
- solve numericals based on the above concepts

Introduction

Why do semiconductors have the electrical characteristics that allow scientists and engineers to develop such interesting devices and circuits? How do we predict the reaction of semiconductors to different stimuli such as heat, light, and injected current? The answer to these questions lies in the specific band structure of semiconductors. In this chapter we will look at the origin of bands in semiconductors and then discuss the specific band structures of some common semiconductors. Nothing interesting happens till the charge carriers in a semiconductor move. The two most common physical processes responsible for the movement of charge carriers are drift and diffusion. The salient features of both these processes are also discussed in this chapter.

3.1 Bonding Force and Formation of Energy Bands

A crystal consists of a collection of atoms. As atoms are brought together, various interactions start occurring between the neighbouring atoms. Some of these interactions and the resulting bonding forces have been discussed in Chapter 2. The system consisting of a collection of atoms is in many ways completely different from the individual atoms. One of the fundamental functions characterizing an electron in an atom is its *radial probability density*, which is defined as the probability of finding the electron at a particular distance from the nucleus. The radial probability density function for the lowest electron energy state of a single, non-interacting hydrogen atom is shown in Fig. 3.1(a). The parameter a_0 is equal to the Bohr radius. The probability density functions for two atoms in close proximity are shown in Fig. 3.1(b).

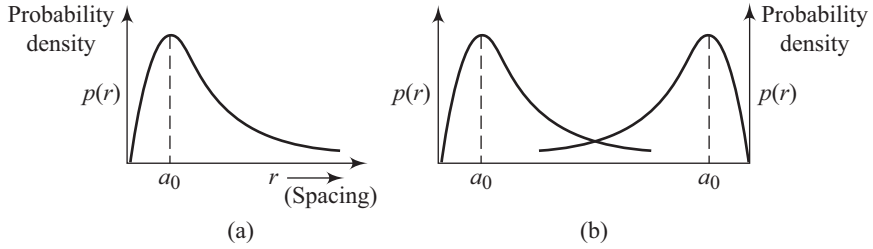


Fig. 3.1 Radial probability density function for (a) one hydrogen atom, (b) two hydrogen atoms in close proximity

One can easily infer from Fig. 3.1(b) that the probability functions of the two electrons overlap, indicating an interaction. Thus the two electrons now form one system and, therefore, cannot have the same energy level according to the Pauli's exclusion principle. The discrete quantized energy level of the individual atom therefore splits into two states as shown in Fig. 3.2.

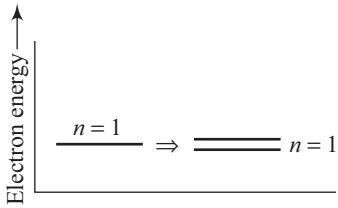


Fig. 3.2 Splitting of discrete quantized energy level

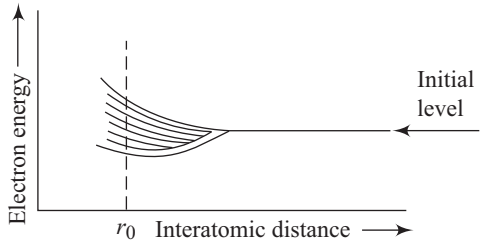


Fig. 3.3 Formation of bands of allowed energy

Suppose now we generalize the above discussion and bring a large number of hydrogen-type atoms in close proximity in a periodic arrangement. Consistent with the Pauli's exclusion principle, the individual initial quantized levels of individual atoms would now split into a band of discrete energy levels as shown schematically in Fig. 3.3.

The parameter r_0 represents the equilibrium interatomic distance in the periodic arrangement of atoms.

In this example we have considered a regular periodic arrangement of atoms having one electron. Suppose we now consider a periodic arrangement of atoms containing electrons up to the $n = 3$ energy level. When the atoms are separated by large distances, there is no interaction between individual atoms and each atom has its allowed discrete energy levels. When the atoms are brought together, the outermost electrons ($n = 3$ energy shell) start interacting. This results in the discrete energy levels splitting into a band of allowed energies. When the atoms come still closer, electrons in the $n = 2$ shell start interacting and the corresponding energy levels also go through the same fate. When the atoms get sufficiently closer, the innermost electrons ($n = 1$ shell) also interact. The process of the formation of allowed energy bands is shown in Fig. 3.4.

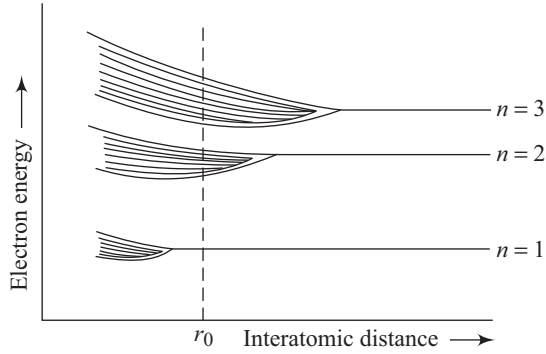


Fig. 3.4 Splitting of three energy states into allowed bands

The parameter r_0 represents the equilibrium interatomic distance. Some interesting characteristics of Fig. 3.4 are as follows:

- Discrete energy levels of individual atoms get split into allowed energy bands.
- Allowed energy bands exist corresponding to every energy level.
- Forbidden bands exist, separating any two allowed energy bands.

Let us now look at the consequences of the band theory of solids (what we have just discussed is part of a more complex subject called band theory of solids) for a semiconductor such as silicon. A silicon atom has 14 electrons, 10 of which occupy deep-lying energy levels close to the nucleus, as shown in Fig. 3.5. The remaining four valence electrons are relatively weakly bound and participate in chemical reactions. These are also shown in Fig. 3.5. The $n = 3$ energy level consists of the 3*s* and the 3*p* states. The 3*s* state corresponds to $n = 3$ and $l = 0$ and contains two quantum states per atom. At $T = 0$ K, the state contains two of the four valence electrons. The 3*p* state corresponds to $n = 3$ and $l = 1$ and contains six quantum states per atom. This state contains the remaining

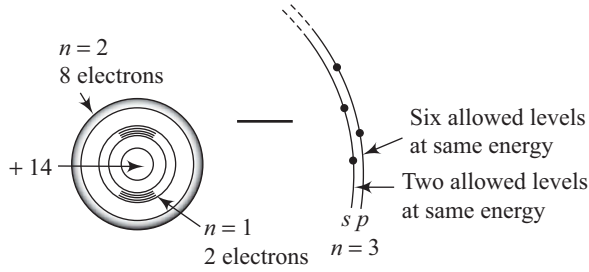


Fig. 3.5 Energy levels of an isolated silicon atom

two electrons of an individual silicon atom. Let us now take a large number *N* of silicon atoms and arrange them in a periodic manner. When we bring the atoms closer, the 3*s* and 3*p* states go through a band splitting as shown in Fig. 3.6.

There are a total of 8*N* states in 3*s*-3*p* levels. After splitting 4*N* states exist in the lower level and 4*N* states in the upper level.

At the equilibrium interatomic distance a_0, we have four quantum states per atom in the lower band and four quantum states per atom in the upper band. At absolute zero, all the valence electrons are in the lower band, and for this reason, this band is often referred to as the *valence band*. The upper band is completely empty at absolute zero and is called the *conduction band*. *The gap between the top of the valence band and the bottom of the conduction band is called the band gap energy* E_g.

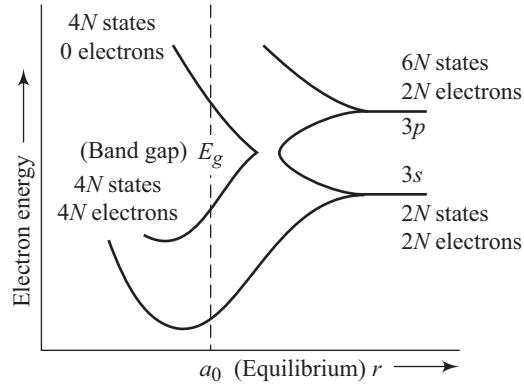


Fig. 3.6 Splitting of 3*s* and 3*p* states of silicon

Formation of energy bands in a semiconductor plays an important role in deciding the electrical characteristics of semiconductors. The concept of energy bands can also be used to differentiate between the three types of solids, namely insulators, semiconductors, and metals. Insulators have a band gap separating a filled valence band and a conduction band. The band structure of insulators is thus quite similar to that of semiconductors, as shown in Fig. 3.7. The energy band gap of insulators is however much larger than semiconductors.

Typically speaking, the energy band gap E_g of semiconductors is of the order of 1 eV, whereas insulators have energy band gaps in the range of 5–15 eV. Metals on the other hand present a completely different type of band structure. The valence band may merge into an empty band, resulting in overlapping bands. The last filled band may overlap the first empty band, resulting in a partially filled band. The high conductivity of metals such as Cu is due to the energy band structure.

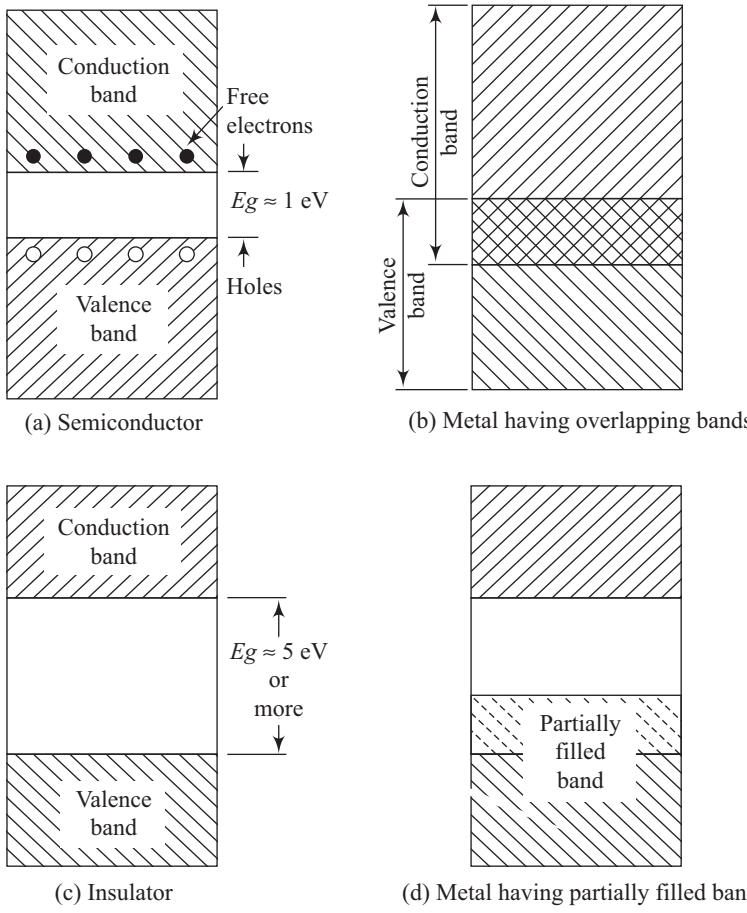


Fig. 3.7 Energy band diagrams of semiconductors, metals, and insulators

3.2 *E*-k Diagrams

The movement of a free electron along the x-axis is described by the Schrödinger equation given by

$$\frac{d^2\psi}{dx^2} + \frac{2m}{\hbar} E\psi = 0 \tag{3.1}$$

where ψ is the wave function (generally a function of both x and t), m is the mass, $\hbar = h/2\pi$, and E is the electron energy.

Since a free electron has only kinetic energy, we have

$$E = \frac{p^2}{2m} \tag{3.2}$$

Using the de Broglie formula for waves associated with the electron, we can write

$$p = \frac{h}{\lambda} = \frac{\hbar}{(\lambda/2\pi)} = \hbar\mathbf{k} \tag{3.3}$$

where **k** is called the wave vector of the electron, having a magnitude equal to $2\pi/\lambda$ and a direction coinciding with the direction of the electron wave propagation. Using p from Eqn (3.3) and putting it in Eqn (3.2) we get

$$E = \frac{\hbar^2}{2m} k^2 \tag{3.4}$$

Equation (3.4) is then the E-**k** relation for a single free electron in one-dimensional motion. The relationship between E and **k** is thus quadratic and the corresponding parabolic curve is shown in Fig. 3.8.

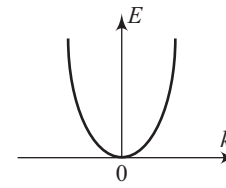


Fig. 3.8 E-**k** diagram for a single free electron

A completely different situation presents itself for an electron travelling through a perfectly periodic lattice of atoms. A few important characteristics of this situation are as follows.

(a) The probability of detecting an electron at a specified point of the crystal is a periodic function of x, since positions displaced from one another by a multiple of the lattice constant a are equiprobable for the electron.

(b) Positions of an electron inside a period a are all different.

Keeping in mind these characteristics, the space-dependent wave function for the electron turns out to be

$$\psi_{\mathbf{k}}(x) = U(\mathbf{k}_x, x) e^{i\mathbf{k}_x x} \tag{3.5}$$

The wave function of the electron is thus assumed to be in the form of a plane wave moving in the x direction with the function $U(\mathbf{k}_x, x)$ modulating the wave function according to the periodicity of the lattice. As a consequence of this the relationship between energy E and wave vector **k** undergoes a change from the

one given in Eqn (3.4). The resulting (E, **k**) diagram has to be plotted for the various crystal directions, since the periodicity of most lattices changes with the direction. In general, the relationship between E and **k** is a complex three-dimensional surface.

The typical band structure for GaAs is shown schematically in Fig. 3.9.

Its conduction band minimum and valence band maximum exist at the same **k** ($\mathbf{k} = 0$). Thus an electron can make a transition, involving smallest energy, from the conduction band to the valence band in GaAs without involving any change in the **k** value. *Such semiconductors are called direct band gap semiconductors*. The E-**k** relationship is symmetric in **k** for the one-dimensional model. Thus no useful purpose is served by giving the E-**k** diagram for both positive and negative axes. Instead, it is normal to plot the [100] direction along the normal **+k** axis and to plot the [111] portion of the diagram in a manner that **+k** points to the left. Figure 3.10 shows the E-**k** diagram for GaAs. The valence band maxima and the conduction band minima both occur at **k** = 0. The minimum conduction band energy and the maximum valence band energy thus occur at the same value of **k**.

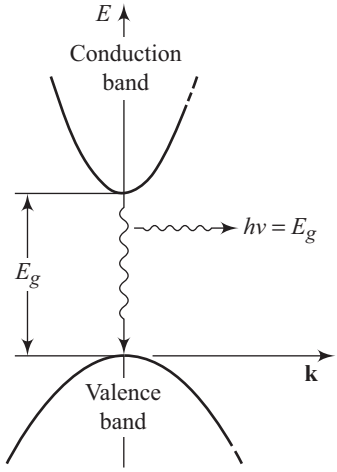


Fig. 3.9 Typical band structure of GaAs

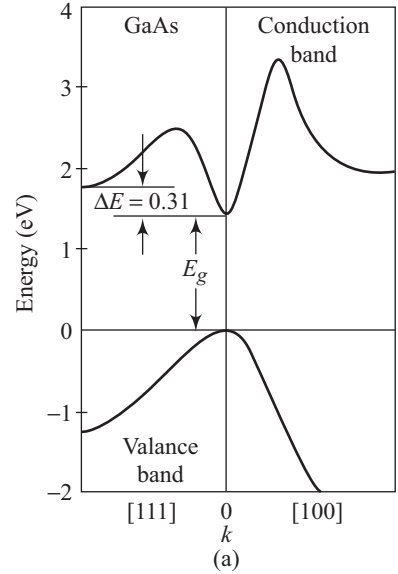


Fig. 3.10 E-**k** diagram for GaAs

The E-**k** diagram for silicon is shown in Fig. 3.11.

For silicon, the maximum valence band energy occurs at $\mathbf{k} = 0$. The minimum conduction band energy occurs at $\mathbf{k} \neq 0$ and along the [100] direction. *A semiconductor for which the maximum valence band energy and the minimum conduction band energy do not occur at the same value of* **k** *is called an indirect band gap semiconductor*.

Suppose an electron is to make a transition between the conduction band and the valence bands of a semiconductor. The law of conservation of momentum must be satisfied in such a transition. For an indirect band gap semiconductor, such a transition must necessarily include an interaction with the crystal so that the *crystal momentum stays conserved*. Examples of indirect semiconductors are Ge, GaP, and AlAs.

This interaction with the lattice can be understood by assuming a particle-like wave called *phonon*. The crystal lattice can be assumed to behave like a mechanical system of masses and springs. Phonons represent the mechanical vibration of the lattice and are quantized just like photons. Phonons are also responsible for carrying energy within the lattice. They can also give up their energy to cause transitions of charge carriers. Phonons associated with long-wavelength vibrations of the lattice are called acoustic phonons. If the frequency of vibrations is in the optical range, then such phonons are called optical phonons. Three-dimensional lattices with one atom per unit cell that include simple cubic, body-centered, or face-centered cubic lattice, have only three acoustic modes of vibration. Three-dimensional lattices with two atoms per unit cell, such as Ge, Si, and GaAs have three acoustic modes and three optical modes of vibration.

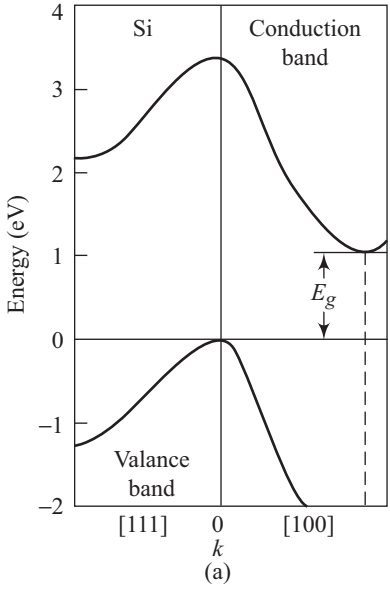


Fig. 3.11 *E*-**k** diagram for Si

Electron effective mass Mass is the property of a particle by virtue of which it resists any change in its state of motion. The relation between the momentum of a free electron and its wave vector **k** is given by the de Broglie formula,

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

From Eqn (3.4) we get

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

The second derivative of energy E leads to

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

i.e.,

$$m = \frac{\hbar^2}{(d^2E/dk^2)} \tag{3.6}$$

So, far, we have considered the electron to be free. It turns out that the expression for momentum of electrons moving in a periodic crystal is also given by

$$p = \hbar \mathbf{k} \tag{3.7}$$

Differentiating the expression for E once with respect to $\mathbf{k}$, we get

$$\frac{dE}{d\mathbf{k}} = \frac{\hbar^2}{m}\mathbf{k} \tag{3.8}$$

which leads to

$$\mathbf{k} = \frac{m}{\hbar^2}\frac{dE}{d\mathbf{k}}$$

Putting this in Eqn (3.7), we get

$$p = \frac{m}{\hbar}\frac{dE}{d\mathbf{k}} \tag{3.9}$$

which yields (for velocity v)

$$v = \frac{1}{\hbar}\frac{dE}{d\mathbf{k}} \tag{3.10}$$

Suppose an external electric field $\mathbf{E}$ is set up in the crystal. The force $\mathbf{F}$ acting on the electron is then given by

$$\mathbf{F} = -q\mathbf{E}$$

The corresponding acceleration $\mathbf{a}_1$ is then

$$a_1 = \frac{dv}{dt} = \frac{1}{\hbar}\frac{d}{dt}\left(\frac{dE}{d\mathbf{k}}\right) = \frac{1}{\hbar}\frac{d^2E}{dk^2}\frac{d\mathbf{k}}{dt} \tag{3.11}$$

Work dW done by force F in a time interval dt is

$$dW = Fvdt = \frac{F}{\hbar}\frac{dE}{d\mathbf{k}}dt$$

This work done increases the electron's energy by an amount dE. Thus,

$$dE = \frac{F}{\hbar}\frac{dE}{d\mathbf{k}}dt$$

which leads to

$$\frac{F}{\hbar} = \frac{d\mathbf{k}}{dt} \tag{3.12}$$

Putting the value of $F/\hbar$ in Eqn (3.11) results in

$$a_1 = \frac{F}{\hbar}\frac{d^2E}{dk^2} = \frac{-qE}{\hbar^2}\frac{d^2E}{dk^2} = \frac{-qE}{m_{\text{eff}}} \tag{3.13}$$

Equation (3.13) is a statement of Newton's second law. Thus, the electron acted upon by an external force moves in a periodic crystal field on the average in the same manner as a free electron would move if its mass is given by

$$m_{\text{eff}} = \frac{\hbar^2}{d^2E/dk^2} \tag{3.14}$$

The mass m_{eff} is called the *effective mass of the electron*. The effective mass may be positive or negative, many times larger or many times smaller in magnitude than the electron's rest mass, m.

3.2.1 Band Structure Modification in Semiconductors

Band structure of a naturally occurring semiconductor is determined by its composition. This band structure in turn decides the semiconductor's optical and electrical characteristics. Material scientists need to therefore modify or tailor the band structure of semiconductors as they come up with new and novel device concepts. This field of band structure modification is also called band gap engineering. There are three main techniques by which the electronic band structure of a semiconductor can be modified. These techniques are:

1. Alloying of two or more semiconductors
2. Quantum confinement or *superlattice* formation using heterostructures
3. Built-in strain developed using lattice-mismatched epitaxy

Many semiconductors form alloys over a large range of concentration. This miscibility allows alloys to be grown over a large range of band structure for specific applications. This field is also called *band gap engineering*. The band gap of the alloy formed using different semiconductors can then be used to fabricate devices with different characteristics. Group III–V compounds have a direct band gap large enough to result in a wavelength covering the visible range 750 nm to 400 nm. This is however possible using the three alloy systems $Ga_{1-x}Al_xAs$, $Ga_{1-x}In_xAs$, and $GaAs_{1-x}P_x$ for different x values. Another interesting example is the II–IV alloy $Hg_{1-x}Cd_xTe$ which has a continuous direct band gap all the way from –0.3 eV (for HgTe) to 1.6 eV (for CdTe). This material is very important for long-wavelength (Eg; ≤ 0.1 eV) device applications.

Suppose a thick layer (several hundred nanometres) of a wide-band gap material such as GaAlAs is first grown followed by a thinner (5–100 nm) layer of a narrower band gap material such as GaAs. A final layer of wide-band gap material is then grown over the GaAs layer. The resultant band diagram is shown in Fig. 3.12.

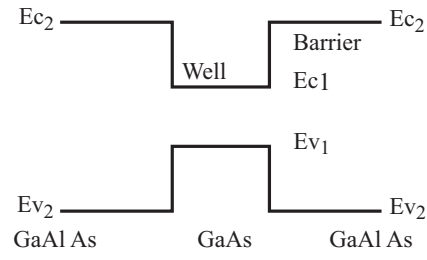


Fig. 3.12 Band diagram for GaAlAs-GaAs-GaAlAs structure

From Fig. 3.12, we can see that a one-dimensional potential well gets created in the conduction and valence band in the narrow band gap material. The electron and hole levels are then the bound states of the well, also known as *subbands*. Transitions taking place between hole and electron subbands then decide the optical absorption or emission characteristics of the quantum well. A large number of such quantum wells kept close to each other is called a *superlattice*. Carriers within the different quantum wells can then tunnel. The process results in the introduction of a new set of energy gaps.

Heterojunctions are not always formed using semiconductors that are lattice matched. If two lattice mismatched semiconductors are used to form a

heterojunction, a mismatched interface results in the formation of strained layers on either side of the interface. GaAs epitaxy on Si substrates could lead to technology resulting in the integration of III–V semiconductor optoelectronic devices and Si-based processing circuitry. There is however a 41 per cent mismatch between the lattice constants of Si and GaAs. Such a large mismatch would produce misfit dislocations. One way of getting rid of the problem is to use extremely thin (< 100 Å) strained layers that can be used as a buffer layer. For Si and GaAs, a strained superlattice such as $Ga_{1-y}In_yAs/GaAs_{1-x}P_x$, where $y = 2x$ can provide the elastic strain to accommodate the required lattice mismatch.

- Electrons close to the bottom of an energy band have a positive effective mass, whereas electrons close to the top of the energy band have a negative effective mass.
- A particle having a negative effective mass gets accelerated in a direction opposite to the acting force.

3.3 Charge Carriers in Semiconductors

Semiconductors owe their special properties to the availability of two types of charge carriers, namely electrons and holes. We will discuss some characteristics of these charge carriers in this section.

3.3.1 Electrons and Holes

Suppose a semiconductor, initially at 0 K is heated to raise its temperature. Some valence band electrons get sufficient thermal energy to cross the energy gap and reach the conduction band. This results in some electrons existing in the earlier empty conduction band and some vacant unoccupied states in the valence band. An empty state in the valence band is called a *hole*. When the hole and the conduction band electron are created due to the excitation of an electron from the valence band to the conduction band, it is called an electron–hole pair (EHP). This process is shown schematically in Fig. 3.13. The electron on reaching the conduction band finds itself surrounded by a large number of unoccupied energy states. For a sample of pure Si maintained at room temperature, there are around 10^{10} EHP/cm^3, whereas the Si atom density is 5×10^{22} atoms/cm^3. The small number of conduction band electrons are thus free to move in the conduction band owing to the availability of the large number of empty states. The charge transport process in the valence band is slightly more complicated to visualize. The wave vector **k** has been shown

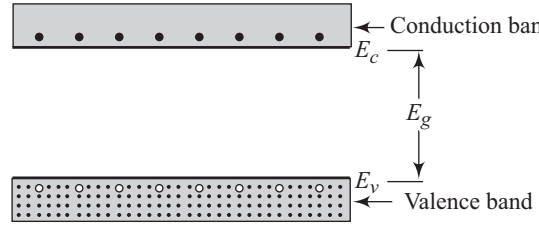


Fig. 3.13 Electron–hole pairs in a semiconductor

to be proportional to electron momentum. In the valence band, all available energy states are occupied. Figure 3.14 shows an electron i with a wave vector $\mathbf{k}_i$ and a corresponding electron i' with a wave vector $-\mathbf{k}_i$. Since the two electrons are oppositely directed, they do not give rise to any net current.

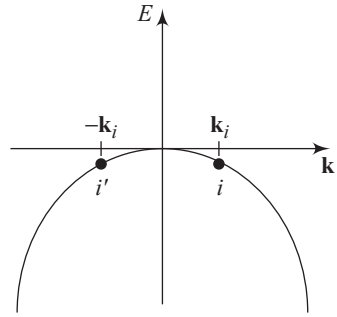


Fig. 3.14 A valence band with all states filled, including states i and i'. The ith electron with wave vector $\mathbf{k}_i$ is matched by an electron at i' with the opposite wave vector $-\mathbf{k}_i$

If the valence band has N electrons/cm^3, then for every electron moving with a particular velocity, there exists another electron whose motion is exactly oppositely directed (the velocity is equal in magnitude but opposite in direction). Thus for a filled valence band, the current density J can be written as

$$J = (-q)\sum_{i=1}^{N} V_i = 0 \tag{3.15}$$

Suppose a particular electron (say, the jth electron) is now given sufficient energy such that it gets excited to the conduction band. The jth electron does not now contribute to the current density, which becomes

$$J = (-q)\sum_{i=1}^{N} V_i - (-q)V_j \tag{3.16}$$

The first term on the RHS is zero from Eqn (3.15). This leads to

$$J = qV_j$$

The current contribution is thus equivalent to that of a positively charged particle of velocity $+V_j$. For mathematical simplicity, *the vacant states in the valence band can therefore be treated as charge carriers with positive charge and positive mass*. One must, however, remember that the contribution to the current density actually arises due to the motion of the uncompensated electron j'.

Current flow in a semiconductor can be accounted for by considering the motion of electrons and holes. We should however remember one important point. *The hole energy increases oppositely to the electron energy because the two carriers have opposite charge.* Thus in Fig. 3.14, the hole energy increases downwards, whereas in the conduction band, the electron energy increases as we go upwards. *Holes thus seek the lowest energy states available at the top of the valence band. Due to the same reason, conduction electrons are generally found at the bottom of the conduction band.*

For routine semiconductor device analysis the E-$\mathbf{k}$ diagrams are not normally used. In such an analysis we use simplified band diagrams, which are plots of the conduction and valence band edges as a function of position. Since the bottom

of the conduction band corresponds to zero electron velocity and the top of the valence band corresponds to zero hole velocity, the simplified band diagram is in effect a plot of the potential energy as a function of position. A simplified band diagram of a typical semiconductor is sketched in Fig. 3.15.

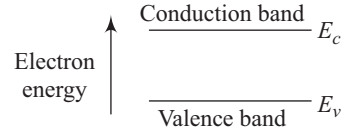


Fig. 3.15 Simplified band diagram

3.3.2 Intrinsic Semiconductor

A perfect, pure semiconductor crystal containing no impurities or lattice defects is called an intrinsic semiconductor. At 0 K, the valence band of such a semiconductor is filled with electrons and the conduction band is empty. Such a material therefore has no charge carriers at 0 K. As the temperature is increased, the valence band electrons acquire energy and get thermally excited across the band gap and start moving to the conduction band. Electron–hole pairs are created, and these EHPs are the only charge carriers in the intrinsic semiconductor. The physical mechanism underlying EHP creation can be understood with the help of the broken bond model of the Si crystal, as shown in Fig. 3.16.

At 0 K, all the covalent bonds are intact. As the temperature increases, some valence electrons acquire sufficient energy to break away from their position in the bonding structure and become free to move about in the lattice. This is how a conduction band electron is created and a hole (broken bond) is left behind in the valence band. *The energy required to break free from a covalent bond is the band gap energy* E_g. One must however emphasize at this stage that the above picture is deceptively simplified and at best qualitative. This model seems to indicate that the free electron and the hole created during the process of EHP generation are localized in the lattice site. The actual picture is not this since the electron and hole are spread out over several lattice spacings.

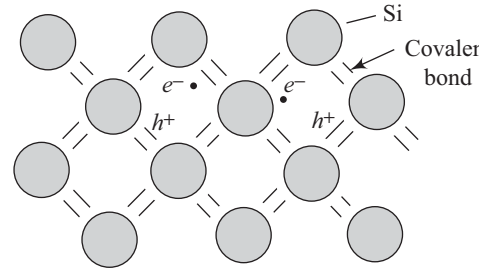


Fig. 3.16 Electron–hole pairs in the covalent bonding model of the Si crystal

All the charge carriers created in an intrinsic semiconductor are due to the process of EHP production. The concentration of electrons n (electrons per cm^3) in the conduction band is therefore equal to the concentration of holes in the valence band p (holes per cm^3). A symbol n_i is used to represent each of these intrinsic carrier concentrations. Thus, for an intrinsic semiconductor,

$$n = p = n_i \tag{3.17}$$

At a given temperature, two competing processes decide the equilibrium concentration of electron–hole pairs. One of them is the generation of EHP through the process of transition of a valence electron to the conduction band.

Annihilation of these pairs takes place through a process, called *recombination*. Recombination occurs when a conduction band electron makes a transition to an empty state (hole) in the valence band. Such transitions can be direct or indirect. Suppose g_i is the generation rate of EHPs (EHP/cm^{3}s) and r_i is the recombination rate. Then at equilibrium, we must have

$$r_i = g_i \tag{3.18}$$

Both g_i and r_i are functions of temperature. If n_0 and p_0 represent the equilibrium concentrations of electrons and holes, respectively, then we can write

$$r_i = \alpha_r n_0 p_0 = \alpha_r n_i^2 = g_i \tag{3.19}$$

where α_r is a constant of proportionality and is dependent upon the particular recombination process.

3.3.3 Extrinsic Semiconductor

We began our discussion on semiconductors by emphasizing the importance of the introduction of controlled amounts of known impurities. This process is called *doping*. *An extrinsic semiconductor is one which has been doped such that the equilibrium carrier concentrations n_0 and p_0 become different from the intrinsic carrier concentration n_i.* There are two types of doped semiconductors, *n*-type and *p*-type. Electrons are the majority charge carriers in the *n*-type semiconductor and holes are the majority charge carriers in the *p*-type semiconductor.

Let us now look at the process of doping using the band diagram picture. Introduction of impurities or lattice defects in a perfect crystal leads to the creation of additional level in the energy band structure, generally within the band gap. If we introduce an impurity from column V of the periodic table (P, As, and Sb), an energy level is created near the conduction band of semiconductors such as Ge or Si. This energy level is filled with electrons at 0 K; but as depicted in Fig. 3.17, very little thermal energy is required to excite these electrons to the conduction band. At temperatures in the region of 50–100 K, practically all the electrons in these additional levels get excited to the conduction band. Such an impurity level is called *donor* level and the corresponding impurities are called donor impurities, since these levels donate electrons to the conduction band. The donor level can

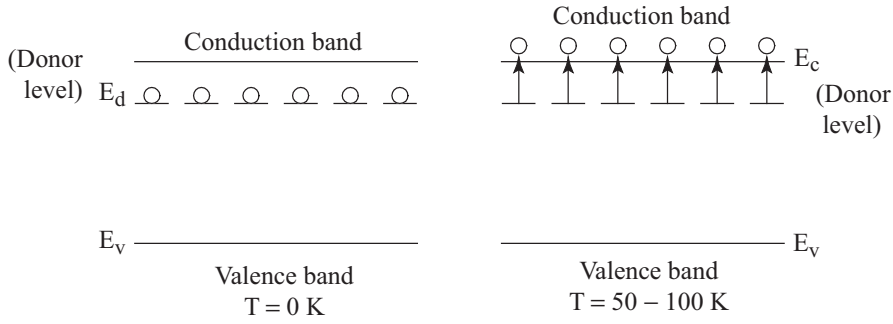


Fig. 3.17 Donation of electrons from donor level

contribute significant concentration of electrons to the conduction band even at temperatures that are low for the appreciable intrinsic EHP concentration. For a semiconductor doped with a considerable concentration of donor atoms, we have at room temperature

$$n_0 \gg (n_i, p_0) \tag{3.20}$$

Such a doped semiconductor is thus an *n*-type semiconductor.

Introduction of impurity atoms from column III (B, Al, Ga, and In) results in the creation of impurity levels near the valence band of semiconductors such as Ge or Si. This is shown schematically in Fig. 3.18. These energy levels are empty at 0 K. At other temperatures (which are still low enough), the available thermal energy is sufficient to excite electrons from the valence band to occupy these levels. These additional levels are called *acceptor* levels and the corresponding impurities are called acceptor impurities since they accept electrons from the valence band. Acceptor impurities can result in an equilibrium hole concentration p_0 much greater than the conduction band electron concentration, n_0. Thus

$$p_0 \gg (n_i, n_0) \tag{3.21}$$

Such an extrinsic semiconductor is called a *p*-type semiconductor.

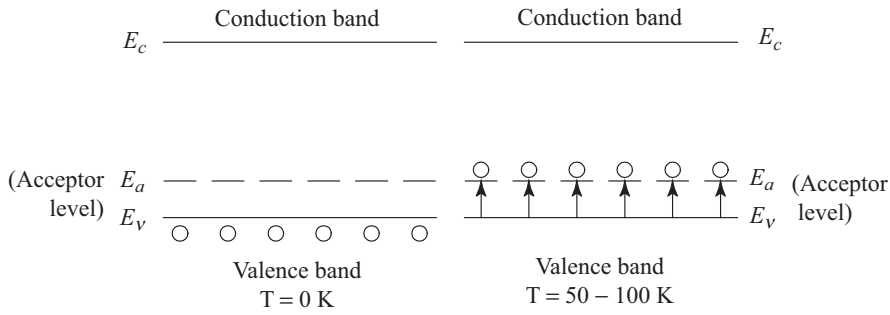


Fig. 3.18 Acceptance of valence band electrons by acceptor level

Figure 3.19 is a schematic representation of the introduction of a donor and an acceptor atom in the Si lattice. The dopant atoms can be seen to occupy substitutional sites. If the dopant atoms do not occupy substitutional sites, they do not behave as donor or acceptor impurities.

The As atom in the Si lattice has an extra electron sharing four of its valence electrons with the neighbouring Si atoms. The fifth electron is loosely bound to the As atom since it does not fit into the bonding structure of the lattice. At 0 K, there is no thermal energy available; but at low enough temperatures, sufficient thermal energy is available to this extra electron to overcome its coulombic binding to the impurity As atom. This extra electron thus gets donated to the crystal and takes part in current conduction. The column III impurity B atom has only three valence electrons. These three electrons participate in covalent

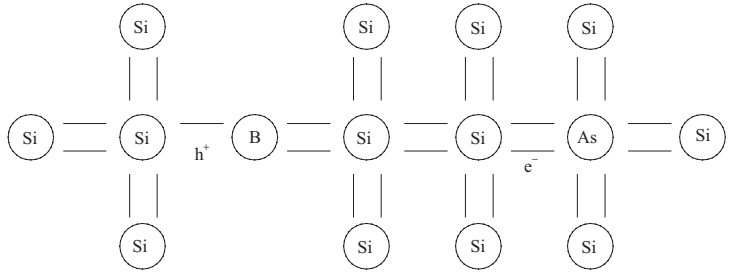


Fig. 3.19 Donor and acceptor atoms in the covalent bond model

bonding with neighbouring Si atoms, but one bond stays incomplete. At low enough temperatures, an electron from an adjacent neighbouring atom *hops* to the incomplete bond at the *B* site, creating an incomplete bond at its original location.

The energy required to excite the fifth electron of a donor atom to the conduction band is called its *binding energy*. For the purpose of an approximate calculation of the binding energy, let us assume that the four covalent bonding electrons are tightly bound and the fifth extra electron is loosely bound to the atom. This situation can be approximated to the case of the loosely bound electron around the tightly bound *core* electrons in a hydrogen-like orbit. The total energy of the electron in the *n*th orbit in accordance with the *Bohr model* is given by

$$E_n = \frac{mq^4}{2K^2n^2\hbar^2}$$

For the ground-state energy, we need to put $n = 1$ in the above equation to yield

$$E = \frac{mq^4}{2K^2\hbar^2} \tag{3.22}$$

The constant K is given by

$$K = 4\pi\varepsilon_0\varepsilon_r \tag{3.23}$$

where ε_r is the relative permittivity of the semiconductor material. The mass m in Eqn (3.22) must be replaced with the conductivity effective mass m_n^* for the semiconductor, where the conductivity effective mass is the harmonic mean of the band curvature effective masses for the band structure of the particular semiconductor.

In Si, the donor and acceptor energy levels lie between 0.03–0.06 eV from the band edges. For Ge, the corresponding donor and acceptor levels are in the region of 0.01 eV from the band edges.

The situation for group III–V semiconductors is different. Group VI impurities such as S, Se, and Te act as donors in GaAs. These impurities substitute the group V atoms As and thereby provide an extra electron. Similarly, group II impurities such as Be, Zn, and Cd substitute for the column III atom Ga in GaAs and form acceptors. An interesting case is when a group III–V compound is doped with a

group IV impurity. If the impurity resides on the group III sublattice of the crystal, it serves as a donor. On the other hand, if it resides on the group V sublattice, it serves as an acceptor. Such impurities are called *amphoteric impurities*. Si or Ge are some examples of amphoteric impurities for group III–V compounds. In GaAs, Si usually occupies Ga sites and thus serves as a donor. If, however, an excess of As vacancies arise due to processing or during growth, Si impurity atoms can occupy As sites, thereby acting as acceptors.

With suitable doping, a semiconductor can become *n*-type or *p*-type. In the *n*-type semiconductor, the conduction band electrons far outnumber the holes. Electrons are therefore called *majority carriers* in the *n*-type semiconductor, whereas holes are called minority carriers in the *n*-type semiconductor. The corresponding majority carriers in the *p*-type semiconductor are holes and minority carriers are electrons.

3.4 Carrier Concentrations in Semiconductors

In the preceding section, we learnt about doping of semiconductors. For standard doping impurities, one majority carrier is obtained from each impurity atom. Thus, for heavily doped semiconductors, the majority carrier concentration can be easily known from the known doping level. We still need to, however, calculate the minority carrier concentration. Carrier concentrations in semiconductors are also temperature dependent. These and many other important characteristics of semiconductors can be obtained by studying the distribution of carriers amongst the available energy states.

3.4.1 Fermi Level

Electrons existing in solids obey Fermi–Dirac statistics. The distribution of electrons over the range of allowed energy levels at thermal equilibrium is given by the *Fermi–Dirac distribution function*, $f(E)$, which is

$$f(E) = \frac{1}{1 + e^{(E - E_F)/kT}} \tag{3.24}$$

where k is Boltzmann's constant ($k = 8.62 \times 10^{-5}$ eV/K $= 1.38 \times 10^{-23}$ J/K). The Fermi–Dirac distribution function, $f(E)$, gives the probability that an available energy state at E will be occupied by an electron at absolute temperature T. The quantity E_F in the function $f(E)$ is called the *Fermi level*. To gain more insight into the physical meaning of E_F, let us put $E = E_F$ in Eqn (3.24). We get

$$f(E_F) = [1 + e^{(E_F - E_F)/kT}]^{-1}$$

implying

$$f(E_F) = \frac{1}{1+1} = \frac{1}{2} \tag{3.25}$$

Thus *the Fermi level E_F represents that energy which has a probability 1/2 of being occupied by an electron*. A plot of $f(E)$ for some temperatures is shown in Fig. (3.20).

A few salient features of the function $f(E)$ are the following:

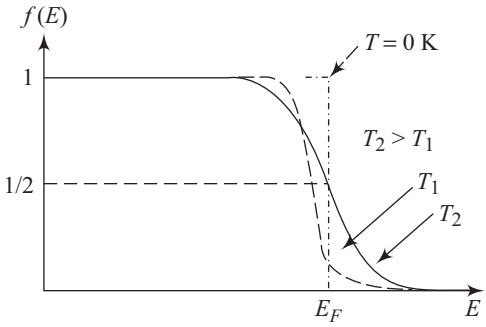


Fig. 3.20 The Fermi–Dirac distribution function for different temperatures

(i) At 0 K, the distribution function $f(E)$ takes a simple rectangular shape. This is because at

$$T = 0 \text{ K},$$

$$f(E) = \frac{1}{(1+0)} = 1 \text{ for } E < E_F$$

and

$$f(E) = \frac{1}{(1+\infty)} = 0 \text{ for } E > E_F$$

The rectangular shape implies that at $T = 0$ K, all states up to E_F are filled with electrons and states above E_F are empty.

(ii) At temperatures greater than 0 K, some probability exists for states above the Fermi level to get filled with electrons. At any temperature, T_1, there is a probability $f(E)$ that states above E_F are filled. A corresponding probability $[1 - f(E)]$ exists that states below E_F are empty.

(iii) The Fermi function is symmetrical about E_F for all temperatures. Thus the probability $f(E_F + \Delta E)$ that a state ΔE above E_F is filled is equal to the probability $[1 - f(E_F - \Delta E)]$ that a state ΔE below E_F is empty.

(iv) Owing to the symmetry of the function $f(E)$ about the Fermi level E_F, the Fermi level becomes a natural choice as a reference point for electron and hole concentrations in semiconductors.

In applying the distribution function $f(E)$ to semiconductors, it must be borne in mind that the existence of a finite value of $f(E)$ does not always ensure an electron's presence at that energy level. The distribution function can have a finite value in the band gap of a semiconductor, but there is no possibility of an electron existing within the band gap. Figure (3.21) shows the Fermi distribution function turned on its side. A little reflection would help us appreciate the relevance of this figure for semiconductors.

For intrinsic semiconductor, the concentration of holes in the valence band is equal to the concentration of electrons in the conduction band. The Fermi level E_F must therefore lie at the middle of the band gap in an intrinsic semiconductor. The function $f(E)$ is symmetric about E_F as already discussed, therefore the electron probability *tail* of $f(E)$ extending into the conduction band [Fig. 3.20(a)] is symmetrical with the hole probability tail $[1 - f(E)]$ in the valence band. Within the band gap, although the distribution function has values, there are no energy

states available. Thus the finite $f(E)$ in the band gap does not result in any actual electron occupancy. Let us now look at the actual magnitudes of the probability of occupancy $f(E)$. For Si at 300 K, $n_i = p_i \cong 10^{10}$ cm^{-3}. The densities of available states at E_v and E_c are of the order of 10^{19} cm^{-3}. Thus the probability of occupancy $f(E)$ for electron occupancy for an individual state in the conduction band is very small. Similarly, the probability of occupancy of a valence band state by a hole $[1 - f(E)]$ is also a very small quantity. Small changes in $f(E)$ can lead to significant changes in the free carrier concentration.

An *n*-type semiconductor has a high concentration of electrons in the conduction band in comparison to the concentration of hole in the valence band. The distribution function $f(E)$ therefore lies above its intrinsic position on the energy scale, as shown in Fig. 3.21(b). The distribution function $f(E)$ retains its shape at any particular temperature. A larger concentration of electrons at E_c in the *n*-type semiconductor leads to a correspondingly smaller hole concentration at E_v. Another interesting observation can be made from Fig. 3.21(b). As E_F moves closer to E_c the value of $f(E)$ for each energy level in the conduction band increases. This also implies an increase in the total electron concentration n_0, as E_F moves towards E_c.

The corresponding situation for the *p*-type semiconductor is shown in Fig. 3.21(c). Here the tail $[1 - f(E)]$ below E_v is larger than the tail of $f(E)$ above E_c.

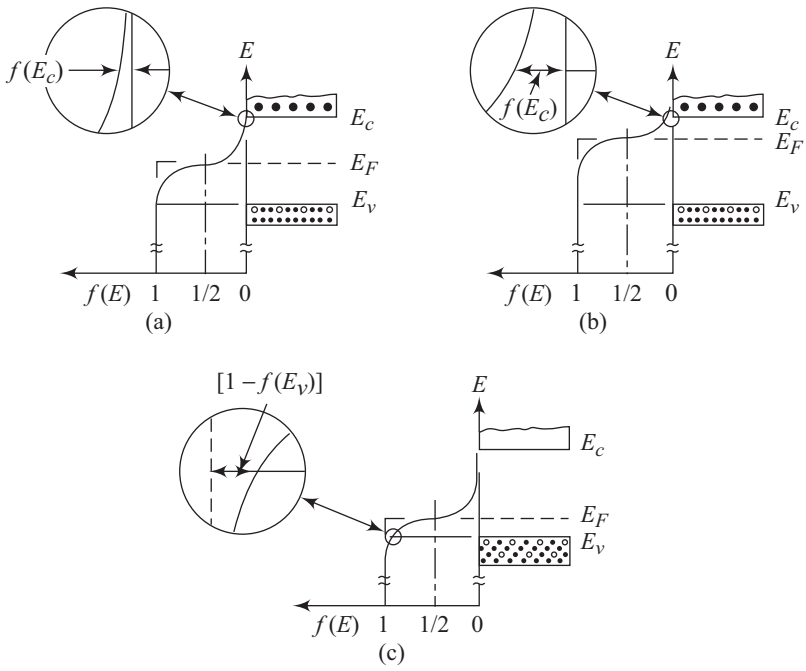


Fig. 3.21 The Fermi distribution function for (a) intrinsic material (b) *n*-type material, (c) *p*-type material

The concentration of holes in the *p*-type material depends upon the difference $(E_F - E_v)$. Thus the position of E_F in a band diagram for a particular temperature is indicative of the electron and hole concentrations.

3.4.2 Equilibrium Electron and Hole Concentrations

To evaluate the electron and hole concentrations in a semiconductor, we need to know the density of states in a particular energy range. Let $N(E)dE$ represent the density of states (in cm^{-3}) in the energy range dE. The electron concentration n_0 in the conduction band at equilibrium is then given by

$$n_0 = \int_{E_c}^{\infty} f(E)N(E)dE \tag{3.26}$$

The function $N(E)$ can be calculated using standard techniques of quantum mechanics and the Pauli's exclusion principle, and can be shown to be given by the equation

$$N(E) = \frac{\sqrt{2}}{\pi^2}\left(\frac{m^*}{\hbar^2}\right)^{3/2} E^{1/2} \tag{3.27}$$

where m^* is the effective mass given by Eqn (3.14). Since $N(E)$ is proportional to $E^{1/2}$, the density of states in the conduction band increases with electron energy. The Fermi function $f(E)$ however, as we have seen, has extremely small values for large energies. The product $f(E)N(E)$ in Eqn (3.26) thus decreases rapidly above the conduction band edge E_c. This is the reason why very few electrons are at energy levels far above the conduction band edge. A similar effect is also observed for holes in the valence band. The probability of finding a hole in the valence band $[1 - f(E)]$ decreases sharply below E_v, and therefore most holes exist near the top of the valence band. Figure 3.22 shows the band diagram, density of states, Fermi–Dirac distribution function, and the carrier concentrations for intrinsic, *n*-type, and *p*-type semiconductors at thermal equilibrium.

We can replace the density of states $N(E)$ in Eqn (3.26) with an *effective density of states* N_c located at the conduction band edge E_c. The conduction band electron concentration n_0 can therefore be written as

$$n_0 = N_c f(E_c) \tag{3.28}$$

where $f(E_c)$ is the probability of occupancy at E_c. The Fermi function $f(E_c)$ is given by

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

If the Fermi level E_F is assumed to lie several kT below the conduction band, then the above equation can be simplified as

$$f(E_c) = \frac{1}{1 + e^{(E_c - E_F)/kT}} \simeq e^{-(E_c - E_F)/kT} \tag{3.29}$$

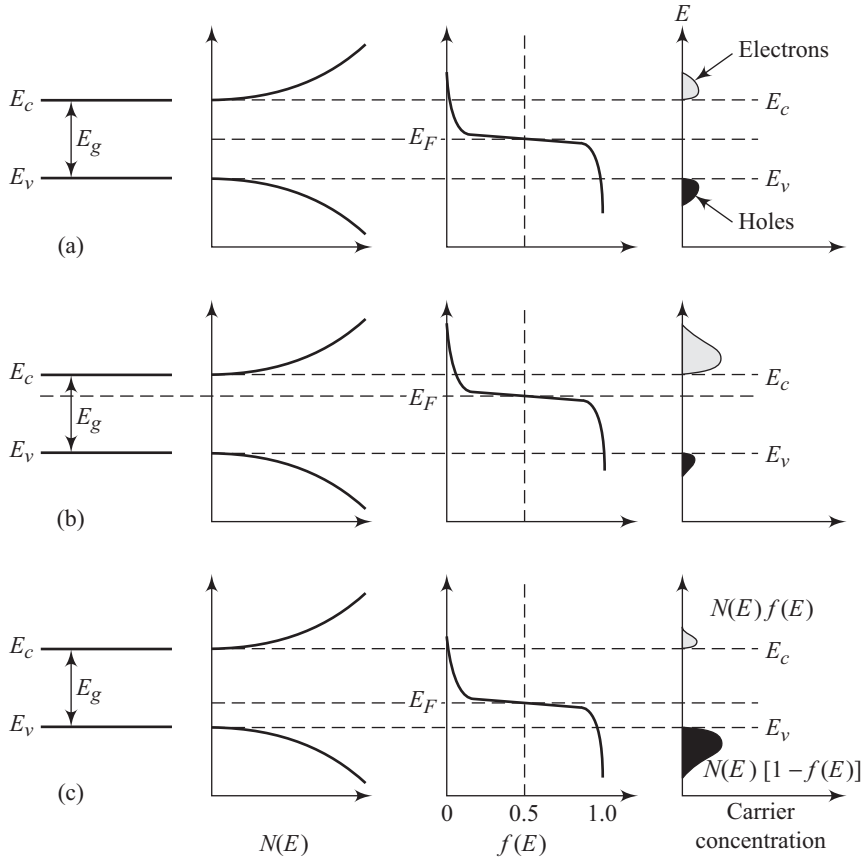


Fig. 3.22 Thermal equilibrium band diagram, density of states, Fermi–Dirac distribution, and the carrier concentrations for (a) intrinsic, (b) *n*-type, and (c) *p*-type semiconductors at thermal equilibrium

Using Eqn (3.29) in Eqn (3.28), we get

$$n_0 = N_c e^{-(E_c - E_F)/kT} \tag{3.30}$$

The effective density of states N_c can be shown to be given by

$$N_c = 2\left(\frac{2\pi m_n^* kT}{h^2}\right)^{3/2} \tag{3.31}$$

where m_n^* is the density of state's effective mass of electrons. This effective mass is different from the band curvature effective mass m^*. This is because in any particular direction in a crystal there are often more than one equivalent conduction band minimas. For Si, for example, there are six equivalent conduction

band minimas along the *x*-direction. Thus we have more than one band curvature to deal with in arriving at the effective mass. There is a longitudinal effective mass along the major axis of the ellipsoid (see Fig. 3.23) and the transverse effective mass m_t along the two minor axes.

By using dimensional equivalence and adding contributions from all six valleys, we get

$$(m_n^*)^{3/2} = 6(m_l m_t^2)^{1/2} \tag{3.32}$$

Thus the density of states effective mass is the geometric mean of the effective masses.

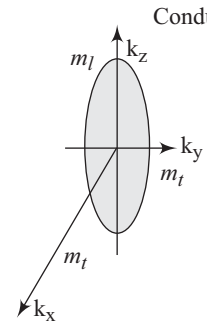


Fig. 3.23 Ellipsoidal constant energy surface for Si, near the 6 conduction band minima along the *x* direction

Similarly, the concentration of holes in the valence band, at equilibrium, is given by

$$p_0 = N_v[1 - f(E_v)] \tag{3.33}$$

where N_v is the effective density of states in the valence band. The probability of finding an empty state (hole) at E_v is given by

$$1 - f(E_v) = 1 - \frac{1}{1 + e^{(E_v - E_F)/kT}} \simeq e^{\frac{-(E_F - E_v)}{kT}} \tag{3.34}$$

if E_F is several kT larger than E_v. Using Eqn (3.34) in Eqn (3.33) leads to

$$p_0 = N_v\, e^{-(E_F - E_v)/kT} \tag{3.35}$$

The effective density of states in the valence band reduced to its band edge is given by

$$N_v = 2\left(\frac{2\pi m_p^* kT}{h^2}\right)^{3/2} \tag{3.36}$$

From Eqn (3.35) it is clear that the hole concentration increases as E_F moves closer to the valence band.

For an intrinsic semiconductor, the equilibrium electron and hole concentration are represented by n_i and p_i, respectively. Using Eqs (3.30) and (3.35), we get

$$n_i = N_c e^{-(E_c - E_i)/kT},\ p_i = N_v e^{-(E_i - E_v)/kT} \tag{3.37}$$

where E_i is the position of the Fermi level E_F for an intrinsic semiconductor. E_i lies near the middle of the bandgap, as shown in Fig. 3.21(a). Multiplying n_i and p_i, we have

$$n_i p_i = (N_c e^{-(E_c - E_i)/kT})(N_v e^{-(E_i - E_v)/kT}) = N_c N_v e^{-(E_c - E_v)/kT}$$

which leads to

$$n_i p_i = N_c N_v e^{-E_g/kT} \tag{3.38}$$

The equilibrium concentrations n_0 and p_0 are in general given by Eqs (3.30) and (3.35). Taking the product of n_0 and p_0 results in

$$n_0 p_0 = (N_c e^{-(E_c - E_F)/kT})(N_v e^{-(E_F - E_v)/kT})$$

which yields

$$n_0 p_0 = N_c N_v e^{-(E_c - E_v)/kT} = N_c N_v e^{-E_g/kT} \tag{3.39}$$

Thus, the product of electron and hole concentrations at equilibrium is a constant for a particular material and temperature.

For an intrinsic semiconductor, $n_i = p_i$, implying

$$n_0 p_0 = n_i^2 \tag{3.40}$$

Using Eqs (3.40) and (3.39) results in

$$n_i = \sqrt{N_c N_v}\, e^{-E_g/2kT} \tag{3.41}$$

From Eqs (3.41) and (3.37) we can conclude that $E_c - E_i = E_g/2$ if the effective densities of states, N_c and N_v, are equal. In general, however, N_c and N_v are slightly different due to some difference between the effective mass for electrons and holes. The displacement of E_i from the middle of the band gap is more for GaAs than for Si or Ge.

From Eqn (3.30), we get

$$n_0 = N_c e^{-(E_c - E_F)/kT} = \left(N_c e^{-(E_c - E_i)/kT}\right) e^{-(E_F - E_i)/kT}$$

which on using Eqn (3.37) results in

$$n_0 = n_i e^{(E_F - E_i)/kT} \tag{3.42}$$

Similarly, we can write

$$p_0 = n_i e^{(E_i - E_F)/kT} \tag{3.43}$$

Equations (3.42) and (3.43) lead to the following two important conclusions:

(i) $n_0 = p_0 = n_i$ when $E_F = E_i$.

(ii) The equilibrium electron concentration n_0 increases exponentially as the Fermi level moves away from E_i towards the conduction band. The equilibrium hole concentration p_0 increases exponentially as the E_F moves away from E_i towards the valence band.

3.4.3 Temperature Dependence of Carrier Concentrations

The equilibrium electron and hole concentrations n_0 and p_0 are dependent upon temperature as given by Eqs (3.42) and (3.43). The dependence of n_i on temperature is given by Eqn (3.41). The Fermi level E_F is also a function of temperature. From Eqn (3.41), we can write for n_i,

$$n_i = \sqrt{N_c N_v}\; e^{-E_g/2kT}$$

Putting the expressions for N_c and N_v from Eqs (3.31) and (3.36) in the above Eqn for n_i leads to

$$n_i = 2\left(\frac{2\pi kT}{h^2}\right)^{3/2} (m_n^* m_p^*)^{3/4} e^{-E_g/2kT} \tag{3.44}$$

The temperature dependence of n_i is dominated by the exponential term. The variation of n_i with temperature, neglecting the $T^{3/2}$ term and the slight variation in E_g with temperature, is shown in Fig. 3.24.

For a given temperature T, n_i can be determined from Fig. 3.24. If the Fermi level E_F position with respect to E_i is known, then the carrier concentrations can be determined by using Eqs (3.42) and (3.43).

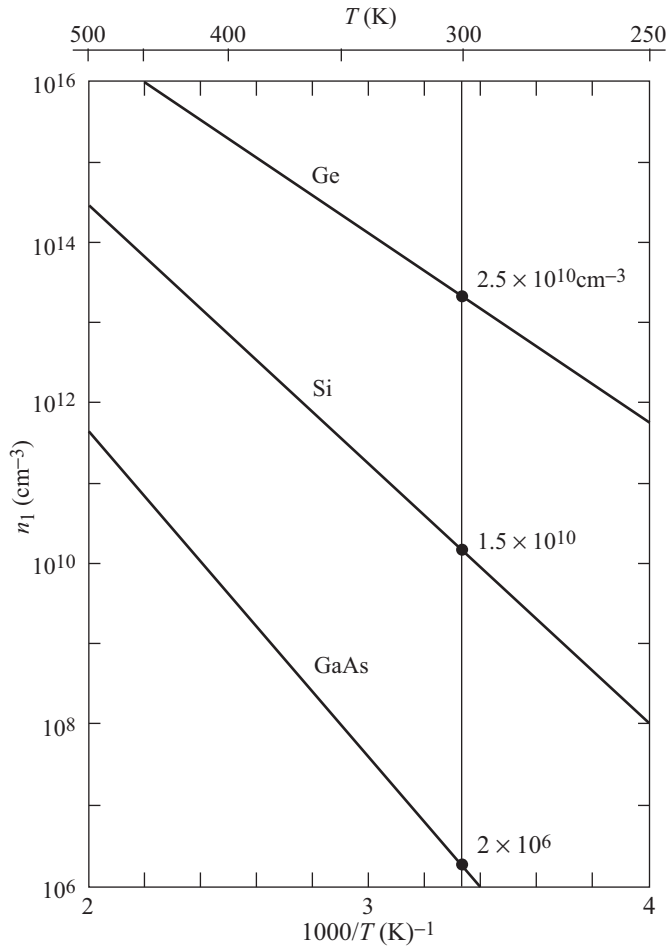


Fig. 3.24 Intrinsic carrier concentration for Ge, Si, and GaAs as a function of inverse temperature

The temperature dependence of an *n*-type Si with a donor concentration N_d of 10^{15} cm^{-3} is shown schematically in Fig. 3.25.

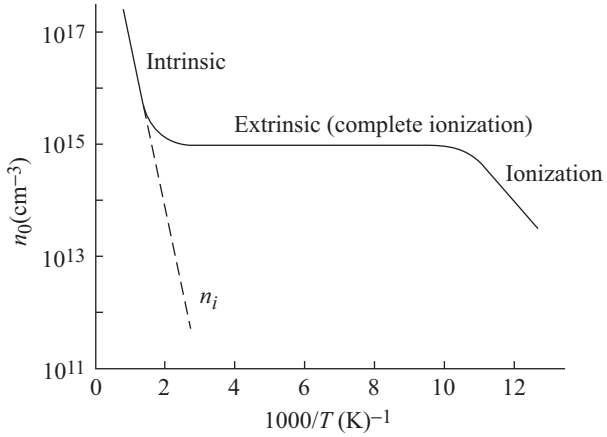


Fig. 3.25 Carrier concentration vs inverse temperature for doped Si ($N_d = 10^{15}$ cm^{-3})

The curve in Fig. 3.25 has the following characteristics:

(i) At very low temperatures (high $1/T$ values), negligible EHP exist. Due to low thermal energy, the donor electrons are bound to donor atoms.

(ii) As the temperature increases, electrons are donated to the conduction band. This process is called *ionization*.

(iii) At around 100 K, i.e., $1000/T = 10$, all donor atoms are ionized. The region of temperature where the donor atoms contribute electrons to the conduction band is called the *ionization region*.

(iv) $n_0 \simeq N_d = 10^{15}$ cm^{-3} when all donor atoms are ionized.

(v) The carrier concentration is constant in the extrinsic region till n_i starts becoming comparable to N_d.

(vi) At higher temperatures the carrier concentration is dominated by n_i, which increases with temperature.

(vii) Semiconductor devices generally use doping levels that ensure an extrinsic region beyond the highest temperature of operation.

3.4.4 Compensation

So far, in our discussions involving extrinsic semiconductors, we have considered either the donor type or the acceptor type doping. Very often a semiconductor may contain both donors and acceptors. In such a situation, two possibilities arise. One of the dopant types may be in excess of the other type or the two dopant types may be comparable. Figure 3.26 is a schematic diagram of a band of a semiconductor with both donor and acceptor type impurities with $N_d > N_a$.

Since the material is having a larger number of donors, it behaves as an *n*-type semiconductor. The Fermi level E_F is therefore more towards the conduction band. The acceptor states are filled with valence band electrons, and this results in the creation of holes in the valence band. The holes are in turn filled with conduction band electrons by the process of *recombination*. Thus the net concentration of electrons in the conduction band is $N_d - N_a$. This process of partial neutralization of one type of dopant by another type of dopant is called *compensation*. Suppose we start with an *n*-type semiconductor and dope it with acceptor impurities till $N_d = N_a$, the effective density of donated electrons becomes zero. In such a compensated semiconductor, $n_0 = p_0 = n_i$.

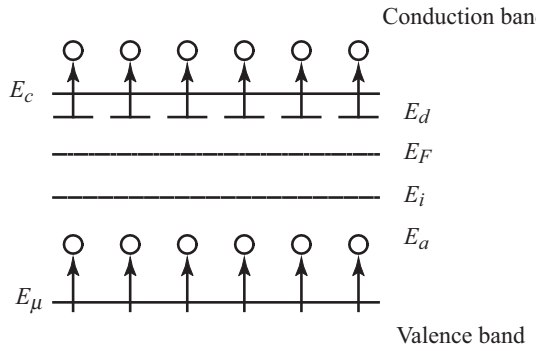


Fig. 3.26 Compensation effect in *n*-type semiconductor ($N_d > N_a$)

Any semiconductor material is electrostatically neutral due to the requirement of *space charge neutrality*. Thus the sum total of the positive charges (holes and ionized donor atoms) must equal the sum of the negative charges (electrons and ionized acceptor atoms). This implies

$$p_0 + N_d^+ = n_0 + N_a^- \tag{3.45}$$

For a compensated semiconductor with $N_d > N_a$, we have

$$n_0 = p_0 + (N_d^+ - N_a^-) \tag{3.46}$$

If the donor concentration is far greater than the acceptor concentration, we can assume

$$n_0 \gg p_0$$

which leads to, using Eqn (3.46),

$$n_0 = (N_d^+ - N_a^-) \tag{3.47}$$

3.5 Carrier Drift

Let us now try and understand what happens when an electric field is applied across a semiconductor. The charge carriers present in the semiconductor would experience a force due to this field and this would lead to a net acceleration. This net movement of charge carriers due to the presence of electric field is called *drift*. Charge carriers in a semiconductor are generally scattered by lattice vibrations, impurities, other electrons, and defects. An individual electron therefore goes through a random motion as shown in Fig. 3.27.

Although the electron has a very small probability of returning to its starting point after any finite time t, any large enough group of electrons does not undergo any net motion over any period of time. Thus, there is no preferred direction of motion for a group of electrons and hence no current flows. If an electric field is applied, the situation undergoes a change. Individual electrons do not undergo an appreciable change in their random motion, but when the motion of a group of electrons is averaged, a net motion does take place in the direction opposite to the electric field direction. This is shown schematically in Fig. 3.28.

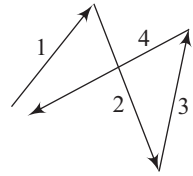


Fig. 3.27 Typical random behaviour of an electron in a semiconductor without applied field

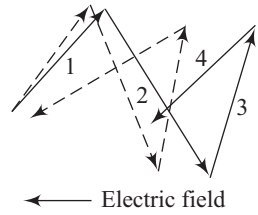


Fig. 3.28 Typical random behaviour of an electron in a semiconductor with applied electric field

3.5.1 Mobility and Conductivity

Suppose p_x is the x component of the total momentum of a group of n electrons present per cm^3. If E_x is the applied electric field in the x-direction, then each of the n electrons experiences a force $-qE_x$ due to the electric field. For the group, we can write

$$-nqE_x = \left.\frac{dp_x}{dt}\right|_{\text{Electric field}} \tag{3.48}$$

Electrons are at the same time decelerated due to the collision processes. In steady state the acceleration due to the electric field is exactly balanced by the deceleration due to collisions. The net rate of change of momentum is therefore zero under steady state current flow.

For random collision, there will be a constant probability of collision at any time for each electron. Let us assume that the group consists of N_0 electrons at time $t = 0$ and let $N(t)$ represent the number of electrons that have not suffered a collision in time t. The rate of decrease of $N(t)$ is proportional to $N(t)$, i.e.,

$$-\frac{dN(t)}{dt} \propto N(t)$$

which implies

$$-\frac{dN(t)}{dt} = \frac{N(t)}{\bar{t}_f} \tag{3.49}$$

where $\bar{t}_f$, the constant of proportionality, is called the *mean free time*. It gives the mean time between consequent scattering/collision events. The solution to Eqn (3.49) is of the form

$$N(t) = N_0 e^{-t/\bar{t}_f} \tag{3.50}$$

The probability that an electron has a collision in a time interval dt is given by $dt/\bar{t}_f$. The differential change in momentum dp_x due to collisions in time dt is thus

$$dp_x = -p_x \frac{dt}{\bar{t}_f} \tag{3.51}$$

Resulting in

$$\left.\frac{dp_x}{dt}\right|_{\text{Collisions}} = \frac{-p_x}{\bar{t}_f} \tag{3.52}$$

At equilibrium the net rate of change of momentum must be zero. Using Eqs (3.48) and (3.52) leads to

$$\frac{-p_x}{\bar{t}_f} - nqE_x = 0 \tag{3.53}$$

The average momentum per electron is then

$$< p_x > = \frac{p_x}{n} = -q\,\bar{t}_f E_x \tag{3.54}$$

The average velocity $< v_x >$ in the negative x-direction is, therefore

$$< v_x > = \frac{< p_x >}{m^*_{nc}} = -\frac{q\,\bar{t}_f}{m^*_{nc}} E_x \tag{3.55}$$

where m^*_{nc} is the conductivity effective mass. Conductivity effective mass is different from the density of states effective mass. One must remember that the individual electrons constituting the group move in many directions by thermal motion during a given time interval. A net drift however takes place of an average electron due to the applied electric field, and it is this average velocity that is given by Eqn (3.55). The average drift velocity is much smaller than the random speed on account of thermal motion. The current density J_x due to this drift is given by the number of electrons crossing a given unit area per unit time multiplied by the charge carried by an electron. Thus

$$J_x = -qn < v_x > \tag{3.56}$$

Using Eqn (3.55) for $< v_x >$ in (3.56) leads to

$$J_x = \frac{nq^2\,\bar{t}_f}{m^*_{nc}} E_x \tag{3.57}$$

From Ohm's law, we know

$$J_x = \sigma E_x \tag{3.58}$$

From Eqs (3.57) and (3.58), we can conclude

$$\sigma = \frac{nq^2 \overline{t}_f}{m^*_{nc}} \tag{3.59}$$

σ is called the conductivity and has units of $(\Omega\ \text{cm})^{-1}$. Conductivity, σ, can also be written as

$$\sigma = nq\mu_n \tag{3.60}$$

Using Eqn (3.59), we can see that

$$\mu_n = \frac{q\overline{t}_f}{m^*_{nc}} \tag{3.61}$$

Comparing Eqs (3.61) and (3.55) leads to

$$\mu_n = -\left(\frac{< v_x >}{E_x}\right) \tag{3.62}$$

Thus, mobility is the average carrier drift velocity per unit electric field. It is expressed in units of cm^2/Vs. Since $< v_x >$ and E_x are oppositely directed, μ_n is a positive quantity.

The current density J_x given by Eqn (3.56) can be written as

$$J_x = nq\mu_n E_x \tag{3.63}$$

We have so far restricted ourselves to the x component of current density and electric field. The general expressions for mobility and current density for electrons are

$$\mu_n = \frac{-< v_n >}{E} \tag{3.64}$$

and

$$J_n = nq\mu_n E \tag{3.65}$$

The corresponding equations for holes are

$$\mu_p = \frac{< v_p >}{E} \tag{3.66}$$

and

$$J_p = pq\mu_p E \tag{3.67}$$

When both electrons and holes contribute to the conduction process, the total current density becomes

$$J = (nq\mu_n + pq\mu_p)E \tag{3.68}$$

The hole and electron mobilities for some typical semiconductors at room temperature are given in Table 3.1.

The term m^*_{nc} used in Eqn (3.61) is the conductivity effective mass of electrons. Silicon has six equivalent conduction band minima along the x-direction. The band curvature longitudinal effective mass along the major axis of the ellipsoid is represented by m_e, and m_t represents the transverse effective mass along the

Table 3.1 Typical mobility values at $T = 300$ K and low doping concentrations in common semiconductors

	μ_n(cm^2/Vs)	μ_p(cm^2/Vs)
Silicon	1350	480
Gallium arsenide	8500	400
Germanium	3900	1900

two minor axes. Equation (3.61) for mobility involves the reciprocal of m_{nc}^* and considering dimensional equivalence we can write

$$\frac{1}{m_{nc}^*} = \frac{1}{3}\left(\frac{1}{m_l} + \frac{2}{m_t}\right) \tag{3.69}$$

The conductivity effective mass is thus the *harmonic mean of the band curvature effective masses*.

The carrier mobility in a semiconductor is decided by the associated scattering processes. There are two such dominant scattering processes in a semiconductor, namely, *phonon or lattice scattering* and *ionized impurity scattering*. The atoms in any semiconductor crystal possess thermal energy at temperatures above the absolute zero. This causes the atoms to vibrate randomly about their mean lattice positions. These lattice vibrations lead to disruptions in the perfect periodic potential function within the crystal. These disruptions lead to an interaction between the charge carriers and the vibrating atoms. This scattering is referred to as lattice or phonon scattering. The random vibration of the atom increase as the temperature of the semiconductor increases. This leads to an increase in the rate at which scattering takes place and results in a decrease in the corresponding mobility. If μL represents the mobility due to lattice scattering, then a detailed theoretical treatment shows that to first order we can write

$$\mu_L \propto T^{-3/2} \tag{3.70}$$

Lattice scattering dominates in lightly doped semiconductors. Figure 3.29 shows the temperature dependence on electron and hole mobilities for silicon. The temperature dependence of mobility at low doping levels is found to obey T^{-n} dependence with $n \neq 3/2$. This is because the relationship given in Eqn (3.70) is not exact and is based on the first-order scattering theory.

We have learnt that extrinsic semiconductors are created by doping impurity atoms. At reasonable temperatures, the impurity atoms are ionized. A coulomb interaction takes place between the charge carriers and the ionized impurities. This coulombic interaction results in scattering and alters the velocity of the charge carriers. This interaction that affects the carrier mobility is called *ionized impurity scattering* and the corresponding mobility is denoted by μ_I. When the temperature of the semiconductor increases, the random thermal velocity of the charge carriers increases. The charge carriers then spend lesser time in the vicinity of the ionized impurity, lowering the scattering effect. An increase in the

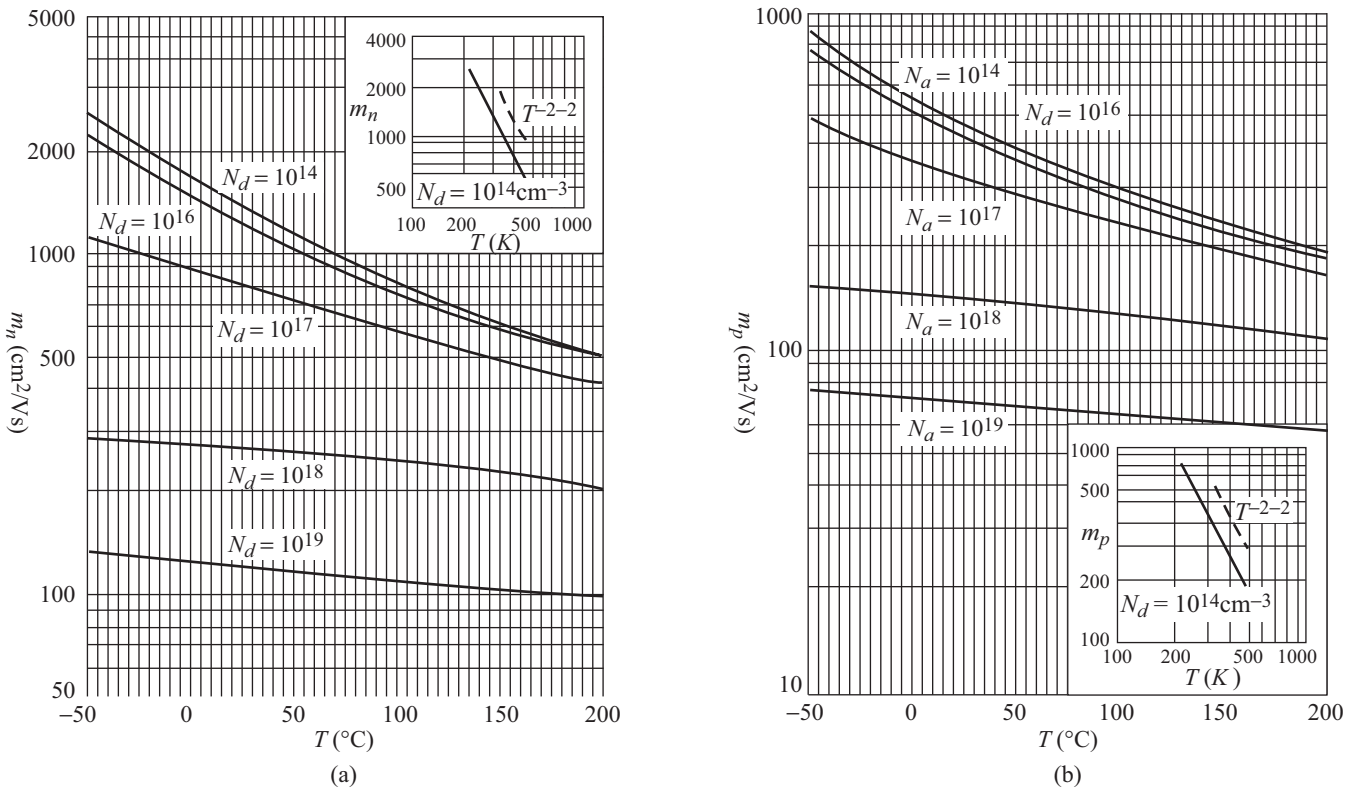


Fig. 3.29 (a) Electron and (b) hole mobilities (in silicon) verses temperature of various doping concentrations. Insets show temperature dependence for almost intrinsic silicon

temperature of a semiconductor therefore results in an increase in mobility μ_I as a result of ionized impurity scattering. According to the first-order theory,

$$\mu_I \propto \frac{T^{3/2}}{N_I} \tag{3.71}$$

where $N_I = N_d^+ + N_a^-$ is the total ionized impurity concentration. As expected, a higher N_I leads to an increase in the probability of a charge carrier undergoing scattering with an ionized impurity centre, resulting in a smaller μ_I. Figure 3.30 is a plot of electron and hole mobilities for some common semiconductors at $T = 300$ K as a function of impurity concentration. At higher impurity concentration values, decrease in mobility with impurity concentration can be seen to be more pronounced.

The net mobility μ involves both the scattering mechanisms. Thus

$$\frac{1}{\mu} = \frac{1}{\mu_I} + \frac{1}{\mu_L} \tag{3.72}$$

The net mobility is thus lower than the lowest contribution.

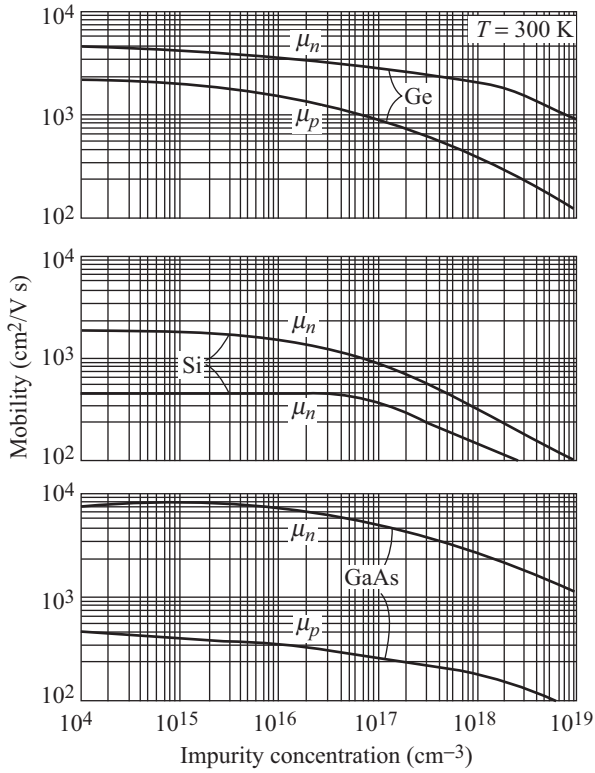


Fig. 3.30 Electron and hole mobilities versus impurity concentration for Ge, Si, and GaAs at $T = 300$ K

3.5.2 High-field Effect

In the treatment presented in Section 3.5.1, it has been assumed that the drift current is proportional to electric field, with the proportionality constant not showing any dependence on field E. This is in fact due to the validity of Ohm's law assumed in that section. This assumption is indeed valid over a wide range of values of electric field. The total velocity of a charge carrier is the sum of the random thermal velocity and drift velocity. At $T = 300$ K, the average random thermal velocity can be found using the equation

$$\frac{1}{2} m v_{\text{th}}^2 = \frac{3}{2} kT = \frac{3}{2} (0.0259) = 0.03885 \text{ eV} \tag{3.73}$$

For an electron in silicon, this leads to $v_{\text{th}} \cong 10^7$ cm/s. Let us take a typical value of $\mu_n = 1400$ cm^2/Vs in low-doped silicon and consider an applied electric field of 100 V/cm. Using the definition of μ_n, we can write

$$V_{\text{drift}} = \mu_n E = 1400 \times 100 = 1.4 \times 10^5 \text{ cm/s}$$

Thus, at such fields the energy due to drift velocity is small as compared to that due to thermal velocity. As the drift velocity approaches the mean thermal velocity, the added energy imparted by the field starts getting transferred to the lattice rather than increase the carrier velocity. A constant scattering limited velocity is then reached by the charge carrier. Figure 3.31 shows plots of average drift velocity as a function of applied electric field for electrons and holes in some common semiconductors.

A linear variation of velocity with electric field is observed at low electric fields. The slope of this linear region gives the mobility of the charge carrier. At high electric fields, a substantial deviation from the linear behaviour is observed

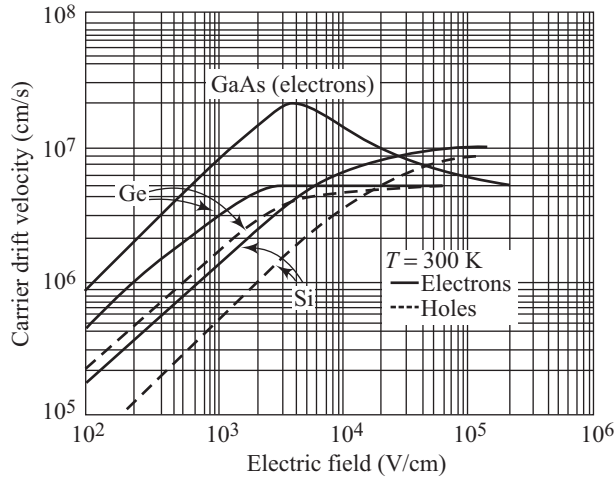


Fig. 3.31 Carrier drift velocity versus electric field for high-purity Si, Ge, and GaAs

from Fig. 3.31. For electrons in silicon, with electric field in the region of 30 kV/cm, the drift velocity of electrons saturates at approximately 10^7 cm/s. The drift current density J_n due to electrons is given by

$$J_n = nq\mu_n E \tag{3.74}$$

which on using the definition of μ_n leads to

$$J_n = nq\, v_{\text{drift}} \tag{3.75}$$

A saturated drift velocity thus also leads to a saturation of drift current density and makes it independent of electric field.

The plot of carrier drift velocity versus electric field for electrons in GaAs is however distinctly different. At low electric field values, the drift velocity varies linearly with electric field, the slope giving the low-field electron mobility. The typical low-field electron mobilities for gallium arsenide are around 8500 cm^2/Vs, a value substantially higher than that for silicon. As the electric field magnitude increases, the electron drift velocity reaches a peak and then decreases. *We may define the instantaneous slope of the v_{drift} versus E plot to be the differential mobility of the charge carrier*. Thus from Fig. 3.30 it is clear that GaAs displays a negative differential mobility. This negative differential mobility leads to a negative differential resistance which can be used to design oscillators.

The cause for the negative differential mobility can be inferred from the *E*-**k** diagram for GaAs shown in Fig. 3.32.

In the band diagram shown, two different valleys are present. The density-of-states effective mass of electrons in the lower valley is $m_n^* = 0.067\, m_0$, where m_0 is the rest mass of the electrons. Electrons have a density-of-states effective mass

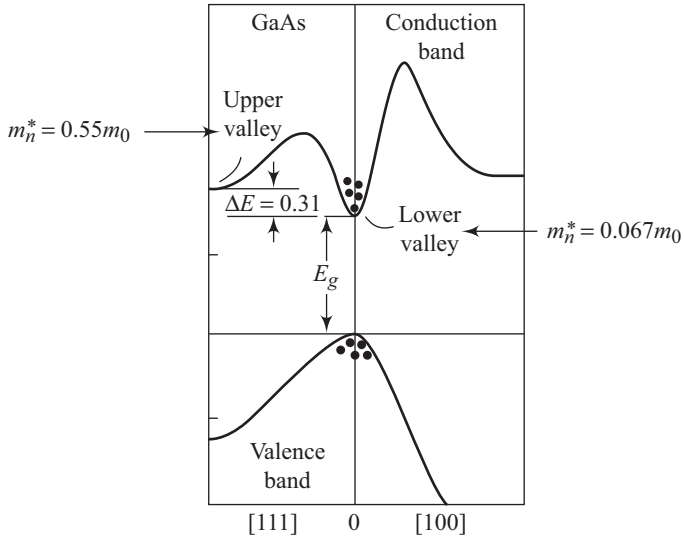


Fig. 3.32 Energy band structure for GaAs showing the upper valley and the lower valley in the conduction band

of $0.55m_0$ in the upper valley. If the applied electric field is increased, the energy of the electrons increases. These more energetic electrons then get scattered from the lower valley to the upper valley, where the larger effective mass leads to a smaller mobility. Due to this transfer of electrons from the lower valley to the upper valley, the average drift velocity of electrons starts decreasing with increasing electric field, as shown in Fig. 3.31.

3.5.3 Hall Effect

Moving charges experience forces in the presence of applied electric and magnetic fields. One interesting consequence of these forces is the *Hall effect.* Simply put, *Hall effect is the process leading to the development of a voltage across one of the faces of a semiconductor slab when a crossed electric and magnetic field is applied across the other two pairs of faces of the slab*. Hall effect can be used to distinguish between *n*-type and *p*-type semiconductors, measure majority carrier concentration, and majority carrier mobility. The effect can also be used to develop a magnetic probe for circuit application.

Figure 3.33 shows a schematic diagram to explain the Hall effect. A current I_x flows through the semiconductor in the *x*-direction. A magnetic field B_Z is applied in the *z*-direction. Charge carriers present in the semiconductors experience a force due to the magnetic field.

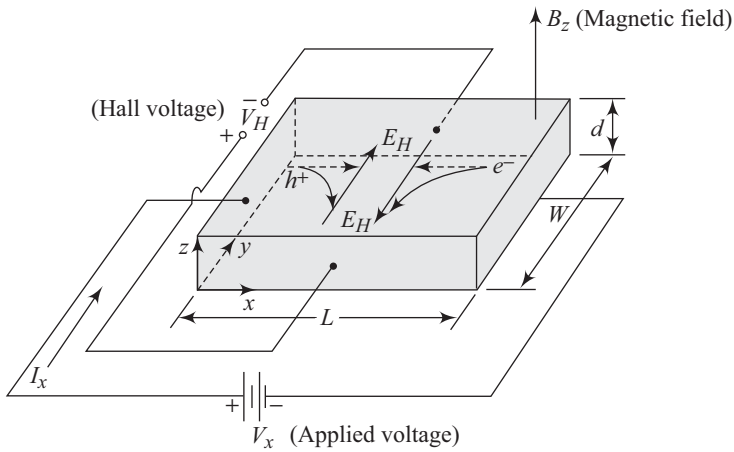


Fig. 3.33 Typical set-up for observing the Hall effect

This magnetic force is felt in the (–*y*)-direction by both electrons and holes. In an *n*-type semiconductor ($n_0 > p_0$), a build-up of electrons would take place on the $y = 0$ surface. For a *p*-type semiconductor ($p_0 > n_0$), the build-up would be of holes on the $y = 0$ surface. This charge build-up would result in an induced electric field in the *y*-direction. The resulting force on the charge carriers due

to this induced electric field would be in steady state exactly balanced by the magnetic force. Thus, at equilibrium we can write

$$qE_y = qv_xB_Z \tag{3.76}$$

where the induced electric field E_y is called the *Hall field* and the corresponding voltage is called the *Hall voltage*. Designating the Hall voltage as V_H and the Hall field as E_H, we can see from Fig. 3.33,

$$V_H = E_HW = E_yW \tag{3.77}$$

where E_H is assumed to be positive in the $+y$-direction and V_H is considered positive if it has the polarity as shown in the figure.

For a *p*-type semiconductor, the Hall voltage would be positive; for an *n*-type semiconductor, it would be negative. *Thus the polarity of the Hall voltage can be used to determine the type of the extrinsic semiconductor.*

Using Eqn (3.76) in Eqn (3.77) leads to

$$V_H = v_xWB_Z \tag{3.78}$$

For a *p*-type semiconductor, we can write for current density J_x,

$$J_x = epv_{dx}$$

which leads to

$$v_{dx} = \frac{J_x}{ep} = \frac{I_x}{ep(Wd)} = v_x \tag{3.79}$$

where v_{dx} is the drift velocity and the product (Wd) represents the relevant face area and e is the magnitude of the electron charge. Putting the expression for v_x from Eqn (3.79) into Eqn (3.78), we get

$$V_H = \frac{I_xB_z}{epd} = \frac{I_xB_z}{d}R_H \tag{3.80}$$

where $R_H = 1/ep$ is called the Hall coefficient. For an *n*-type semiconductor,

$$R_H = -\frac{1}{en}$$

which yields

$$p = \frac{I_xB_z}{edV_H} \tag{3.81}$$

Thus, the majority carrier concentration can be determined from the knowledge of current, magnetic field, and Hall voltage using Eqn (3.81). The corresponding equations for the *n*-type semiconductor are

$$V_H = -\frac{I_xB_z}{ned} \tag{3.82}$$

and

$$n = -\frac{I_xB_z}{edV_H} \tag{3.83}$$

Since the Hall voltage is negative for electrons, the electron concentration arrived at using Eqn (3.83) is positive as expected. The current density, J_x, can be expressed as

$$J_x = ep\mu_p E_x \tag{3.84}$$

where μ_p is the low-field hole mobility. Equation (3.84) can be rewritten as

$$\frac{I_x}{Wd} = \frac{ep\mu_p V_x}{L}$$

which leads to

$$\mu_p = \frac{I_x L}{epV_x Wd} \tag{3.85}$$

The corresponding equation for μ_n would be

$$\mu_n = \frac{I_x L}{enV_x Wd} \tag{3.86}$$

Equations (3.85) and (3.86) can be used to calculate the low-field carrier mobilities.

3.6 Carrier Diffusion

Drift as a process leading to the flow of current through a semiconductor has been discussed in the preceding section. Diffusion is another process by virtue of which current can flow through a semiconductor. *Diffusion is a process that leads to the flow of particles from a region of high concentration towards a region of low concentration. The resulting current is called diffusion current.*

3.6.1 Diffusion Current Density

To understand the diffusion process in detail, let us assume that a particular piece of an *n*-type semiconductor has an electron concentration as given in Fig. 3.34.

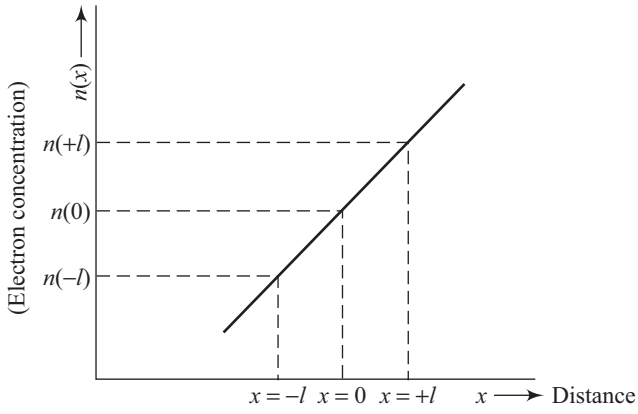


Fig. 3.34 Electron concentration versus distance leading to diffusion current

In this figure, l has been assumed to be one mean-free path of the electron, i.e., the mean distance travelled by an electron between two collisions ($l = v_{th}\ \tau_{cn}$, where τ_{cn} is the mean time between collisions, for electrons). Let us assume that the semiconductor has a uniform temperature, resulting in a uniform average thermal velocity v_{th}. Let us evaluate the net flow of electrons per unit time per unit area crossing the plane at $x = 0$. On an average, electrons moving towards the right at $x = -l$ and electrons moving towards the left at $x = l$ will be able to cross the plane at $x = 0$. At any given instant of time, one half of the electrons at $x = l$ can be assumed to be travelling towards the right. Similarly, on an average, one half of the electrons at $x = +l$ can be assumed to be travelling towards the left at any given instant of time. The net rate of electron flow, ϕ_{nf} in the $+x$-direction at $x = 0$ is given by

$$\phi_{nf} = \frac{1}{2} n(-l) v_{th} - \frac{1}{2} n(+l) v_{th} = \frac{1}{2} v_{th} [n(-l) - n(l)] \tag{3.87}$$

where $n(-l)$ and $n(l)$ represent the electron concentration at $(-l)$ and $(+l)$, respectively. Expanding the electron concentration in a Taylor series about $x = 0$ and keeping only the first two terms, we get from Eqn (3.87)

$$\phi_{nf} = \frac{1}{2} v_{th} \left\{ \left[n(0) - l \frac{dn}{dx} \right] - \left[n(0) + l \frac{dn}{dx} \right] \right\}$$

This leads to

$$\phi_{nf} = -v_{th} l \frac{dn}{dx} \tag{3.88}$$

Since the electrons carry a charge $(-e)$, the corresponding current density $J_{nx/diff}$ can be written as

$$J_{nx/diff} = -e\phi_{nf} = e v_{th} l \frac{dn}{dx}$$

which can be rewritten in the form

$$J_{nx/diff} = eD_n \frac{dn}{dx} \tag{3.89}$$

where D_n is called the *electron diffusion coefficient,* and it has units of cm^2/s.

The diffusion of electrons from a region of high concentration to a region of low concentration leads to a net flow of electrons in the negative x-direction in the specific case considered. The conventional current flow is in the positive x-direction. Figure 3.35(a) shows a schematic diagram of the electron diffusion current.

The corresponding current density expression for the diffusion current due to holes is given by

$$J_{px/diff} = -eD_p \frac{dp}{dx}$$

where D_p is the hole diffusion coefficient. The situation is shown in Fig. 3.35(b). Holes diffuse from a region of high concentration to a region of low concentration, producing a hole flux and conventional diffusion current density in the negative x-direction.

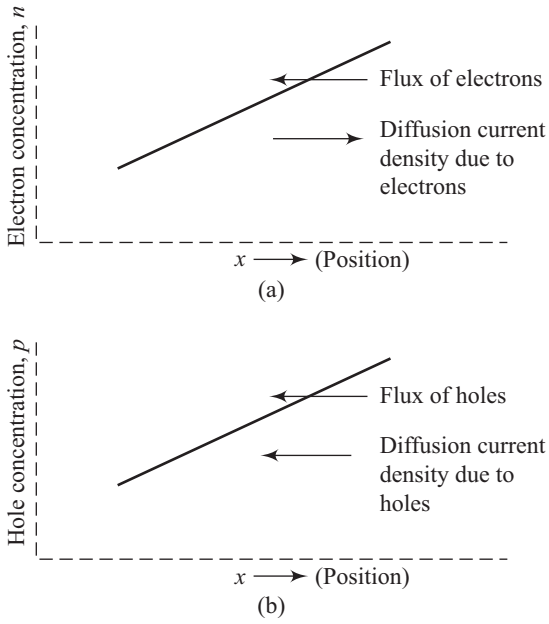


Fig. 3.35 (a) Diffusion of electrons due to a density gradient (b) Diffusion of holes due to a density gradient

3.6.2 Total Current Density

In general, there can be four possible different current sources in a semiconductor. These are

- electron drift
- electron diffusion
- hole drift
- hole diffusion

The total current density in the most general case can therefore be written as

$$J = en\mu_n E_x + ep\mu_p E_x + eD_n \frac{dn}{dx} - eD_p \frac{dp}{dx} \tag{3.90}$$

for the case of an electric field and concentration gradient in the x-direction. Equation (3.90) can be generalized in three dimensions to read

$$J = en\mu_n E + ep\mu_p E + eD_n \nabla n - eD_p \nabla p \tag{3.91}$$

Mobility and diffusion coefficients are however not independent quantities but are related to each other through the *Einstein relation*, which we shall derive in a later section.

3.7 Graded Impurity Distribution

In our discussion in this chapter so far, we have assumed a uniformly doped semiconductor. In many circumstances, there may exist regions in a semiconductor having non-uniform doping. The thermal equilibrium state of a non-uniformly doped semiconductor would be discussed in this section.

3.7.1 Induced Field

Let us suppose that a semiconductor is doped with an *n*-type impurity such that the doping concentration decreases in the positive *x*-direction as shown in Fig. 3.36.

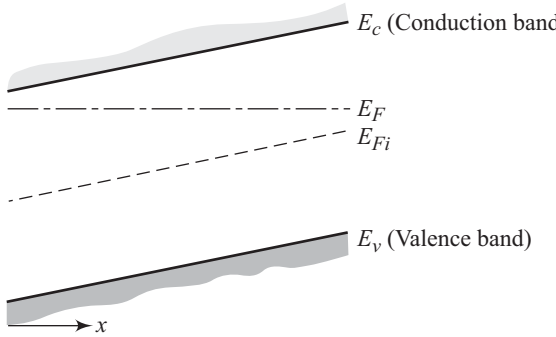


Fig. 3.36 Energy band diagram for a semiconductor in thermal equilibrium with a non-uniform donor impurity concentration

As we have seen, the closeness of the Fermi level E_F to the conduction band gives a measure of the donor concentration. At thermal equilibrium, the Fermi level is constant throughout the semiconductor. The Fermi level E_F is therefore moving away from the conduction band as one moves in the positive *x*-direction. The intrinsic Fermi level is shown in Fig. 3.36 using a symbol E_{Fi}. A diffusion of majority carrier electrons takes place in the $+x$-direction. The diffusing electrons leave behind ionized donors, which are positively charged. The separated opposite charges create an electric field in the $+x$-direction. This induced electric field exerts a force on the electrons in the $-x$-direction. *At equilibrium, the induced electric field stops any further diffusion of electrons*. The electron concentration at equilibrium is thus not exactly equal to the fixed impurity concentration. Since, in most cases of practical interest, the space charge induced by the diffusion process is small in comparison to the impurity concentration, the difference between the electron concentration at equilibrium and the fixed impurity concentration is however too small.

The potential ϕ across the semiconductor is given by

$$\phi = \frac{1}{e}(E_F - E_{Fi}) \tag{3.92}$$

For a one-dimensional situation, Eqn (3.92) leads to

$$E_x = -\frac{d\phi}{dx} = \frac{1}{e}\frac{dE_{Fi}}{dx} \tag{3.93}$$

This equation uses the invariance of E_F with distance at thermal equilibrium. Let us now assume the quasi-neutrality condition, i.e., the electron concentration is nearly equal to the donor impurity concentration. In this case, we can write

$$n_0 = n_i \exp\left[\frac{E_F - E_{Fi}}{kT}\right] \approx N_d(x) \tag{3.94}$$

which leads to

$$E_F - E_{Fi} = kT \ln\left[\frac{N_d(x)}{n_i}\right] \tag{3.95}$$

Taking the derivatives of both sides results in

$$-\frac{dE_{Fi}}{dx} = \frac{kT}{N_d(x)}\frac{dN_d(x)}{dx} \tag{3.96}$$

Using Eqn (3.93) in Eqn (3.96) yields

$$E_x = -\left(\frac{kT}{e}\right)\frac{1}{N_d(x)}\frac{dN_d(x)}{dx} \tag{3.97}$$

Thus a non-uniform doping leads to a built-in electric field and a built-in potential difference within the semiconductor.

3.7.2 Einstein Relation

Consider a non-uniformly doped semiconductor kept in the open circuit condition at thermal equilibrium. Then the individual electron and hole currents must be zero. Thus we can write

$$J_n = 0 = en\mu_n E_x + eD_n \frac{dn}{dx} \tag{3.98}$$

Assume the semiconductor to be represented by Fig. 3.36. If quasi-neutrality is valid, then $n \approx N_d(x)$. Now using Eqn (3.98), we get

$$0 = e\mu_n N_d(x) E_x + eD_n \frac{dN_d(x)}{dx} \tag{3.99}$$

Substituting the expression for E_x from Eqn (3.97) into Eqn (3.99) leads to

$$0 = -e\mu_n N_d(x)\left(\frac{kT}{e}\right)\frac{1}{N_d(x)}\frac{dN_d(x)}{dx} + eD_n \frac{dN_d(x)}{dx} \tag{3.100}$$

Equation (3.100) is valid, provided

$$\frac{D_n}{\mu_n} = \frac{kT}{e} \tag{3.101}$$

Considering the hole current and equating it also to zero leads to

$$\frac{D_p}{\mu_p} = \frac{kT}{e} \tag{3.102}$$

Combining Eqs (3.101) and (3.102) yields

$$\frac{D_n}{\mu_n} = \frac{D_p}{\mu_p} = \frac{kT}{e} \tag{3.103}$$

Equation (3.103) is known as the Einstein relation. It emphasizes the interdependence of the diffusion coefficient and mobility. Table 3.2 lists the diffusion coefficients and mobility values of some common semiconductors.

Table 3.2 Typical mobility and diffusion coefficient values at $T = 300$ K ($\mu = \text{cm}^2/\text{V s}$ and $D = \text{cm}^2/\text{s}$)

	μ_n	D_n	μ_p	D_p
Silicon	1350	35	480	12.4
Gallium arsenide	8500	220	400	10.4
Germanium	3900	101	1900	49.2

One must however remember that the diffusion coefficients and mobilities are themselves functions of temperature and therefore Eqn (3.103) oversimplifies the real temperature dependence of the diffusion coefficient and mobility.

- In addition to the scattering mechanisms, other mechanisms such as intra-valley and inter-valley scattering also affect the electron mobility.
- Under strong magnetic fields, a significant increase is observed in resistivity due to the magneto-resistance effect.
- For GaAs the high-field velocity approaches 6×10^6 cm/s.

Solved Problems

3.1 The velocity of an electron initially travelling with a velocity of 5×10^7 cm/s increases by a value of 2 cm/s. Calculate the corresponding increase in the kinetic energy of the electron.

Solution

The increase in kinetic energy, ΔE, is given by

$$\Delta E = \frac{1}{2} m \left(v_f^2 - v_i^2 \right) \tag{3.1.1}$$

where v_f and v_i represent the final and initial velocities, respectively. Let us assume

$$v_f = v_i + \Delta v$$

Then we get

$$v_f^2 = (v_i + \Delta v)^2 = v_i^2 + 2v_i \Delta v + (\Delta v)^2$$

Since Δv is a small quantity, the term $(\Delta v)^2$ can be neglected. This gives

$$v_f^2 = v_i^2 + 2v_i\Delta v$$

which implies

$$v_f^2 - v_i^2 = 2v_i\Delta v \tag{3.1.2}$$

Using Eqn (3.1.2) in Eqn (3.1.1) yields

$$\Delta E \approx \frac{1}{2}m(2v_i\Delta v) = mv_i\Delta v \tag{3.1.3}$$

Putting the given values in Eqn (3.1.3), we get

$$\Delta E = (9.11\times10^{-31})(5\times10^5)(0.02) = 9.11\times10^{-27}\text{ J}$$

Then the increase in kinetic energy ΔE in eV is

$$\Delta E = \frac{9.11\times10^{-27}}{1.6\times10^{-19}} = 5.7\times10^{-8}\text{ eV}$$

The energy change involved between adjacent energy states within an allowed band is typically $\sim 10^{-19}$ eV. Such changes are very small indeed. Thus an allowed energy band can be treated as a quasicontinuous-distribution.

3.2 Evaluate the approximate donor binding energy for GaAs. Assume $\varepsilon_r = 13.2$ and $m_n^* = 0.067m_0$.

Solution

Using Eqs (3.22) and (3.23), we get

$$E = \frac{m_n^* q^4}{8(\varepsilon_0\varepsilon_r)^2 h^2} \tag{3.21}$$

Substituting the given values of m_n^*, ε_r, and the known values of other quantities in Eqn (3.2.1) leads to

$$E = \frac{(0.067)(9.11\times10^{-31})(1.6\times10^{-19})^4}{8(8.85\times10^{-12}\times13.2)^2(6.63\times10^{-34})^2}$$

yielding

$$E = 8.34\times10^{-22}\text{ J} = 0.0052\text{ eV}$$

Thus an energy of 5.2 meV is required to excite the donor electron.

3.3 Determine the density-of-states effective mass of electrons in silicon. Assume $m_l = 0.98m_0$ and $m_t = 0.19m_0$, where m_0 is the rest mass of electrons.

Solution

From Eqn (3.32) we get

$$(m_n^*)^{3/2} = 6(m_l m_t^2)^{1/2}$$

This gives

$$m_n^* = 6^{2/3}(m_l m_t^2)^{1/3} \tag{3.3.1}$$

Substituting the given values in Eqn (3.3.1) results in

$$m_n^* = 6^{2/3}[0.98(0.19)^2]^{1/3}\, m_0 = 1.1 m_0$$

3.4 A Si wafer is doped with 10^{16} P atoms/cm^3. Calculate the equilibrium hole concentration p_0 at 300 K. How is E_F located relative to E_i? Sketch the resulting band diagram. Take $n_i = 1.5 \times 10^{10}$ cm^{-3}.

Solution

For $N_d >> n_i$, we can assume $n_0 = N_d$. This leads to

$$p_0 = \frac{n_i^2}{n_0} = \frac{2.25 \times 10^{20}}{10^{16}} = 2.25 \times 10^4 \text{ cm}^{-3}$$

From Eqn (3.42), we can write

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

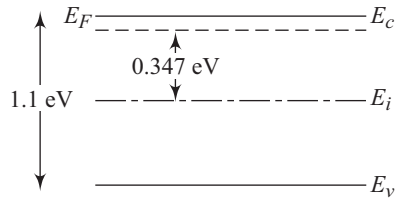
This gives

$$E_F - E_i = kT \ln \frac{n_0}{n_i} \tag{3.4.1}$$

Putting the given values in Eqn (3.4.1), we get

$$E_F - E_i = 0.0259 \ln\left(\frac{10^{16}}{1.5 \times 10^{10}}\right) = 0.347 \text{ eV}$$

The resulting band diagram would be



3.5 Determine the conductivity effective mass of electrons in Si. Use the m_l and m_t values given in solved example 3.3.

Solution

From Eqn (3.69), the conductivity effective mass m_{nc}^* is given by

$$\frac{1}{m_{nc}^*} = \frac{1}{3}\left(\frac{1}{m_l} + \frac{2}{m_t}\right) \tag{3.5.1}$$

Putting the given values in Eqn (3.5.1) results in

$$\frac{1}{m_{nc}^*} = \frac{1}{3}\left(\frac{1}{0.98 m_0} + \frac{2}{0.19 m_0}\right)$$

This yields

$$\frac{1}{m^*_{\text{nc}}} = 0.26m_0$$

3.6 A particular sample of germanium has a donor density $N_d = 10^{14}$ atoms/cm^3. Assuming all donor atoms to be ionized, calculate the resistivity of the sample.

Solution

For $n >> p$, the conductivity, σ, is given by

$$\sigma = ne\mu_n$$

This implies

$$\rho = \frac{1}{\sigma} = \frac{1}{ne\mu_n} \tag{3.6.1}$$

From Table 3.1, $\mu_n = 3900$ cm^2/V s for Ge at 300 T. Thus resistivity ρ, using (3.6.1), is

$$\rho = \frac{1}{10^{14} \times 1.6 \times 10^{-19} \times 3900} = 16.03\ \Omega\ \text{cm}$$

3.7 A sample of Si is doped with As to a level of 5×10^{16} atoms/cm^3. Calculate the resistivity of the sample. Determine the Hall coefficient and the Hall voltage if the thickness of the sample is 200 μm , $I_x = 2$ mA, and $B_Z = 5$ kG $= 5 \times 10^{-5}$ Wb/cm^2. Assume $\mu_n = 800$ cm^2/Vs.

Solution

$$\text{Resistivity } \rho = \frac{1}{\sigma} = \frac{1}{q\mu_n n_0} \tag{3.7.1}$$

Putting the given values in Eqn (3.7.1) results in

$$\rho = \frac{1}{1.6 \times 10^{-19} \times 800 \times 5 \times 10^{16}} = 0.156\ \Omega\ \text{cm}$$

Hall coefficient, R_H, for an n-type semiconductor is given by

$$R_H = -\frac{1}{en} \tag{3.7.2}$$

Putting the given value in Eqn (3.7.2) yields

$$R_H = -\frac{1}{en} = -\frac{1}{1.6 \times 10^{-19} \times 5 \times 10^{16}} = -125\ \text{cm}^3/\text{c}.$$

The Hall voltage V_H using Eqn (3.82) is given by

$$V_H = \frac{I_x B_z}{d} R_H \tag{3.7.3}$$

where d is the sample thickness. Putting appropriate values in Eqn (3.7.3) results in

$$V_H = \frac{(2\times10^{-3})(5\times10^{-5})(-125)}{(2\times10^{-2})}$$
$$= -62.5\times10^{-5}\ \text{V}$$

3.8 Two possible conduction and valence bands are shown in the E versus k diagrams given in Fig. 3.8.1.

(i) Identify the particles for which each of these diagrams is valid.

(ii) Which band will have heavier particle mass in each of the figures and why?

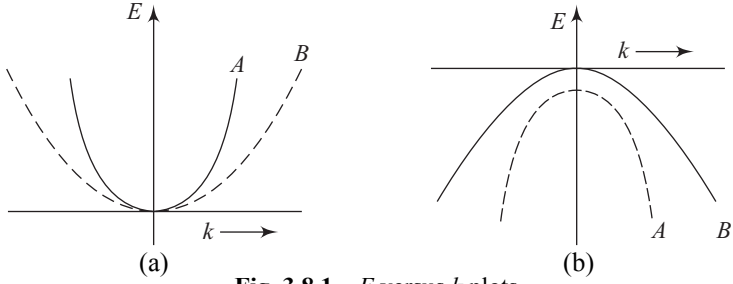


Fig. 3.8.1 E versus k plots

Solution

(i) Figure (a): Electrons (because, we are showing a conduction band)
Figure (b): Holes (because, we are showing a valence band)

(ii) Using Eqn (3.6) we get

$$\frac{d^2E}{dk^2} = \frac{\hbar^2}{m} \tag{3.8.1}$$

This means that effective mass of a particle is inversely proportional to the curvature of the energy band. Therefore, for Fig. 3.8.1(a) we get
Curvature of A > Curvature of B
Hence effective mass of a particle at A < effective mass of a particle at B
For Fig. 3.8.1(b),
Curvature of A > Curvature of B
Hence effective mass of a particle at A < effective mass of a particle at B

3.9 A silicon crystal is known to contain 10^{-4} atomic per cent of Sb (antimony) as an impurity. It is then uniformly doped with 3×10^{16} P (phosphorus) atoms/cm^3 and then with 10^{18} B (boron) atoms/cm^3. A thermal annealling treatment then completely activates all impurities.

(i) What is the conductivity type of this silicon crystal?

(ii) What is the density of the majority carriers?

Solution

Sb is a group V impurity, and, therefore, acts as a donor.
Since silicon has 5×10^{22} atoms/cm^3, 10^{-4} atomic per cent implies that the silicon is doped to a concentration of

$$5\times10^{22}\times10^{-6} = 5\times10^{16}\ \text{Sb atoms/cm}^3$$

Added doping of 3×10^{16} P atoms/cm^3 increases the donor doping of the crystal to 8×10^{16} cm^{-3}.
Additional doping 10^{18} atoms/cm^3 of B (a group III impurity) converts the silicon from *n*-type to *p*-type because the density of acceptors now exceeds the density of donors.
(i) Hence the silicon is *p*-type.
(ii) Density of majority carriers (holes) = $10^{18} - 8 \times 10^{16} = 9.2 \times 10^{17}$ cm^{-3}

3.10 In an intrinsic semiconductor (band gap = 1.1 e V) 4000 EHPs/cm^3 are created under an applied electric field. The valence band contribution to the current density is about 14.14×10^{-10} A/m^2. Assuming that all the holes are moving in the same direction, calculate the drift velocity of one of the group of holes (out of four groups, each with 1000 holes), if the values for the remaining three groups are 3×10^5 m/s, 5×10^6 m/s, and 6×10^6 m/s.

Solution

From Eqn (3.15) we get

$$J = q\sum_i v_i \tag{3.10.1}$$

We have

$$J = 14.14 \times 10^{-10} \text{ A/m}^2$$
$$= 14.14 \times 10^{-14} \text{ A/cm}^2$$

The drift velocities of three groups of holes are given as 3×10^5 m/s, 5×10^6 m/s, and 6×10^6 m/s. Hence, using Eqn 3.10.1 we get

$$14.14 \times 10^{-14} \text{ A/cm}^2 = 1.6 \times 10^{-19} \times 1000\,(3 \times 10^7 + 5 \times 10^8 + 6 \times 10^9 + x)$$

where x = unknown velocity of fourth group of holes

$$14.14 \times 10^{-14} = 1.6 \times 10^{-19} \times [1000(3 \times 10^7 + 5 \times 10^8 + 6 \times 10^8 + x)]$$

$$\frac{14.14 \times 10^{-14}}{1.6 \times 10^{-19} \times 1000} = (3 \times 10^7) + (5 \times 10^8) + (6 \times 10^8) + x$$

$$\Rightarrow 8.8375 \times 10^2 = 10^7 \left[3 + 50 + 60 + \frac{x}{10^7}\right]$$

$$\Rightarrow \frac{883.75}{10^7} = 113 + \frac{x}{10^7}$$

$$\Rightarrow x = -11.3 \times 10^8 \text{ cm s}^{-1}$$

From here $x \sim -11.3 \times 10^8$ cm/sec or -11.3×10^6 ms^{-1}. The negative sign indicates that the velocity of at least one group of holes is in the opposite direction to produce the given current density.

3.11 For a piece of GaAs semiconductor (E_g = 1.43 e V):
(a) Determine the minimum frequency of an incident photon that can interact with a valence electron and elevate the electron to the conduction band.

(b) What is the corresponding wavelength?

Solution

(a) $E = h\nu$

$\Rightarrow 1.43 = (4.14 \times 10^{-15})\nu$

$$\nu = \frac{1.43}{4.14} \times 10^{15}$$
$$= 3.454 \times 10^{14} \text{ Hz}$$

(Here the value of Planck's constant, h, has been taken to be 4.14×10^{-15} e V s)
Therefore, the minimum frequency = 3.45×10^{14} Hz

(b) $\lambda = \frac{c}{\nu}$

$$= \frac{3 \times 10^8 \text{ ms}^{-1}}{3.45 \times 10^{14} \text{ Hz}}$$
$$= 8.69 \times 10^{-7} \text{ m}$$

Therefore, the wavelength = 8.69×10^{-7} m.

3.12 Design a semiconductor resistor of resistance 10 KΩ which is to be operated at 300 K. The resistor should be able to handle a current density of 50 A/cm^2 when a potential of 5 V is applied across it. Given that the semiconductor material has been initially doped with a donor concentration of $N_d = 5 \times 10^{15}$ cm^{-3} and acceptors are to be added to form a compensated p-type material.

Solution

For 5 V applied to a 10 KΩ resistor, the total current is given as

$$I = V/R$$
$$= 5/(10 \times 10^3)$$
$$= 0.5 \text{ mA}$$

According to design constraints, the current density is supposed to be limited to 50 A/cm^2. The cross sectional area of the resistor is then given by

$$A = I/J$$
$$= 0.5 \times 10^{-3}/50$$
$$= 10^{-5} \text{ cm}^2$$

The next thing to be calculated for designing the resistor is its length. Now to calculate the length of the resistor we must fix a limit for the value of electric field, which can be handled by the resistor to be designed. Let us limit this electric field to 100 V/cm, then the length of the resistor can be calculated as

$$L = V/E$$
$$= 5/100$$
$$= 5 \times 10^{-2} \text{ cm}$$

Now, we know that

$$R = \frac{\rho L}{A} \tag{3.12.1}$$

Substituting the values for A, L, and R in Eqn (3.12.1) we get

$$10\ \text{K}\Omega = \frac{\rho \times (5 \times 10^{-2}\ \text{cm})}{(10^{-5}\ \text{cm}^2)}$$

$$\Rightarrow \rho = \frac{10 \times 10^3 \times 10^{-5}}{5 \times 10^{-2}} = 2\,\Omega\ \text{cm}$$

Now the conductivity, σ, is given as

$$\sigma = \frac{1}{\rho}$$
$$= \frac{1}{2\ \Omega\ \text{cm}} = 0.50\ (\Omega\,\text{cm})^{-1}.$$

Now, for a compensated p-type semiconductor we have

$$\sigma = q\mu_p p = q\mu_p (N_a - N_d) \tag{3.12.2}$$

In Eqn (3.12.2), the values of σ and N_d are known. But, to complete the design of the resistor, we have to choose a value of N_a which will give us a realistic value of hole mobility, μ_p. We know that if the doping is of the order of 10^{16}–10^{17} atoms/cm^3, then the mobility value falls. Therefore, here come in the skills of a professional design engineer, who can always guess what is going to be the value of mobility for a particular doping concentration. So, using trial and error we have
If $N_a = 1.25 \times 10^{16}$ cm^{-3}, then $N_a + N_d = 1.75 \times 10^{16}$ cm^{-3} and the hole mobility ~ 410 cm^2/V s.
Substituting these parameters in Eqn (3.12.2) we get

$$\sigma = 1.6 \times 10^{-19} \times 410 \times (1.25 \times 10^{16} - 5 \times 10^{15})$$
$$= 1.6 \times 10^{-19} \times 410 \times (0.75) \times 10^{16}$$
$$= 0.492(\Omega\ \text{cm})^{-1}\ \text{which is very close to the value we need.}$$

Therefore, the following parameters must be followed for designing the resistor
Limiting electric field, $E = 100$ V/cm
Length of resistor, $L = 5 \times 10^{-2}$ cm
Area of cross-section, $A = 10^{-5}$ cm^2
Acceptor doping concentration, $N_a = 1.75 \times 10^{16}$ cm^{-3}.

3.13 A piece of intrinsic silicon at room temperature is kept at thermal equilibrium. The position of some random level E_x is to be fixed at 0.9 eV above the valence band edge. Probability of capture of an energy state by an electron at E_x is 50%.

(a) Calculate the amount of doping required to satisfy this condition.
(b) Which impurity (name the exact element) should be chosen so that the energy required for transition of 100% doped carriers to conduction band is never more than 4.4% of silicon energy band gap ($E_g = 1.1$ e V)?

(c) Compare the resistivities of the doped and undoped pieces of silicon. Interpret the result obtained. Take $\mu_n = 1400\ \text{cm}^2/\text{V s}$ and $\mu_p = 500\ \text{cm}^2/\text{V s}$.

Given data:

Table 3.13.1 Ionization energies for different impurities in silicon

Impurity	Ionization energy (e V)
Phosphorous	0.045
Arsenic	0.050
Boron	0.045
Aluminum	0.060

Solution

Probability of capture of an energy state by an electron at E_x is 50%. This means that E_x is actually the Fermi level.

(a) The Fermi level is to be 0.9 e V above valence band edge and the band gap of Si is 1.1 e V. So, it means that the Fermi level is 0.2 e V below the bottom of the conduction band edge. And $E_F - E_i = 0.35$ e V

Now

$$n_0 = n_i e^{(E_F - E_i)/kT} \tag{3.13.1}$$

Substituting the known values in Eqn (3.13.1) we get

$$n_0 = 1.5 \times 10^{10} \times e^{0.35/0.0259}$$
$$= 1.109 \times 10^{16}\ \text{cm}^{-3}$$

(b) 4.4 % of Si band gap = 0.0484 e V

From Table 3.13.1 it can be seen that the IE of phosphorus in Si is less than the value calculated above (0.0484 eV). Therefore, phosphorus is the impurity that should be used for doping the given silicon sample.

(c) Conductivity, $\sigma = q(n\mu_n + p\mu_p)$.

Therefore, for the doped substrate

$$\sigma = 1.6 \times 10^{-19}[1.109 \times 10^{16} \times 1400]$$
$$= 2.484\,(\Omega\,\text{cm})^{-1}$$

Now, resistivity, $\rho = \dfrac{1}{\sigma}$

$$= \frac{1}{2.484}$$
$$= 0.4025\ \Omega\,\text{cm}$$

For the intrinsic (undoped substrate), $n = p = n_i$

Thus, conductivity, σ becomes

$$\sigma = 1.6 \times 10^{-19}\,[1.5 \times 10^{10} \times 1400 + 1.5 \times 10^{10} \times 500]$$

yelding

$$\sigma = 45.6 \times 10^{-7}\ (\Omega\ \text{cm})^{-1}$$

resulting in

$$\rho = \frac{1}{\sigma} = \frac{1}{45.6 \times 10^{-7}} \sim 2.1 \times 10^5 \ \Omega\,\text{cm}$$

3.14 For the silicon sample in the previous problem (3.13) the position of energy level, E_x is to set exactly 0.6 e V above the intrinsic level. What concentration of doping is required to satisfy this condition? What are these types of semiconductors called?

Solution

Fermi level 0.6 eV above intrinsic level means that the Fermi level has moved into the conduction band.

$$\begin{aligned} n_0 &= n_i e^{(E_F - E_i)/kT} \\ &= (1.5 \times 10^{10})\, e^{0.6/0.0259} \\ &= 1.725 \times 10^{20}\ \text{cm}^{-3} \end{aligned}$$

Such semiconductors are called degenerate semiconductors.

3.15 The electron concentration in an n-type GaAs semiconductor maintained at T = 300 K varies linearly from 1×10^{18} to 8×10^{17} cm^{-3} over a distance of 0.12 cm. The electron diffusion coefficient for the material is D_n = 225 cm^2/s. Determine the diffusion current density.

Solution

From Eqn (3.89) we get,

$$J_{n/\text{diff}} = eD_n \frac{dn}{dx} \tag{3.15.1}$$

Putting values in Eqn (3.15.1) gives,

$$J_{n/\text{diff}} = \frac{(1.6 \times 10^{-19})(225)(1 \times 10^{18} - 8 \times 10^{17})}{0.12}$$

resulting in,

$$J_{n/\text{diff}} = \frac{(1.6 \times 10^{-19})(225)(2 \times 10^{17})}{0.12}$$

giving,

$$J_{n/\text{diff}} = 60\ \text{A/cm}^2$$

3.16 The electron concentration in n-type silicon at T = 300 K is given by,

$$n(x) = 2 \times 10^{15}\, e^{(-x/\ln)} \quad \text{for} \quad x \geq 0$$

Calculate the electron diffusion current density at x = 0. Assume ln = 10^{-4} cm and D_n = 25 cm^2/s.

Solution

Eqn (3.89) gives,

$$J_{n/\text{diff}} = eD_n \frac{dn}{dx} \tag{3.16.1}$$

Putting values in Eqn (3.16.1) we get,

$$J_{n/\text{diff}} = eD_n \times 2 \times 10^{15} \times \left(-\frac{1}{\ln}\right) e^{\left(\frac{-x}{\ln}\right)}$$

At $x = 0$, we have,

$$\begin{aligned} J_{n/\text{diff}} &= (1.6 \times 10^{-19})(25)(2 \times 10^{15})\left(-\frac{1}{10^{-4}}\right) \\ &= -(1.6 \times 25 \times 2)\, 10^{-19+15+4} \\ &= -80 \text{ A/cm}^2 \end{aligned}$$

Recapitulation

- Individual quantized levels of individual atoms split into a band of discrete energy levels for a collection of atoms in close proximity.
- Splitting of energy states results in the formation of valence and conduction bands.
- The E - **k** relation for a single free electron is given by

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

- Electron transitions in direct band semiconductors do not involve change in **k**.
- The effective mass m_{eff} of an electron in a crystal is given by

$$m_{\text{eff}} = \frac{\hbar^2}{d^2E/dk^2}$$

- For an intrinsic semiconductor,

$$n = p = n_i$$

- For an n-type semiconductor, $n_0 >> (n_i, p_0)$ and for a p-type semiconductor, $p >> (n_i, n_0)$.
- The ground-state energy of an electron according to the Bohr Model is given by

$$E = \frac{mq^4}{2K^2\hbar^2} \text{ with } K = 4\pi\, \varepsilon_0 \varepsilon_r .$$

- The Fermi–Dirac distribution function $f(E)$ is given by

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

where, E_F is the Fermi level.

- Conduction band electron concentration n_0 is given by

$$n_0 = N_c e^{-(E_c - E_F)/kT}$$

where N_c is the effective density of states in the conduction band.

- $n_i p_i = n_0 p_0 = N_c N_v e^{-E_g/kT}$, where N_v is the effective density of states in the balance band and E_g is the band gap.
- Equilibrium electron and hole concentration n_0 and p_0 are given by

$$n_0 = n_i e^{(E_F - E_i)/kT} \text{ and } p_0 = n_i e^{(E_i - E_F)/kT}$$

respectively, where E_i is the intrinsic Fermi level.

- For a compensated semiconductor with $N_d > N_a$,

$$n_0 = N_d - N_a$$

- Conductivity σ for an *n*-type semiconductor is given by

$$\sigma = nq\mu_n$$

where μ_n is electron mobility expressed as v/ε.

- Total current density J for a semiconductor is

$$J = (nq\mu_n + pq\mu_p)E$$

- Mobility due to lattice scattering is $\mu_L \propto T^{-3/2}$ and due to impurity scattering is $\mu_I \propto T^{3/2}$.
- Hall voltage V_H is expressible as

$$V_H = \frac{I_x B_z}{epd} = \frac{I_x B_z}{d} R_H$$

for a *p*-type semiconductor. For an *n*-type semiconductor,

$$R_H = -\frac{1}{en}$$

- Total current density J including diffusion is given by

$$J = en\mu_n E_x + ep\mu_p E_x + eD_n \frac{dn}{dx} - eD_p \frac{dp}{dx}$$

for an applied field in the x direction and a concentration gradient in the x direction.

- $D_n/\mu_n = D_p/\mu_p = kT/e$ is called the Einstein relation.
- Phonons associated with long-wavelength vibrations of the lattice are called acoustic phonons.
- Phonons associated with vibrations in the optical frequency range are called optical phonons.
- Band gap engineering involves the modification of band structure.
- Large number of quantum wells kept close to each other constitute a *superlattice*.

Exercises

Review Questions

3.1 Define radial probability density.
3.2 What is Pauli's exclusion principle?
3.3 Why do discrete quantized energy levels associated with individual atoms split up when atoms are brought in close proximity?
3.4 Explain the formation of forbidden bands with the help of a suitable diagram.
3.5 Describe the process of formation of valence band and conduction band for silicon.
3.6 Distinguish between insulator, semiconductor, and metal on the basis of energy band diagrams.
3.7 Write down the Schrödinger equation for a free electron moving along the x-axis.
3.8 Sketch the shape of the E-**k** relation for a single free electron.
3.9 What are direct band gap semiconductors? Give one example.
3.10 What is the role of momentum conservation in electron transitions in an indirect semiconductor?
3.11 Define electron effective mass.
3.12 What does a negative electron effective mass imply?
3.13 Describe the process of EHP generation using a suitable diagram.
3.14 Compare the electron and hole energy depiction in a band diagram.
3.15 Explain the process of EHP generation using the broken bond model.
3.16 What is an extrinsic semiconductor? How is it produced from an intrinsic semiconductor?
3.17 What are amphoteric impurities? Give some examples.
3.18 Describe the Fermi–Dirac distribution function and schematically show its dependence on temperature.
3.19 Draw diagrams indicating the Fermi–Dirac distribution function for (a) an intrinsic semiconductor (b) an n-type semiconductor, and (c) a p-type semiconductor.
3.20 Write an expression giving the conduction band electron concentration as a function of temperature.
3.21 Prove that the equilibrium electron concentration is given by

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

3.22 Plot n_0 versus $\frac{1000}{T}$ for moderately doped silicon.
3.23 Explain the process of compensation with the help of a suitable diagram.
3.24 Show that conductivity σ is given by the relation

$$\sigma = \frac{nq^2 \overline{t}_f}{m_{\text{hc}}^*}$$

3.25 Distinguish between lattice scattering and ionized impurity scattering.
3.26 Why does the carrier drift velocity saturate at high electric fields?
3.27 Sketch the E-$\boldsymbol{k}$ diagram for GaAs.
3.28 Derive an expression for Hall voltage in terms of other experimental conditions in a typical Hall set-up.
3.29 Give a one-dimensional derivation for the relation

$$J = eD_n \frac{dn}{dx}$$

3.30 Derive the Einstein relation.
3.31 Explain the process of formation of energy bands using suitable sketches.
3.32 Derive the E-$\boldsymbol{k}$ relation for a free electron.
3.33 Define (a) band curvature effective mass, (b) density-of-states effective mass, and (c) conductivity effective mass.
3.34 For GaAs, the conduction band equi-energy surfaces are spherical. What is the relationship between (a), (b), and (c) of Question (3) for GaAs?
3.35 Show that $m_{\text{eff}} = \dfrac{\hbar^2}{d^2E/dk^2}$ from simple application of Newton's second law.
3.36 Draw a graph of $f(E)$ versus E and explain its salient features.
3.37 Show that $n_i = N_c \exp[-(E_c - E_i)/\text{kT}]$
3.38 Explain the temperature dependence of n_i using a suitable graph.
3.39 Draw a graph showing the temperature dependence of n_0 and explain the distinctive regions.
3.40 Differentiate between intrinsic semiconductor and compensated semiconductor of very high resistivity.
3.41 Derive the one-dimensional expression $\left\langle v_x \right\rangle = -\dfrac{q\,\overline{t}_f}{m^*_{\text{nc}}} E_x$.
3.42 Explain the temperature dependence of μ_L and μ_I.
3.43 Explain the origin of negative differential mobility of electrons in GaAs.
3.44 Derive an expression for Hall voltage.
3.45 Establish the dimensional consistency of the Einstein relation.
3.46 What are phonons?
3.47 Differentiate between acoustic and optical phonons. What are three main techniques by which electronic band structure of a semiconductor can be modified?
3.48 Find the band gap range for $\text{Hg}_{(1-x)}\text{Cd}_x\text{Te}$.
3.49 Sketch a quantum well schematically.
3.50 What is a superlattice?
3.51 What are strained layers?

Problems

3.1 Calculate the resistivity of an *n*-type germanium sample at 300 K. The sample has a donor density $N_d = 10^{20}$ atoms/m^3. Assume all donors to be ionized and take $\mu_n = 0.38$.

$$\left[\text{Hint: } \rho = \frac{1}{\sigma} = \frac{1}{nq\mu_n}\right]$$

Ans. 0.164 Ωm.

3.2 A particular sample of *n*-type germanium has a resistivity of 0.1 Ω m at 300 K. Calculate the donor concentration.

$$\left[\text{Hint: } n = N_d = \frac{\sigma}{e\mu_n}\right]$$

Ans. 1.64×10^{20}/m^3

3.3 Mobilities of free electrons and holes in pure Ge are 0.38 and 0.18 m^2/V s, respectively. Assume n_i for Ge = 2.5×10^{19} m^{-3}. Calculate the intrinsic conductivity.

[Hint: $\sigma_i = en_i(\mu_n + \mu_p)$]

Ans. 0.446 Ω m

3.4 Find the intrinsic resistivity of silicon from the following data:
$n_i = 1.5 \times^{16}$/m^3; $\mu_e = 0.13$, $\mu_h = 0.05$ m^2/V s.

$$\left[\text{Hint: } \rho_I = \frac{1}{\sigma_I}\right]$$

Ans. 2.314×10^3 Ω m

3.5 Energy of photons is being used to excite electrons from the conduction band to the valence band in semiconductors. Calculate the maximum wavelength of incident photons that can result in EHP generation for (a) silicon ($E_g = 1.12$ eV) and (b) Diamond ($E_g = 7$ eV)

$$\left[\text{Hint: } \lambda_{\max} = \frac{hc}{E_g}\right]$$

Ans. (a) 11080 Å; (b) 1774 Å

3.6 A sample of *n*-type silicon material has a donor concentration of $N_d = 2.5 \times 10^{20}$ m^{-3}. Calculate the temperature at which the Fermi level coincides with the edge of the conduction band. Assume the effective mass of electron to be equal to its rest mass.

$$\left[\text{Hint: (a) } E_F = E_c \text{ if } N_d = N_c\right.$$

$$\left.\text{(b) } N_c = 2\left(\frac{2\pi m_n^* kT}{h^2}\right)^{3/2} = 4.82 \times 10^{21}\left[\frac{m_n^*}{m_0}\right]^{3/2} T^{3/2} m^{-3}\right]$$

Ans. $T = 0.14$ K

3.7 A Si sample is doped with As to a concentration of 10^{16} atoms/cm^3. The thickness of the sample is 500 μm. A current $I_x = 1$ mA is made to flow in the x-direction and $B_Z = 5 \times 10^{-4}$ Wb/cm^2. Calculate the Hall voltage.

[Hint: $V_H = \dfrac{I_x B_z}{d} R_H$, where d is the thickness in the direction of the applied magnetic field

Ans. −6.25 mV

3.8 The donor concentration in an n-type semiconductor held at $T = 300$ K has an x dependence given by

$N_d(x) = 10^{15} - 10^{18}x$ (cm^{-3})

where x is in centimetres and varies between $0 \le x \le 1$ μm. Calculate the resulting field E_x at $x = 0$.

[Hint: Use Eqn (3.97) for E_x with

$$\frac{dN_d(x)}{dx} = -10^{18},$$

Thus,

$$E_x = -\frac{(0.0259)(-10^{18})}{10^{15}}]$$

Ans. 25.9 V/cm

3.9 In a particular sample of n-type semiconductor, the donor concentration dependence on x at 300 K is given by

$$N_d(x) = 10^{16} - 10^{20}x,$$

with x in the range of $0 \le x \le 1.5$ μm. Calculate the resulting field E_x at $x = 0.5$ μm.

$$\left[\text{Hint: } E_x = -\frac{kT}{e}\frac{1}{N_d(x)}\frac{d}{dx}N_d(x)\right]$$

Ans. 518 V/cm

3.10 A semiconductor has a mobility of 500 cm^2/V s at $T = 300$ K. Calculate the diffusion coefficient.

$$\left[\text{Hint: } D = \left(\frac{kT}{e}\right)\mu\right]$$

Ans. 12.95 cm^2/s

3.11 Figure 3.P11.1 shows the carrier concentration versus inverse temperature plot for silicon with about 10^{15} donors/cm^3.

(i) Write the names of three regions mentioned as Region I, Region II, and Region III in Fig. 3.P11.1(a).

(ii) Out of the four values given below, what minimum value of doping concentration must be chosen so that the device made out of this silicon works up to a maximum temperature of 400 K? Justify your choice of doping concentration briefly. Use Fig. 3.P11.1(b) for performing calculations and for values of any required parameters.

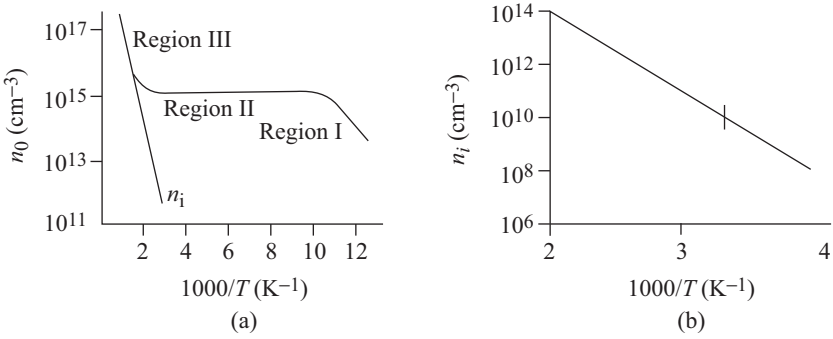


Fig. 3.P11.1 Carrier concentration versus inverse temperature

(a) ~ 10^{11} cm^{-3} (b) ~ 10^{12} cm^{-3} (c) ~ 10^{13} cm^{-3} (d) 10^{10}cm^{-3}

Ans. (i) Region I—Ionization region
Region II—Extrinsic region
Region III—Intrinsic region
(ii) (c)

3.12 (a) Find the equilibrium concentrations of electrons and holes, if a GaAs sample at 300 K is doped with 10^{12} silicon atoms/cm^3. 90% of these atoms replace the Ga atoms and 10% replace the As atoms. Assume that all the silicon impurities are ionized.

(b) Calculate the resistivity of the sample, if mobilities of electrons and holes are given as 8500 cm^2/V s and 400 cm^2/V s, respectively.

(c) Indicate the position of the Fermi level after doping with the help of an energy band diagram.

Ans. (a) 8×10^{11} cm^{-3} electrons and 5 cm^{-3} holes (b) 919.11 Ωcm

3.13 Consider a silicon sample at $T = 300$ K with doping concentrations of $N_a = 0$ and $N_d = 10^{16}$ cm^{-3}. Assuming complete ionization, calculate the drift current density, if the applied electric field is $E = 10$ V/cm . Given $\mu_n = 1350$ cm^2/ V s and $\mu_p = 480$ cm^2/V s.

Ans. 21.60 A/cm^2

3.14 A specimen of Si has a square cross section of 2×2 cm^2 and length of 2 cm. The current is mainly due to electrons, which have a mobility of 1300 cm^2/ V s. An applied dc voltage of 1 V across the bar produces a current of 8 mA in it.

Calculate (a) concentration of free electrons (b) drift velocity of electrons. Also briefly discuss, what overall effect do you foresee on the performance of this specimen, if you increase the doping concentration (at constant temperature) while still retaining its non degenerate nature.

Ans. (a) 1.92×10^{13} cm^{-3} (b) 6.5 m/s

3.15 A silicon crystal is known to contain 10^{-4} atomic per cent of As as an impurity. It then receives a uniform doping of 3×10^{16} P atoms/cm^3 and a subsequent uniform doping of 10^{18} B atoms/cm^3. A thermal annealling treatment then completely activates all impurities.

(a) What is the conductivity type of this silicon crystal?

(b) What is the density of the majority carriers?
(c) Sketch the band diagram.
[*Hint:* Refer to the case of compensated semiconductors]

Ans. (a) *n*-type (b) 4×10^{18} cm^{-3}

3.16 A silicon sample at $T = 300$ K contains a donor impurity concentration of 10^{16} atoms/cm^3. Determine the concentration of impurity to be added so that the Fermi level is just 0.2 eV above the valence band edge.

Ans. 2.1×10^{16} atoms/cm^3

3.17 A silicon sample is doped with 6×10^{15} cm^{-3} donors and 2×10^{15} cm^{-3} acceptors. Find the position of Fermi level with respect to E_i at 300 K.

Ans. 0.3235 eV

3.18 An intrinsic piece of silicon has been doped with a certain number of arsenic impurity atoms. Draw the resulting energy band diagram clearly indicating the position of the donor/acceptor level. What happens to the position of donor energy level, if the original semiconductor is replaced by a piece of GaAs. Comment on the result obtained.

Take ε_r(Si) = 11.8, ε_r(GaAs) = 13.2, $\varepsilon_0 = 8.85 \times 10^{-12}$ F/m., m_n^* (Si) = $0.26m_0$, m_n^*(GaAs) = $0.067m_0$.

[Hint: Use the following relation (obtained from Eqn 3.22):

$$E = \frac{m_n^* q^4}{8(\varepsilon_0 \varepsilon_r)^2 h^2}$$

For Si, the donor doping level will set somewhere close to 0.03 eV to 0.06 eV from a band edge. Similarly, in case of donor doping level in case of GaAs will settle somewhere close to 0.0052 eV. This means that lower amount of energy is required to excite the electron from the donor energy level to the conduction band in case of GaAs as compared to Si.]

3.19 The electron concentration in an *n*-type GaAs sample held at $T = 300$ K varies linearly from 3×10^{18} to 5×10^{17} cm^{-3} over a distance of 0.1 cm. Calculate the diffusion current density assuming, $D_n = 225$ cm^2/s.

$$\left[\text{Hint } J_{n/\text{diff}} = eD_n \frac{dn}{dx}\right]$$

Ans. 900 A/cm^2

3.20 The electron concentration in a particular sample of *n*-type silicon maintained at $T = 300$ K is given by

$$n(x) = 8 \times 10^{14} e^{(-x/\ln)} \quad \text{for} \quad x \geq 0$$

Assuming ln = 10^{-4} cm and D_n = 25 cm^2/s, determine the electron diffusion current density at $x = 0$.

$$\left[\text{Hint } J_{n/\text{diff}} = eD_n \frac{dn}{dx}\right]$$

Ans. –32 A/cm^2

4 Excess Carriers in Semiconductors

- Semiconductor in equilibrium
- Excess carrier generation and recombination
- Optical absorption
- Lifetime of excess minority carriers
- Shockley-Read-Hall theory
- Continuity equation
- Haynes-Shockley experiment
- Quasi-Fermi energy levels
- Surface effects
- Surface recombination velocity

Learning Objectives

After going through this chapter the student will be able to

- appreciate the concepts of generation and recombination rates for a semiconductor in equilibrium
- derive the relation giving the intensity dependence of incident radiation on its position within a semiconductor
- derive relations for recombination rates of excess carriers
- understand the Shockley–Read–Hall theory
- derive expression for recombination rates of electrons and holes using SRH theory
- solve numericals based on excess carrier concentration dependence on time
- solve numerical based on recombination rates of electrons and holes
- derive the time-dependent continuity equations for electron and holes
- understand the Haynes–Shockley experiment
- understand the concept of quasi-Fermi energy levels for electrons and holes
- understand the effect of semiconductor surfaces on excess carrier concentration
- define surface recombination velocity
- solve numericals based on quasi-Fermi energy level of electrons and holes
- solve numericals based on time-dependent continuity equation
- solve numericals based on Haynes–Shockley experiment

Introduction

The concept of extrinsic semiconductors having electrons and holes as charge carriers was discussed in Chapter 3. Charge carrier concentrations were to be considered under equilibrium conditions. In a variety of applications, semiconductor devices operate under non-equilibrium conditions. This can happen, for example, when excess carriers are created by incident light or applied bias. This chapter deals with the behaviour of the excess carriers under non-equilibrium conditions. The mechanisms involved in the process of the decay of these excess carriers over a period of time are discussed in detail. We will also revisit the topic of diffusion, but this time we consider non-equilibrium circumstances. The study of excess carriers in semiconductors help us to understand the classical experiment conducted by Haynes and Shockley at the Bell Telephone Laboratories. The Fermi level introduced in the preceding chapter assumed thermal equilibrium. In this chapter we introduce the concepts of the quasi-Fermi level for electrons and holes to deal with the behaviour of excess carriers under non-equilibrium conditions. This chapter also deals with the continuity equation, which will be very helpful in the later chapter. Semiconductor surfaces play an important role in the operation of semiconductor devices. In this chapter, we will study the effects of different surface conditions and introduce parameters that will help us to identify different surface conditions.

4.1 Semiconductor in Equilibrium

In this section we will recollect some important characteristics of a semiconductor in equilibrium. A semiconductor, as we have understood, has a conduction band and a valence band, with a band gap separating the two. Free electrons exist in the conduction band and holes exist in the valence band. At any given temperature, electrons are being continually excited from the valence band to the conduction band. This happens due to the random nature of the thermal processes involved. These electrons move randomly through the crystal in the conduction band. At times the electrons come across empty states (holes) in the valence band. The free electrons fall into these empty states in the valence band and in the process annihilate themselves and the holes corresponding to the empty states into which they fall. This process of annihilation of an EHP is called *recombination*. At thermal equilibrium, the net carrier concentrations are independent of time, therefore the rate of generation of electrons and holes must equal the rate of recombination. The processes of generation and recombination are shown schematically in Fig. 4.1.

Suppose G_{n0} and G_{p0} represent the thermal generation rates of electrons and holes, respectively, in units of $cm^{-3}\ s^{-1}$. Since electrons and holes are generated in pairs in a typical EHP process, we get

$$G_{n0} = G_{p0} \tag{4.1}$$

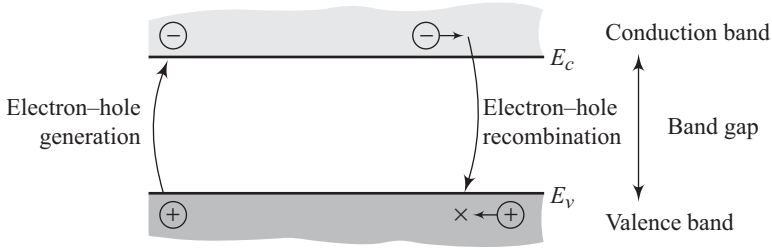


Fig. 4.1 Electron–hole generation and recombination processes

Suppose the corresponding recombination rates of electrons and holes in thermal equilibrium are represented by R_{n0} and R_{p0} respectively. R_{n0} and R_{p0} are expressed in units of $cm^{-3}\ s^{-1}$. If we consider direct band-to-band recombination, then electron and holes always recombine in pairs, leading to

$$R_{n0} = R_{p0} \tag{4.2}$$

Under thermal equilibrium conditions, the electron and hole concentrations are independent of time, implying

$$G_{n0} = G_{p0} = R_{n0} = R_{p0} \tag{4.3}$$

4.2 Excess Carrier Generation and Recombination

Excess charge carriers can be generated in a semiconductor by a variety of physical processes. A semiconductor with excess carriers is said to be in non-equilibrium. One of the important methods of creating excess carrier is optical absorption.

4.2.1 Optical Absorption

Suppose a beam of photons with a range of energies is incident on the surface of a semiconductor material. These photons can interact with the electrons present in the valence band of the material. Photons with energy hv greater than the band gap E_g will get absorbed in the semiconductor, whereas photons with energies less than E_g will be transmitted through the semiconductor. Absorption of photons with energy greater than E_g will lead to the transfer of a valence band electron to the conduction band and will thus result in an EHP creation. One must however remember that it is possible for an electron in the valence band to acquire energy greater than the band gap energy E_g. Such electrons initially reach the conduction band with energies greater than the average conduction band electrons as shown in Fig. 4.2. Subsequent scattering events with the lattice ultimately result in electron velocities reaching the thermal equilibrium velocity of other conduction band electrons. These electron-hole pairs created are *excess charge carriers* that alter the conductivity of the semiconductor material.

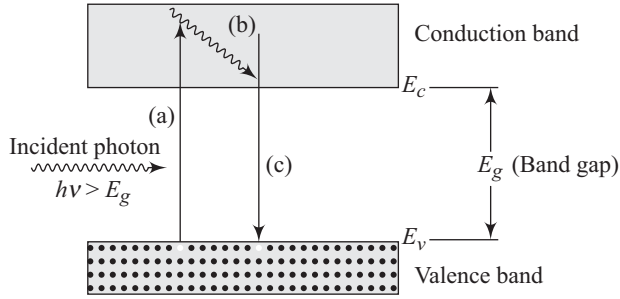


Fig. 4.2 Optical absorption of a photon with $h\nu > E_g$: (a) an EHP is created during photon absorption, (b) the excited electron gives up energy to the lattice by scattering events, (c) the electron recombines with a hole in the valence band

Suppose a photon beam of intensity I_0 (photons/cm^{2}s) is incident on a semiconductor sample of finite thickness. Let us further assume that the photons are of a fixed wavelength λ. If $h\nu > E_g$ photons get absorbed within the semiconductor. The photon that survives up to a depth x has no memory of the distance travelled without absorption, therefore the probability of absorption in any incremental depth dx is constant. The rate at which the intensity falls within the semiconductor, $-dI(x)/dx$, is proportional to the intensity remaining at x; thus

$$\frac{-dI(x)}{dx} = \alpha I(x) \tag{4.4}$$

where α is the proportionality constant. Equation (4.4) has solutions given by

$$I(x) = I_0\, e^{-\alpha x} \tag{4.5}$$

The coefficient α is called the absorption coefficient and has the unit of cm^{-1}. It is a function of the wavelength of the incident photons. A typical plot of α as a function of wavelength is given in Fig. 4.3.

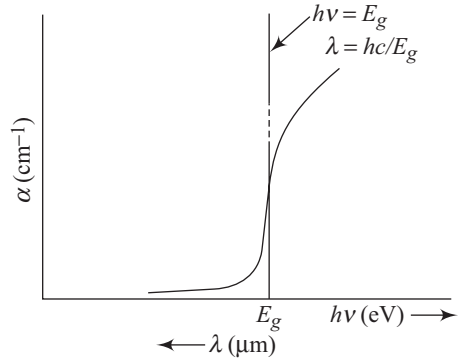


Fig. 4.3 Dependence of absorption coefficient α on the wavelength λ

Considerable absorption can be concluded from Fig. 4.3. The photon energy E is given by $h\nu = hc/\lambda$. If λ is expressed in μm, the photon energy in units of eV is given by $E = 1.24/\lambda$.

Appendix D gives the energy band gaps of some common semiconductors. The figure also gives the corresponding wavelengths of photons which will have their energies equal to the band gap. Any of these semiconductors will absorb photons with energies equal to or larger than the band gap. The most used semiconductor Si will, for example, absorb photons with wavelengths less than or equal to 1 μm.

4.2.2 Excess Minority Carrier Lifetime

Suppose g_n' is the excess electron generation rate and g_p' is the excess hole generation rate. Both g_n' and g_p' have units of $cm^{-3}\ s^{-1}$. If we consider only direct band-to-band generation of excess electrons and holes, then we must have

$$g_n' = g_p' \tag{4.6}$$

If δn and δp represent the excess electron and hole concentrations, then for the total electron and hole concentrations n and p we can write.

$$n = n_0 + \delta n \tag{4.7}$$

and

$$p = p_0 + \delta p \tag{4.8}$$

where n_0 and p_0 are the corresponding equilibrium concentrations. It is also clear that under non-thermal equilibrium conditions,

$$np \neq n_0 p_0 = n_i^2 \tag{4.9}$$

The process of creation of electron–hole pairs by optical absorption is shown schematically in Fig. 4.4.

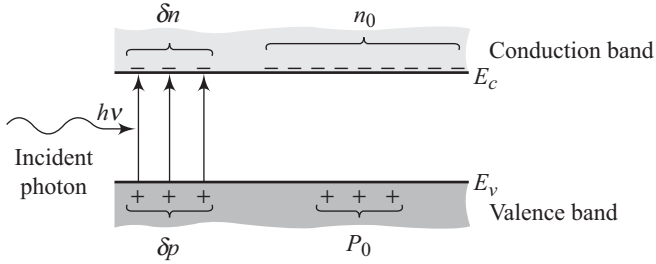


Fig. 4.4 Creation of excess electron and hole densities by optical absorption

In additon to the excess electron–hole generation, there is a competing process of excess electron–hole recombination. In this process, an excess electron in the conduction band falls to the valence band and occupies the vacancy representing a hole. Let R_n' and R_p' be the recombination rates of excess electrons and holes in units of $cm^{-3}\ s^{-1}$. The process of recombination of excess carriers is shown schematically in Fig. 4.5.

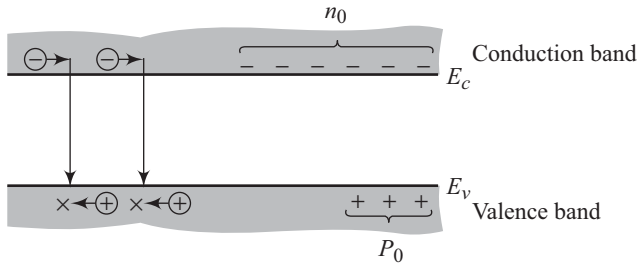


Fig. 4.5 Recombination process of excess carriers re-establishing thermal equilibrium

Once again, for direct band-to-band recombinations, we must have

$$R_n' = R_p' \tag{4.10}$$

The net rate of change in the electron concentration can be written as

$$\frac{dn(t)}{dt} = \alpha_r \left[n_i^2 - n(t)p(t) \right] \tag{4.11}$$

where the net electron concentration $n(t)$ is given by

$$n(t) = n_0 + \delta n(t) \tag{4.12}$$

A similar expression can be written for $p(t)$,

$$p(t) = p_0 + \delta p(t) \tag{4.13}$$

Obviously,

$$\delta n(t) = \delta p(t) \tag{4.14}$$

Since n_0 and p_0 are time independent, Eqn (4.11) can be rewritten as

$$\frac{d}{dt}\left(\delta n(t)\right) = \alpha_r \left\{ n_i^2 - [n_0 + \delta n(t)][p_0 + \delta p(t)] \right\}$$

which can be simplified to

$$\frac{d}{dt}\left(\delta n(t)\right) = -\alpha_r \delta n(t) \left[(n_0 + p_0) + \delta n(t) \right] \tag{4.15}$$

Let us now assume *low-level injection*, which simply means that the excess carrier concentration is much smaller than the thermal equilibrium majority carrier concentration. For an *n*-type semiconductor, we have

$$n_0 \gg p_0 \tag{4.16}$$

The low-level injection thus implies

$$\delta p\ (t) \ll n_0 \tag{4.17}$$

Similarly, for a *p*-type semiconductor we must have

$$\delta n\ (t) \ll p_0 \tag{4.18}$$

Thus for a *p*-type material under low injection, Eqn (4.15) reduces to

$$\frac{d}{dt}\left(\delta n(t)\right) = -\alpha_r p_0 \delta n(t) \tag{4.19}$$

Equation (4.19) permits solutions of the form

$$\delta n\ (t) = \delta n(0)\ e^{-\alpha_r p_0 t} \tag{4.20}$$

We now define a quantity $\tau_{n0} = 1/\alpha_r p_0$ to be the excess minority carrier lifetime. Equation (4.20) can then be expressed as

$$\delta n(t) = \delta n(0)\ e^{-t/\tau_{n0}} \tag{4.21}$$

τ_{n0} is the mean time available to an excess minority carrier before it recombines. The recombination rate R_n' using Eqn (4.19) is then

$$R_n' = -\frac{d}{dt}\left(\delta n(t)\right) = \alpha_r p_0 \delta n(t) = \frac{\delta n(t)}{\tau_{n0}} \tag{4.22}$$

where the negative sign is owing to the fact that R_n' is a positive quantity, whereas the rate of change of excess minority carrier concentration is a negative quantity.

For the case of direct band-to-band recombination, we have

$$R_n' = R_p' = \frac{\delta n(t)}{\tau_{n0}} \tag{4.23}$$

For a *p*-type material we have

$$R_n' = R_p' = \frac{\delta p(t)}{\tau_{p0}} \tag{4.24}$$

In general, the generation and recombination rates are functions of space coordinates and time.

4.3 Carrier Lifetime (General Case)

In the preceeding section, we introduced the excess carrier lifetime for an ideal semiconductor having no electronic energy states within the forbidden-energy band gap. A perfect single-crystal semiconductor material does display this behaviour. In reality, however, all semiconductors contain defects within the crystal. For a reasonable number of defects, the result is the creation of discrete energy states within the forbidden-energy band. These allowed-energy levels within the band gap play a dominant role in determining the mean carrier lifetime. Some impurities result in energy states that are far away from the energy band edges. Such energy states are called *deep impurity levels*. Some common examples are gold, copper, manganese, iron, etc. in silicon and germanium. These deep impurity levels play an important role in recombination processes involving non-equilibrium charge carriers. An exact analysis of the mean carrier lifetime can be undertaken using the Shockley–Read–Hall theory of recombination, which we will discuss next.

4.3.1 Shockley–Read–Hall Theory

An allowed energy state within the forbidden gap is called a *trap*. A trap under some circumstances can act as a *recombination centre* if it captures electrons and holes with equal probability, leading to equal capture cross sections. The Shockley–Read–Hall theory assumes the existence of a single trap or recombination centre at an energy E_t within the forbidden gap. The trap is called an *acceptor trap* if it is negatively charged and contains an electron. A trap is neutral when it does not contain an electron. We will assume the trap to be an acceptor trap.

The four possible processes involving the trap are shown schematically in Fig. 4.6.

These four processes are as follows.

Process 1 An electron is captured from the conduction band by an initially neutral empty trap.

Process 2 A trapped electron is emitted back into the conduction band.

Process 3 A hole is captured from the valence band by a trap already containing an electron. This process can also be viewed as emission of an electron from the trap to the valence band.

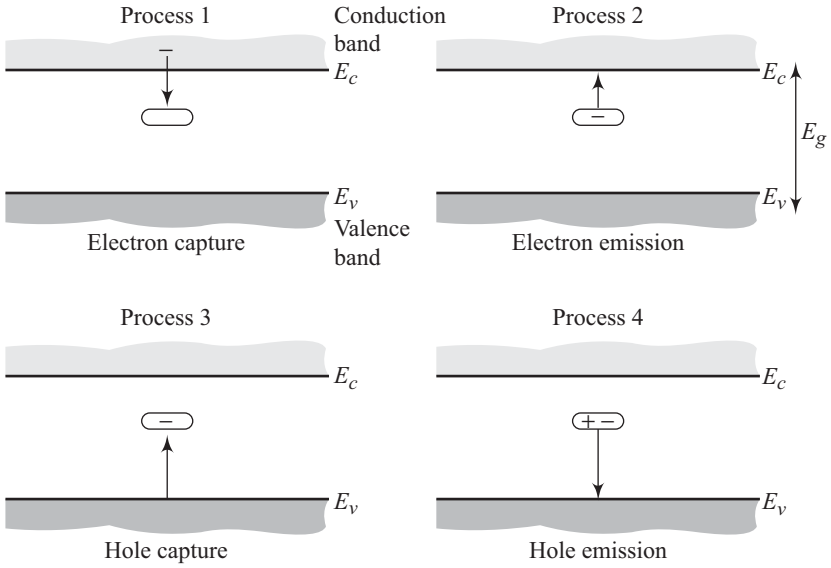


Fig. 4.6 Schematic representation of the four basic trapping and emission processes for the case of an acceptor-type trap

Process 4 It involves emission of a hole from a neutral trap into the valence band. This process can also be viewed as a capture of a valence band electron.

The rate at which electrons in the conduction band are captured by the traps (process 1) is proportional to the density of electrons in the conduction band and is also proportional to the density of empty trap states. If N_t is the concentration of trapping centres in the semiconductor, then the concentration of unoccupied trapping centres is given by $N_t\,[(1 - f_F\,(E_t)]$ where $f_F\,(E_t)$ is the Fermi function of the trap energy E_t. Thus the capture rate R_{cn} is given by

$$R_{cn} = C_n\,N_t\,[1 - f_F\,(E_t)]n \tag{4.25}$$

where n is the electron concentration and

$$f_F\,(E_t) = \frac{1}{1 + \exp\left(\dfrac{E_t - E_F}{kT}\right)} \tag{4.26}$$

In Eqn (4.25), C_n is a constant proportional to the electron-capture cross section. In fact, $C_n = v_{\text{th}}\,\sigma_n$ where, v_{th} is the thermal velocity of the carriers and σ_n is the trap-capture cross section. A degeneracy factor of 1 has been assumed in Eqn (4.26).

Similarly, the rate R_{en} with which electrons are emitted from the filled traps (process 2) is given by

$$R_{en} = E_n\,N_t f_F\,(E_t) \tag{4.27}$$

where the proportionality constant E_n is called *emission probability*.
R_{cn} in Eqn (4.25) and R_{en} in Eqn (4.27) have units of $\text{cm}^{-3}\ \text{s}^{-1}$.

At thermal equilibrium, we must have

$$R_{cn} = R_{en} \tag{4.28}$$

Using Eqs (4.25) and (4.27) in Eqn (4.28) yields

$$E_n N_t f_{F0}(E_t) = C_n N_t [1 - f_{F0}(E_t)] n_0 \tag{4.29}$$

where f_{F0} is the thermal-equilibrium Fremi function and n_0 is thermal-equilibrium electron concentration given by

$$n_0 = N_c e^{-(E_c - E_F)/kT} \tag{4.30}$$

Using Eqn (4.30) in Eqn (4.29) yields

$$E_n N_t f_{F0}(E_t) = C_n N_t [1 - f_{F0}(E_t)] N_c e^{-(E_c - E_F)/kT} \tag{4.31}$$

the Fermi function for energy E is given by

$$f_F(E) = \frac{1}{1 + e^{(E - E_F)/kT}} \tag{4.32}$$

For $(E - E_F) \gg kT$, Eqn (4.32) can be rewritten as

$$f_F(E) \approx e^{-(E - E_F)/kT} \tag{4.33}$$

Equation (4.33) is known as the Maxwell–Boltzmann approximation. Using Eqs (4.33) and (4.30) in Eqn (4.31) results in

$$E_n = N_c e^{-(E_c - E_t)/kT} C_n \tag{4.34}$$

which can be written in the form

$$E_n = n_t C_n \tag{4.35}$$

where

$$n_t = N_c e^{-(E_c - E_t)/kT} \tag{4.36}$$

Clearly, n_t is equivalent to the electron concentration existing in the conduction band if the trap energy level E_t coincide with the Fermi energy level E_F.

Under non-equilibrium conditions, $R_{cn} \neq R_{en}$ and the net rate of capture of electrons is then

$$R_n = R_{cn} - R_{en} \tag{4.37}$$

Using Eqs (4.25) and (4.27) in Eqn (4.37) yields

$$R_n = \{C_n N_t n[1 - f_F(E_t)]\} - [E_n N_t f_F(E_t)] \tag{4.38}$$

where n is the total electron concentration and Fermi energy in the Fermi function would have to be replaced with quasi-Fermi energy, which we shall discuss later on in this chapter. At this stage, it is sufficient to recollect that Fermi energy can no longer be defined under non-equilibrium conditions. The corresponding equivalent energies are called quasi-Fermi levels and are different for electrons and holes. Using Eqn (4.35) in Eqn (4.38) leads to

$$R_n = C_n N_t \{n[1 - f_F(E_t)] - n_t f_F(E_t)\} \tag{4.39}$$

considering processes 3 and 4 from Fig. (4.6) an equation similar to Eqn (4.39) can be derived for the rate of capture of holes. The final equation is

$$R_p = C_p N_t \{ p f_F(E_t) - p_t [1 - f_F(E_t)]\} \tag{4.40}$$

where

$$p_t = N_v \exp\left[-\frac{E_t - E_v}{kT}\right] \tag{4.41}$$

If the trap density is not very large, the excess electron and excess hole concentrations can be considered nearly equal. Under these circumstances, $R_n = R_p$. Equating Eqs (4.39) and (4.40) leads to

$$C_n N_t \{n[1 - f_F(E_t)] - n_t f_F(E_t)\} = C_p N_t \{p f_F(E_t) - p_t [1 - f_F(E_t)]\}$$

yielding

$$f_F(E_t) = \frac{C_n n + C_p p_t}{C_n(n + n_t) + C_p(p + p_t)} \tag{4.42}$$

Using Eqs (4.36) and (4.41), we can write

$$n_t p_t = N_c N_v \exp\left[-\frac{(E_c - E_v)}{kT}\right] = n_i^2 \tag{4.43}$$

Using Eqs (4.43) and (4.42) in Eqn (4.39) or (4.40) leads to

$$R_n = R_p = \frac{C_n C_p N_t (np - n_i^2)}{C_n(n + n_t) + C_p(p + p_t)} = R \tag{4.44}$$

where R is the common recombination rate for electrons and holes.

At thermal equilibrium $np = n_0 p_0 = n_i^2$, and Eqn (4.44) implies

$$R_n = R_p = R = 0$$

Drawing a parallel with Eqs (4.23) and (4.24), we may write Eqn (4.44) in the form

$$R = \frac{\delta n}{\tau} \tag{4.45}$$

where δn represents the excess carrier concentration and τ, the mean excess carrier lifetime.

4.3.2 Low Injection

For an n-type semiconductor under low injection, we have

$$n_0 \gg p_0, \quad n_0 \gg \delta p, \quad n_0 \gg n_t, \quad n_0 \gg p_t \tag{4.46}$$

where δp is the excess hole concentration. The assumptions $n_0 \gg n_t$ and $n_0 \gg p_t$ are valid if the trap energy level is close to the mid gap so that n_t and p_t are not much different from the intrinsic carrier concentration. Under the low-injection conditions given in Eqn (4.46), Eqn (4.44) reduces to

$$R = C_p N_t \delta p \tag{4.47}$$

Thus the recombination rate of excess carriers in the n-type semiconductor is proportional to the parameter C_p which is dependent upon the minority carrier hole-capture cross section σ_p. Comparing Eqs (4.45) and (4.47) leads to

$$R = \frac{\delta n}{\tau} = C_p N_t \delta p = \frac{\delta p}{\tau_{p0}} \tag{4.48}$$

where τ_{p0} is the excess minority carrier hole lifetime and is given by

$$\tau_{p0} = \frac{1}{C_p N_t} \tag{4.49}$$

From Eqn (4.49) we can conclude that the minority carrier lifetime decreases with increasing N_t. This is only to be expected, because the probability of excess carrier recombination increases as the trap concentration increases.

A similar treatment for a *p*-type semiconductor under low injection assumes

$$p_0 \gg n_0, \quad p_0 \gg \delta n, \quad p_0 \gg n_t, \quad p_0 \gg p_t \tag{4.50}$$

This leads to

$$\tau_{n0} = \frac{1}{C_n N_t}$$

Thus the excess-carrier lifetime for an extrinsic semiconductor reduces to the minority-carrier lifetime under low-injection conditions. *As we move from extrinsic to intrinsic semiconductors, the number of majority carriers available to recombine with excess minority carriers decreases, leading to an increase in the mean lifetime.*

Another type of recombination process is called the Auger recombination process. It is a non-radiative recombination process. Two variants of Auger recombination process are shown in Fig. 4.7.

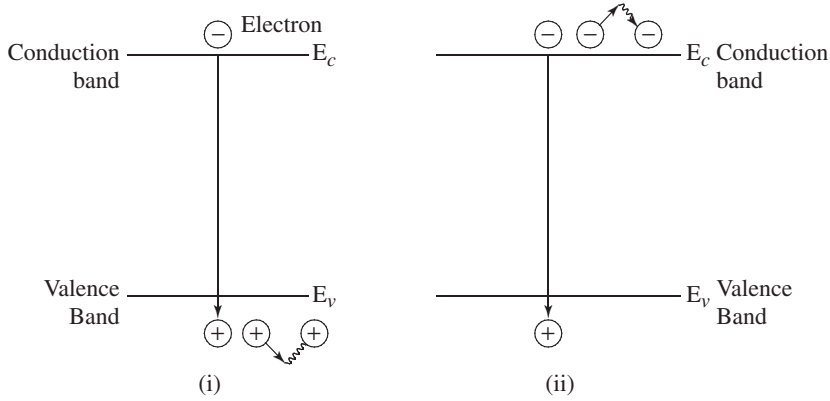


Fig. 4.7 Two variants of Auger recombination process

Figure 4.7 (i) shows a process in which an electron and a hole recombine and the resultant energy is transferred to another free hole. The second free hole ultimately loses its energy to the lattice releasing heat. The variant (ii) shows the recombination of an electron and a hole and the resultant energy being transferred to another free electron. This electron would eventually lose its energy to the lattice in the form of heat. Auger recombination becomes important for direct band gap materials at high doping concentrations. Variant (i) would occur in heavily doped *p*-type material, whereas variant (ii) would occur in heavily doped *n*-type material.

- The capture cross section of a trap is a measure of how close the electron has to come to the trap to be captured.
- Steady state does not imply equilibrium.
- The term $pn - n_i^2$ can be identified to be the *driving force* for recombination.

4.4 Diffusion and Recombination

We discussed diffusion in a semiconductor due to a concentration gradient in Chapter 3. Where, we neglected recombination effects in treatment. We will now present a comprehensive discussion on the conduction process.

4.4.1 Continuity Equation

Let us consider a differential volume element of side *dx*, *dy*, and *dz* as shown in Fig. 4.8. Also assume that a hole flux ϕ_{px}^{+} enters the differential element at *x* and leaves it at $(x + dx)$.

ϕ_{px}^{+} has units of holes/cm^2 s. The *x*-component of the particle current density has a dependence of the form

$$\phi_{px}^{+}(x + dx) = \phi_{px}^{+}(x) + \frac{\partial \phi_{px}^{+}}{\partial x} dx \tag{4.51}$$

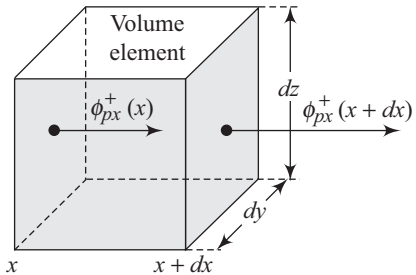


Fig. 4.8 Differential volume indicating the component of hole-particle flux

In writing Eqn (4.51) we have assumed the element *dx* to be a small quantity and therefore used the first two terms of the Taylor expansion of $\phi_{px}^{+}(x + dx)$.

The net rate of increase in holes within the volume element is given by

$$\frac{\partial p}{\partial t} dx\, dy\, dz = [\phi_{px}^{+}(x) - \phi_{px}^{+}(x + dx)] dy\, dz = \frac{-\partial \phi_{px}^{+}}{\partial x} dx\, dy\, dz \tag{4.52}$$

Thus the hole concentration within the elemental volume element increases when the incoming hole flux is more than the outgoing hole flux. Equation (4.52) does not take into account the generation and recombination rates. Let us assume that the generation rate of holes is represented by G_p, and τ_p represents the hole lifetime. Taking generation and recombination into account, Eqn (4.52) can be rewritten in the form

$$\frac{\partial p}{\partial t} dx\, dy\, dz = \frac{-\partial \phi_{px}^{+}}{\partial x} dx\, dy\, dz + G_p\, dx\, dy\, dz - \frac{p}{\tau_p} dx\, dy\, dz \tag{4.53}$$

Dividing both sides of Eqn (4.53) by *dx dy dz*, we get

$$\frac{\partial p}{\partial t} = \frac{-\partial \phi_p^{+}}{\partial x} + G_p - \frac{p}{\tau_p} \tag{4.54}$$

Similarly, for electrons we can write

$$\frac{\partial n}{\partial t} = \frac{-\partial \phi_n^-}{\partial x} + G_n - \frac{n}{\tau_n} \tag{4.55}$$

Equations (4.54) and (4.55) are called continuity equation for holes and electrons, respectively. The charge carrier flux is expressed as the number of charge carriers/cm^2 s. In Chapter 3 we derived the hole and electron current densities to be

$$J_p = e\mu_p pE - eD_p \frac{\partial p}{\partial x} \tag{4.56}$$

and

$$J_n = e\mu_n nE + eD_n \frac{\partial n}{\partial x} \tag{4.57}$$

The hole flux ϕ_p and electron flux ϕ_n are then

$$\phi_p^+ = \frac{J_p}{(+e)} = \mu_p pE - D_p \frac{\partial p}{\partial x} \tag{4.58}$$

and

$$\phi_n^- = \frac{J_n}{(-e)} = -\mu_n nE - D_n \frac{\partial n}{\partial x} \tag{4.59}$$

From Eqs (4.58) and (4.59) we obtain

$$\frac{\partial \phi_p}{\partial x} = \mu_p \frac{\partial (pE)}{\partial x} - D_p \frac{\partial^2 p}{\partial x^2} \tag{4.60}$$

and

$$\frac{\partial \phi_n}{\partial x} = -\mu_n \frac{\partial (nE)}{\partial x} - D_n \frac{\partial^2 n}{\partial x^2} \tag{4.61}$$

Substituting the expressions for $\partial\phi_p/\partial x$ and $\partial\phi_n/\partial x$ from Eqs (4.60) and (4.61) into Eqs (4.54) and (4.55) yields

$$\frac{\partial p}{\partial t} = -\mu_p \frac{\partial (pE)}{\partial x} + D_p \frac{\partial^2 p}{\partial x^2} + G_p - \frac{p}{\tau_p} \tag{4.62}$$

and

$$\frac{\partial n}{\partial t} = +\mu_n \frac{\partial (nE)}{\partial x} + D_n \frac{\partial^2 n}{\partial x^2} + G_n - \frac{n}{\tau_n} \tag{4.63}$$

We also have

$$\frac{\partial (pE)}{\partial x} = p\frac{\partial E}{\partial x} + E\frac{\partial p}{\partial x} \tag{4.64}$$

and

$$\frac{\partial (nE)}{\partial x} = n\frac{\partial E}{\partial x} + E\frac{\partial n}{\partial x} \tag{4.65}$$

Substituting Eqs (4.64) and (4.65) into Eqs (4.62) and (4.63) leads to

$$\frac{\partial p}{\partial t} = D_p \frac{\partial^2 p}{\partial x^2} - \mu_p \left(E\frac{\partial p}{\partial x} + p\frac{\partial E}{\partial x} \right) + G_p - \frac{p}{\tau_p} \tag{4.66}$$

and

$$\frac{\partial n}{\partial t} = D_n \frac{\partial^2 n}{\partial x^2} + \mu_n \left(E\frac{\partial n}{\partial x} + n\frac{\partial E}{\partial x} \right) + G_n - \frac{n}{\tau_n} \tag{4.67}$$

Equations (4.66) and (4.67) are then the time-dependent diffusion equation for holes and electrons, respectively. The carrier concentrations n and p include the thermal-equilibrium concentrations n_0 and p_0 and the excess concentrations. For a homogeneous semiconductor, n_0 and p_0 are independent of position, whereas the excess concentrations δn and δp can be position dependent. We can then write

$$\frac{\partial(\delta p)}{\partial t} = D_p \frac{\partial^2(\delta p)}{\partial x^2} + \mu_p \left(\frac{E\partial(\delta n)}{\partial x} + p\frac{\partial E}{\partial x} \right) + G_p - \frac{p}{\tau_p} \tag{4.68}$$

and,

$$\frac{\partial(\delta n)}{\partial t} = D_n \frac{\partial^2(\delta n)}{\partial x^2} + \mu_n \left(\frac{E\partial(\delta n)}{\partial x} + n\frac{\partial E}{\partial x} \right) + G_n - \frac{n}{\tau_n} \tag{4.69}$$

4.4.2 Haynes–Shockley Experiment

J.R. Haynes and W. Shockley carried out a classic experiment at the Bell Telephone Laboratories in 1951. This experiment can be used to evaluate minority carrier mobility μ and diffusion coefficient D independently. Figure 4.9 illustrates the basic principle behind the experiment.

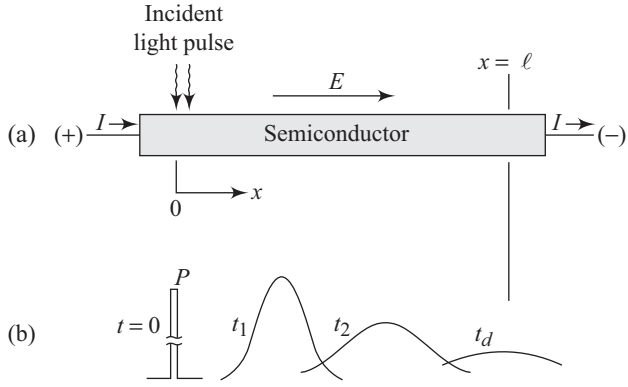


Fig. 4.9 Drift and diffusion of a hole pulse in an *n*-type semiconductor bar: (a) sample geometry; (b) position and shape of the pulse for different times during its drift along the bar

A pulse of holes is created at $x = 0$ using a light flash in an *n*-type semiconductor (remember $n_0 \gg p_0$ for the *n*-type semiconductor). The *n*-type semiconductor has an externally applied electric field E. The pulse of holes drifts in the field and spreads out by diffusion. The excess hole concentration is then monitored at another point $x = \ell$ using the set-up shown in Fig. 4.10(a).

The carrier concentration is monitored on an oscilloscope using a suitable detector such as a reverse-biased junction (we will study a *p*-*n* junction in Chapter 5). We assume that the electron concentration changes negligibly, whereas a significant change takes place in the hole concentration.

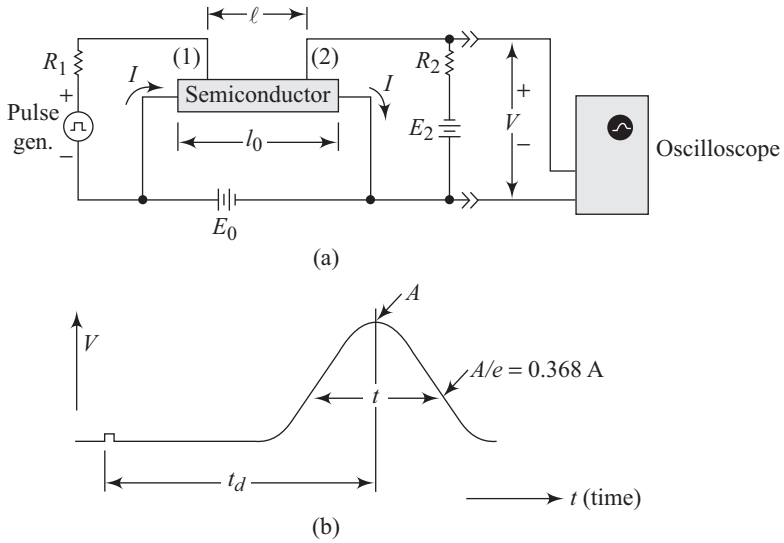


Fig. 4.10 (a) Schematic diagram of Haynes–Shockley experiment, (b) A typical oscilloscope trace

If t_d is the drift time for the hole pulse to reach $x = \ell$, we can write for drift velocity v_d,

$$v_d = \frac{\ell}{t_d} \tag{4.70}$$

The hole mobility μ_p is given by

$$\mu_p = \frac{v_d}{E} \tag{4.71}$$

One must at this stage mention that the mobility evaluated from Eqn (4.71) is the *minority carrier mobility*, whereas the *Hall effect* described in Chapter 3 is used to obtain the *majority* carrier mobility.

Neglecting drift, recombination (τ_p is long compared with time involved in diffusion), and also the generation term, the time dependence of excess carrier concentration is given by

$$\frac{\partial}{\partial t}\delta p(x,t) = D_p \frac{\partial^2}{\partial x^2}\delta p(x,t) \tag{4.72}$$

Equation (4.72) has solutions of the form

$$\delta p(x,t) = \left[\frac{\Delta p}{2\sqrt{\pi D_p t}}\right] e^{-x^2/(4D_p t)} \tag{4.73}$$

where Δp represents the number of holes per unit area created over a negligibly small distance at $t = 0$. Thus the peak value of the pulse (at $x = 0$) decreases with time, as indicated by the term within the brackets in (4.73). The spread of the pulse in the positive and negative x-directions is expressed in the exponential factor of Eqn (4.73) and is plotted in Fig. 4.11.

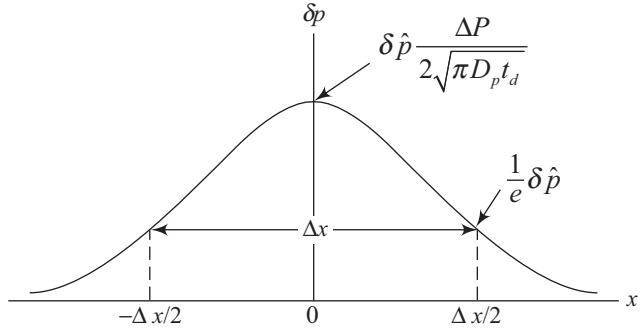


Fig. 4.11 Calculation of D_p from the shape of the δp distribution after time t_d. Drift or recombination is neglected

If $\delta \hat{p}$ is the peak value at time t_d, then at $(\Delta x/2)$ δp becomes $1/e$ of its peak value $\delta \hat{p}$. Using Eqn (4.73), we can write

$$\frac{\delta \hat{p}}{e} = \delta \hat{p}\, e^{-(\Delta x/2)^2/4D_p t_d} \tag{4.74}$$

which leads to

$$D_p = \frac{(\Delta x)^2}{16 t_d} \tag{4.75}$$

The set-up described in Fig. 4.9(a) is used to display the detected pulse as shown in Fig. 4.9(b). The pulse width Δt is measured. Using the calculated value of v_d, Δx is given by

$$\Delta x = \Delta t\, v_d = \Delta t \left(\frac{\ell}{t_d} \right) \tag{4.76}$$

The diffusion coefficient is then calculated using Eqn (4.75).

4.5 Quasi-Fermi Energy Levels

We have already seen that the thermal-equilibrium electron and hole concentration are given by

$$n_0 = n_i \exp\left(\frac{E_F - E_{Fi}}{kT} \right) \tag{4.77}$$

and

$$p_0 = n_i \exp\left(\frac{E_{Fi} - E_F}{kT} \right) \tag{4.78}$$

where n_i is the intrinsic carrier concentration. E_F is the Fermi energy, and E_{Fi} is the intrinsic Fermi energy. Typical energy band diagram for *n*- and *p*-type semiconductors are shown schematically in Fig. 4.12.

From Fig. 4.12 we can see that for the *n*-type semiconductor, $E_F > E_{Fi}$. This according to Eqs (4.77) and (4.78) leads to

$$n_0 > n_i \quad \text{and} \quad p_0 < n_i \tag{4.79}$$

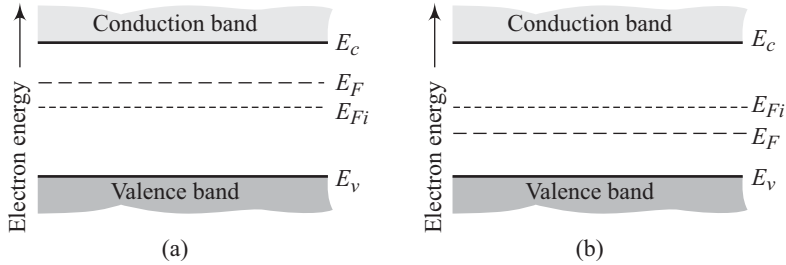


Fig. 4.12 Thermal-equilibrium energy band diagrams for (a) n-type semiconductor and (b) p-type semiconductor

Similarly, for the *p*-type semiconductor, $E_F < E_{Fi}$ and we have

$$p_0 > n_i \quad \text{and} \quad n_0 < n_i \tag{4.80}$$

When excess carriers are generated, the semiconductor is no longer in thermal equlibirium. The concept of Fermi energy is now not strictly defined. Instead we now define the quasi-Fermi level for electrons and hole according to equations

$$n_0 + \delta n = n_i \exp\left(\frac{E_{Fn} - E_{Fi}}{kT}\right) \tag{4.81}$$

and

$$p_0 + \delta p = n_i \exp\left(\frac{E_{Fi} - E_{Fp}}{kT}\right) \tag{4.82}$$

where δn and δp are the excess electron and hole concentrations, respectively, and E_{Fn} and E_{Fp} are the quasi-Fermi energy levels for electrons and holes, respectively. The majority carrier concentration does not change appreciably under low-injection conditions. The quasi-Fermi level associated with the majority carriers is not much different from the thermal equilibrium Fermi level. The quasi-Fermi energy level for the minority carriers, however, differs considerably from the thermal-equilibrium Fermi energy level.

From Eqn (4.57) we have,

$$J_n(x) = e\mu_n n(x)\, E(x) + eD_n \frac{dn}{dx} \tag{4.83}$$

From Eqn (4.81) we can write,

$$\frac{dn}{dx} = \frac{d}{dx}\left[n_i \exp\left(\frac{E_{Fn} - E_{Fi}}{kT}\right)\right]$$

resulting in

$$\frac{dn}{dx} = \frac{n_i}{kT}\left[\frac{dF_{Fn}}{dx} - \frac{dE_{Fi}}{dx}\right]\exp\left(\frac{E_{Fn} - E_{Fi}}{kT}\right)$$

which on using Eqn (4.81) can be written in the form,

$$\frac{dn}{dx} = \frac{n(x)}{kT}\left[\frac{dF_{Fn}}{dx} - \frac{dE_{Fi}}{dx}\right] \tag{4.84}$$

Using Eqn (4.84) in Eqn (4.83) yields,

$$J_n(x) = e\mu_n n(x)E(x) + \frac{eD_n n(x)}{kT}\left[\frac{dF_{Fn}}{dx} - \frac{dE_{Fi}}{dx}\right] \tag{4.85}$$

Einstein's relation given by Eqn (3.103) is

$$\frac{eD_n}{kT} = \mu_n \tag{4.86}$$

Using Eqn (4.86) in Eqn (4.85) results in,

$$J_n(x) = e\mu_n n(x)E(x) + \mu_n n(x)\left[\frac{dE_{Fn}}{dx} - \frac{dE_{Fi}}{dx}\right] \tag{4.87}$$

$E(x)$ can be written as

$$E(x) = \frac{-dv(x)}{dx} \tag{4.88}$$

where $v(x)$ is the electrostatic potential. Using E_{Fi} as a reference, we can rewrite Eqn (4.88) as,

$$E(x) = \frac{-d}{dx}\left[\frac{E_{Fi}}{(-e)}\right] = \frac{1}{e}\frac{dE_{Fi}}{dx} \tag{4.89}$$

Using Eqs (4.89) and (4.87) leads to,

$$J_n(x) = \mu_n n(x)\frac{dE_{Fi}}{dx} + \mu_n n(x)\frac{dE_{Fn}}{dx} - \mu_n n(x)\frac{dE_{Fi}}{dx}$$

yielding,

$$J_n(x) = \mu_n n(x)\frac{dE_{Fn}}{dx} \tag{4.90}$$

Thus the spatial variation of the quasi-Fermi level leads to the current density due to electron drift and diffusion.

4.6 Surface Effects

In our discussion on semiconductors, we have so far neglected the presence of surfaces. We have thus assumed the semiconductors to be infinite in extent. In actual applications, surfaces do exist between semiconductors and the adjacent medium.

4.6.1 Surface States

The perfect periodic arrangement of an ideal single-crystal lattice gets abruptly terminated at the surface. This disruption in periodicity results in the creation of allowed-electronic-energy states within the forbidden energy gap. Thus, at the surface of a semiconductor, a distribution of allowed-energy states exists along with the discrete energy states within the band gap present in the bulk of the semiconductor. This is shown schematically in Fig. 4.13.

We have dealt with the Shockley–Read–Hall recombination in an earlier section of this chapter. According to this theory, the excess minority carrier lifetime is inversely proportional to the density of trap states [see Eqn (4.49)]. The density of traps at the surface is higher than that in the bulk, therefore *the excess minority carrier lifetime at the surface is smaller than the corresponding bulk lifetime*. The recombination rate of excess hole in the bulk of an extrinsic *n*-type semiconductor is given by

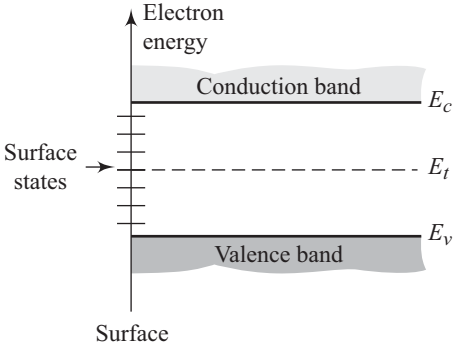


Fig. 4.13 Distribution of surface states within the forbidden band gap

$$R = \frac{\delta p}{\tau_{p0}} = \frac{\delta p_B}{\tau_{p0}} \tag{4.91}$$

where δp_B is the excess minority carrier hole concentration in the bulk of the *n*-type semiconductor. The corresponding recombination rate R_S at the surface is given by

$$R_S = \frac{\delta p_S}{\tau_{p0S}} \tag{4.92}$$

where δp_S and τ_{p0S} are the excess minority carrier hole concentration and minority carrier hole lifetime at the surface, respectively.

Let us assume that the generation rate of excess carriers is constant throughout the semiconductor. At steady state, for a homogeneous and infinite semiconductor, the generation rate has been shown to be equal to the recombination rate. Thus, for a constant generation rate we must have

$$R = R_S \tag{4.93}$$

Since $\tau_{p0S} < \tau_{p0}$, from Eqs (4.93) and (4.92) we can conclude

$$\delta p_S < \delta p_B \tag{4.94}$$

The excess-carrier concentration at the semiconductor surface is lower than that in the bulk as shown is Fig. 4.14.

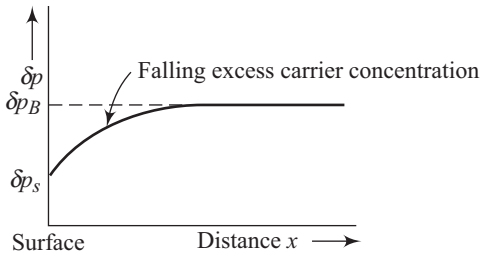


Fig. 4.14 Steady state excess hole concentration versus distance from a typical semiconductor surface

4.6.2 Surface Recombination Velocity

An excess carrier concentration gradient exists near the surface of a semiconductor as shown in Fig. 4.13. Excess carriers diffuse towards the surface under the influence of this gradient. This diffusion towards the surface can be described using the equation

$$-D_p\left[\frac{\hat{n}\cdot d(\delta p)}{dx}\right]\Bigg|_S = s\delta p\Big|_S \tag{4.95}$$

where $\hat{n}$ is a unit vector directed outward and normal to the surface, and s is called the *surface recombination velocity* and has units of cm/s. For the geometry sketched in Fig. 4.14, $\hat{n}$ is along the $-x$ direction and is therefore negative, $[d(\delta p)]/dx$ is along the $+x$ direction. The surface recombination velocity is therefore positive. With $\hat{n} = -1$, Eqn (4.95) simplifies to

$$D_p \frac{d(\delta p)}{dx}\Bigg|_S = s\delta p\Big|_S \tag{4.96}$$

Two special cases deserve a mention at this stage.

Case A $\boldsymbol{s = 0}$ The excess minority carrier concentration at the surface is the same as that in the bulk. The surface has no effect.

Case B $\boldsymbol{s = \infty}$ The excess minority carrier concentration at the surface is zero. The excess minority carrier lifetime at the surface is also zero.

- The transition of an electron from the conduction band to the valence band is possible by way of emission of a photon (radiative process) or through the transfer of energy to another free carrier (Auger process).
- Gold in silicon is an example of an efficient recombination centre.
- $L = \sqrt{D\tau}$ where L is called the diffusion length.

Solved Problems

4.1 Excess electrons are generated in a semiconductor to a concentration of $\delta n(0) = 10^{16}$ cm^{-3}. The excess carrier lifetime in the semiconductor is 5×10^{-6} s. The source generating the excess carriers is switched off at $t = 0$. Calculate the excess electron concentration at $t = 1$ μs.

Solution

The excess electron concentration at any time t is given by

$$\delta n(t) = \delta n(0)\, e^{-t/\tau_{n0}} \tag{4.1.1}$$

Putting the given values of $\delta n(0)$ and τ_{n0} in the above equation, we get

$$\delta n(t) = 10^{16} \times e^{-1/5} = 8.19 \times 10^{15} \text{ cm}^{-3}$$

4.2 For the situation presented in solved numerical 4.1, calculate the recombination rate of excess electrons for $t = 5\ \mu s$.

Solution

The excess electron concentration $\delta n(t)$ at $t = 5\ \mu s$ can be evaluated using the Eqn,

$$\delta n\ (t) = \delta n\ (0)\ e^{-t/\tau_{n0}} \tag{4.2.1}$$

Putting the given values in Eqn (4.2.1), we get

$$\delta n\ (5\ \mu s) = 10^{16}\ e^{-1} = 3.68 \times 10^{15} \tag{4.2.2}$$

The recombination rate R_n' for excess electrons is given by

$$R_n' = \frac{\delta n(t)}{\tau_{n0}} \tag{4.2.3}$$

Putting $\delta n\ (5\ \mu s)$ from Eqn (4.2.2) and the given value of τ_{n0} in Eqn (4.2.3) yields

$$R_n' = \frac{3.68 \times 10^{15}}{5 \times 10^{-6}} = 0.74 \times 10^{21}\ \text{cm}^{-3}\ \text{s}^{-1}$$

4.3 A sample of silicon maintained at $T = 300$ K has dopant concentrations of $N_d = 10^{15}$ cm^{-3} and $N_a = 0$. Assume $\tau_{p0} = \tau_{n0} = 10 \times 10^{-7}$ s and $n_t = p_t = n_i$. Calculate the recombination rate of excess carriers if $\delta n = \delta p = 10^{14}$ cm^{-3}.

Solution

Recombination rate R is given by

$$R = \frac{C_n C_p N_t (np - n_i^2)}{C_n (n + n_t) + C_p (p + p_t)} \tag{4.3.1}$$

Putting $\tau_{p0} = 1/C_p N_t$ and $\tau_{n0} = 1/C_n N_t$ in Eqn (4.3.1) yields

$$R = \frac{(np - n_i^2)}{\tau_{p0}(n + n_t) + \tau_{n0}(p + p_t)} \tag{4.3.2}$$

Here

$$n = (n_0 + \delta n) \text{ and } p = (p_0 + \delta p)$$

Thus

$$(np - n_i^2) = [(n_0 + \delta n)\ (p_0 + \delta p) - n_i^2\ ]$$

This yields

$$np - n_i^2 = n_0 p_0 + n_0 \delta p + \delta n p_0 - n_i^2 = n_0 \delta p + \delta n p_0 + \delta n \delta p + \delta n \delta p \tag{4.3.3}$$

Also,

$$n_0 \cong N_d = 10^{15}, p_0 = \frac{n_i^2}{N_d} = \frac{(1.5 \times 10^{10})^2}{10^{15}} = 2.25 \times 10^5$$

and

$$n = n_0 + \delta n = 10^{15} + 10^{14} = 1.1 \times 10^{15},\ p = p_0 + \delta p = 2.25 \times 10^5 + 10^{14} \cong 10^{14} \tag{4.3.4}$$

Using Eqn (4.3.4) in Eqn (4.3.3) results in

$$np - n_i^2 = 10^{15} \times 10^{14} + 10^{14} \times 2.25 \times 10^5 + 10^{28}$$

which gives

$$np - n_i^2 \cong 10^{29} + 10^{28} = 1.1 \times 10^{29} \tag{4.3.5}$$

Putting Eqn (4.3.5) in Eqn (4.3.2) results in

$$R = \frac{1.1 \times 10^{29}}{10 \times 10^{-7}(1.1 \times 10^{15} + 1.5 \times 10^{10}) + 10 \times 10^{-7}(10^{14} + 1.5 \times 10^{10})}$$

This leads to

$$R \cong \frac{1.1 \times 10^{29}}{10 \times 10^{-7}(1.1 \times 10^{15} + 10^{14})} = \frac{1.1 \times 10^{29}}{10 \times 10^{-7}(1.2 \times 10^{15})}$$

implying

$$R = 9.17 \times 10^{19} \text{ cm}^{-3}\text{ s}^{-1}.$$

4.4 A particular sample of an *n*-type semiconductor at T = 300 K has a carrier concentration $n_0 = 5 \times 10^{15}$ cm^{-3}, $n_i = 10^{10}$ cm^{-3}, and $p_0 = 2 \times 10^4$ cm^{-3}. A source creates excess carriers in the semiconductor such that $\delta n = \delta p = 5 \times 10^{13}$ cm^{-3}. Calculate the position of the Fermi level for thermal equilibrium and positions of quasi-Fermi levels for electrons and holes at non-equilibrium conditions with respect to the intrinsic Fermi level.

Solution.

The position of the Fermi level at thermal equilibrium is given by

$$E_F - E_{Fi} = kT \ln\left(\frac{n_0}{n_i}\right) \tag{4.4.1}$$

Putting $kT = 0.026$ eV and the given values of n_0 and n_i in Eqn (4.4.1) leads to

$$E_F - E_{Fi} = 0.026 \ln\left(\frac{5 \times 10^{15}}{10^{10}}\right)$$

This yields

$$E_F - E_{Fi} = 0.3412 \text{ eV}$$

The quasi-Fermi level for electrons in non-equilibrium is given by

$$E_{Fn} - E_{Fi} = kT \ln\left(\frac{n_0 + \delta n}{n_i}\right) \tag{4.4.2}$$

where $\delta n = 5 \times 10^{13}$ cm^{-3}. Thus, Eqn (4.4.2) results in

$$E_{Fn} - E_{Fi} = 0.026 \ln\left(\frac{5 \times 10^{15} + 5 \times 10^{13}}{10^{10}}\right)$$

which implies

$$E_{Fn} - E_{Fi} = 0.3414 \text{ eV}$$

One can notice that E_{Fn} is not much different from E_F. This is because δn does not change the total electron concentration significantly.

The quasi-Fermi level for holes in non-equilibrium is given by

$$E_{Fi} - E_{Fp} = kT \ln\left(\frac{p_0 + \delta p}{n_i}\right) \tag{4.4.3}$$

with $\delta p = 5 \times 10^{13}$ cm^{-3}

Equation (4.4.3) thus yields

$$E_{Fi} - E_{Fp} = 0.026 \ln\left(\frac{2 \times 10^4 + 5 \times 10^{13}}{10^{10}}\right)$$

which implies

$$E_{Fi} - E_{Fp} = 0.2214 \text{ eV}$$

E_{Fp} is thus significantly different from E_{Fi}.

4.5 An *n*-type semiconductor sample is illuminated with light to create electron–hole pairs uniformly throughout the sample. The generation rate of EHP pairs is G cm^{-3} s^{-1}. Boundary conditions include $E = 0$ and $[\partial^2(\delta p)]/\partial x^2 = [\partial(\delta p)]/\partial x = 0$. Derive an expression for the excess minority carrier concentration dependence on time lapsed at steady state conditons. [This is the main idea behind the Stevenson–Keyes method for measuring minority-carrier lifetime. The method, however, incorporates a small variation by studying the photoconductivity decay.]

Solution

The time-dependent continuity equation for the situation presented is

$$G - \frac{\delta p}{\tau_{p0}} = \frac{d(\delta p)}{dt}$$

This equation has solution of the form

$$\delta p(t) = G\, \tau_{p0}\, (1 - e^{-t/\tau_{p0}})$$

This solution assumes a low-injection condition.

4.6 An experimental set-up based on Haynes–Shockley experiment uses an *n*-type Ge sample (see Fig. 4.9). The length of the sample is 2 cm and the probes 1 and 2 are kept at a distance of 1.8 cm. The source battery voltage E_0 is 3 V. The time it takes for a pulse injected at 1 to reach point 2 is 0.6 ms and the pulse width is $\Delta t = 236$ μs. Evaluate the hole mobility and diffusion coefficient for holes. Use the calculated values to check the validity of the Einstein relation.

Solution

Hole mobility μ_p is given by

$$\mu_p = \frac{v_d}{E} \tag{4.6.1}$$

where

$$v_d = \frac{1.8}{0.6 \times 10^{-3}} = 3000 \text{ cm/s}$$

Putting the value of v_d into Eqn (4.6.1) results in

$$\mu_p = \frac{3000}{3/2} = 2000 \text{ cm}^2/\text{ Vs} \tag{4.6.2}$$

Diffusion coefficient D_p can be calculated using the expression

$$D_p = \frac{(\Delta x)^2}{16 t_d} = \frac{(\Delta t \ell)^2}{16 t_d^3} \tag{4.6.3}$$

Putting the given values in Eqn (4.6.3) yields

$$D_P = \frac{(236 \times 10^{-6})^2 (1.8)^2}{16 \times (0.6 \times 10^{-3})^3} = 52.22 \text{ cm}^2/\text{s}$$

Also,

$$\frac{D_p}{\mu_p} = \frac{52.22}{2000} = 0.026 = \frac{kT}{q}$$

Thus, the Einstein relation is valid.

4.7 A silicon sample is doped with 10^{15} donors/cm^3 and has a hole-lifetime of 0.5 μs. Assuming all the donors to be ionized, determine:

(a) the photo generation rate, which will produce 4×10^{14} excess EHP in steady state.

(b) the sample resistivity before and after illumination and the percentage of conductivity due to minority carriers. Assume $\mu_n = 3\mu_p = 1200$ cm^2/Vs, T = 300 K.

(c) how far are the electron and hole populations from their equilibrium values while under illumination?

Solution

$$g_{op} \tau_p = \delta p = 4 \times 10^{14}/\text{cm}^3$$

$$g_{op} = 4 \times 10^{14}/0.5 \times 10^{-6} = 8 \times 10^{20} \text{ EHPs/cm}^3 \text{ s}$$

(b) Before illumination:

$$\rho_0 = (q \, \mu_n \, n_0)^{-1} = 5.21 \ \Omega \text{ cm}$$

After illumination:

$$n = n_0 + \delta n = 10^{15} + 4 \times 10^{14} = 1.4 \times 10^{15} \text{ cm}^{-3}$$

$$p = p_0 + \delta p = 4 \times 10^{14} \text{ cm}^{-3}$$

Now we know

$$\rho = \frac{1}{q(n\mu_n + p\mu_p)}$$

$$= \frac{1}{1.6 \times 10^{-19}[1.4 \times 10^{15} \times 1200 + 4 \times 10^{14} \times 400]}$$

$$= \frac{10^{19}}{1.6(1680 \times 10^{15} + 1600 \times 10^{14})}$$

$$= \frac{10^{19}}{1.6 \times 10^{14}(16800 + 1600)}$$

$$= \frac{10^{5}}{1.6 \times 18400}$$

$$= 3.396\ \Omega\ \text{cm}$$

Conductivity due to minority carriers = $(q\ \mu_p\ p_0)$

$$= 1.6 \times 10^{-19} \times 400 \times 4 \times 10^{14}$$

$$= 25.6 \times 10^{-3}\ \text{mho cm}^{-1}$$

% of conductivity = (Conductivity due to minority carriers/Total conductivity) $\times 100$

$$= \left(\frac{25.6 \times 10^{-3}}{1/3.396}\right) \times 100$$

$$= 8.69\%$$

(c) To find how far the electron and hole populations are from their equilibrium values, we have to find the position of quasi Fermi levels.

$$n = n_i \exp\left[\frac{E_{Fn} - E_{Fi}}{kT}\right]$$

where E_{Fn} is the position of the quasi Fermi level due to electrons.

$$\therefore \quad 1.4 \times 10^{15} = 1.5 \times 10^{10} \exp\left[\frac{(E_{Fn} - E_{Fi})}{0.0259}\right]$$

$$\Rightarrow \quad (E_{Fn} - E_{Fi}) = 0.0259 \ln\left[\frac{1.4 \times 10^{15}}{1.5 \times 10^{10}}\right]$$

$$E_{Fn} - E_{Fi} = 0.296\ \text{e V}$$

$$\Rightarrow \qquad E_{Fn} = E_{Fi} + 0.296\ \text{e V}$$

Now, to find the position of the quasi Fermi level corresponding to holes, we use the relationship

$$p = n_i \exp[(E_{Fi} - E_{Fp})/kT]$$

(where E_{Fp} is the position of the quasi Fermi level due to holes)

$$\therefore \quad 1.4 \times 10^{15} = 1.5 \times 10^{10} \exp[(E_{Fi} - E_{Fp})/0.0259]$$

$$\Rightarrow \quad E_{Fi} - E_{Fp} = 0.0259 \ln \frac{1.4 \times 10^{14}}{1.5 \times 10^{10}}$$

$$\Rightarrow \qquad E_{Fp} = E_{Fi} - 0.264\ \text{e V}$$

4.8 An *n*-type silicon sample with block geometry of unit volume is doped with 10^{16}cm^{-3} donors. At 300 K, under equilibrium conditions, this sample experiences thermal generation of EHPs at the rate of $2.25 \times 10^{10}\text{cm}^{-3}\text{s}^{-1}$. Assuming that all the donor atoms are ionized at the given temperature, the given sample is uniformly illuminated by a radiation to generate excess carriers at the rate of $10^{21}\ \text{cm}^{-3}\text{s}^{-1}$. Based on this information answer the questions that follow:

(a) Determine the lifetimes of both types of carriers, taking into consideration only thermal generation and assuming that only simple band-to-band recombination is allowed.
(b) What is the order of concentration of minority carriers under equilibrium and steady state? Are they equal? Explain. Assume carrier lifetime due to optical generation to be equal to carrier lifetime due to thermal generation.
(c) This sample is to be used as a photoconductor in which the optically induced change in current ΔI, is dominated only by the mobility μ_n, and lifetime for electrons. If the transit time of electrons drifting down the length of the bar is 2.5 ms then find the value of ΔI for this semiconductor.
(d) Does the value of ΔI obtained in (c) affect the steady state value of the current density in the semiconductor?

Solution

Here $N_d = 10^{16}$ cm^{-3} (Donors are added to silicon to make it *n*-type.)
Volume of the sample = 1 cm^3

(a) The sample is maintained at 300 K under equilibrium, therefore
$G(T) = G(300\text{ K})$ = Thermal generation rate of carriers under equilibrium = 2.25×10^{10}cm^{-3} s^{-1} (This is the generation rate of thermal EHPs provided and not the concentration of thermally generated EHPs at 300 K)
Now under equilibrium

$$G(T) = \alpha_r n_i^2 \quad (\text{where } n_i = 1.5 \times 10^{10}\text{cm}^{-3}) \tag{4.8.1}$$

$$\Rightarrow \quad 2.25 \times 10^{10} = \alpha_r (1.5 \times 10^{10})^2$$

$$\therefore \quad \alpha_r = 10^{-10}\text{ cm}^3/\text{s}$$

Hence $\alpha_r = 10^{-10}$cm^3/sec (this value of proportionality constant implies that to establish equilibrium, there are processes (such as scattering and velocity saturation) which try to preserve the intrinsic carrier concentration value at a given temperature against the thermal generation rate.)
Now, we have

$$\tau_p = \{\alpha_r (n_0 + p_0)\}^{-1} \tag{4.8.2}$$

Now semiconductor being *n*-type; $n_0 \gg p_0$.
Therefore, neglecting p_0 in Eqn (4.8.2)

$$\tau_p = \{\alpha_r n_0\}^{-1}$$

$$= \frac{1}{10^{-10} \times 10^{16}}$$

$$= 10^{-6}\text{ s}$$

$$= 1\ \mu\text{s}$$

(In the above expression we use $n_0 = 10^{16}$cm^{-3} because it is provided that all donors are ionized at 300 K.)
Now as only simple band-to-band recombination is allowed, we have

$$\tau_p = \tau_n$$

Hence lifetime of electrons in the sample is also 1 μs.

(b) $\delta n = \delta p = g_{op}\tau_n$ gives excess carrier concentration of 10^{15}cm^{-3} (As $\delta n = \delta p = 10^{21} \times 10^{-6} = 10^{15}$ cm^{-3})
Hence the equilibrium value of minority carriers is of the order of 10^4cm^{-3} while the steady state values are of the order of 10^{15}cm^{-3}.

(c) We know that optically induced change in conductivity of a semiconductor is given by

$$\Delta\sigma = q\delta n\mu_n + q\delta p\mu_p$$

Now, $\delta n = \delta p$ and $\mu_n >> \mu_p$ as given
Hence $\Delta\sigma = q\delta n\mu_n$
Now for a given potential across the semiconductor

$$\Delta J = (\Delta\sigma)E.$$

$$\therefore \quad \Delta J = q\delta n\, \mu_n E.$$

Now by definition of excess carriers

$$\delta n = g_{op}\tau_n$$

and drift velocity of electrons is given as

$$v_d = \mu_n E$$

$$\therefore \quad \Delta J = q\, g_{op}\tau_n v_d$$

Now $\Delta J = \Delta I/A$ implies that

$$\Delta I = A\, \Delta J$$

$$\therefore \quad \Delta I = qAg_{op}\tau_n v_d$$

Also for a given potential

$$v_d = L/\tau_t$$

where τ_t = transit time of electrons across the length of the block

$$\therefore \quad \Delta I = q\, Ag_{op}\tau_n\, L/\tau_t$$

The volume of block, $V = A \times L$ and therefore

$$\Delta I = qVg_{op}\tau_n/\tau_t \qquad (4.8.3)$$

Putting all the values in Eqn (4.8.3) gives

$$\Delta I = \frac{1.6 \times 10^{-19} \times 1 \times 10^{21} \times 10^{-6}}{2.5 \times 10^{-3}}$$

$$= 0.064 \text{ A}$$

(d) This change in value of current does affect the value of steady state current but not the equilibrium current which is used in part (c) above.

4.9 An experiment was conducted on a very small piece of silicon (E_g = 1.1 e V) of length 100 μm at room temperature. Energy band diagram for this silicon piece is shown in Fig. 4.9.1

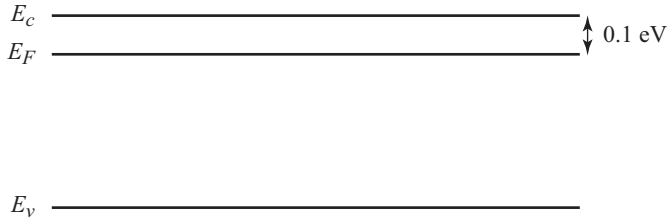


Fig. 4.9.1 Band diagram for the given silicon sample

(a) Engineers tried to apply a potential difference varying from about 1 V to about 10000 V across the length of this sample and observed corresponding values of current densities as recorded below in Table 4.9.1. Find the values of current density denoted by '??' field in the table.

Table 4.9.1 Variation of current density with voltage

Voltage (V)	1	10	100	1000	10000
Current density (A/cm^2)	16.96×10^3	16.96×10^4	67.84×10^4	8.48×10^5	??

(b) Electron mobility of an intrinsic piece of silicon at room temperature is given as 1400 cm^2/Vs. Can we take the same value of mobility for making some useful calculations for the silicon sample which has the energy band diagram as shown in Fig. 4.9.1? Justify your answer.

(c) The given silicon sample is uniformly excited at room temperature, such that 10^{19} EHPs are generated per cubic centimeter per second. How far are the electron and hole populations from their equilibrium values under the effect of illumination? (Given: electron and hole life times = 10 μs, D_p = 12 cm^2/s)

Solution

(a) At 10000 V the corresponding electric field is given as

$$E = 10000 \text{ V}/100 \text{ μm}$$
$$= 10^6 \text{ V/cm}$$

At such a high value of electric field, the corresponding value of v_d becomes saturated.

Now,

$$J = -q\, n\, v_d$$

Therefore, the current density for a 10000 V potential will be the same value as that for 1000 V.

$$\therefore \quad ?? = 8.48 \times 10^5 \text{ A/cm}^2$$

(b) From the energy band diagram given in Fig. 4.9.1 we can find the doping concentration as

$$n_0 = n_i \exp\left[\frac{E_F - E_i}{kT}\right]$$

$$= 1.5 \times 10^{10} \exp\left[\frac{0.1}{0.0259}\right]$$

$$= 5.3 \times 10^{17} \text{ cm}^{-3}$$

At such a high doping concentration, mobility gets drastically reduced due to the effect of impurity scatterings. For example, electron mobility of intrinsic silicon at 300 K is 1350 cm^2/V s. However, at donor doping concentration of 10^{17} cm^{-3}, mobility of electrons is only 700 cm^2/ V s.

(c) Using Eqs (4.79) and (4.80) we get

$$n = n_i e^{(E_{Fn} - E_{Fi})/kT} \text{ and } p = n_i e^{(E_{Fi} - E_{Fp})/kT} \tag{4.9.1}$$

Also, the excess carrier concentrations, δn and δp, are given as

$$\delta n = \delta p = g_{\text{op}} \tau$$

It is given that

$$g_{\text{op}} = 10^{19} \text{ cm}^{-3} \text{ s}^{-1} \text{ and } \tau = 10 \text{ μs}$$

$$\therefore \quad \delta n = \delta p = g_{\text{op}} \times \tau = (10^{19})(10^{-5})$$

$$= 10^{14} \text{ cm}^{-3} \text{ (which is less than the dopant concentration } = 5.3 \times 10^{17} \text{ cm}^{-3})$$

$$n = (5.3 \times 10^{17}) + 10^{14}$$

$$= 5.3 \times 10^{17} \text{ cm}^{-3}$$

$$p_0 + \delta p = n_i^2/n_0 + \delta p$$

$$= 2.25 \times 10^{20}/[(5.3 \times 10^{17})] + 10^{14} \sim 10^{14} \text{ cm}^{-3}$$

Now using Eqn (4.9.1) we get

$$E_{Fn} - E_{Fp} = kT \ln\left(\frac{np}{n_i^2}\right)$$

$$= 0.0259 \times \ln(5.3 \times 10^{17} \times 10^{14})/2.25 \times 10^{20}$$

$$= 0.6781 \text{ e V}$$

Recapitulation

- At thermal equilibrium,

  $$G_{n0} = G_{p0} = R_{n0} = R_{p0}$$

  where G represents thermal generation rates and R represents the recombination rates.
- For optical absorption, intensity at any depth $I(x)$ is given by

  $$I(x) = I_0 e^{-\alpha x}$$

  where α is called the absorption coefficient.
- When excess carriers are created in a semiconductor,

  $$np \neq n_0 p_0 = n_i^2$$
- Excess minority carrier decay is given by

  $$\delta n(t) = \delta n(0) e^{-t/\tau_{n0}}$$

where τ_{n0} is excess minority carrier lifetime.

- Electron capture, electron emission, hole capture, and hole emission are the four possible processes in the presence of traps.
- According to Shockley–Read–Hall theory,

$$R_n = R_p = \frac{C_n C_p N_t (n_p - n_i^2)}{C_n (n + n_t) + C_p (p + p_t)} = R$$

- The continuity equation for holes and electron is

$$\frac{\partial p}{\partial t} = D_p \frac{\partial^2 p}{\partial x^2} - \mu_p \left(E \frac{\partial p}{\partial x} + p \frac{\partial E}{\partial x} \right) + G_p - \frac{p}{\tau_p}$$

and

$$\frac{\partial n}{\partial t} = D_n \frac{\partial^2 n}{\partial x^2} - \mu_n \left(E \frac{\partial n}{\partial x} + n \frac{\partial E}{\partial x} \right) + G_n - \frac{n}{\tau_n}$$

- The Haynes-Shockley experiment can be used to determine mobility and diffusion coefficients. These quantities are evaluated using the relations

$$\mu = \frac{v_d}{E} \quad \text{and} \quad D_p = \frac{(\Delta x)^2}{16 t_d}$$

where t_d is the delay time of the input pulse.

- Quasi-Fermi energy levels E_{Fn} and E_{Fp} for electrons and holes are defined through the equations

$$n_0 + \delta n = n_i \exp\left(\frac{E_{Fn} - E_{Fi}}{kT} \right)$$

and

$$p_0 + \delta p = n_i \exp\left(\frac{E_{Fi} - E_{Fp}}{kT} \right)$$

- The excess-carrier concentration at the semiconductor surface is lower than that in the bulk.
- Surface recombination velocity s for hole can be evaluated using the relation

$$D_p \left. \frac{d(\delta p)}{dx} \right|_S = s \, \delta p \big|_S$$

- $s = 0$ implies that the excess minority carrier concentration at the surface is the same as that in the bulk.
- $s = \infty$ implies that the excess minority carrier concentration at the surface is zero.
- $J_n(x) = \mu_n n(x) \frac{d E_{fn}}{dx}$
- In auger recombination, an electron and a hole recombine and the resultant energy is transferred to another free hole.

Exercises

Review Questions

4.1 What is recombination? Give its measurement units.

4.2 Give the relationship between recombination rate and generation rate at thermal equilibrium conditions.

4.3 Derive the relation $I(x) = I_o e^{-\alpha x}$. What is the measurement unit of α?

4.4 A semiconductior absorbs photons with energies equal to or larger than the band gap. Justify this statement.

4.5 Show that for direct band-to-band recombination under low-injection conditions,

$$R_n' = R_p' = \frac{\delta n(t)}{\tau_{n0}}$$

for an *n*-type semiconductor.

4.6 Show the four possible processes involving a trap schematically and describe each of them.

4.7 Show that according to Shockley–Read–Hall theory,

$$R_n = R_p = R = \frac{C_n C_p N_t (np - n_i^2)}{C_n (n + n_t) + C_p (p + p_t)}$$

4.8 Give the conditions for low-level injection in a *p*-type semiconductor and show that

$$\tau_{n0} = \frac{1}{C_n N_t}$$

4.9 Derive the one-dimensional continuity equation.

4.10 Write down the time-dependent diffusion equations for hole and electrons and explain each term.

4.11 Describe a typical experimental set-up for carrying out the Haynes–Shockley experiment.

4.12 What is the difference between the mobility determined using the Haynes–Shockley experiment and the mobility determined using the Hall effect.

4.13 Describe a procedure for calculating the diffusion coefficient using the Haynes–Shockley experiment.

4.14 What are quasi-Fermi energy levels? In what way are they different from Fermi energy levels?

4.15 Sketch the typical distribution of surface states within the forbidden band gap of a semiconductor.

4.16 The excess minority carrier lifetime at the surface is smaller than that in the bulk. Justify this statement.

4.17 Define surface recombination velocity and give its units.

4.18 Show that at thermal equilibrium conditions

$$G_{n0} = G_{p0} = R_{n0} = R_{p0}$$

4.19 What are the assumptions made in deriving the relation?

$$I(x) = I_0\, e^{-\alpha x}$$

4.20 Justify the relation:

$$n\mathrm{p} \neq n_0\, p_0 = n_i^2$$

4.21 Show that the time dependence of excess minority carrier concentration δn of electrons is given by

$$\delta n(t) = \delta n(0)\, e^{-t/\tau_{n0}}$$

4.22 Show the four processes involving a trap in the forbidden gap.

4.23 Explain the terms C_n, C_p, and n_t in the expression

$$R_n = R_p = \frac{C_n C_p N_t (n_p - n_i^2)}{C_n (n + n_t) + C_p (p + p_t)} = R$$

4.24 Using the expressions for R_n and R_p show that at low-injection conditions

$$\tau_{p0} = \frac{1}{C_p N_t} \quad \text{and} \quad \tau_{n0} = \frac{1}{C_n N_t}$$

4.25 Derive the time-dependent continuity equation for excess holes.

4.26 Describe the Haynes–Shockley experiment with suitable sketches.

4.27 Why do we have to define the quasi-Fermi levels for electrons and holes separately?

4.28 Explain the role of surface in deciding the excess-carrier concentration profile in a semiconductor.

4.29 What is the physical significance of surface recombination velocity, S = (i) 0 and (ii) ∞?

4.30 How is the derivative of quasi-Fermi level related to current density?

4.31 What are deep impurity levels?

4.32 Show a schematic representation of Auger recombination process.

Problems

4.1 Excess holes are generated in a semiconductor to an initial concentration of $\delta p(0) = 5 \times 10^{16}$ cm^{-3}. The excess-carrier lifetime in the semiconductor is known to be 10 μs. The source used to generate the excess-carriers is switched off at $t = 0$. Determine the excess hole concentration at $t = 2$ μs.
[*Hint:* $\delta p(t) = \delta p(0)\, e^{-t/\tau_{p0}}$]

Ans. 4.09×10^{16}

4.2 Excess electrons are generated in a semiconductor. The initial concentration is N cm^{-3}. The excess-carrier lifetime in the semiconductor is τ μs. The source generating the excess-carriers is switched off at $t = 0$. Calculate the excess electron concentration at (i) $t = \tau$, (ii) $t = 2\tau$, and (iii) $t = 3\tau$.
[*Hint:* $\delta n(t) = N\mathrm{e}^{-t/\tau}$]

Ans. (i) 0.37 N (ii) 0.135 N (iii) 0.05 N

4.3 Excess electrons are generated in a semiconductor to an initial concentration of $\delta n(0) = 5 \times 10^{15}$ cm^{-3}. Excess-carrier lifetime in the semiconductor is 10 μs. Calculate the recombination rate of excess electrons for $t = 5$ μs.

$$\left[\textit{Hint: } R'_n = \frac{\delta n(t)}{\tau_{n0}}\right]$$

Ans. 0.304×10^{21} cm^{-3} s^{-1}

4.4 A sample of silicon is maintained at $T = 300$ K. It has a dopant concentration $N_a = 10^{16}$ cm^{-3} and $N_d = 0$. The minority carrier lifetime $\tau_{p0} = \tau_{n0} = 5 \times 10^{-7}$ s and $n_t = p_t = n_i$. The excess-carrier concentrations are given by $\delta n = \delta p = 10^{14}$ cm^{-3}. Calculate the recombination rate of excess-carriers using the Shockley–Read–Hall theory. Assume $n_i = 1.5 \times 10^{10}$ cm^{-3}.

$$\left[\textit{Hint: } p_0 \cong N_a,\ n_0 = \frac{n_i^2}{N_a},\ \text{and } R = \frac{(n_p - n_i^2)}{\tau_{p0}(n + n_t) + \tau_{n0}(p + p_t)}\right]$$

Ans. 1.98×10^{20} cm^{-3} s^{-1}

4.5 Excess EHP is being created in a piece of silicon at the rate of 10^{13} EHP/cm^3 every 6 microseconds. For the sample, it is also given that $n_0 = 10^{14}$ cm^{-3} and $\tau_n = \tau_p = 3$ μs. Calculate the position of electron quasi-Fermi level with respect to E_i at room temperature. Assume $kT = 0.026$ eV.

$$\left[\textit{Hint: } \text{steady state excess electron concentration}\right.$$

$$\delta n = g_{0p}\tau_n = \frac{10^{13}}{10^{-6}} \times 3 \times 10^{-6} = 3 \times 10^{13} \text{ cm}^{-3}$$

where g_{0p} = optical generation rate.

$$\left.E_{Fn} - E_i = kT \ln\left(\frac{n_0 + \delta n}{n_i}\right)\right]$$

Ans. 0.236 eV

4.6 Show that the condition $np \neq n_i^2$ is valid for the situation presented in Problem 4.5.

$$\left[\textit{Hint: } \delta n = \delta p = 3 \times 10^{13}\ \ p_0 = \frac{n_i^2}{n_0}\right]$$

4.7 A sample of an n-type semiconductor at $T = 300$ K has a carrier concentration of $n_0 = 5 \times 10^{16}$ cm^{-3} and $n_i = 2 \times 10^{10}$ cm^{-3}. A source creates excess carriers in the semiconductor, such that $\delta n = \delta p = 10^{13}$ cm^{-3}. Determine the positions of quasi-Fermi levels for electrons and holes with respect to the intrinsic Fermi level.

$$\left[\textit{Hint: } E_{Fn} - E_{Fi} = 0.026 \ln\left(\frac{n_0 + \delta n}{n_i}\right)\right.$$

$$\left.E_{Fi} - E_{Fp} = 0.026 \ln\left(\frac{p_0 + \delta p}{n_i}\right)\right]$$

Ans. $E_{Fn} - E_{Fi} = 0.383$ eV, $E_{Fi} - E_{Fp} = 0.1616$ eV

4.8 A sample of an *n*-type silicon has a dopant concentration of $N_d = 10^{16}$ cm^{-3} at $T = 300$ K. Assume $\tau_{p0} = 5 \times 10^{-7}$ s and a generation rate $G = 10^{21}$ cm^{-3} s^{-1}. Derive an expression for the excess hole concentraion dependence on time.
[*Hint:* $\delta p(t) = G\tau_{p0}(1 - e^{-t/\tau_{p0}})$]

Ans. $\delta p(t) = 5 \times 10^{14}\,[1 - e^{-t/5\times10^{-7}}]$ cm^{-3}

4.9 Use the expression derived in unsolved numerical (8) to evaluate the excess hole concentration at $t = \tau_{p0}$. What is the excess hole concentration at $t = \infty$?
[*Hint:* for $t = \infty$; $\delta p(t) = G\tau_{p0}$ [see Problem (4.5)]

Ans. 3.16×10^{14}, 5×10^{14}

4.10 An *n*-type semiconductor sample is used in the Haynes–Shockley experiment. The length of the sample is 1.5 cm and probes (1) and (2) are at a distance of 1.3 cm from each other. The supply battery voltage is $E_0 = 3$ V. The delay time between the injection of the pulse at point (1) and its reaching point (2) is $\Delta t = 200$ μs. Calculate the hole mobility and diffusion coefficient for holes. Also check the validity of the Einstein relation.

$$\left[\textit{Hint: } \mu_p = \frac{v_d}{\varepsilon},\ D_p = \frac{(\Delta t\ell)^2}{16\, t_d^3}\right]$$

Ans. $\mu_p = 1300$ cm^2 V^{-1}s^{-1}, $D_p = 33.8$ cm^2/s.

4.11 Boron is diffused into an intrinsic Si sample, resulting in the acceptor distribution shown in Fig. 4.P11.1. Sketch the equilibrium band diagram and show the direction of the resulting electric field, for $N_a(x) \gg n_i$. Repeat for phosphorous with $N_d(x) \gg n_i$.
[*Hint:* Write the equation for current densities due to holes. At equilibrium $J_p(x) = 0$. So, find an expression for rate of variation of electric field with respect to x. If it is positive, then electric field will be in the positive x direction and if it is negative then the field will be in the negative x direction. The energy band diagram will get tilted accordingly.]

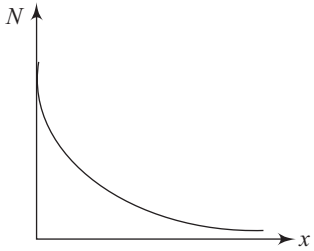


Fig. 4.11.1 Acceptor distribution in silicon doped with boron

4.12 A new semiconductor was brought in the lab. Its band gap was determined to be 2.45 eV. When light of wavelength 0.55 μm was used, the sample exhibited a 5% increase in conductivity. Is this possible? Justify your answer.
[*Hint:* Yes. It is possible with the help of traps in forbidden energy band gap. Discuss with your teacher in detail.]

4.13 A device engineer tried to use a sample of silicon doped with phosphorous impurities of the order of 10^{17} cm^{-3} at room temperature . While performing the experiment and making calculations, the value of diffusion coefficient for electrons was required. After some theoretical calculations it was found to be

34.965 cm^2/s. But it was discovered that the experimental setup was predicting a different value of diffusion coefficient for electrons, which was less than what had been theoretically calculated. What went wrong? Does the experimental setup require revision? If yes, then suggest a good experimental set-up to find the diffusion coefficient for electrons in this *n*-type sample of silicon. Otherwise provide a justification for the experimental result obtained by the device engineer.

[*Hint:* Find the value of mobility at given values of doping levels and discuss the answer with your teacher in detail.]

4.14 A semiconductor device engineer wants to design a photoconductive cell which should be able to respond to pulses of frequency as high as 1 Ghz. The device engineer decides to cut down the size of the device, so as to reduce the drift time of the carriers. Will this compromise the performance of this photoconductive device on some other account? Justify your answer. [*Hint:* Yes, the optical sensitivity of the device will be compromised.]

4.15 A silicon sample has $n_0 = 5 \times 10^{15}$ cm^{-3} and $\tau_p = 1 \times 10^{-6}$ s at 300 K. The sample is illuminated uniformly with penetrating light which generates 10^{19} EHP per cm^3 per second.

(a) Determine the steady state electron and hole concentrations in the sample and predict whether it is a case of low or high level injection.

(b) Determine the photoconductivity change of the sample if $\mu_n = 3\mu_p = 1260$ cm^2/V s.

(c) Determine the position of the electron and hole quasi-Fermi levels in the sample and show these levels on the energy band diagram.

4.16 Excess electrons have been generated in a semiconductor to a concentration of $\delta n(0) = 10^{15}$ cm^{-3}. The excess carrier lifetime is $\tau_{n0} = 10^{-6}$ s. The forcing function generating the excess carriers turns off at $t = 0$, so that semi- conductor is allowed to return to an equilibrium condition for $t > 0$. Calculate the excess electron concentration for

(a) $t = 0$,

(b) $t = 1$ μs

(c) $t = 4$ μs

Ans. (a) 10^{15} cm^{-3} (b) 3.68×10^{14} cm^{-3} (c) 1.83×10^{13} cm^{-3}

4.17 Excess electrons are generated in a semiconductor to a concentration level of $\delta n(0) = 10^{17}$ cm^{-3}. The source generating the excess carriers is switched off at $t = 0$. At $t = 2$μs the excess carrier concentration becomes 7×10^{16} cm^{-3}. Calculate the excess carrier lifetime.

Ans. 5.6 μs

4.18 The thermal equilibrium fermi level for an *n*-type semiconductor is given by $(E_F - E_{Fi}) = 0.4$ ev at T = 300K. The intrinsic carrier concentration at 300K is $n_i = 2 \times 10^{10}$ cm^{-3}. Calculate the equilibrium carrier concentration, n_0.

Ans. 9.6×10^{16} cm^{-3}

5 p-n Junction

- Fabrication of *p-n* junction
- Thermal oxidation
- Diffusion
- *p-n* junction with no bias
- Reverse-biased *p-n* junction
- Junction with non-uniform doping
- Varactor diode
- Junction breakdown
- Tunnel diode

Learning Objectives

After going through this chapter the student will be able to

- understand different methods of fabricating *p-n* junctions
- understand the process of thermal oxidation
- understand predeposition and drive-in steps
- derive expressions for built-in voltage and space charge width of an unbiased junction
- solve numericals based on predeposition and drive-in steps of diffusion
- solve numericals based on built-in potential of unbiased *p-n* junctions
- derive expressions for space charge capacitances of a *p-n* junction under applied reverse bias and for a linearly graded junction
- understand hyperabrupt junctions
- understand different junction breakdown processes
- derive conditions for avalanche breakdown
- understand the operation of tunnel diode
- solve numericals based on space charge region width of *p-n* junctions with reverse bias
- solve numericals based on junction capacitance of *p-n* junction under reverse bias

Introduction

We have so far discussed different aspects related to individual intrinsic and extrinsic semiconductors. The real immense developments in the area of semiconductor electronics started with the understanding of the mechanisms in play when a *p*-type semiconductor is paired with an *n*-type semiconductor. Such a pair is called a *p-n* junction. Almost all useful applications involving semiconductors use such a junction or a combination of such junctions. Although, simply put, a *p-n* junction can be created by putting a *p*-type semiconductor on an *n*-type semiconductor, the actual process of fabrication of a *p-n* junction is a lot more complicated. Some of the steps/procedures involved in the fabrication processes are presented in this chapter. Very dramatic differences arise as one changes the polarity of the applied voltage across a *p-n* junction. These differences hold the key to the secret of the operation of many devices. Some of these devices are discussed in detail in this chapter. In fact, to put it briefly, our serious interaction with the basics of semiconductor electronics starts with this chapter.

5.1 Fabrication of *p-n* Junctions

As explained above, a *p-n* junction requires some specialized processing techniques. In this section we learn about some common techniques used to form *p-n* junctions.

5.1.1 *p-n* Junction Formation

A number of different techniques exist for obtaining *p-n* junctions in semiconductors. The particular process sequence decides the choice of the technique to be used for realizing the *p-n* junction. In this section we will briefly discuss the commonly used techniques.

Alloyed junction

In the alloying technique, the impurity material–semiconductor system is first heated to a temperature slightly higher than the eutectic temperature. The temperature is then lowered, leading to a recrystallization of the semiconductor, which is saturated with the impurity atoms. Figure 5.1 shows the alloy junction formation for the Al–Si system. A small pellet of aluminium is first placed on an *n*-type ⟨111⟩-oriented silicon wafer. The system is then heated to a temperature slightly higher than the eutectic temperature (~580°C) for the Al–Si system. This leads to the formation of a small puddle of molten Al–Si mixture.

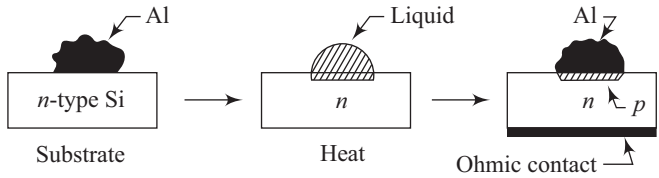


Fig. 5.1 Alloy junction formation

As the temperature is lowered further, the puddle begins to solidify. This results in a recrystallized portion saturated with the acceptor impurities. The recrystallized part maintains the crystal orientation of the original substrate. A heavily doped (p^+) region is thus formed on the *n*-type substrate. The location of the junction formed by the alloy method depends critically on the temperature–time cycle of alloy formation. This makes a precise control difficult.

Solid state diffusion

In this process, a diffusion of a suitable dopant is first carried out. Portions of the surface are then protected (using wax or metal contacts), and the rest of the exposed surface is etched to form mesa. A solid state diffusion method, involving diffusion of boron (in the form BBr_3) into an *n*-type silicon substrate, is shown schematically in Fig. 5.2.

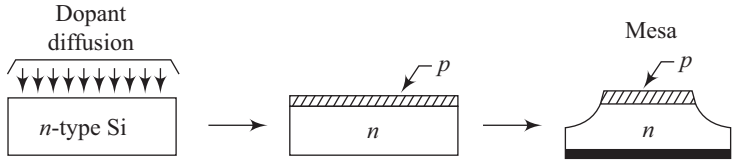


Fig. 5.2 Solid state diffusion method

Planar process

One of the most popular methods of fabricating semiconductor devices and integrated circuits is the planar process, shown schematically in Fig. 5.3.

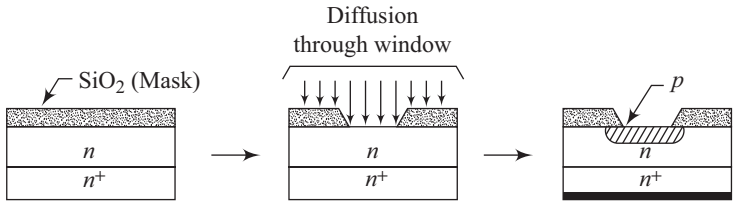


Fig. 5.3 Planar process

In this process, a thin layer of thermally grown silicon dioxide (typically 1 μm) is first obtained. Standard lithographic techniques are then used to open windows in the oxide. Impurities are then allowed to diffuse through the exposed silicon surface, thereby leading to the formation of *p*-*n* junctions in this region. The substrate used in the process is heavily doped (n^+) to reduce the series resistance. An epitaxial (derived from Greek words *epi* and *taxis*, meaning 'on' and 'arrangement' respectively) layer is used to grow the active layer. The epitaxial growth ensures a lattice structure identical to the substrate.

Ion implantation

A technique that gives excellent control over the impurity profile is *ion implantation*. The technique replaces the high-temperature diffusion process. This technique can

be used to fabricate complicated semiconductor devices. A beam of impurity ions is first accelerated to kinetic energies in the range of several keV to several MeV. The energetic beam is then directed towards the crystal, as shown in Fig. 5.4.

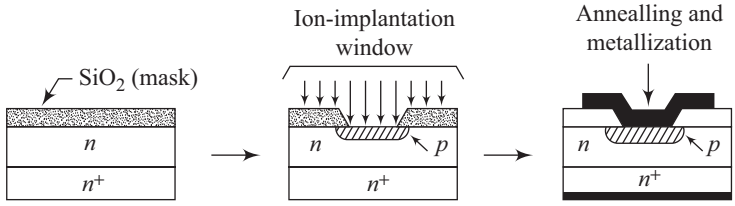


Fig. 5.4 Ion implantation process

The impurity atoms come to rest at some average penetration depth, called the *projected range*. The projected range depends upon the impurity and its implantation energy. It generally ranges between a few hundred angstroms to a few microns. The implanted dose of ϕ ions/cm^2 is distributed approximately in accordance with the Gaussian formula

$$N(x) = \frac{\phi}{\sqrt{2\pi}\ \Delta R_p} \exp\left[-\frac{1}{2}\left(\frac{x - R_p}{\Delta R_p} \right)^2 \right] \tag{5.1}$$

where $N(x)$ represents the concentration at a particular depth and ΔR_p is the half-width of the distribution at $e^{-1/2}$ of the peak, as shown in Fig. 5.5. ΔR_p is also called *straggle*.

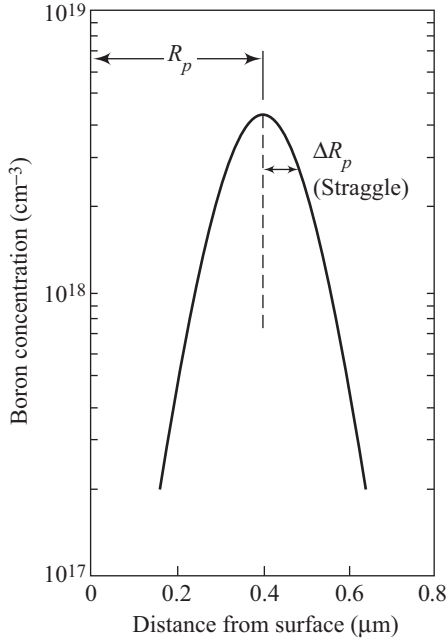


Fig. 5.5 Gaussian distribution of boron atoms about a projected range R_p. Dose is 10^{14} B atoms/cm^2 implanted at 140 keV.

R_p and ΔR_p increase with increasing implantation energy. By carrying out several implantations at different chosen energies, one can, in principle, create a uniformly doped region, as shown in Fig. 5.6.

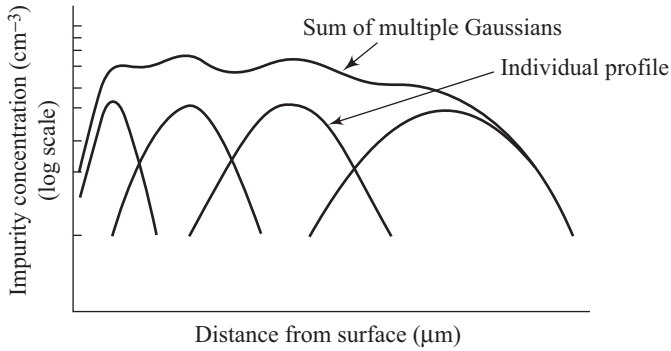


Fig. 5.6 Summation of four Gaussians leading to a flat impurity distribution

A typical ion implantation system consists of the following components:

Source A gas containing the desired impurity in the ionized form.

Acceleration tube It provides acceleration to the ions to increase their kinetic energy.

Mass separator It selects the desired ion species.

Drift tube This helps selected ions drift towards the target.

Target chamber The wafer to be ion implanted is kept in this chamber. This chamber also has a facility to carry out repetitive scanning in a raster pattern.

Using ion implantation, desired impurities can be introduced into a semiconductor at relatively low temperatures. This has a very interesting and useful consequence. Any previously doped profile within the semiconductor does not get much disturbed during the ion implantation process. The areas of the semiconductor surface where no ion implantation is desired can be protected using metal or photoresist layers. The technique is very adapted to very shallow (~ 0.1 μm) doped regions. A very precise control of the doping concentration is also possible using this technique. Another important advantage of ion implantation is its ability to introduce impurities that are not easy to diffuse into semiconductors.

The technique, however, has one major disadvantage. The energetic ions collide with atoms of the semiconductor lattice. This leads to crystal damage. The damage can, though, be removed by annealling the semiconductor at high temperatures. Silicon can be annealled at temperatures around 1000°C without any problem. Compound semiconductors such as GaAs can, however, dissociate at such high temperatures. The evaporation of As from GaAs surface can be minimized by using a thin layer of silicon nitride. Another technique of annealling that is gaining popularity is the rapid thermal processing (RTP) or rapid thermal

annealling technique. In this technique, the ion implanted wafer is subject to a high temperature for a very short duration (~ 10 s). Irrespective of the method of annealling, some unwanted diffusion does take place during annealling. The post-anneal impurity profile is given by

$$N(x) = \frac{\phi}{\sqrt{2\pi}(\Delta R_p^2 + 2Dt)^{1/2}} \exp\left[-\frac{1}{2}\left(\frac{(x - R_p)^2}{\Delta R_p^2 + 2Dt}\right)\right] \tag{5.2}$$

where t is the annealling time.

5.1.2 Thermal Oxidation

The planar process described earlier is one of the most popular methods to realize semiconductor devices. The insulating silicon dioxide layer forms an integral part of this processing technique. In this section, we will learn about the *thermal oxidation technique* to realize this silicon dioxide layer. The semiconductor wafer to be oxidized is placed in a suitable quartz (clean silica) boat inside a tube of clean silica, which is heated to very high temperatures (~ 800–1100°C) in a furnace. Excellent temperature control (~ ± 0.1°C) is possible with modern control electronics. An oxygen-containing gas such as dry O_2 or H_2O is made to flow through the tube at atmospheric pressure from one end and to come out of the tube at the other end. Both horizontal and vertical furnaces are commonly employed. In vertical furnaces, gases flow in from the top and flow out at the bottom, ensuring a uniform flow. A schematic set-up of the furnace in the horizontal format is shown in Fig. 5.7.

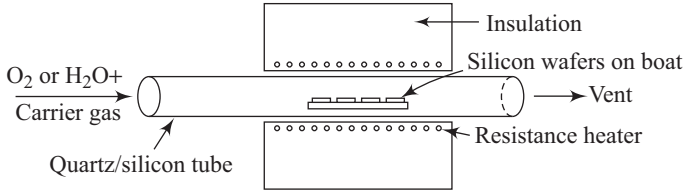


Fig. 5.7 Thermal oxidation set-up

Oxidation follows the following chemical reactions:

$$\text{Si(solid)} + O_2 \rightarrow SiO_2\,\text{(solid)} \qquad \text{[dry oxidation]}$$

$$\text{Si(solid)} + 2H_2O \rightarrow SiO_2\,\text{(solid)} + 2H_2 \qquad \text{[wet oxidation]} \tag{5.3}$$

Both the oxidation processes consume Si from the surface of the substrate. From the densities and molecular weights of silicon, it can be shown that 0.44 μm of Si is consumed for every micron of grown SiO_2. During oxidation, the oxidant (O_2 or H_2O) molecules diffuse through the already grown oxide to the Si–SiO_2 interface. The reactions indicated in Eqn (5.3) take place at this interface. The ability to obtain a stable thermal oxide holds the key to the success of the planar process in realizing the semiconductor devices. The silicon dioxide, as we have seen, thus grows from inside out (the detailed treatment is called the Deal–Grove

model). The growth rate of oxide is determined by the slower process out of the two competing processes of *reaction* and *diffusion*.

Figure 5.8 shows the plots of the oxide thickness as a function of time at different temperatures for dry and wet oxidation of Si (100) .

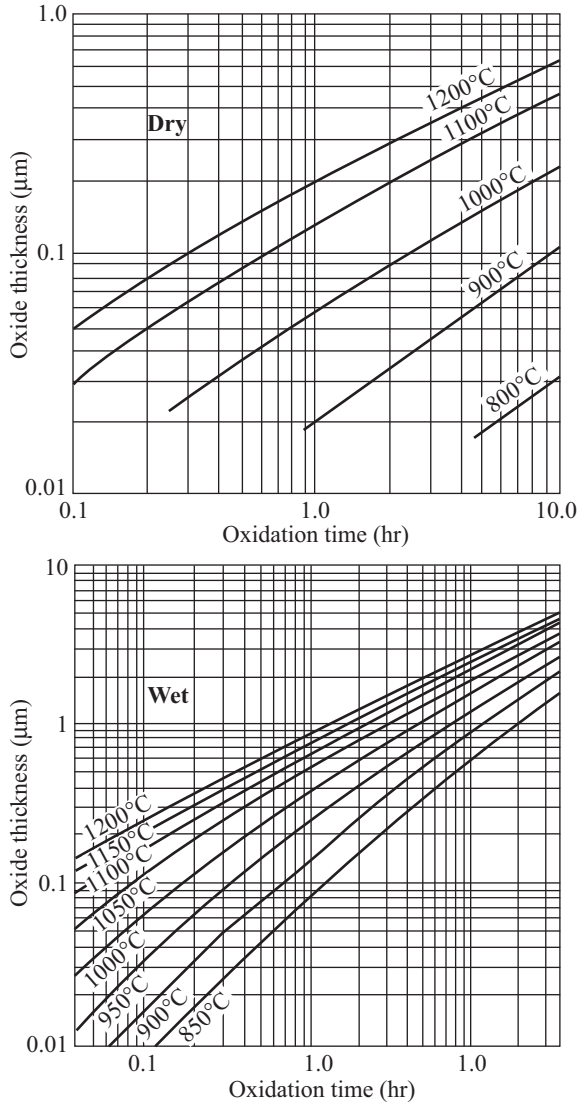


Fig. 5.8 Dry and wet thermal oxide thickness grown on Si (100)

5.1.3 Diffusion

Another important processing step involved in the planar process is *diffusion*. The oxidation process described in the preceding section is first used to obtain an

oxide layer over the substrate. A process, called *photolithography*, is then used to open windows into the oxide. Dopants such as B, P, or As are then introduced through these windows. These impurities are introduced at temperatures ~800–1100°C using high-temperature furnaces, through a process called *diffusion*. Gas or vapour sources are generally used. In diffusion, dopants are transported from a region of high concentration near the surface to the surface where the concentration is lower. The process of diffusion of dopants is similar to the one involving carriers as discussed in Chapter 4. The maximum number of impurities that can be dissolved in a given semiconductor surface is given by the solid solubility of the particular impurity in the semiconductor. The solid solubility of various impurities in Si as a function of temperature is shown in Fig. 5.9.

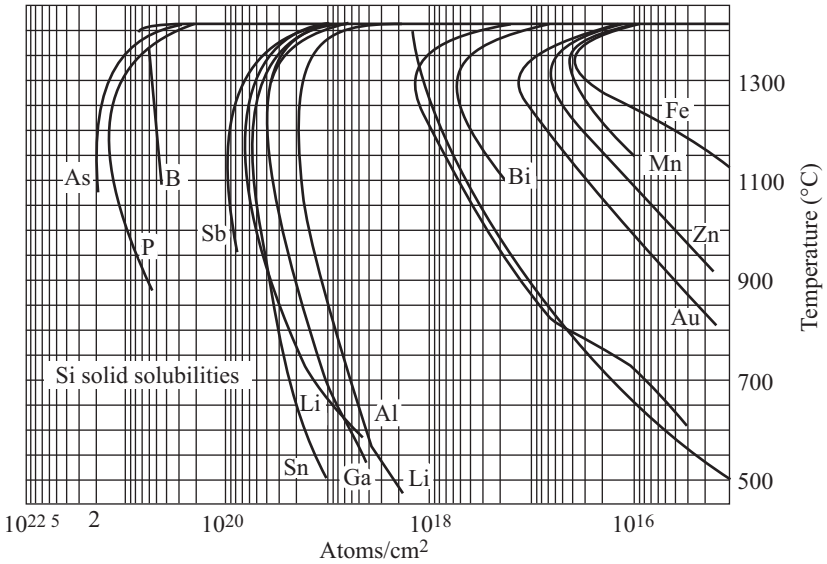


Fig. 5.9 Solid solubilities of impurities in Si

Diffusivity, D, of dopants in solids has a temperature dependence given by

$$D = D_0 \exp(-E_A/kT) \tag{5.4}$$

where D_0 is a material-and dopant-dependent constant and E_A is called the activation energy. Such a dependence on temperature is called *Arrhenius dependence* and leads to high enough diffusivities only at high temperatures, enabling appreciable diffusions only at such temperatures. Another important conclusion that can be drawn from Eqn (5.4) is the fact that excellent control of temperature is required to result in control of impurity profiles obtained using diffusion. The oxide over the silicon surface serves as a mask against the dopants since they have a low diffusivity in oxide. Figure 5.10 shows the diffusivities of some common dopants in Si and SiO_2 as a function of temperature.

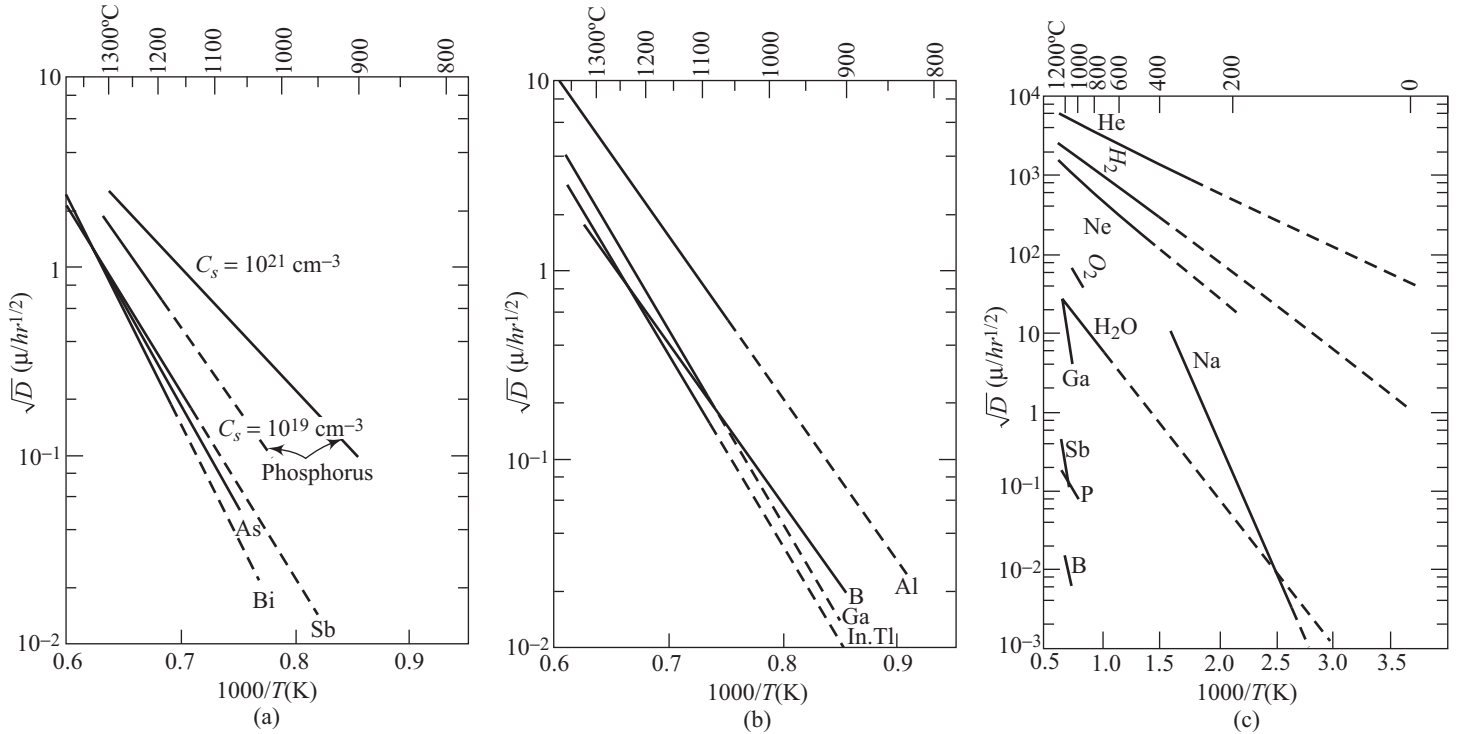


Fig. 5.10 (a) Diffusivity of common donor impurities in silicon versus T (b) Diffusivity of acceptor impurities in silicon versus T (c) Diffusivities in amorphous SiO_2

The impurity distribution during the diffusion process can be calculated as a function of time from the solution of the diffusion equation given by

$$\frac{\partial C}{\partial t} = D\frac{\partial^2 C}{\partial x^2} \tag{5.5}$$

where C represents the impurity concentration. This equation assumes a zero electric field and a constant value of diffusivity. Appropriate boundary conditions need to be used alongwith this equation.

Case (i) The predeposition step In this step, the semiconductor sample is kept in a furnace through which an inert gas containing the desired impurities flows. Temperatures in the range of 800°C to 1200°C are generally used. The surface concentration, C_S, of the impurity is maintained constant during the predeposition process. The diffusion equation [Eqn (5.5)] has to be solved under the boundary conditions

$$C(0, t) = C_S \tag{5.6}$$

and

$$C(\infty, t) = 0 \tag{5.7}$$

with the initial condition

$$C(x, 0) = 0 \tag{5.8}$$

The solution of Eqn (5.5) under the conditions given by Eqs (5.6), (5.7), and (5.8) is the *complementary error function* given by

$$C(x, t) = C_S \text{erfc}\frac{x}{2\sqrt{Dt}} \tag{5.9}$$

The resulting concentration distribution using the above simplified theory is shown in logarithmic scale in Fig. 5.11.

Case (ii) The drive-in step The predeposition step indicated above results in a surface concentration equal to the solid solubility for the particular impurity. It is often desirable to lower the surface impurity concentration to values below the solid solubility value. Additionally, there always exists the need to move the impurities deeper into the bulk of the semiconductor without modifying the total number of impurity atoms within the semiconductor. These tasks can be accomplished by a high-temperature heat treatment in a gas that does not contain any impurities. This additional heat treatment is called the drive-in diffusion step. It is generally carried out in an oxidizing ambient. Thus, the boundary conditions for the drive-in step are

$$\left.\frac{\partial C}{\partial x}\right|_{(0, t)} = 0 \tag{5.10}$$

and

$$C(\infty, t) = 0 \tag{5.11}$$

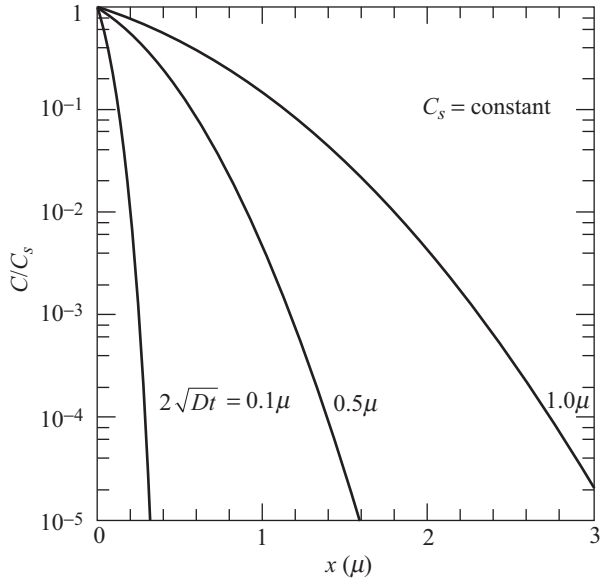


Fig. 5.11 Normalized concentration versus distance for successive times assuming constant surface concentration

The total amount of impurity, Q, is constant. The initial condition is given by

$$C(x, 0) = C_s \text{erfc} \frac{x}{2\sqrt{(Dt)_{\text{predep}}}} \tag{5.12}$$

This is because the impurity distribution at the start of the drive-in diffusion step is the one that existed at the conclusion of the predeposition step. The solution of Eqn (5.5) under the conditions given by Eqs (5.10), (5.11), and (5.12) is difficult to obtain. However, in most practical situations, $\sqrt{Dt}$ for the drive-in step is much larger than $\sqrt{Dt}$ for the predeposition step. Thus, the extent of the penetration of the impurity profile during the predeposition step can be neglected in comparison to the profile obtained at the end of the drive-in diffusion step. Under such an assumption, the predeposition profile can be mathematically represented by a *delta function*.

The impurity concentration distribution after the drive-in diffusion step under this assumption is given by

$$C(x, t) = \frac{Q}{\sqrt{\pi Dt}} e^{-x^2/4Dt} \tag{5.13}$$

Equation (5.13) is the well-known *Gaussian distribution*. Figure 5.12 shows the Gaussian profile normalized to Q in a logarithmic scale.

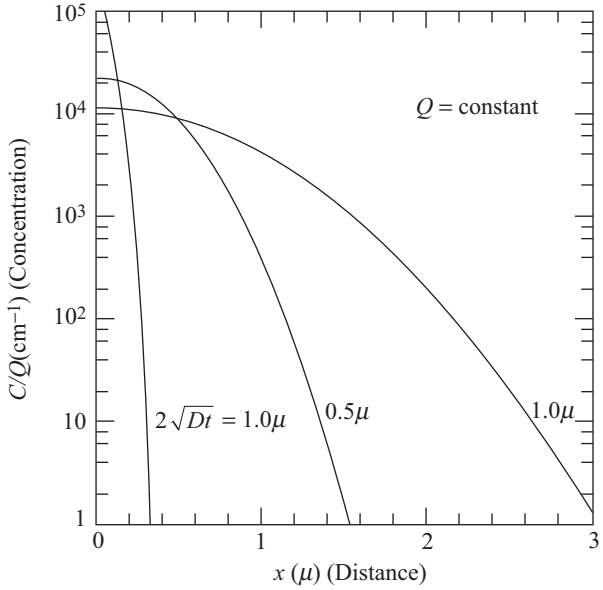


Fig. 5.12 Normalized concentration versus distance for successive times assuming constant amount of impurity

5.2 Basic *p-n* Junction

In Section 5.2, we looked at the various methods for fabricating *p-n* junctions. We now take a closer look at some important characteristics of *p-n* junctions with and without any applied bias.

5.2.1 Basic Structure

A *p-n* junction, as we have seen, is fabricated by creating an interface which separates *n*- and *p*-type semiconductors using any of the common techniques described in Section 5.1. Such an interface that separates the *n* and *p* regions is called a *metallurgical junction* and is shown schematically in Fig. 5.13(a). Figure 5.13(b) shows the corresponding simplified doping profile.

As can be seen from Fig. 5.13(b), the simplified picture presented assumes a constant dopant concentration on either side of the metallurgical junction. Furthermore, the dopant type has been assumed to change abruptly from one type to the other. Such a

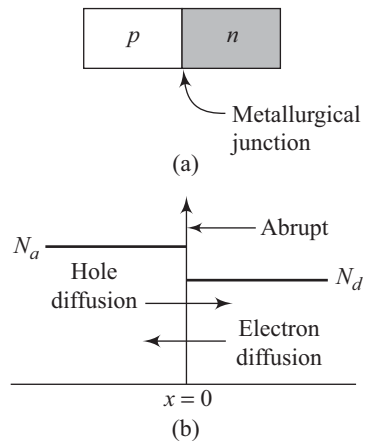


Fig. 5.13 Schematic diagram of a *p-n* junction, (b) doping profile of a uniformly doped *p-n* junction

junction is called a *step junction*. The metallurgical junction also has a very large gradient in the electron and hole concentrations. Majority carrier electrons in the *n*-region start diffusing into the *p*-region, whereas the majority carrier holes in the *p*-region start diffusing into the *n*-region. The process of diffusion, however, does not continue indefinitely. As electrons diffuse from the *n*-region, they leave behind positively charged donor atoms. Similarly, the holes diffusing from the *p*-region leave behind negatively charged acceptor atoms. The positive charge due to the donor atoms and the negative charge due to the acceptor atoms induce an electric field in the region near the metallurgical junction, (Fig. 5.14).

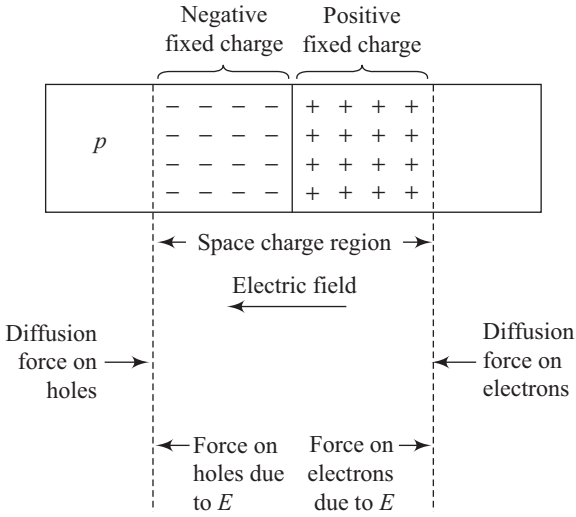


Fig. 5.14 Space charge region, electric field, and forces acting on charged carriers

The electric field is directed from the positive to the negative charge, i.e., from the *n*- to the *p*-region.

The two regions shown in Fig. 5.14, having a net positive and negative charge, are collectively referred to as the *space charge region*. The space charge region does not contain free electrons and holes and is, therefore, also referred to as the *depletion region*. A density gradient exists in the majority carrier concentration at the edges of the space charge region. The majority carrier concentration gradient produces a *diffusion force* that pushes the majority carriers towards the space charge region. The built-in electric field in the space charge region, on the other hand, produces a force on the electrons and the holes in the direction opposite to the *diffusion force.* At thermal equilibrium conditions, the diffusion force is balanced out by the force due to the built-in electric field.

5.2.2 No Applied Bias

We still continue with the assumption of no applied external bias as well as the absence of any other external excitation. The *p-n* junction under such circumstances

is at thermal equilibrium and the Fermi energy level is constant throughout the entire system. The energy band diagram for a *p-n* junction at thermal equilibrium is shown in Fig. 5.15.

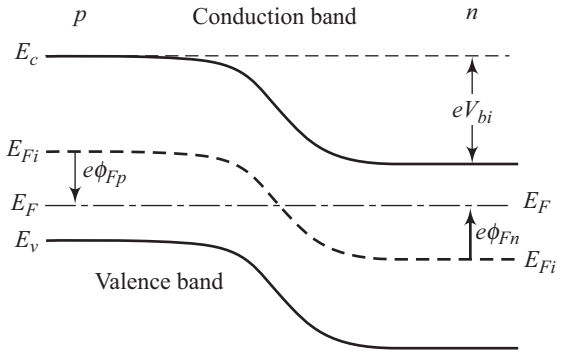


Fig. 5.15 Thermal equilibrium energy band diagram for a *p-n* junction

The relative positions of the conduction and valence bands with respect to the Fermi energy change as we go from the *p*-type to the *n*-type of semiconductor. This leads to a bending of the conduction and valence band energies as we go from one type of semiconductor to the other. As can be seen from Fig. 5.15, the electrons in the conduction band of the *n*-region encounter a potential barrier as they try to move over into the conduction band of the *p*-region. This potential barrier is called the *built-in potential barrier* and the symbol V_{bi} is used for it. The built-in potential barrier is responsible for maintaining an equilibrium between the majority carrier electrons in the *n*-region and the minority carrier electrons in the *p*-region. One must, however, remember that no current flows through the junction under thermal equilibrium conditions. Unlike the emf of a cell, built-in potential of a *p-n* junction cannot be measured by using any voltage-measuring instrument.

The intrinsic Fermi level, E_{Fi}, in Fig. 5.15 is equidistant from the conduction band edge through the junction. It is also clear from Fig. 5.15 that the built-in potential barrier is simply the difference between the intrinsic Fermi levels in the *p* and *n* regions. Thus, if ϕ_{Fn} and ϕ_{Fp} represent the difference between the intrinsic Fermi level and the Fermi levels in the *n* and *p* regions, respectively, then we can write

$$V_{\text{bi}} = |\phi_{Fn}| + |\phi_{Fp}| \tag{5.14}$$

The electron concentration, n_0, in the conduction band for the *n*-region is given by

$$n_0 = N_c \exp\left[\frac{-(E_c - E_F)}{kT}\right] \tag{5.15}$$

We also know that

$$n_0 = n_i = N_c \exp\left[\frac{-(E_c - E_{Fi})}{kT}\right] \tag{5.16}$$

Combining Eqs (5.15) and (5.16), we get

$$n_0 = n_i \exp\left[\frac{E_F - E_{Fi}}{kT}\right] \tag{5.17}$$

The potential ϕ_{Fn} can be defined as

$$e\phi_{Fn} = E_{Fi} - E_F \tag{5.18}$$

Using Eqn (5.18) in Eqn (5.17) yields

$$n_0 = n_i \exp\left[\frac{-(e\phi_{Fn})}{kT}\right] \tag{5.19}$$

leading to

$$\phi_{Fn} = \frac{-kT}{e} \ln\left(\frac{n_0}{n_i}\right) \tag{5.20}$$

Putting $n_0 = N_d$ = donor concentration in Eqn (5.20) results in

$$\phi_{Fn} = \frac{-kT}{e} \ln\left(\frac{N_d}{n_i}\right) \tag{5.21}$$

Similarly, the hole concentration, p_0, in the p-region is given by

$$p_0 = N_a = n_i \exp\left[\frac{E_{Fi} - E_F}{kT}\right] \tag{5.22}$$

Once again, we can define the potential ϕ_{Fp} in the form

$$e\phi_{Fp} = E_{Fi} - E_F \tag{5.23}$$

Using Eqn (5.23) in Eqn (5.22) and rearranging, we get

$$\phi_{Fp} = \frac{kT}{e} \ln\left(\frac{N_a}{n_i}\right) \tag{5.24}$$

Putting Eqs (5.21) and (5.24) in Eqn (5.14) for V_{bi} leads to

$$V_{bi} = \frac{kT}{e} \ln\left(\frac{N_a N_d}{n_i^2}\right) = V_t \ln\left(\frac{N_a N_d}{n_i^2}\right) \tag{5.25}$$

where the factor $V_t = kT/e$ is referred to as the thermal voltage. The terms N_a and N_d represent the net acceptor and donor concentrations in the p and n regions, respectively.

5.2.3 Built-in Electric Field

The depletion region has a separation of positive and negative space charge densities. Such a region also has an associated electric field. We will now derive an expression for this electric field. Figure 5.16 shows a schematic diagram of the volume charge density

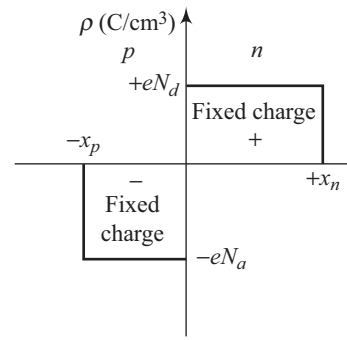


Fig. 5.16 Volume charge density in a uniformly doped p-n junction following abrupt junction approximation

disribution. The diagram assumes a uniform doping and an abrupt junction. The space charge region extends from $x = -x_p$ in the *p*-region to $x = x_n$ in the *n*-region. For a one-dimensional analysis, Poisson equation is given by

$$\frac{d^2\phi(x)}{dx^2} = \frac{-\rho(x)}{\varepsilon_s} = \frac{-dE(x)}{dx} \tag{5.26}$$

where $\phi(x)$ represents the position-dependent potential; $E(x)$, the position-dependent electric field; $\rho(x)$, the position-dependent charge density; and ε_s represents the permittivity of the semiconductor. From Fig. 5.16 we can infer

$$\rho(x) = -eN_a \text{ for } -x_p < x < 0 \tag{5.27}$$

and

$$\rho(x) = eN_d \text{ for } 0 < x < x_n \tag{5.28}$$

Integrating Eqn (5.26) leads to

$$E = \int \frac{\rho(x)}{\varepsilon_s} dx \tag{5.29}$$

For the *p*-region, we can write

$$E = -\int \frac{eN_a}{\varepsilon_s} dx = \frac{-eN_a x}{\varepsilon_s} + C_p \tag{5.30}$$

where C_p is a constant of integration. At thermal equilibrium, no currents flow and, therefore, the neutral *p*-region is not expected to have any electric field. Putting $E = 0$ at $x = -x_p$ in Eqn (5.30) results in

$$0 = \frac{eN_a x_p}{\varepsilon_s} + C_p$$

which implies

$$C_p = \frac{-eN_a x_p}{\varepsilon_s} \tag{5.31}$$

Substituting the expression for C_p into Eqn (5.30), we get for the electric field in the *p*-region,

$$E = \frac{-eN_a}{\varepsilon_s}(x + x_p) \text{ for } -x_p \le x \le 0 \tag{5.32}$$

Similarly, the electric field in the *n*-region is given by

$$E = \int \frac{(eN_d)\, dx}{\varepsilon_s} = \frac{eN_d x}{\varepsilon_s} + C_n \tag{5.33}$$

The constant of integration C_n is evaluated by putting $E = 0$ at $x = x_n$, since the electric field can be assumed to be zero in the neutral *n*-region. One must also remember that *E* is a continuous function. Thus

$$E = 0 = \frac{eN_d}{\varepsilon_s} x_n + C_n \tag{5.34}$$

which leads to

$$C_n = \frac{-eN_d\ x_n}{\varepsilon_s} \tag{5.35}$$

Putting Eqn (5.35) in Eqn (5.33) results in

$$E = \frac{-eN_d}{\varepsilon_s}(x_n - x) \tag{5.36}$$

The electric field must be continuous at the metallurgical junction, i.e., at $x = 0$. Equating Eqn (5.32) and Eqn (5.36) at $x = 0$ leads to

$$N_a x_p = N_d x_n \tag{5.37}$$

Thus the number of negative charges per unit area in the *n*-region equals the number of positive charges in the *p*-region.

The electric field in the depletion region is plotted in Fig. 5.17.

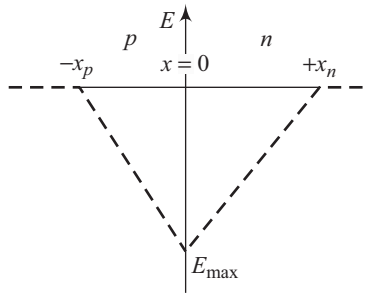


Fig. 5.17 Electric field in the depletion region for a uniformly doped *p*-*n* junction

The electric field has the following important characteristics:

(a) It is directed from the *n*- to the *p*-region.

(b) For the uniformly doped condition assumed, the electric field is a linear function of the distance through the junction.

(c) Its maximum value occurs at the metallurgical junction.

(d) The depletion region has an in-built electric field even in the absence of any applied voltage across the junction.

The potential across the junction can be evaluated by integrating the expressions for electric field. Using Eqn (5.32), we get for the potential in the *p* region,

$$\phi(x) = -\int E(x)dx = \int \frac{eN_a}{\varepsilon_s}(x + x_p)\,dx \tag{5.38}$$

This leads to

$$\phi(x) = \frac{eN_a}{\varepsilon_s}\left(\frac{x^2}{2} + x_p x\right) + C_p' \tag{5.39}$$

where C_p' is a constant of integration, which we will now evaluate. Since we are more interested in the potential difference across the junction, nothing is lost by assuming the potential to be zero at $x = x_p$, i.e., at the edge of the depletion region in the *p*-region. Thus,

$$\phi(-x_p) = 0 = \frac{eN_a}{\varepsilon_s}\left(\frac{x_p^2}{2} - x_p^2\right) + C_p'$$

leading to

$$C'_p = \frac{eN_a}{2\varepsilon_s} x_p^2 \tag{5.40}$$

Using Eqn (5.40) in Eqn (5.39) yields

$$\phi(x) = \frac{eN_a}{2\varepsilon_s}(x + x_p)^2 \quad \text{for } -x_p \le x \le 0 \tag{5.41}$$

Similarly, the potential in the *n*-region can be evaluated by using

$$\phi(x) = \int \frac{eN_d}{\varepsilon_s}(x_n - x)\, dx \tag{5.42}$$

This yields

$$\phi(x) = \frac{eN_d}{\varepsilon_s}\left(x_n x - \frac{x^2}{2}\right) + C'_n \tag{5.43}$$

Potential $\phi(x)$ is a continuous function, therefore, Eqn (5.43) must become equal to Eqn (5.41) at the metallurgical junction, i.e., at $x = 0$. Thus

$$C'_n = \frac{eN_a}{2\varepsilon_s} x_p^2 \tag{5.44}$$

Putting Eqn (5.44) in Eqn (5.43) yields

$$\phi(x) = \frac{eN_d}{\varepsilon_s}\left(x_n x - \frac{x^2}{2}\right) + \frac{eN_a}{2\varepsilon_s} x_p^2 \tag{5.45}$$

for $0 \le x \le x_n$

A plot of the potential function is shown in Fig. 5.18 displaying a quadratic dependence.

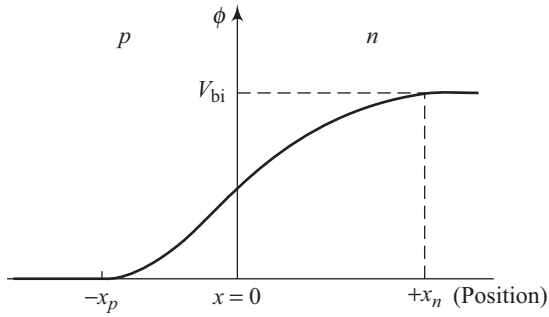


Fig. 5.18 Electric potential in the space charge region for a uniformly doped *p-n* junction

The built-in potential, V_{bi}, is the value of $\phi(x)$ at $x = x_n$ and is given by

$$V_{bi} = |\phi(x = x_n)| = \frac{e}{2\varepsilon_s}(N_d x_n^2 + N_a x_p^2) \tag{5.46}$$

Electron potential energy PE = $-e\phi$, which is also clearly a quadratic function of the distance through the space charge region.

5.2.4 Space Charge Region Width

Using Eqn (5.37), we can write

$$x_p = \frac{N_d x_n}{N_a} \tag{5.47}$$

Putting the expression for x_p in Eqn (5.46), we get

$$V_{\text{bi}} = \frac{e}{2\varepsilon_s}\left[N_d x_n^2 + N_a\left(\frac{N_d}{N_a}\right)^2 x_n^2\right]$$

This implies

$$V_{\text{bi}} = \frac{e}{2\varepsilon_s}\left[N_d x_n^2 + \frac{N_d^2}{N_a} x_n^2\right]$$

resulting in

$$V_{\text{bi}} = \frac{e}{2\varepsilon_s}\left(\frac{N_d}{N_a}\right) x_n^2 [N_a + N_d]$$

This leads to

$$x_n = \left\{\frac{2\varepsilon_s V_{\text{bi}}}{e}\left(\frac{N_a}{N_d}\right)\left[\frac{1}{(N_a + N_d)}\right]\right\}^{1/2} \tag{5.48}$$

If we chose to write x_n in terms of x_p using Eqn (5.37) and solved for x_p, we would have obtained

$$x_p = \left\{\frac{2\varepsilon_s V_{\text{bi}}}{e}\left(\frac{N_d}{N_a}\right)\left[\frac{1}{(N_a + N_d)}\right]\right\}^{1/2} \tag{5.49}$$

The total space charge region width, W, is given by the addition of Eqs (5.48) and (5.49), which leads to

$$W = x_n + x_p = \left\{\frac{2\varepsilon_s V_{\text{bi}}}{e}\left[\frac{N_a + N_d}{N_a N_d}\right]\right\}^{1/2} \tag{5.50}$$

- High-intensity laser radiations can be used to remove damages associated with ion implantation.
- For a one-sided abrupt junction, Eqn (5.50) reduces to

$$W = \sqrt{\frac{2\varepsilon_s V_{\text{bi}}}{eN_d}} \quad \text{if } N_a \gg N_d$$

- A more accurate treatment considering the majority carrier contribution in addition to the impurity concentration leads to

$$W = \sqrt{\frac{2\varepsilon_s}{eN_d}\left(V_{\text{bi}} - \frac{2kT}{e}\right)}$$

for a one-sided abrupt junction with $N_a \gg N_d$.

- At thermal equilibrium, depletion-layer widths of abrupt junctions in silicon are about $8L_D$, where L_D is the Debye length given by

$$L_D = \sqrt{\frac{\varepsilon_s kT}{e^2 N_d}} \text{ for } N_a \gg N_d$$

5.3 Reverse-biased *p-n* Junction

In the preceding section, the *p-n* junction did not have any externally applied voltage or bias across it. We now consider the situation where an external bias is applied across a *p-n* junction in such a manner that the positive end of the external source is connected to the *n*-region and the negative end is connected to the *p*-region of the *p-n* junction.

5.3.1 Energy Band Diagram

With an externally applied potential, the *p-n* junction is no longer in equilibrium. The Fermi level is, therefore, no longer constant throughout the system. Figure 5.19 shows the energy band diagram of a *p-n* junction in reverse bias.

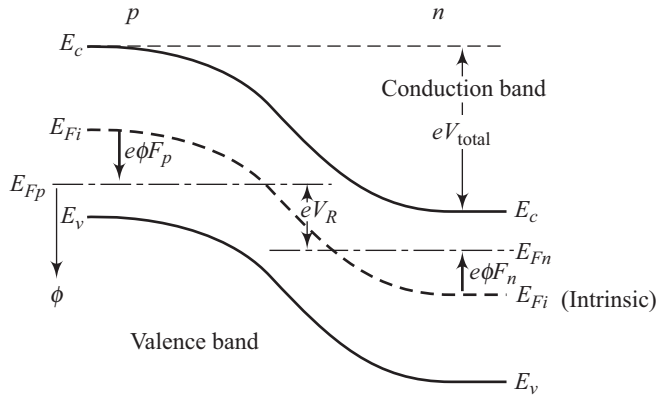


Fig. 5.19 Energy band diagram for a *p-n* junction in reverse bias

The Fermi energy level on the *n*-side is below the Fermi energy level on the *p*-side since the positive potential is downward in the band diagram. The

difference between the Fermi level on the *p*-side and the Fermi level on the *n*-side is equal to the applied reverse voltage, V_R, in units of energy.

The total potential barrier, V_{TB}, is given by

$$V_{TB} = |\phi_{Fn}| + |\phi_{Fp}| + V_R \tag{5.51}$$

where V_R represents the applied reverse-bias voltage. From Eqn (5.51), it is clear that the potential barrier increases with the application of reverse voltage.

Using Eqn (5.14), we can rewrite Eqn (5.51) in the form

$$V_{TB} = V_{bi} + V_R \tag{5.52}$$

5.3.2 Space Charge Width and Electric Field

A *p-n* junction with applied reverse bias is shown schematically in Fig. 5.20.

The electric field in the space charge region is also shown in the figure. The electric fields in the neutral *n*- and *p*-region can be considered to be negligible. Thus, under the application of the reverse bias, the electric field in the space charge region increases above the thermal equilibrium value. We know that electric field originates from positive charges and terminates on negative charges. An increased electric field, therefore, implies an increase in the number of positive and negative charges. Since the impurity doping concentration is fixed, the number of positive and negative charges can increase only with an increase in the space charge width, *W*. Thus, with an applied reverse-bias voltage, V_R, the depletion region width, *W*, increases above its thermal-equilibrium value.

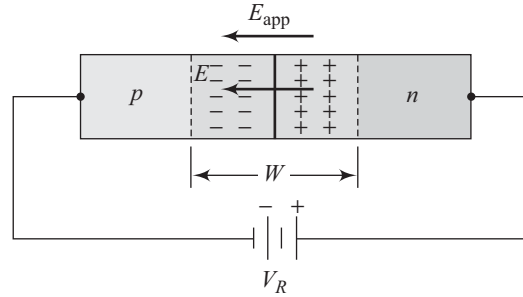


Fig. 5.20 A *p-n* junction, with an applied reverse-bias voltage

The total space charge width, *W*, with an applied reverse bias can be obtained from Eqn (5.50) by replacing V_{bi} with $(V_{bi} + V_R)$. Thus, we get

$$W = \left\{ \frac{2\varepsilon_s (V_{bi} + V_R)}{e} \left[\frac{N_a + N_d}{N_a N_d} \right] \right\}^{1/2} \tag{5.53}$$

From Eqn (5.53), we can conclude that the total space charge width increases with the application of an external reverse bias.

The electric field in the depletion region increases with the application of reverse bias, as already discussed. Fields are still given by Eqs (5.30) and (5.33) but require the replacement of x_n and x_p values appropriate for a junction with an applied reverse bias. These can be obtained using Eqs (5.48) and (5.49) by replacing V_{bi} with $(V_{bi} + V_R)$. The maximum electric field E_{max} at the

metallurgical junction can be obtained from Eqs (5.48) and (5.49) by putting $x = 0$. Thus

$$E_{\max} = \frac{-eN_d x_n}{\varepsilon_s} = \frac{-eN_a x_p}{\varepsilon_s} \tag{5.54}$$

leading to

$$E_{\max} = -\left\{ \frac{2e(V_{\text{bi}} + V_R)}{\varepsilon_s} \left(\frac{N_a N_d}{N_a + N_d} \right) \right\}^{1/2} \tag{5.55}$$

Using Eqn (5.53) in Eqn (5.55), we can conclude that

$$E_{\max} = \frac{-2(V_{\text{bi}} + V_R)}{W} \tag{5.56}$$

where W represents the total space charge width.

5.3.3 Depletion Capacitance

Any separation of positive and negative charges with a medium in between has a capacitance associated with it. A *p-n* junction, therefore, has a capacitance associated with the depletion region. Figure 5.21 illustrates the charge density distribution across an abrupt junction with uniform doping.

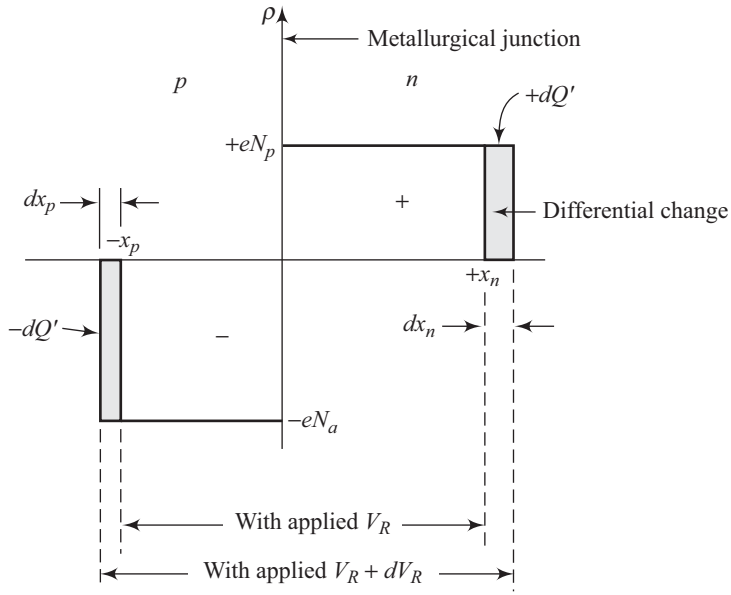


Fig. 5.21 Differential change in the space charge width with a differential change in reverse-bias voltage

An increase in the applied reverse-bias voltage from V_R to $(V_R + dV_R)$ leads to an increase in the positive charge in the *n*-region and negative charge in the *p*-region. The depletion region capacitance C_{dep} is then given by

$$C_{dep} = \frac{dQ'}{dV_R} \tag{5.57}$$

where dQ' is the differential charge (in C/cm^2) due to the incremental reverse-bias increase dV_R.
Clearly

$$dQ' = eN_d\, dx_n = eN_a\, dx_p \tag{5.58}$$

From Eqn (5.57), we get

$$C_{dep} = eN_d \frac{dx_n}{dV_R} = eN_a \frac{dx_p}{dV_R} \tag{5.59}$$

From Eqn (5.48), with V_{bi} replaced by $(V_{bi} + V_R)$, we can write

$$x_n = \left\{ \frac{2\varepsilon_s (V_{bi} + V_R)}{e} \left(\frac{N_a}{N_d} \right) \left[\frac{1}{(N_a + N_d)} \right] \right\}^{1/2} \tag{5.60}$$

Using Eqn (5.60) in Eqn (5.59) leads to

$$C_{dep} = \left\{ \frac{e\varepsilon_s N_a N_d}{2(V_{bi} + V_R)(N_a + N_d)} \right\}^{1/2} \tag{5.61}$$

The same equation can be obtained by using the equation for x_p instead of x_n. Using Eqs (5.53) and (5.61), we can write

$$C_{dep} = \frac{\varepsilon_s}{W} \tag{5.62}$$

One must, however, remember that C_{dep} given by Eqs (5.61) and (5.62) is the depletion region capacitance per unit area (in F/cm^2). A little thinking reveals the similarity between Eqn (5.62) and the equation for capacitance per unit area of a parallel-plate capacitor. Furthermore, the depletion capacitance is a function of applied bias through W.

5.3.4 One-sided Abrupt Junction

If one of the two regions in a *p-n* junction is heavily doped in comparison to the other, such a junction is called one-sided abrupt junction. Suppose $N_a \gg N_d$ in a *p-n* junction, then such a junction is referred to as a p^+-*n* junction. The opposite situation leads to an n^+-*p* junction. For a p^+-*n* junction, the total depletion region width W, using Eqn (5.53), is given by

$$W \approx \left\{ \frac{2\varepsilon_s (V_{bi} + V_R)}{eN_d} \right\}^{1/2} \tag{5.63}$$

For a p^+-n junction,

$$x_p \ll x_n \tag{5.64}$$

since the space charge width decreases with the increasing dopant concentration. Thus

$$W \approx x_n \tag{5.65}$$

From Eqn (5.65), one can conclude that for a one-sided abrupt junction, the depletion region is almost entirely into the low-doped region of the junction.

The depletion capacitance for a p^+-n junction is similarly given by

$$C_{\text{dep}} \approx \left\{ \frac{e\varepsilon_s N_d}{2(V_{\text{bi}} + V_R)} \right\}^{1/2} \tag{5.66}$$

From Eqn (5.66), we can obtain

$$\left(\frac{1}{C_{\text{dep}}} \right)^2 = \frac{2(V_{\text{bi}} + V_R)}{e\varepsilon_s N_d} \tag{5.67}$$

A plot of inverse capacitance squared as a function of reverse-bias voltage is a straight line as shown in Fig. 5.22.

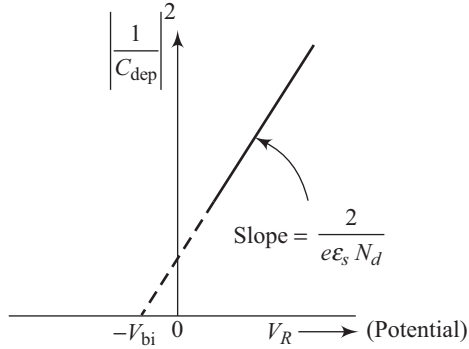


Fig. 5.22 $\left(\frac{1}{C_{\text{dep}}} \right)^2$ versus V_R for a uniformly doped p-n junction

5.4 Junctions With Nonuniform Doping

In the discussion about p-n junctions, so far, we have assumed the constituent semiconductor regions to be uniformly doped. This is an ideal situation and does not exist in actual p-n junction structures. Moreover, nonuniform doping profiles are sometimes deliberately introduced to obtain special characteristics.

5.4.1 Linearly Graded Junctions

Suppose we take a uniformly doped n-type semiconductor and diffuse p-type impurities into it. The impurity doping concentrations then looks typically like Fig. 5.23. The point $x = x_J$ corresponds to the metallurgical junction, i.e., the point at which the donor and acceptor concentrations become equal.

As can be seen from this figure, the net *p*-type concentration on the *p*-side of the metallurgical junction is approximately linear. Similarly, the net *n*-type concentration on the *n*-side of the metallurgical junction is also approximately linear. It is, therefore, worthwhile to study the linearly graded *p-n* junction in some detail. Figure 5.24 is a schematic diagram of the space charge density in the depletion region of a linearly graded junction. Assuming the metallurgical junction to be at $x = 0$, we can write for the space charge density $\rho(x)$

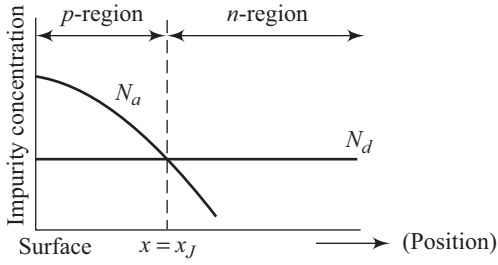


Fig. 5.23 Impurity concentrations within a *p-n* junction with a non-uniformly doped *p*-region

$$\rho(x) = eax \tag{5.68}$$

where a represents the gradient of the net impurity concentration.

We can write the relevant Poisson equation as

$$\frac{dE}{dx} = \frac{\rho(x)}{\varepsilon_s} = \frac{eax}{\varepsilon_s} \tag{5.69}$$

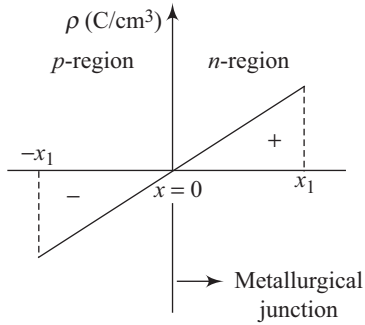


Fig. 5.24 Volume space charge density for a linearly graded *p-n* junction

Equation (5.69) can be integrated to yield the electric field

$$E = \int \frac{eax}{\varepsilon_s} dx = \frac{ea}{2\varepsilon_s}(x^2 - x_1^2) \tag{5.70}$$

Thus the electric field for a linearly graded junction is a quadratic function of distance. We should recollect at this stage that the electric field for a uniformly doped junction is a linear function of distance. The electric field attains its maximum value at the metallurgical junction. From Eqn (5.70) it is also clear that the electric field is zero at $x = +x_1$ and $x = -x_1$. In reality the electric field in a non-uniformly doped semiconductor does not become zero, but becomes vanishingly small. Therefore, for all practical purposes, E can be assumed to be zero in the bulk of the semiconductor.

The electric potential can be determined by integrating the equation for electric field

$$\phi(x) = -\int E\, dx \tag{5.71}$$

Assuming $\phi = 0$ at $x = -x_1$, we get for the potential across the junction,

$$\phi(x) = -\frac{ea}{2\varepsilon_s}\left(\frac{x^3}{3} - x_1^2 x\right) + \frac{ea}{3\varepsilon_s}x_1^3 \tag{5.72}$$

Obviously the value of $\phi(x)$ at $x = +x_1$ must equal the built-in potential V_{bi}, *i.e.*

$$\phi(x_1) = \frac{2}{3}\frac{eax_1^3}{\varepsilon_s} = V_{bi} \tag{5.73}$$

We can obtain another equation for V_{bi} for a non-uniformly doped junction by suitably modifying Eqn (5.25). This yields

$$V_{bi} = \frac{kT}{q} \ln\left[\frac{N_d(x_1)\, N_a(-x_1)}{n_i^2}\right] \tag{5.74}$$

where $N_d(x_1)$ N_a $(-x_1)$ represents the donor and acceptor concentrations at the space charge region edges, respectively. From Eqn (5.68), we can conclude

$$N_d(x_1) = ax_1 \tag{5.75a}$$

and

$$N_a(-x_1) = ax_1 \tag{5.75b}$$

Substituting Eqs (5.75a) and (5.75b) into Eqn (5.74) leads to

$$V_{bi} = \frac{kT}{q} \ln\left(\frac{ax_1}{n_i}\right)^2 \tag{5.76}$$

Under reverse bias, the potential barrier increases and is obtained by adding the reverse-bias voltage V_R to V_{bi}. Then the total potential barrier becomes $V_{bi} + V_R$.

Using Eqn (5.73), we get for x_1 for a reverse-biased linearly graded junction,

$$x_1 = \left\{\frac{3}{2}\frac{\varepsilon_s}{ea}(V_{bi} + V_R)\right\}^{1/3} \tag{5.77}$$

Figure 5.25 is a schematic representation of the differential change in the space charge width with a differential change in applied reverse bias.

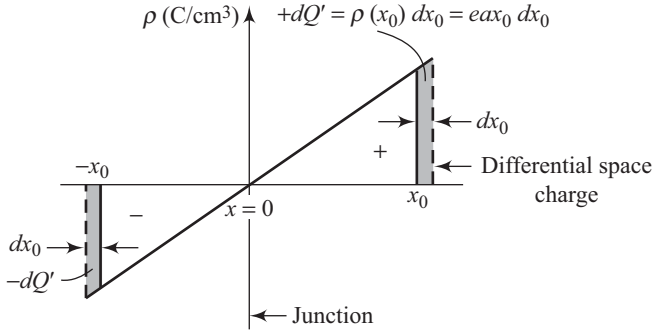


Fig. 5.25 Differential change in space charge width with a differential change in reverse-bias voltage for a linearly graded *p-n* junction

The depletion capacitance C_{dep} is then given by

$$C_{dep} = \frac{dQ'}{dV_R} = (eax_1)\frac{dx_1}{dV_R} \tag{5.78}$$

Using Eqn (5.77) in Eqn (5.78) results in

$$C_{\text{dep}} = \left\{ \frac{ea\varepsilon_s^2}{12(V_{\text{bi}} + V_R)} \right\}^{1/3} \tag{5.79}$$

Thus C_{dep} is proportional to $(V_{\text{bi}} + V_R)^{-1/3}$ for a lineraly graded junction in comparison to $(V_{\text{bi}} + V_R)^{-1/2}$ for a uniformly doped junction. This leads to a slower dependence of C_{dep} on reverse-bias voltage for a linearly graded junction.

5.4.2 Hyper-abrupt Junctions

For a generalized one-sided p^+-n junction, the n-type doping concentration is neither uniform nor linearly graded. Figure 5.26 shows generalized doping profiles for a one-sided p^+-n junction. For $x > 0$, the n-type doping concentration can be written as

$$N = Bx^m \tag{5.80}$$

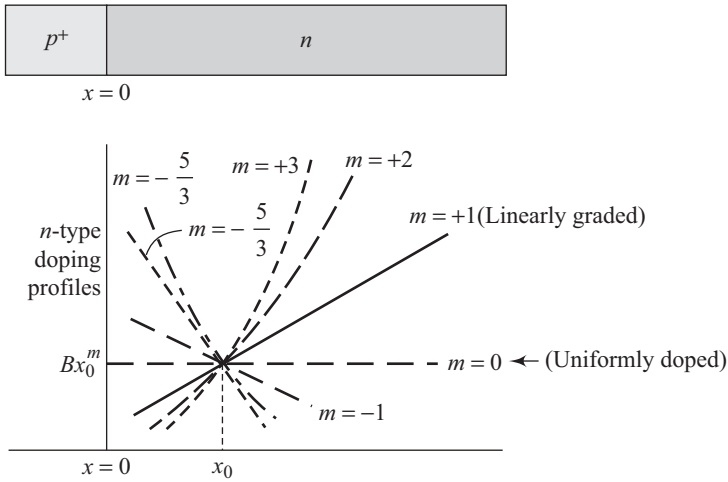


Fig. 5.26 Generalized doping profiles for a one-sided p^+-n junction

From Eqn (5.80), it is clear that $m = 0$ leads to a uniformly doped junction, whereas $m = +1$ corresponds to a linearly graded junction.

A negative value of m implies that the n-type doping is larger near the metallurgical junction than in the bulk of the semiconductor. Such a junction is called a *hyper-abrupt junction*. One must appreciate that the n-type doping at $x = 0$ cannot be evaluated using Eqn (5.80) when m is a negative quantity. Equation (5.80) is, therefore, used to find the n-type doping near $x = x_1$, the edge of the space charge region.

Using an analysis similar to the one given in the earlier section, the junction capacitance can be shown to be

$$C_{\text{dep}} = \left[\frac{eB\varepsilon_s^{(m+1)}}{(m+2)\,(V_{\text{bi}} + V_R)} \right]^{1/(m+2)} \tag{5.81}$$

Thus C_{dep} is a very strong function of reverse-bias when m acquires negative values.

5.5 Varactor Diode

Varactor is the shortened form of a combination of the two words *variable* and *reactor*. It is a device whose reactance can be varied in a controlled manner with applied bias. We have already seen in the preceding section that the depletion capacitance of a *p-n* junction is dependent upon the applied reverse-bias. A varactor diode uses this voltage-variable property of the junction capacitance.

From Eqn (5.81), we can see that C_{dep} can be written in the form

$$C_{\text{dep}} = C_1\,(V_{\text{bi}} + V_R)^{-1/(m+2)} \tag{5.82}$$

For $V_R \gg V_{\text{bi}}$, Eqn (5.82) leads to

$$C_{\text{dep}} \cong C_1 (V_R)^{-1/(m+2)} \tag{5.83}$$

Putting $m = -3/2$, as a special case of hyper-abrupt junction, in Eqn (5.83), yields

$$C_{\text{dep}} \cong C_1 (V_R)^{-2} \tag{5.84}$$

Suppose such a capacitor is used with an inductor L in a resonant circuit. The resonant frequency, ω_r, can then be written as

$$\omega_r = \frac{1}{\sqrt{LC}} \frac{1}{\sqrt{V^{-2}}} V_R \tag{5.85}$$

Thus the resonant frequency of such a circuit would be linearly dependent upon the applied reverse bias.

Varactor diodes are required for different applications. Specific requirements can be fulfilled by using suitable doping profiles.

5.6 Junction Breakdown

A reverse-biased *p-n* junction diode develops a small constant current when reverse biased. At a particular reverse-bias voltage, called the *breakdown voltage*, the reverse-bias current increases rapidly with the applied voltage. We will study the junction breakdown phenomenon in this section. Reverse-bias breakdown in a *p-n* junction arises due to two physical mechanisms. These two physical mechanisms are the *Zener effect* and the *avalanche effect.* These mechanisms result in the corresponding breakdowns.

5.6.1 Zener Breakdown

Suppose we take a heavily doped *p*-*n* junction and apply reverse bias across it. At a reasonably low voltage, a crossing of the bands takes place. At this point, a large number of empty states in the *n*-side conduction band get aligned opposite to the many filled states on the *p*-side valence band. Since the *p*-*n* junction is heavily doped, the barrier separating these two bands is very narrow. This leads to tunnelling of electrons from the valence band on the *p*-side into the conduction band on the *n*-side. This mechanism is called the *Zener effect*. The resultant breakdown, called the Zener breakdown, leads to a rapidly increasing reverse-bias current, as shown in Fig. 5.27(c). The crossing of bands is shown schematically in this figure.

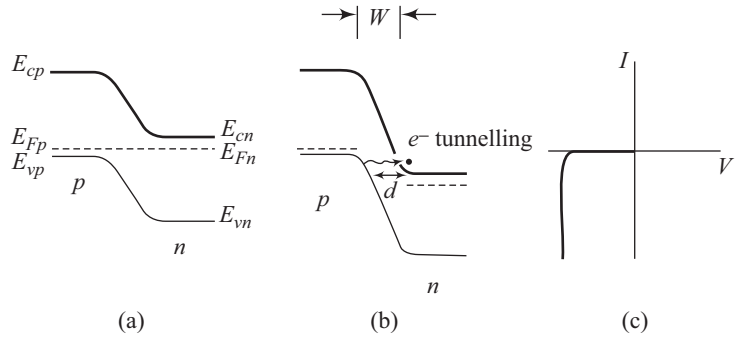


Fig. 5.27 The Zener effect: (a) heavily doped *p*-*n*-junction (b) reverse bias with electron tunnelling from *p* to *n* (c) *I*-*V* characteristic.

For a reasonable tunnelling current, there must exist a collection of electrons separated from empty state in large numbers by a narrow barrier of finite height. The tunnelling probability is dependent upon the width, *d*, of the barrier. This requires a sharp metallurgical junction with high doping, resulting in a short transition region *W* extending on either side of the junction [see Fig. 5.27(b)]. Initially, as the bands cross, at a few tenths of a volt for a heavily doped junction, the tunnelling distance, *d*, is too large to result in appreciable tunnelling. As the reverse bias increases, *d* becomes smaller, because increasing electric fields result in steeper slopes for band edges. This argument assumes that the transition region width, *W*, does not increase appreciably with increased reverse bias. This necessitates low applied voltages and heavy doping on either side of the junction.

Zener breakdown also can be visualized using the simple covalent bonding. The reverse bias applied across a heavily doped junction results in a large electric field within the transition region *W*. At a certain critical electric field strength, the electrons in the covalent bonds are torn apart from the bonds by the field and then accelerated to the *n*-side of the junction. This process is also called *field ionization* and occurs at electric field strengths of the order of 10^6 V/cm.

5.6.2 Avalanche Breakdown

The electrons and holes moving across the space charge region acquire energy due to the electric field existing in the region. For high enough reverse bias, the energy acquired by the charge carriers is sufficient to create electron–hole pairs due to collisions with atomic electrons. The charge carriers created due to this collision process, in turn, themselves acquire sufficient energy to ionize further atoms. The whole sequence thus generates an avalanche of carriers as shown in Fig. 5.28.

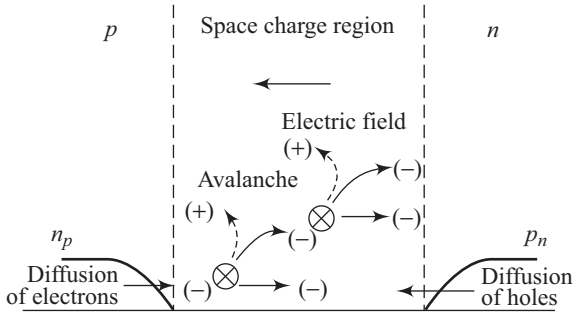


Fig. 5.28 Sequence generating the avalanche process

It is important to realize that in most *p-n* junctions the avalanche breakdown is the dominant breakdown mechanism.

Suppose a reverse-bias electron current, I_{n0}, enters at one end of the depletion region at $x = 0$, as shown in Fig. 5.29.

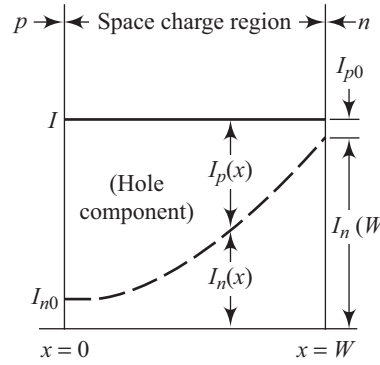


Fig. 5.29 Electron and hole current components in the space charge region during avalanche process

We have qualitatively understood that the electron number density and hence electron current increases as we travel through the depletion region, due to the avalanche process. If we assume the depletion region width to be W, we can write for electron current I_n at $x = W$

$$I_n(W) = M_n I_{n0} \tag{5.86}$$

where M_n represents a multiplication factor. The corresponding hole current reaches its maximum value at $x = 0$, since it enters the depletion region from the *n*-region to the *p*-region.

Let α_n and α_p represent the electron and hole ionization rates, respectively. Here α_n represents the number of electron–hole pairs generated per unit length by an electron. The incremental electron current at an arbitrary point x in the depletion region can then be written as

$$d I_n(x) = I_n(x)\, \alpha_n dx + I_p(x)\, \alpha_p\, dx \tag{5.87}$$

Rewriting Eqn (5.87) leads to

$$\frac{dI_n(x)}{dx} = I_n(x)\alpha_n + I_p(x)\alpha_p \tag{5.88}$$

The total current, I, is a constant and is given by

$$I = I_n(x) + I_p(x) \tag{5.89}$$

This leads to

$$I_p(x) = I - I_n(x) \tag{5.90}$$

Substituting $I_p(x)$ from Eqn (5.90) into Eqn (5.88) yields

$$\frac{dI_n(x)}{dx} = I_n(x)\,\alpha_n + [I - I_n(x)]\;\alpha_p \tag{5.91}$$

giving

$$\frac{dI_n(x)}{dx} + (\alpha_p - \alpha_n)\,I_n(x) = \alpha_p I \tag{5.92}$$

Let us assume that the electron and hole ionization rates are equal, i.e.,

$$\alpha_n = \alpha_p = \alpha \tag{5.93}$$

Using Eqn (5.93) in Eqn (5.92) leads to

$$\frac{dI_n(x)}{dx} = \alpha I \tag{5.94}$$

which, after integration across the space charge region, yields

$$I_n(W) - I_n(0) = I\int_0^W \alpha\, dx \tag{5.95}$$

Using Eqn (5.86) in Eqn (5.95) results in

$$\frac{M_n I_{n0} - I_n(0)}{I} = \int_0^W \alpha\, dx \tag{5.96}$$

Also

$$M_n I_{n0} \approx I \;\text{ and }\; I_n(0) = I_{n0} \tag{5.97}$$

Using Eqn (5.97) in Eqn (5.96) implies

$$1 - \frac{1}{M_n} = \int_0^W \alpha\, dx \tag{5.98}$$

Obviously, M_n approaches infinity at the avalanche breakdown voltage. Thus, the avalanche breakdown condition is then given by

$$\int_0^W \alpha\, dx = 1 \tag{5.99}$$

One must, however, remember that the ionization rates are strongly dependent upon electric field. Moreover, since electric field is not constant throughout the space charge region, the actual evaluation of Eqn (5.99) is not easy.

The magnitude of maximum electric field, $E_{\max}$, for a one-side p^+-n junction is, from Eqn (5.54), given by

$$E_{\max} = \frac{eN_d x_n}{\varepsilon_s} \tag{5.100}$$

The corresponding depletion width, x_n, is given by using Eqs (5.63) and (5.65)

$$x_n \simeq \left\{ \frac{2\varepsilon_s V_R}{eN_d} \right\}^{1/2} \tag{5.101}$$

where V_R has been assumed to be much higher than the built-in potential V_{bi}.

Equating $E_{\max}$ to the critical field E_B at breakdown in Eqn (5.100) and putting $V_R = V_B$ (V_B = breakdown voltage), we get [from Eqs (5.100) and (5.101)]

$$E_B = \frac{eN_d x_n}{\varepsilon_s} \tag{5.102}$$

and

$$x_n \simeq \left[\frac{2\varepsilon_s V_B}{eN_d} \right]^{1/2} \tag{5.103}$$

Equation (5.102) can be rewritten as

$$x_n = \frac{\varepsilon_s E_B}{eN_d} \tag{5.104}$$

Equations (5.103) and (5.104) result in

$$\frac{2\varepsilon_s V_B}{eN_d} = \frac{\varepsilon_s^2 E_B^2}{e^2 N_d^2} \tag{5.105}$$

Solving for V_B, we get

$$V_B = \frac{\varepsilon_s E_B^2}{2eN_d} \tag{5.106}$$

where N_d is the doping in the low-doped region of the one-sided p^+-n junction. If this region is designated as the base region and the doping in this region is represented by N_B, then the general form of Eqn (5.106) becomes

$$V_B = \frac{\varepsilon_s E_B^2}{2eN_B} \tag{5.107}$$

The dependence of critical field E_B as a function of doping concentration is shown schematically in Fig. 5.30.

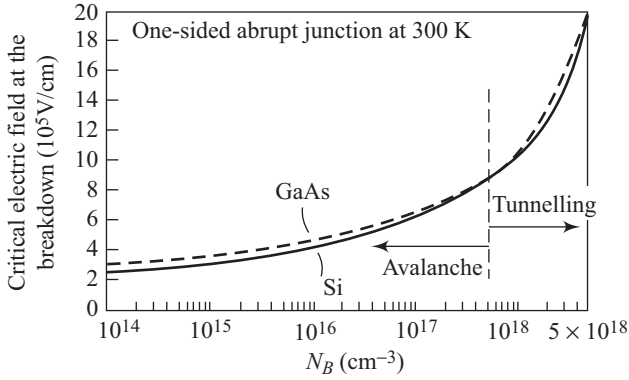


Fig. 5.30 Critical electric field at breakdown in a one-sided junction as a function of impurity doping concentrations for Si and GaAs ($T = 300$ K)

The breakdown voltages for linearly graded junctions would be lower than those for a one-sided abrupt junction. The curvature of a diffused junction would bring down the breakdown voltage further. Figure 5.31 shows plots of breakdown voltages for a one-sided abrupt junction and a linearly graded junction.

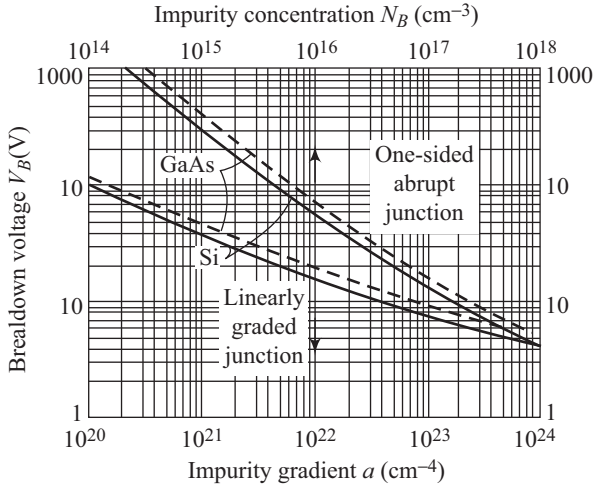


Fig. 5.31 Breakdown voltage versus impurity concentration in uniformly doped and linearly graded junctions for Si and GaAs ($T = 300$ K)

5.7 Tunnel Diode

Interesting dynamic characteristics can be realized by using *p-n* junctions with suitable doping profiles. One such interesting device is a *tunnel diode*. A tunnel

diode consists of a *p-n* junction with both the *n* and *p* regions degenerately doped. The device exhibits a negative differential resistance in a region of its current–voltage characteristics and is a proof of the existence of the phenomenon of tunnelling.

The Fermi level in a degenerately doped *n*-type semiconductor is in the conduction band, and in a degenerately doped *p*-type semiconductor the Fermi level is in the valence band. Thus electrons will exist in the conduction band of an *n*-type material even at $T = 0$ K and a similar statement is true for holes in the valence band of a *p*-type material at $T = 0$ K. Figure 5.32 is a schematic band diagram of a *p-n* junction in thermal equilibrium when both its regions are degenerately doped.

We have already seen that the depletion region width decreases with increasing doping level of the constituent regions. The potential barrier can be approximated by a triangular shape shown in Fig. 5.33.

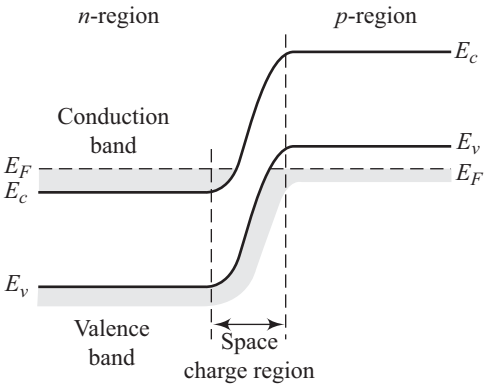


Fig. 5.32 Energy band diagram for a *p-n* junction in thermal equilibrium with *n* and *p* regions degenerately doped

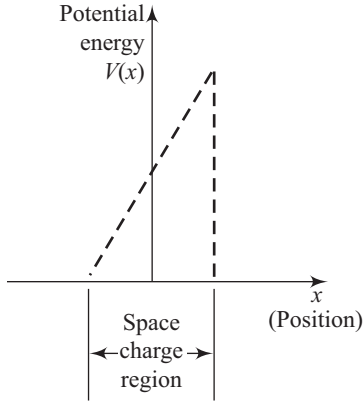


Fig. 5.33 Triangular potential barrier approximation for the tunnel diode

As the barrier width becomes small, the electric field becomes quite large, resulting in a finite probability that an electron can *tunnel* through the forbidden band from one side of the *p-n* junction to the other. The modification of the band diagram as forward bias is increased across the junction is shown in Fig. 5.34.

The following salient features can be noted from having a look at the figure:

(i) A small applied forward bias [Fig. 5.34(b)] results in the electrons in the conduction band of the *n*-region aligning opposite the empty states in the valence band of the *p*-region. A finite probability exists for these electrons to tunnel directly into the empty states, resulting in a small forward-bias tunnelling current as shown.

(ii) At a higher forward bias [Fig. 5.34(c)], the maximum number of electrons in the *n*-region align with the maximum number of empty states in the *p*-region. This leads to a maximum tunnelling current as shown.

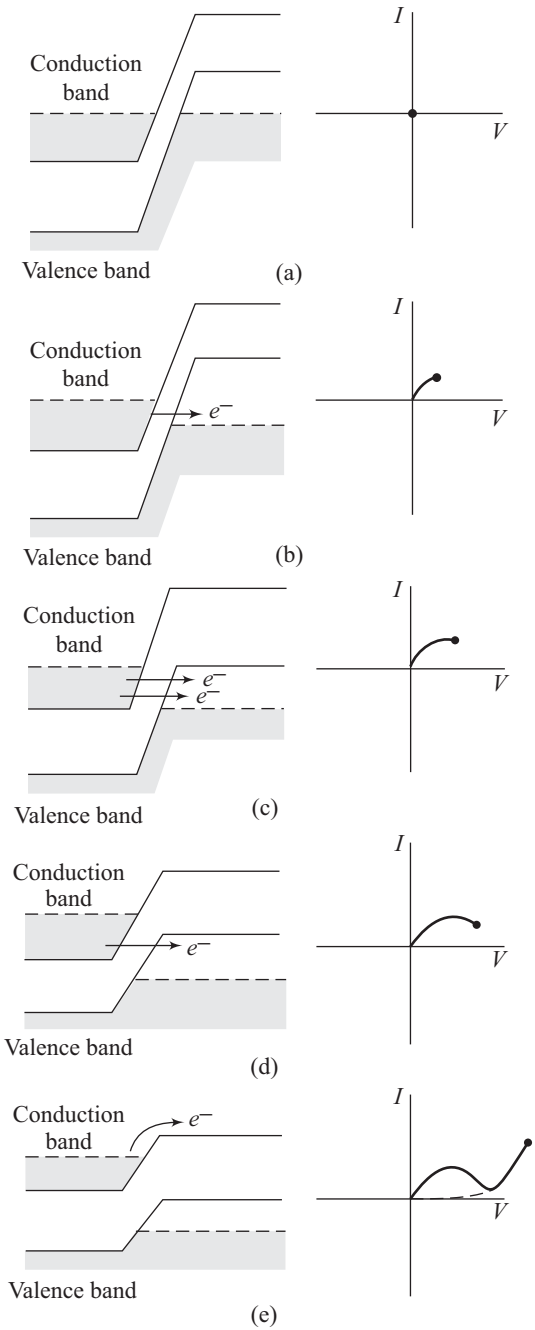


Fig. 5.34 Simplified energy band diagrams and *I-V* characteristics of the tunnel diode at (a) zero bias; (b) small forward bias; (c) a forward bias producing maximum tunnelling current; (d) a higher forward bias showing less tunnelling current; (e) a forward bias for which the diffusion current dominates

(iii) As the forward-bias voltage increases further, the number of electrons on the *n*-side directly aligned with the empty states on the *p*-side, decreases, as shown in Fig. 5.34(d). This results in a decrease in the forward-biased tunnelling current.

(iv) Ultimately, at a high enough forward bias, no electrons on the *n*-side exist directly opposite the available empty states on the *p*-side. The tunnelling current then reduces to zero. The normal diffusion current (discussed in Chapter 6) associated with a forward-biased *p-n* junction however *c*ontinues to exist.

(v) It is thus clear that a certain portion of the forward-biased tunnel diode *I-V* characteristics shows a decrease in current with an increase in applied voltage. This region of the *I-V* characteristic, therefore, demonstrates a *differential negative resistance*. Since the voltage and current ranges resulting in a differential negative resistance are small, any oscillator designed using this property also has low power associated with it. The behaviour of the tunnel diode under applied reverse bias is shown schematically in Fig. 5.35.

Electrons in the valence band on the *p*-side now align directly opposite the empty states in the conduction band on the *n*-side of the *p-n* junction. These electrons can tunnel from the *p*-region into the *n*-region giving rise to a large reverse-bias tunnelling current. This tunnelling-based reverse-bias current keeps on increasing monotonically with increasing reverse-bias voltage.

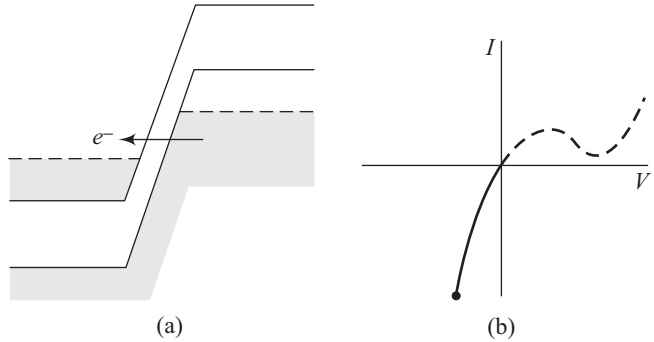


Fig. 5.35 (a) Energy band diagram (b) *I-V* characteristic of a tunnel diode with a reverse-bias voltage

- Thermal instability can also lead to junction breakdown. Reverse current causes heat dissipation, resulting in increase in junction temperature.
- Junction curvature effects have to be taken into account for calculating the junction breakdown voltage of a planar junction.

Solved Problems

5.1 Boron diffusion is to be carried out in an *n*-type silicon substrate of 10 Ω cm resistivity. For achieving this, the substrate is subjected to a constant surface concentration of 5×10^{18} cm^{-3}. The desired *p-n* junction should have a depth of 1 μm. At what temperature should the diffusion be carried out if it is to be completed in 1 hour? Assume electron mobility to be 1300 cm^2/V s.

Solution

Carrier concentration $n = \sigma/\mu_n e$. Substituting the values, we get

$$n = \frac{1}{(10)(1300)(1.6 \times 10^{-19})} = 4.81 \times 10^{14} \text{ cm}^{-3}$$

Since the surface concentration is maintained constant, we have from Eqn (5.9)

$$C(x, t) = C_s \text{ erfc} \frac{x}{2\sqrt{Dt}} \tag{5.1.1}$$

At the junction

$$C(x, t) = n$$

which yields

$$n = C_s \text{ erfc} \frac{x}{2\sqrt{Dt}}$$

This in turn leads to

$$\frac{n}{C_s} = \frac{4.81 \times 10^{14}}{5 \times 10^{18}} = 9.62 \times 10^{-5} \tag{5.1.2}$$

Using Eqn (5.1.2) in Eqn (5.1.1), we get

$$\text{erfc} \frac{x}{2\sqrt{Dt}} = 9.62 \times 10^{-5} \tag{5.1.3}$$

Figure 5.1.1 is a plot of erfc(y) as a function of y.
From the figure, we can see that

$$\text{erfc } (y) = 9.62 \times 10^{-5}$$

$\Rightarrow \quad y = 2.75$

Thus from Eqn (5.1.3), we have

$$\frac{x}{2\sqrt{Dt}} = 2.75$$

which results in

$$\sqrt{D} = \frac{x}{2 \times 2.75 \times \sqrt{t}} = \frac{1}{2 \times 2.75 \times 1} \text{ μm}/\sqrt{\text{h}}$$

yielding

$$\sqrt{D} = 0.18 \text{ μm}/\sqrt{\text{h}}$$

From Fig. 5.10 (b), we get $T \cong 1100°\text{C}$

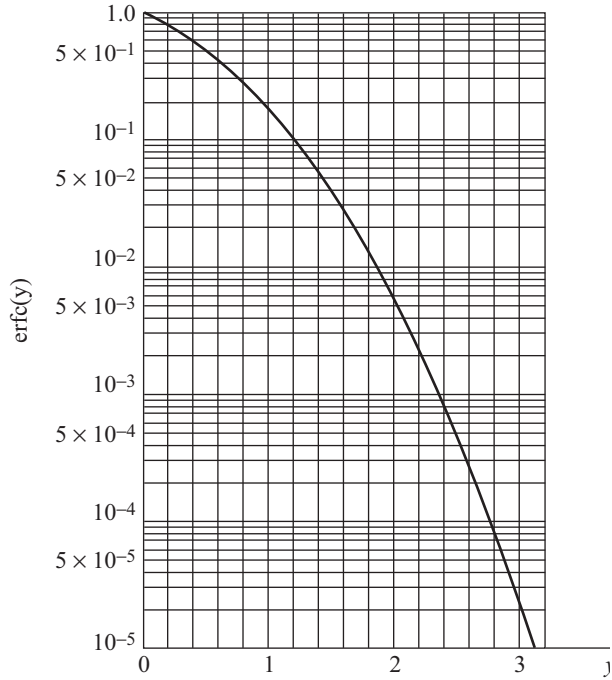


Fig. 5.1.1 The complementary error function plotted in semilogarithmic scale

5.2 A *p-n* junction has doping densities $N_a = 5 \times 10^{18}$ cm^{-3} and $N_d = 5 \times 10^{15}$ cm^{-3} in the two regions. Assuming $n_i = 1.5 \times 10^{10}$ cm^{-3}, calculate the built-in potential at 300 K.

Solution

From Eqn (5.25), we can write

$$V_{\text{bi}} = \frac{kT}{e} \ln\left(\frac{N_a N_d}{n_i^2}\right) \tag{5.2.1}$$

Putting the given values in Eqn (5.2.1) and using the value of $kT/e = 0.026$ at 300 K, we get

$$V_{\text{bi}} = (0.026) \ln\left[\frac{(5\times10^{18})(5\times10^{15})}{(1.5\times10^{10})^2}\right]$$

This leads to

$$V_{\text{bi}} = 0.838 \text{ V}$$

5.3 Calculate the total space charge width for the situation presented in Solved Problem 5.2. Assume the dielectric constant of silicon to be 11.7 and $\varepsilon_0 = 8.85 \times 10^{-14}$ F/cm.

Solution

From Eqn (5.50), we can write

$$W = \left\{ \frac{2\varepsilon_s V_{\text{bi}}}{e} \left[\frac{N_a + N_d}{N_a N_d} \right] \right\}^{1/2} \tag{5.3.1}$$

Putting the given values into Eqn (5.3.1) leads to

$$W = \left[\frac{2(11.7)(8.85 \times 10^{-14})\,(0.838)}{1.6 \times 10^{-19}} \left\{ \frac{(5 \times 10^{18}) + (5 \times 10^{15})}{(5 \times 10^{18}) \times (5 \times 10^{15})} \right\} \right]^{1/2}$$

This gives

$$W = 0.466\ \mu\text{m}$$

5.4 A silicon *p-n* junction has doping concentrations of $N_a = 5 \times 10^{18}$ cm^{-3} and $N_d = 5 \times 10^{15}$ cm^{-3}. Determine the space charge region width of the junction at $T = 300$ K and at $V_R = 4$ V. Assume $\varepsilon_s = 11.7$ and $n_i = 1.5 \times 10^{10}$ cm^{-3}.

Solution

We have calculated the built-in potential of the *p-n* junction in Solved Problem 5.2 to be $V_{\text{bi}} = 0.838$ V.

Depletion region W is given by

$$W = \left\{ \frac{2\varepsilon_s (V_{\text{bi}} + V_R)}{e} \left[\frac{N_a + N_d}{N_a N_d} \right] \right\}^{1/2} \tag{5.4.1}$$

Putting the given values in Eqn (5.4.1), we get

$$W = \left\{ \frac{2 \times (11.7) \times (8.85 \times 10^{-14})\,(0.838 + 4)}{1.6 \times 10^{-19}} \left[\frac{5 \times 10^{18} + 5 \times 10^{15}}{(5 \times 10^{18} \times 5 \times 10^{15})} \right] \right\}^{1/2}$$

yielding

$$W = \left\{ \frac{2 \times (11.7) \times (8.85 \times 10^{-14})\,(4.838)}{1.6 \times 10^{-19}} \left[\frac{5.005 \times 10^{18}}{25 \times 10^{33}} \right] \right\}^{1/2}$$

$$= 1.12 \times 10^{-4}\ \text{cm} = 1.12\ \mu\text{m}$$

5.5 For the *p-n* junction assumed in Solved Problem 5.4, calculate the junction capacitance for $V_R = 3$ V and area A of the junction equal to 5×10^{-4} cm^2.

Solution

The junction capacitance, C_{dep}, per unit area is given by

$$C_{\text{dep}} = \left\{ \frac{e\varepsilon_s N_a N_d}{2(V_{\text{bi}} + V_R)(N_a + N_d)} \right\}^{1/2} \tag{5.5.1}$$

Putting the given values in Eqn (5.5.1) results in

$$C_{\text{dep}} = \left\{ \frac{1.6 \times 10^{-19} \times (11.7) \times (8.85 \times 10^{-14}) (5 \times 10^{18}) (5 \times 10^{15})}{2(0.838 + 3) (5 \times 10^{18} + 5 \times 10^{15})} \right\}^{1/2}$$

yielding

$$C_{\text{dep}} = 1 \times 10^{-8} \text{ F/cm}^2$$

The depletion capacitance C'_{dep} then becomes

$$C'_{\text{dep}} = C_{\text{dep}} A = 1 \times 10^{-8} \times 5 \times 10^{-4} = 5 \times 10^{-12} \text{ F} = 5 \text{ pF}$$

5.6 The contour for an electric field in the space charge region of doped *p-n* junction is given in Fig. 5.6.1. What will you do to make the distance from *O* to $-x_p$ equal to the distance from *O* to $+x_n$?

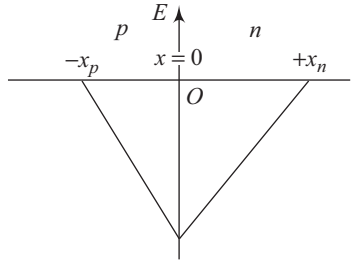


Fig. 5.6.1 Contour of electric field

Solution

Make the doping concentrations on both sides of the *p-n* junction to be equal.

5.7 Under the application of a bias, in a typical *p-n* junction structure, the hole current contribution is around 40% of the total current. As a designer, what will you do to increase this hole current contribution to 80%? Keeping in mind the diode equation, suggest two possible approaches.

Solution

(i) Increase the *p* dopant concentration to double of its existing value.
(ii) Reduce the *n* dopant concentration to half of its existing value.

5.8 The directions of motion of particles in a typical *p-n* junction are shown in Fig. 5.8.1. Draw the direction of currents due to the movements of these carriers.

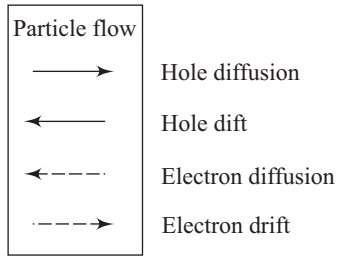
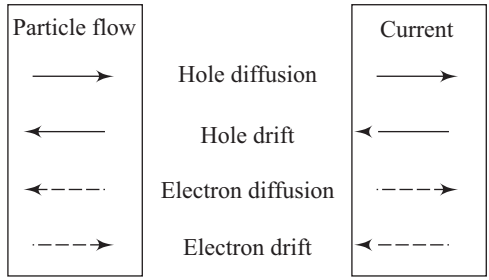


Fig. 5.8.1 Particle flow directions

Solution



5.9 Draw (qualitatively only) the charge density contour within the transition region of an abrupt *p*-*n* junction, when *p*-side is doped to a concentration of 10^{14} cm^{-3} and *n*-side is doped to a concentration of 10^{17} cm^{-3}. Label the contour properly showing all the important quantities.

Solution

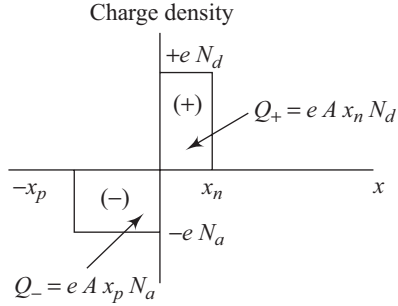
As *p*-side is lightly doped, the formed space charge region will extend more toward *p*-type.

Actually, for a sample of cross-sectional area *A* the total uncompensated charge on either side of junction would be

$$eAx_pN_a = eAx_nN_d$$

Cancellation of *e* and *A* on both sides will lead to the result that the space charge region will extend more toward the lightly doped side.

Therefore, the contour diagram will look like:



5.10 We have a symmetric *p*-*n* silicon junction ($N_a = N_d = 10^{17}$ cm^{-3}). If the peak electric field in the junction at breakdown is 5×10^5 V/cm, what is the reverse breakdown voltage for this junction?

Solution

As is obvious from the statement, a symmetric junction is one in which doping concentration on both *n* and *p* sides is equal.

So, if $N_a = N_d$ and if *W* is the total width of the space charge region, then

$$x_n = x_p = W/2.$$

We also know from Eqn 5.30 that

$$E = \frac{-eN_a}{\varepsilon_s} x = \frac{-eN_a}{\varepsilon_s} \cdot \frac{W}{2} \tag{5.10.1}$$

$$\varepsilon_s = \varepsilon_0 \cdot \varepsilon_{Si} = 8.85 \times 10^{-14} \times 11.8 = 88.76 \times 10^{-14}$$

Also, break-down voltage (V_{br}) is

$$\int_{-x_p}^{x_n} E\, dx = \text{Area under the triangle}$$

$$= \frac{1}{2}(E_{max}\, W) \text{ (refer to Fig. 5.17 in the text)}$$

$$E_{max} = 5 \times 10^5 \text{ V/cm (given)}$$

$$\therefore \quad \int_{-x_p}^{x_n} E\, dx = \frac{1}{2}(5 \times 10^5 \times W) \tag{5.10.2}$$

Now, from Eqn (5.10.1) we get

$$W = \frac{2\varepsilon_s E}{eN_a} \tag{5.10.3}$$

Here, we have neglected the negative sign used in Eqn (5.10.1) because it just depicts the direction.

Now, substituting Eqn (5.10.3) in Eqn (5.10.2) we get

$$\int_{-x_p}^{x_n} E\, dx = \frac{1}{2} \frac{(5 \times 10^5 \times 2 \times E \times \varepsilon_s)}{eN_a}$$

$$= \frac{(5 \times 10^5)^2 \times 88.76 \times 10^{-14}}{1.6 \times 10^{-19} \times 10^{17}}$$

$$= \frac{25 \times 10^{10} \times 88.76 \times 10^{-14}}{1.6 \times 10^2}$$

$$= 13.86 \text{ V}$$

Hence, the breakdown voltage for the given junction is 13.86 V.

5.11 An abrupt p^+-n junction is formed in silicon with a donor doping of $N_d = 10^{15}$ cm^{-3} and acceptor doping of $N_a = 10^{19}$ cm^{-3}. If the length of the n-region is 22 μm and the junction is operated at a reverse bias of 300 V (which is also the breakdown voltage for the junction),

Out of the following options, which mode will the device work in?

(i) Avalanche breakdown (ii) Zener breakdown

(iii) Punch-through (iv) Ohmic region

Justify your answer with the help of suitable set of calculations, wherever required.

Solution

Given $N_d = 10^{15}$ cm^{-3}

$N_a = 10^{19}$ cm^{-3}

For finding out the mode in which the device will work, we must find out the extension of the space charge region towards the *n*-side. If it comes out to be less than the width of the *n*-region, then this is a case of avalanche breakdown, else of punch-through. The other two options, i.e. Zener breakdown and ohmic region are simply not possible because both sides of the junction are not doped heavily enough. Ohmic region is not even a valid term for *p*-*n* junction operation.

So, we can find out x_n using Eqn (5.48)

$$x_n = \left\{ \frac{2\varepsilon_s V_{\text{bi}}}{e} \left(\frac{N_a}{N_d} \right) \left[\frac{1}{N_a + N_d} \right] \right\}^{\frac{1}{2}}$$

$$\varepsilon_S = \varepsilon_{\text{Si}} \times \varepsilon_0$$

$$= 88.76 \times 10^{-14}$$

$V_{\text{bi}} \ll$ Applied voltage (say V)

Hence V_{bi} should be replaced by applied voltage, i.e., 300 V

Putting these values in the expression for x_n

$$x_n = \left\{ \frac{2 \times 88.76 \times 10^{-14} \times 300}{1.6 \times 10^{-19}} \left(\frac{10^{19}}{10^{15}} \right) \left(\frac{1}{10^{19} + 10^{15}} \right) \right\}^{\frac{1}{2}}$$

$$= \left\{ 33285 \times 10^{-14+19+\cancel{19}-15-\cancel{19}} \right\}^{\frac{1}{2}}$$

$$= \left\{ 33285 \times 10^{-10} \right\}^{\frac{1}{2}}$$

$$= 182.44 \times 10^{-5} \text{ cm}$$

$$x_n = 18.24 \ \mu\text{m}$$

As x_n is less than given length of the *n*-region, the device will only have avalanche breakdown.

5.12 A typical Si *p*-*n* junction is shown in Fig. 5.12.1. At 300 K the doping concentrations on the *p* and *n* sides are $N_a = 10^{15}$ cm^{-3} and $N_d = 10^{17}$ cm^{-3}, respectively. The mobility of electrons in the *p*-region is 1400 cm^2/Vs and the carrier lifetime = 10 μs.

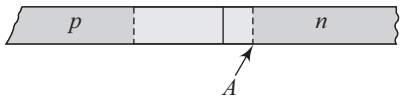


Fig. 5.12.1 Si *p*-*n* junction

(a) Assuming a case of low level injection, find the *total concentration* of holes at A, when the supply of excess carriers at A is continuously replenished with time, under the application of a forward bias voltage of 0.5 V.

(b) In addition to the situation presented in part (a), how much additional voltage is required to be applied to the junction in the form of reverse bias to facilitate the Zener breakdown of this junction?

Solution

(a) Using the relationship

$p_n = p_{n0} \exp(eV/kT)$

where p_{n0} is the hole concentration in the n-region at equilibrium. Here, $V = 0.5$ V and p_{n0} can be found out by applying the law of mass action on the n-side.

$$p_{n0} = \frac{n_i^2}{n_{n0}} = \frac{2.25 \times 10^{20}}{10^{17}}$$

$$= 2.25 \times 10^3 \text{ cm}^{-3}$$

$$\therefore \quad p_n = (2.25 \times 10^3) \exp\left(\frac{0.5e}{0.0259}\right)$$

$$= 5.45 \times 10^{11} \text{ cm}^{-3}$$

(b) For the Zener breakdown the valence band of p-type should be above the conduction band of n-type. This means that $V_{\text{bi}} + V_r$ should at least be equal to the bandgap of silicon, i.e., 1.1 eV. The forward bias of 0.5 V of part (a) implies that there is already a barrier of $V_{\text{bi}} - V_f$. We can find V_{bi} from the expression

$$V_{\text{bi}} = \frac{kT}{e} \ln \frac{N_a N_d}{n_i^2}$$

From this we get

$$V_{\text{bi}} = 0.6949 \text{ V}$$

$\therefore$ the potential already present $= 0.6949 - 0.5$

$= 0.1949$ V

So, for Zener breakdown we need to apply $(1.1 - 0.1949)$ V

$= 0.9051$ V

5.13 For a sharp graded p^+-n junction of area 10^{-4} cm^2, the measured capacitance, under a reverse bias of 20 V, is 12 pF/cm^2. Calculate the donor concentration.

Solution

$$C_j = \frac{A}{2}\left[\frac{2e\varepsilon}{V_0 - V} N_d\right]^{1/2}$$

For $V = -V_r \gg V_0$

$$C_j = \frac{A}{2}\left[\frac{2e\varepsilon}{V_r} N_d\right]^{1/2}$$

$$\frac{2e\varepsilon}{V_r} N_d = \left(\frac{2C_j}{A}\right)^2$$

This gives

$$N_d = \left(\frac{2C_j}{A}\right)^2 \frac{V_r}{2e\varepsilon}$$

$$= \left(\frac{2 \times 12 \times 10^{-12}}{10^{-4}}\right)^2 \frac{20}{2 \times 1.6 \times 10^{-19} \times 11.8 \times 8.85 \times 10^{-14}}$$

$$= 3.447 \times 10^{18} \text{ cm}^{-3}$$

5.14 A silicon *p-n* junction has a built-in potential of 0.65 V and the acceptor concentration on the *p*-side is 100 times greater than the donor concentration on the *n*-side. Find the value of depletion capacitance per unit area when a reverse bias of 10 V is applied across it. Also find the width of the depletion region.
Given: $\varepsilon_{Si} = 11.7$, $\varepsilon_0 = 8.854 \times 10^{-14}$ F/cm

Solution

$$V_{bi} = \frac{kT}{e} \ln \frac{N_a N_d}{n_i^2} \tag{5.14.1}$$

On substituting $N_a = 100 \times N_d$ and other values in Eqn (5.14.1) we get

$$0.65 = 0.0259 \ln \frac{100 N_d^2}{2.25 \times 10^{20}}$$

$\therefore$ $N_d = 4.22 \times 10^{14}$ cm^{-3}

$\Rightarrow$ $N_a = 4.22 \times 10^{16}$ cm^{-3}

$$W = \left[\frac{2\varepsilon(V_{bi} - V)}{e}\left(\frac{N_a + N_d}{N_a N_d}\right)\right]^{1/2}$$

where V_{bi} is the built-in potential. On substituting the values in the above expression for W we get

$$W = \left[\frac{2 \times 11.7 \times 8.854 \times 10^{-14} \times (0.65 + 10)}{1.6 \times 10^{-19}}\left(\frac{4.22 \times 10^{14} + 4.22 \times 10^{16}}{(4.22)^2 \times 10^{30}}\right)\right]^{1/2}$$

$$= 5.76 \times 10^{-4} \text{ cm}$$

Now, junction capacitance (C_j) =

$$C_j = \frac{\varepsilon A}{W}$$

$$\frac{C_j}{A} = \frac{\varepsilon}{W}$$

$$\frac{C_j}{A} = \frac{8.85 \times 10^{-14} \times 11.7}{5.76 \times 10^{-4}}$$

$$\frac{C_j}{A} = 1.797 \text{ nF/cm}^2$$

Recapitulation

- *p-n* junctions can be formed using various techniques such as alloying, diffusion, and ion implantation.
- The diffusivity D of dopants in solids has a temperature dependence given by

  $$D = D_0 \exp(-E_A/kT)$$

- The predeposition step results in a concentration profile given by

  $$C(x, t) = C_s \operatorname{erfc} \frac{x}{2\sqrt{Dt}}$$

- The drive-in step gives a concentration profile expressible as

  $$C(x, t) = \frac{Q}{\sqrt{\pi Dt}} e^{-x^2/4\,Dt}$$

- Built-in potential V_{bi} is given by the expression

  $$V_{\text{bi}} = \frac{kT}{e} \ln\left(\frac{N_a N_d}{n_i^2}\right)$$

- The total space charge width, W, is given by

  $$W = x_n + x_p = \left\{\frac{2\varepsilon_s V_{\text{bi}}}{e}\left[\frac{N_a + N_d}{N_a N_d}\right]\right\}^{1/2}$$

  for a *p-n* junction with no bias and by the expression

  $$W = \left\{\frac{2\varepsilon_s (V_{\text{bi}} + V_R)}{e}\left[\frac{N_a + N_d}{N_a N_d}\right]\right\}^{1/2}$$

  for a *p-n* junction with an applied reverse bias V_R.

- Depletion capacitance per unit area, C_{dep}, can be evaluated using

  $$C_{\text{dep}} = \left\{\frac{e\varepsilon_s N_a N_d}{2(V_{\text{bi}} + V_R)(N_a + N_d)}\right\}^{1/2}$$

- For a p^+-n junction, C_{dep} takes the form

  $$C_{\text{dep}} \approx \left\{\frac{e\varepsilon_s N_d}{2(V_{\text{bi}} + V_R)}\right\}^{1/2}$$

  and thus a plot of $1/(C_{\text{dep}})^2$ versus V_R is a straight line.

- For a linearly graded junction, C_{dep} is proportional to $(V_{bi} + V_R)^{-1/3}$.
- A hyper-abrupt junction with $m = -3/2$ can be used in a resonant circuit to result in a resonant frequency proportional to V_R.
- Zener breakdown occurs due to tunnelling of electrons from the valence band on the *p*-side into the conduction band on the *n*-side.
- An avalanche breakdown results when charge carriers gain sufficient energy to create electron–hole pairs due to collision with atomic electrons.
- The avalanche breakdown condition is given by

$$\int_0^W \alpha dx = 1$$

- A tunnel diode demonstrates a differential negative resistance.

Exercises

Review Questions

5.1 Explain the following processes for obtaining *p*-*n* junctions:
(a) Alloyed junction (b) Solid state diffusion
(c) Planar process (d) Ion implantation

5.2 Use a schematic diagram to illustrate how multiple Gaussians obtained by ion implantation can be used to realize a near-flat impurity concentration.

5.3 Explain the process of thermal oxidation with the help of a suitable schematic diagram and give the appropriate chemical reactions.

5.4 What factors decide the growth rate of thermally grown oxides?

5.5 How does the diffusivity of a dopant depend upon temperature? Why does silicon dioxide serve as a mask against the common dopants in silicon?

5.6 Explain predeposition and drive-in steps in a diffusion process.

5.7 Explain the process of depletion region formation in an unbiased *p*-*n* junction.

5.8 Derive expressions for the following:
(a) Built-in potential (b) Built-in electric field
(c) Depletion region width

5.9 Sketch a band diagram for a reverse-biased *p*-*n* junction.

5.10 Derive expressions for space charge width and depletion capacitance of a reverse-biased *p*-*n* junction.

5.11 What is a one-sided abrupt junction? Show that the plot of $1/C^2$ versus V for a reverse-biased one-sided abrupt junction is a straight line.

5.12 Derive an expression for the depletion capacitance of a linearly graded *p*-*n* junction.

5.13 What are hyper-abrupt junctions?

5.14 How can one realize a resonant circuit with resonant frequency that is linearly dependent upon applied reverse bias?

5.15 Explain the two phenomena that can give rise to junction breakdown.
5.16 Show that the integral $\int \alpha dx$ is equal to 1 for avalanche breakdown conditions.
5.17 Justify the *I-V* characteristic of a tunnel diode with suitable band diagrams.
5.18 Explain the solid state diffusion process of realizing *p-n* junctions, using a suitable sketch.
5.19 How can ion implantation be used to tailor flat impurity distribution? What is the role of sintering in this process?
5.20 Differentiate between dry and wet thermal oxidation processes.
5.21 Why is drive-in an important processing step?
5.22 Check the dimensional consistency of the expression for built-in potential for the *p-n* junction diode with no applied bias.
5.23 Derive an expression for the total space charge width for a *p-n* junction with (a) no applied bias and (b) with applied reverse bias V_R.
5.24 Derive the expression for depletion capacitance for a *p-n* junction under applied reverse bias V_R. What important parameters can be determined from the plot of $\frac{1}{(C_{\text{dep}})^2}$ versus V_R?
5.25 Write an expression for the depletion capacitance of a hyper-abrupt junction.
5.26 Differentiate between Zener breakdown and avalanche breakdown. Which one of the two takes place at a lower voltage?
5.27 Derive the avalanche breakdown condition.
5.28 Explain the existence of negative differential resistance using the *I-V* characteristics of a tunnel diode.

Problems

5.1 Phosphorus is diffused into a *p*-type silicon substrate of resistivity 10 Ω cm. To realize this, the substrate is exposed to a constant surface concentration of 10^{19} cm^{-3}. The resulting *p-n* junction should have a junction depth of 0.5 μm. The diffusion is carried out at 1100°C. Calculate the diffusion time assuming $\mu_h = 500$ cm^2/V s.

$$\left[Hint: C(x,t) = p = \frac{\sigma}{\mu_h e} = 10^{19} \operatorname{erfc} \frac{x}{2\sqrt{Dt}} \right]$$

Ans. ~ 5 min

5.2 Boron diffusion is first carried out in an *n*-type silicon substrate of 0.5 Ω cm resistivity to result in a boron doping level of 1×10^{17} cm^{-3} at $x = 3$ μm. The sample is then subjected to phosphorus diffusion at 1100°C with a constant surface concentration of 10^{21} cm^{-3} to result in a *p-n* junction formation at 3 μm. Determine the phosphorus doping profile.

$$\left[Hint: 1 \times 10^{17} = 10^{21} \operatorname{erfc} \frac{3}{2\sqrt{Dt}} \right]$$

Ans. $N_p = 10^{21}$ erfc $(0.9x)$

5.3 A germanium *p-n* junction has a donor concentration of 10^{21} m^{-3} and an acceptor concentration of 10^{23} m^{-3}. Calculate the built-in potential at 300 K. Assume $n_i = 2.5 \times 10^{19}$ m^{-3}.

$$\left[Hint: V_{\text{bi}} = \frac{kT}{e} \ln \left(\frac{N_a N_d}{n_i^2} \right) \right]$$

Ans. $N_p = 0.312$ V

5.4 The *n*-type and *p*-type regions of a germanium *p-n* junction have conductivities of 5×10^4 Ω^{-1} m^{-1} and 10^2 Ω^{-1} m^{-1}, respectively, at 300 K. Calculate the built-in potential of the junction at 300 K, assuming $\mu_n = 0.38$ m^2/V s and $\mu_p = 0.18$ m^2/V s. Assume $n_i = 2.5 \times 10^{19}$ m^{-3}.

$$\left[Hint: \sigma_n \approx ne\mu_n, \sigma_p \approx pe\mu_p, \text{ and } V_{\text{bi}} = \frac{kT}{e} \ln \left(\frac{N_a N_d}{n_i^2} \right) \right]$$

Ans. 0.399 V

5.5 Calculate the total space charge width for the germanium *p-n* junction given in Problem 5.3 assuming the dielectric constant ε_s of Ge to be 16. Take $\varepsilon_0 = 8.85 \times 10^{-12}$ F/m.

$$\left[Hint: W = \left\{ \frac{2\varepsilon_s V_{\text{bi}}}{e} \left[\frac{N_a + N_d}{N_a N_d} \right] \right\}^{1/2} \right]$$

Ans. 0.746 μm

5.6 A *p-n* junction fabricated using silicon has $N_d = 10^{17}$ cm^{-3} and $N_a = 10^{16}$ cm^{-3}. No applied bias exists on the junction. Calculate x_n (space charge width on the *n*-side) and $|E_{\max}|$. Assume $T = 300$ K and $\varepsilon_s = 11.7$.

$$\left[Hint: x_n = \left\{ \frac{2\varepsilon_s V_{\text{bi}}}{e} \left[\frac{N_a}{N_d} \right] \left[\frac{1}{(N_a + N_d)} \right] \right\}^{1/2} \text{ and } |E_{\max}| = \frac{eN_d x_n}{\varepsilon_s} \right]$$

Ans. $x_n = 2.99 \times 10^{-6}$ cm and $E_{\max} = 4.62 \times 10^4$ V/cm

5.7 Suppose the *p-n* junction mentioned in Problem 5.6 is fabricated using GaAs. Calculate x_n and $|E_{\max}|$ for the *p-n* junction. Assume $T = 300$ K, $n_i = 1.8 \times 10^6$ cm^{-3} and $\varepsilon_s = 13.1$.

Ans. $x_n = 4.03 \times 10^{-6}$ cm; $|E_{\max}| = 5.56 \times 10^4$ V/cm

5.8 A *p-n* junction formed of Ge has doping concentration of $N_a = 5 \times 10^{18}$ cm^{-3} and $N_d = 5 \times 10^{15}$ cm^{-3}. Calculate the space charge width of the *p-n* junction at $T = 300$ K and at $V_R = 3$ V. Assume $\varepsilon_s = 16$ and $n_i = 2.4 \times 10^{13}$ cm^{-3}.

$$\left[Hint: W = \left\{ \frac{2\varepsilon_s (V_{\text{bi}} + V_R)}{e} \left[\frac{N_a + N_d}{N_a N_d} \right] \right\}^{1/2} \right]$$

Ans. 1.1 μm

5.9 Determine the total space charge width for the *p-n* junction in Problem 5.3 at a reverse bias of 4 V .

$$\left[Hint: W = \left\{ \frac{2\varepsilon_s (V_{\text{bi}} + V_R)}{e} \left[\frac{N_a + N_d}{N_a N_d} \right] \right\}^{1/2} \right]$$

Ans. 2.8 μm

5.10 Calculate the junction capacitance for the *p-n* junction given in Problem 5.8 at a reverse bias of 4 V. Assume the area A of the junction to be 10^{-4} cm^2.

$$\left[Hint: C_{\text{dep}} = \left\{ \frac{e\varepsilon_s N_a N_d}{2(V_{\text{bi}} + V_R)(N_a + N_d)} \right\}^{1/2} \text{ and } C'_{\text{dep}} = C_{\text{dep}} A \right]$$

Ans. 1.1 pF

5.11 Abrupt p^+-n junctions were made out of two different materials, one with energy band gap (E_g) = 2.26 eV and other one with E_g = 1.1 eV. Assuming everything to be same, which of these two materials will require a larger value of reverse voltage for avalanche breakdown to occur? Justify your answer with the help of a suitable energy band diagram.

Ans. Material with E_g = 2.26 eV
Discuss with your teacher.

5.12 While characterizing a typical *p-n* junction (with breakdown voltage = 38.7 V) for its breakdown behavior, the following set of observations were obtained for n = 3:

Applied bias (V)	–10	–20	–30	–38.3	–38.6
Multiplication factor, M	1.0175	1.1601	1.872	32.58	??

(a) Predict the value of M for the '??' entry in the above table.

Ans. M = 129.33

(b) Comment on the result obtained.

(c) What will happen to the value of M ('??' entry only), if the value of n was 2?

Ans. M = 193.75

5.13 In a typical *p-n* junction structure which is to be used for varactor diode applications, what kind of doping profile will you choose on the *n*-side so as to have a square root dependence of the capacitance on the applied reverse bias? Explain briefly. ***Ans.*** Abrupt. Discuss with your teacher

5.14 Consider a *p-n* junction made of GaAs at 300 K. Let the doping densities be $N_a = 10^{17}$/cm^3 and $N_d = 10^{16}$/cm^3. Sketch the charge, field, and band diagrams under the depletion approximation. Label it with the calculated values of all

important quantities such as x_n, x_p, W, E_{max} and V_0. Use natural units like nm for length, V/cm for fields, and volts for potential. (Given $\varepsilon_r = 13.2$ for GaAs, $n_i = 2 \times 10^6$ cm^{-3}).

5.15 Suppose we have a silicon substrate of thickness 800 μm. An arbitrary energy level denoted by X, (where the probability of occupancy by an electron is 50%) is found to be 0.3 eV below the conduction band edge. Now, with the help of an unknown fabrication process, some gallium atoms are impinged into this sample so as to form an abrupt junction exactly at half of the thickness of original Si substrate. In the first half of the original Si substrate, level X is 0.3 eV above the VB edge and the rest of Si still behaves in the original way. (Use $\varepsilon_s = 1.04 \times 10^{-12}$ F/cm, atomic no. of Ga = 31, $\mu_n = 1350$ cm^2/V s and $\mu_p = 500$ cm^2/V s, $T =$ 27°C). Assume minority carrier concentration in respective opposite halves of the structure to be almost nil.

(a) Draw the overall energy band diagram of the resulting structure.
(b) Will there be any space charge region in this resulting structure?

If no, then explain the reason behind it briefly. If yes, then find the value for the same.

(c) Will there be some point of maximum electric field? If no then explain why. If yes, then find the value for the same. Will there be any points of zero electric field? Use the 'depletion approximation' to justify your answer.
(d) Draw the corresponding shape of the resulting charge density contour and electric field contour (only qualitatively).
(e) Can we use the resulting structure as a voltage sensitive switch? If no, then explain the reason behind it and if yes, then find the value of minimum potential that will make the switch *on*.

5.16 An abrupt *p-n* junction is made of silicon, where the resistivities of the two sides are 2 ohm cm (*p*-side) and 1 ohm cm (*n*-side). Compute the contact potential and the total width of the depletion regions for zero applied voltage. Take the value of relative permittivity for Si as 11.8.
[*Hint:* Use the expressions for resistivity:

$$\rho = \frac{1}{en\mu_n}, \rho = \frac{1}{ep\mu_p}$$

Use suitable values of mobility of electrons and holes and get the corresponding values of n and p on both sides of the *p-n* junction. Then use direct relation (Eqn 5.25) for contact potential to get the value of contact potential. Then use Eqn 5.50 to find the value of W, i.e., depletion region width.]

5.17 The built-in potential for a *p-n* junction is 0.7 V at 300k. If $N_a = 2 \times 10^{18}$ cm^{-3} and $n_i = 2 \times 10^{10}$ cm^{-3}, calculate N_d.

Ans. 9.8×10^{13} cm^{-3}

5.18 A graded $p + n$ junction has an area of 5×10^{-5} cm^2 and a donor concentration of 5×10^{18} cm^{-3}. Calculate the junction capacitance at a reverse bias of 15 V. Assume t =11.8.

Ans. 8.33 pf

6 *p-n* Junction Current

- *p-n* junction current flow
- Junction current in an ideal *p-n* junction
- Small-signal model of *p-n* junction
- Diffusion resistance, diffusion capacitance, generation–recombination currents, junction diode switching times

Learning Objectives

After going through this chapter the student will be able to

- understand the mechanism of charge flow in a *p-n* junction
- derive the ideal current–voltage characteristics of a *p-n* junction
- derive an expression for carrier current density for a short diode
- solve numericals based on minority carrier concentration at the edges of space charge region
- solve numericals based on saturation current density and total current density
- define diffusion conductance
- derive an expression for diffusion capacitance and diffusion conductance
- understand the complete equivalent circuit of a *p-n* junction
- derive an expression for reverse-bias generation current and forward-bias recombination current
- understand switching characteristics of *p-n* junction diodes
- solve numericals based on diffusion capacitance and diffusion conductance
- solve numericals based on generation–recombination currents
- solve numericals based on switching characteristics of *p-n* junction diodes

Introduction

The preceding chapter dealt with the structure, band diagram, and the detailed reverse-bias characteristics of the *p-n* junction. We will discuss the behaviour of the *p-n* junction under forward-bias in this chapter. One of the most powerful properties of the *p-n* junction is rectification. The appreciable difference between the conduction of a reverse-biased *p-n* junction and the conduction of a forward-biased junction holds the key to this property. We introduced the concept of potential barrier across an unbiased junction in Chapter 5. Under an applied forward bias, this potential barrier gets lowered, leading to a flow of electrons and holes across the space charge region. Electrons and holes react to this lowering of the potential barrier by rushing through the depletion region into the opposite region. Electrons and holes thus become minority carriers and are subjected to physical processes such as minority carrier diffusion, drift, and recombination. The *p-n* junction diode can be used as a circuit element in circuits like linear amplifiers. In such applications, the *p-n* junction is subjected to time-varying signals alongwith dc voltages. The response of the *p-n* junction can be understood by using the small-signal model of the *p-n* junction, which will be discussed in this chapter. Another important characteristic of a *p-n* junction is the process of switching between two conducting states (forward and reverse). The junction diode switching phenomenon will be discussed in detail in this chapter.

6.1 *p-n* Junction Current Flow

In this section, we will look at the physical mechanisms involved when a *p-n* junction is forward-biased. We will also develop a mathematical relation that gives the current–voltage characteristics of a *p-n* junction.

6.1.1 Charge Flow in a *p-n* Junction

Figure 6.1 shows the band diagram of a *p-n* junction under different biasing conditions. Figure 6.1(a) shows a *p-n* junction with no applied bias. The *p-n* junction can be seen to have a potential barrier V_{bi}, which stops the electrons from the *n*-region and holes from the *p*-region from flowing across the junction. The reverse-biased *p-n* junction is shown in Fig. 6.1(b). The potential barrier can be seen to increase in a reverse-biased *p-n* junction in comparison to an unbiased junction. This increased potential barrier ensures that a reverse-biased junction allows negligible current to flow through it. The energy band diagram of a *p-n* junction under forward-bias is shown in Fig. 6.1(c). The Fermi level in the *p*-region is lower than the Fermi level in the *n*-region. This leads to a lowering of the potential barrier, ensuring the lowering of the depletion region electric field. This lowered potential barrier leads to diffusion of holes from the *p*-region across the space charge region. A corresponding diffusion of electrons takes place from

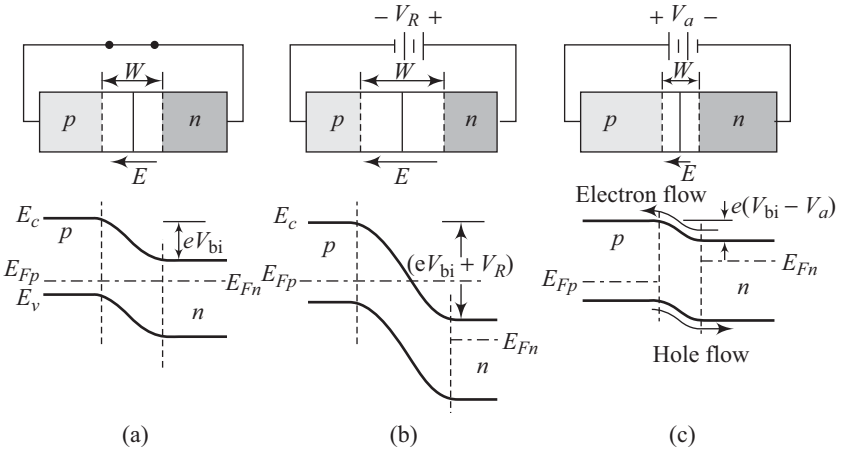


Fig. 6.1 A *p-n* junction and its energy band diagram for (a) zero bias, (b) reverse bias, and (c) forward bias

the *n*-region of the *p-n* junction. This flow of charge carriers across the junction leads to the forward-biased current in the *p-n* junction.

The holes are majority carriers in the *p*-region, but when they cross the space charge region and reach the *n*-region, they become minority carriers. Similarly, electrons crossing from the *n*-region become minority carriers on reaching the *p*-region. These excess minority carriers are subject to diffusion and recombination when they flow through the *p-n* junction.

6.1.2 Ideal Current–Voltage Characteristics

We will now derive an expression for the current–voltage characteristics of a *p-n* junction under ideal conditions. These ideal conditions mean the following assumptions:

1. The abrupt depletion-layer approximation is valid. This implies that the space charge regions have abrupt boundaries and the semiconductor is neutral outside the depletion region.
2. Carriers satisfy the Maxwell–Boltzmann approximation. Thus the Fermi–Dirac distribution function can be written as

$$f_F(E) \approx \exp\left[\frac{-(E - E_F)}{kT}\right] \tag{6.1}$$

3. The low injection approximation is valid, i.e., the excess carrier concentration is much smaller than the thermal-equilibrium majority carrier concentration.
4. The total current is constant throughout the *p-n* junction diode and the individual carrier currents are continuous throughout the *p-n* diode.
5. The depletion region has constant electron and hole currents. This implies that no generation current exists throughout the depletion layer.

6.1.3 Boundary Conditions

In Chapter 5, the built-in potential barrier, V_{bi}, for a *p-n* junction was shown to be represented by the equation

$$V_{bi} = \frac{kT}{e} \ln\left(\frac{N_a N_d}{n_i^2}\right) \tag{6.2}$$

This equation can be rewritten in the form

$$\frac{n_i^2}{N_a N_d} = \exp\left(\frac{-eV_{bi}}{kT}\right) \tag{6.3}$$

In the *n*-type semiconductor, assuming complete ionization, we can write for the equilibrium majority carrier concentration, n_{n0},

$$n_{n0} \approx N_d \tag{6.4}$$

In the *p*-region, we can write for the thermal equilibrium concentration, n_{p0}, of minority carriers,

$$n_{p0} \approx \frac{n_i^2}{N_a} \tag{6.5}$$

Substituting for N_d and N_a from Eqs (6.4) and (6.5) into Eqn (6.3) yields

$$n_{p0} = n_{n0} \exp\left(\frac{-eV_{bi}}{kT}\right) \tag{6.6}$$

Equation (6.6) is a relation between the minority carrier concentration on the *p*-side and the majority carrier concentration on the *n*-side of the *p-n* junction.

As already discussed, an applied forward-bias voltage reduces the potential barrier. Figure 6.2 shows a *p-n* junction with an applied forward bias *V*. Figure 6.2(a) shows a physical schematic diagram, whereas Fig. 6.2(b) shows an energy band diagram.

The entire applied voltage appears across the *p-n* junction depletion region. For an ideal *p-n* junction, the bulk *p* and *n* regions do not have appreciable electric fields. As can be seen from Fig. 6.2, the electric field induced due to the applied forward bias is opposite to the electric field existing in the space charge region in thermal equilibrium. The net electric field in the space charge region thus gets reduced. The

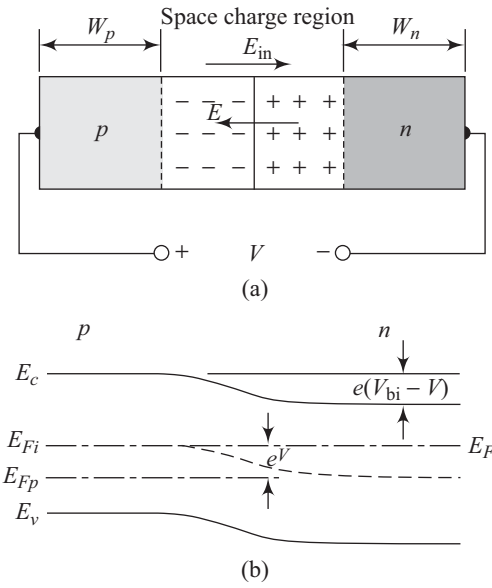


Fig. 6.2 (a) A *p-n* junction with an applied forward-bias voltage, (b) energy band diagram of the forward-biased *p-n* junction

thermal equilibrium space charge region is created by the equalization of the diffusion and electric field forces. Since under forward bias the force due to the net electric field reduces, majority carrier electrons from the *n*-side and majority carrier holes from the *p*-side are injected across the depletion region into the opposite sides, acquiring the status of minority carriers in their new regions. Thus a forward bias leads to current flow across the junction.

Under forward bias, Eqn (6.6) gets modified to

$$n_p = n_{n0} \exp\left[\frac{-e(V_{bi} - V)}{kT}\right] = n_{n0} \exp\left(\frac{-eV_{bi}}{kT}\right) \exp\left(\frac{+eV}{kT}\right) \tag{6.7}$$

where n_p represents the total minority carrier concentration in the *p*-region.

Using Eqn (6.6), Eqn (6.7) can be rewritten in the form

$$n_p = n_{p0} \exp\left(\frac{eV}{kT}\right) \tag{6.8}$$

Under forward-bias, a *p-n* junction is not in thermal equilibrium. *The total minority carrier electron concentration in the p-region can be seen to be greater than the thermal-equilibrium minority carrier electron concentration*, as can be inferred from Eqn (6.8). One must remember that this increased minority carrier concentration is due to the increased injection of majority carriers from the opposite side. The excess minority carrier electrons are subjected to processes such as diffusion and recombination. Thus, Eqn (6.7) or (6.8) essentially gives the minority carrier electron concentration at the *p*-region edge of the space charge region.

Majority carrier holes in the *p*-region are also subjected to the same process. For the minority carrier hole concentration at the *n*-region edge of the space charge region, we can write

$$p_n = p_{n0} \exp\left(\frac{eV}{kT}\right) \tag{6.9}$$

where p_n once again represents the total concentration of minority carrier holes. These results are indicated schematically in Fig. 6.3. The increase in the total minority carrier concentration at the edges of the space charge region is clear from Fig. 6.3.

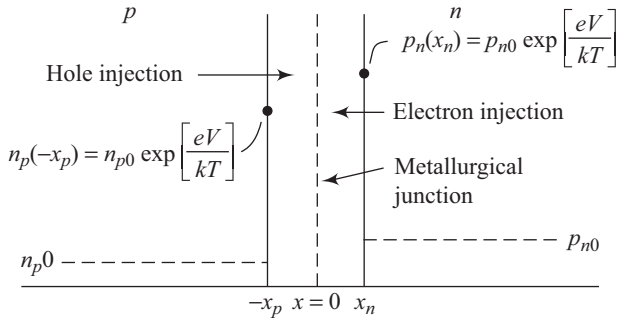


Fig. 6.3 Excess minority carrier concentrations at the space charge region edges generated by the forward-bias voltage

Equations (6.8) and (6.9) are valid for a reverse-biased junction also, provided V in the equations is replaced by $(-V)$. *Thus, the reverse-bias condition leads to the minority carrier concentration at the space charge edges to fall below the thermal-equilibrium values*. For reverse-bias voltages greater than a few tenths of a volt, the minority carrier concentrations at the edges of the space charge region essentially reduce to zero.

Equations (6.8) and (6.9) thus represent the bounding conditions for minority carriers under any type of biasing conditions in a *p-n* junction.

6.1.4 Minority Carrier Distribution

The time-dependent diffusion equation for the excess minority carrier hole concentration in one dimension can be written in the form

$$D_p \frac{\partial^2(\delta p_n)}{\partial x^2} - \mu_p E \frac{\partial(\delta p_n)}{\partial x} + G - \frac{\delta p_n}{\tau_{p0}} = \frac{\partial(\delta p_n)}{\partial t} \tag{6.10}$$

where $\delta p_n = p_n - p_{n0}$ represents the excess minority carrier hole concentration, G is the generation rate, and τ_{p0} is the excess minority carrier hole concentration.

We will assume zero electric field in both the neutral p and n regions and also assume steady state, implying $[\partial(\delta_{pn})]/\partial t = 0$. Equation (6.10) then reduces to

$$\frac{d^2(\delta p_n)}{dx^2} - \frac{\delta p_n}{D_p \tau_{p0}} = 0 \ (x > x_n) \tag{6.11}$$

From Fig. 6.3, it can be seen that $x = x_n$ is the space charge region edge in the n-region of the *p-n* junction.

Putting $L_p^2 = D_p \tau_{p0}$ in Eqn (6.11) leads to

$$\frac{d^2(\delta p_n)}{dx^2} - \frac{\delta p_n}{L_p^2} = 0 \ (x > x_n) \tag{6.12}$$

Similarly, the excess minority carrier electron concentration in the p-region beyond the space charge region is given by

$$\frac{d^2(\delta n_p)}{dx^2} - \frac{\delta n_p}{L_n^2} = 0 \ (x < x_p) \tag{6.13}$$

where

$$L_n^2 = D_n \tau_{n0} \tag{6.14}$$

Using the results obtained in the preceding section, the boundary conditions for the total minority carrier concentrations are

$$p_n(x_n) = p_{n0} \exp\left(\frac{eV}{kT}\right) \tag{6.15}$$

$$n_p(-x_p) = n_{p0} \exp\left(\frac{eV}{kT}\right) \tag{6.16}$$

$$p_n(x \to +\infty) = p_{n0} \tag{6.17}$$

and

$$n_p(x \to -\infty) = n_{p0} \tag{6.18}$$

We will now assume that the thicknesses W_n and W_p of the *n* and *p* regions of the *p-n* junction satisfy the conditions

$$W_n \gg L_p \text{ and } W_p \gg L_n \tag{6.19}$$

where L_p and L_n are the diffusion lengths for holes and electrons, respectively. The assumptions indicated in Eqn (6.19) are also referred to as *long p-n junction approximation*. Thus the excess minority carrier concentrations can be assumed to approach zero at large distances from the space charge region.

Equation (6.12) has a general solution of the form

$$\delta p_n(x) = p_n(x) - p_{n0} = Ae^{x/L_p} + Be^{-x/L_p} \; (x \geq x_n) \tag{6.20}$$

Similarly, Eqn (6.13) has a general solution of the form

$$\delta n_p(x) = n_p(x) - n_{p0} = Ce^{x/L_n} + De^{-x/L_n} \; (x \leq -x_p) \tag{6.21}$$

Boundary conditions (6.17) and (6.18) when used in Eqs (6.20) and (6.21) give

$$A = D = 0 \tag{6.22}$$

Thus, Eqn (6.20) reduces to

$$\delta p_n(x) = Be^{-x/L_p} = p_n(x) - p_{n0} \tag{6.23}$$

Using Eqn (6.15) in Eqn (6.23) results in

$$Be^{-x_n/L_p} = p_{n0} \exp\left(\frac{eV}{kT}\right) - p_{n0}$$

which implies

$$B = \frac{p_{n0}\left[\exp(eV/kT) - 1\right]}{\exp[-x_n/L_p]} \tag{6.24}$$

Putting the value of B in Eqn (6.23) yields

$$\delta p_n(x) = p_{n0}\left[\exp\left(\frac{eV}{kT}\right) - 1\right]\exp\left(\frac{x_n - x}{L_p}\right) \text{for } x \geq x_n \tag{6.25}$$

Similarly, for $x \leq -x_p$, we get

$$\delta n_p(x) = n_p(x) - n_{p0} = n_{p0}\left[\exp\left(\frac{eV}{kT}\right) - 1\right]\exp\left(\frac{x_p + x}{L_n}\right) \tag{6.26}$$

The exponential fall in the minority carrier concentration with distance is shown schematically in Fig. 6.4. The figure has been drawn assuming the validity of the long diode approximation. The decay of minority carrier concentration is due to minority carrier recombination with majority carriers.

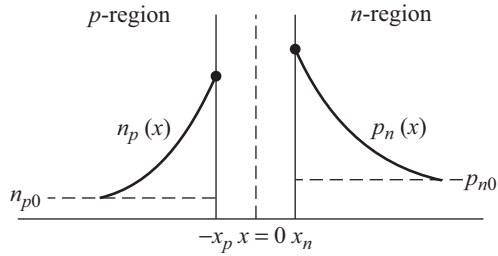


Fig. 6.4 Steady state minority carrier concentrations in a *p-n* junction under forward bias

6.1.5 Junction Current in Ideal *p-n* Junction

We are now in a position to derive an expression for the junction current flowing through an ideal *p-n* junction. As per our assumption for an ideal *p-n* junction, the individual electron and hole currents are constant through the *p-n* junction. Moreover, the electron and hole currents are continuous functions through the *p-n* junction. The total *p-n* junction current is the sum of the minority carrier hole diffusion current at $x = x_n$ and the minority carrier electron diffusion current at $x = -x_p$. The minority carrier drift current can be neglected because the electric field has been assumed to become zero at the space charge edges. This concept is shown schematically in Fig. 6.5.

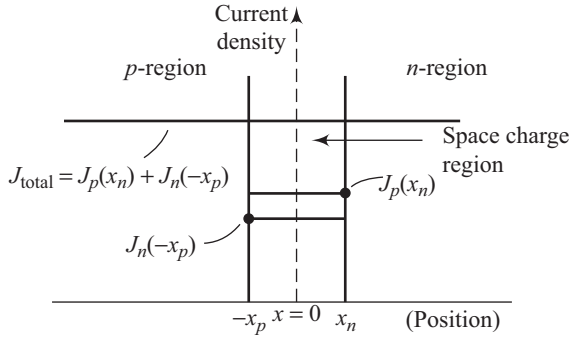


Fig. 6.5 Electron and hole current densities through the space charge region of a *p-n* junction

The minority carrier hole diffusion current density at $x = x_n$ is given by

$$J_p(x_n) = -eD_p \left.\frac{dp_n(x)}{dx}\right|_{x = x_n} \tag{6.27}$$

The thermal-equilibrium carrier concentration being constant, we can rewrite Eqn (6.27) in the form

$$J_p(x_n) = -eD_p \left.\frac{d(\delta p_n(x))}{dx}\right|_{x = x_n} \tag{6.28}$$

Using Eqn (6.25) for $\delta p_n(x)$ and substituting it in Eqn (6.28), we get

$$J_p(x_n) = \frac{eD_p p_{n0}}{L_p}\left[\exp\left(\frac{eV}{kT}\right) - 1\right] \tag{6.29}$$

The hole current density is in the $+x$ direction. The electron diffusion current density at $x = -x_p$, similarly, is given by

$$J_n(-x_p) = eD_n \frac{d}{dx}(\delta n_p(x))\bigg|_{x=-x_p} \tag{6.30}$$

Using Eqn (6.26) for $\delta n_p(x)$ in Eqn (6.30) results in

$$J_n(-x_p) = \frac{eD_n n_{p0}}{L_n}\left[\exp\left(\frac{eV}{kT}\right) - 1\right] \tag{6.31}$$

We might note that the electron current density is also in the $+x$ direction.

The total current density through the *p-n* junction is given by

$$J = J_p(x_n) + J_n(-x_p) = \left[\frac{eD_p p_{n0}}{L_p} + \frac{eD_n n_{p0}}{L_n}\right]\left[\exp\left(\frac{eV}{kT}\right) - 1\right] \tag{6.32}$$

Defining J_s as

$$J_s = \left[\frac{eD_p p_{n0}}{L_p} + \frac{eD_n n_{p0}}{L_n}\right] \tag{6.33}$$

Equation (6.32) can be rewritten in the form

$$J = J_s\left[\exp\left(\frac{eV}{kT}\right) - 1\right] \tag{6.34}$$

Equation (6.34) is called the ideal *p-n* diode equation. This equation expresses the current–voltage characteristics not only under forward-bias but also under reverse-bias conditions. For the reverse-bias condition, the voltage V takes negative values. A complete plot of Eqn (6.34) is shown in Fig. 6.6.

For negative voltage values beyond a few kT/e V, the reverse–bias current density can be seen to saturate in conformity with Eqn (6.34). The parameter J_s is, therefore, referred to as the reverse saturation current density. It is clear from Fig. 6.6 that the *p-n* junction diode conducts very well in the forward bias and offers high resistance (allows low current) in the reverse-bias condition. This accounts for the rectifying property of a *p-n* diode.

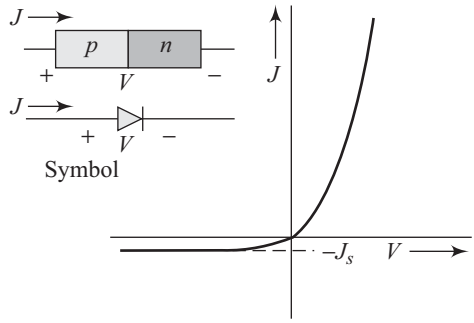


Fig. 6.6 Ideal *I-V* characteristic for a junction diode along with symbolic representation

6.1.6 Short Diode

In the preceding section, the neutral p and n regions considered were longer than the corresponding minority carrier diffusion lengths. We will now relax this condition, at least in one of the two regions. Figure (6.7) shows a p-n junction where the p-region is still long but the length of the n-region, W_n, is much smaller than the minority carrier hole diffusion length, L_p.

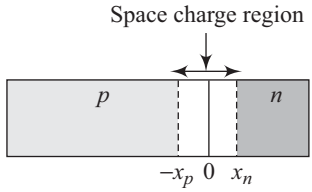


Fig. 6.7 Schematic diagram of a short diode

The steady state excess minority carrier hole concentration in the n-region is given by the equation

$$\frac{d^2(\delta p_n)}{dx^2} - \frac{\delta p_n}{L_p^2} = 0 \tag{6.35}$$

Assuming an ohmic contact at $x = x_n + w_n$ means an infinite surface recombination velocity and, therefore, zero excess minority carrier concentration.

The boundary conditions are, therefore

$$p_n(x_n) = p_{n0} \exp\left(\frac{eV}{kT}\right) \tag{6.36}$$

and

$$p_n(x = x_n + w_n) = p_{n0} \tag{6.37}$$

The general solution of Eqn (6.35) is given by

$$\delta p_n(x) = p_n(x) - p_{n0} = Ae^{x/L_p} + Be^{-x/L_p} \; (x \geq x_n) \tag{6.38}$$

Using the boundary conditions indicated by Eqs (6.36) and (6.37), the solution of Eqn (6.35) is shown to be

$$\delta p_n(x) = p_{n0}\left[\exp\left(\frac{eV}{kT}\right) - 1\right]\frac{\sinh[x_n + w_n - x)/L_p]}{\sinh[w_n/L_p]} \tag{6.39}$$

For $w_n \ll L_p$, the hyperbolic term reduces to

$$\sinh\left(\frac{x_n + w_n - x}{L_p}\right) \approx \left(\frac{x_n + w_n - x}{L_p}\right) \tag{6.40}$$

and

$$\sinh\left(\frac{w_n}{L_p}\right) \approx \left(\frac{w_n}{L_p}\right) \tag{6.41}$$

Substituting Eqs (6.40) and (6.41) into Eqn (6.39) gives

$$\delta p_n(x) = p_{n0}\left[\exp\left(\frac{eV}{kT}\right) - 1\right]\left(\frac{x_n + w_n - x}{w_n}\right) \tag{6.42}$$

Thus *the excess minority carrier hole concentration is linearly dependent on distance*. One must recollect at this stage that the excess minority carrier hole concentration for a long diode is exponentially dependent on the distance from the junction.

The minority carrier hole diffusion current is given by

$$J_p = -eD_p \frac{d}{dx}[\delta p_n(x)] \tag{6.43}$$

Using Eqn (6.42) in Eqn (6.43) yields

$$J_p(x) = \frac{qD_p p_{n0}}{w_n}\left[\exp\left(\frac{eV}{kT}\right) - 1\right] \tag{6.44}$$

A comparison of Eqn (6.44) with Eqn (6.29) reveals that the minority carrier hole current density for a short diode has the length w_n instead of the minority carrier diffusion length L_p. Since $w_n \ll L_p$, *the diffusion current density for a short diode is larger than that for a long diode*. A short diode also implies a constant current density since the minority carrier concentration is a linear function of distance, as can be seen from Eqn (6.42).

- The reverse saturation current is a function of temperature. For a one-sided p^+-n junction,

  $$J_s \approx \frac{eD_p p_{n0}}{L_p} \approx e\sqrt{\frac{D_p}{\tau_p}}\frac{n_i^2}{N_d}$$

  If D_p/τ_p is proportional to T^γ, we get

  $$J_s \sim \left[T^3 \exp\left(\frac{-E_g}{kT}\right)\right] T^{\gamma/2} = T^{(3+\gamma/2)} \exp^{-E_g/kT}$$

  The exponential term generally dominates.
- The ideal diode equation is also called Shockley equation.
- In the reverse direction, $J \sim J_s$. Thus, J increases approximately as $e^{-E_g/kT}$. In the forward direction, $J \sim e^{eV/kT}$ and thus increases approximately as $\exp[-(E_g - eV)/kT]$.

6.2 Small-signal Model of *p-n* Junction

We have so far looked at the response of a *p-n* junction to an applied dc potential. Some applications of a *p-n* junction however require it to have time-varying signals in addition to the dc potential. The simplest case is the superimposition of a sinusoidal signal with dc voltages, and currents. We will discuss the behaviour of the *p-n* junction to such superimposed signals in this section. To make the mathematics simpler, we assume the amplitude of the applied sinusoidal signal to be much smaller than the magnitude of the dc signal. Such a model of the *p-n* junction is called the small-signal model.

6.2.1 Diffusion Resistance

The current–voltage characteristics under ideal conditions were derived in the preceding section. The expression in that section involved current density.

The corresponding relationship in terms of current appears like

$$I_D = I_s\left[\exp\left(\frac{eV}{kT}\right) - 1\right] \tag{6.45}$$

where the subscript D has been used to emphasize that the current is due to the diffusion of carriers.

Suppose a diode is forward-biased with a dc voltage V_0 and a small, low-frequency sinusoidal voltage is superimposed on it. The dc voltage produces a dc current I_0 and the sinusoidal voltage results in a corresponding sinusoidal current. Figure 6.8 shows the behaviour of the *p-n* junction when a dc voltage superimposed on a sinusoidal voltage is applied to it.

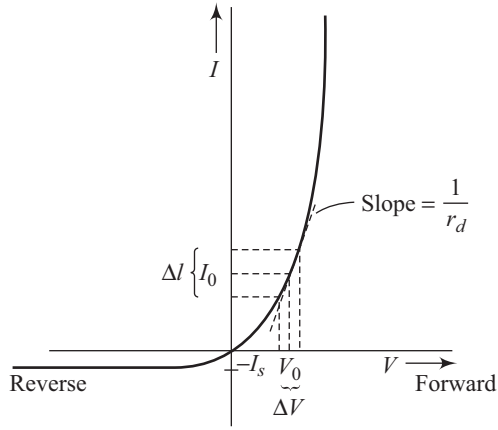


Fig. 6.8 Curve showing the concept of the small-signal diffusion resistance

The ratio of sinusoidal current to sinusoidal voltage is called the *diffusion conductance* g_d. Under small-signal conditions, i.e., in the limit of very small sinusoidal current and voltage, we can write

$$g_d = \left.\frac{dI_D}{dV}\right|_{V = V_0} \tag{6.46}$$

Diffusion resistance r_d is the reciprocal of the diffusion conductance. Thus,

$$r_d = \left.\frac{dV}{dI_D}\right|_{I_D = I_0} \tag{6.47}$$

where I_0 is the dc diode current.

If the *p-n* junction diode is biased with a large enough forward bias, the (–1) term in Eqn (6.45) can be neglected. The diffusion conductance then becomes

$$g_d = \left(\frac{e}{kT}\right) I_s \exp\left(\frac{eV_0}{kT}\right) \approx \frac{eI_0}{kT} \tag{6.48}$$

Then, the corresponding small-signal diffusion resistance is

$$r_d = \frac{kT}{e}\left(\frac{1}{I_0}\right) \tag{6.49}$$

The parameter (kT/e) is sometimes referred to as the thermal voltage and is represented by the symbol V_t. In this notation we get

$$g_d = \frac{I_0}{V_t} \text{ and } r_d = \frac{V_t}{I_0} \tag{6.50}$$

From Fig. 6.8 it is clear that the slope of the current–voltage characteristic is the reciprocal of diffusion resistance. At any given bias current, the incremental diffusion resistance is inversely proportional to the slope of the *I-V* characteristic.

The incremental diffusion resistance decreases as the bias current increases. The diffusion conductance g_d as defined in Eqn (6.50) has units of mho. If current density is used instead of current in the expression for diffusion conductance, its unit becomes mho/cm^2 or mho/m^2, depending upon the unit of area.

6.2.2 Diffusion Capacitance

In Chapter 5, we derived expressions for depletion-layer capacitance for different types of *p-n* junctions. Depletion-layer capacitance is the predominant capacitance under reverse-bias conditions. Under forward bias, a rearrangement of minority carrier density takes place as discussed in the preceding section. This results in a capacitance called the *diffusion capacitance*.

Suppose a dc voltage superimposed on a sinusoidal voltage is applied on a *p-n* junction. This can be expressed mathematically in the form

$$V(t) = V_0 + V_1 e^{j\omega t} \tag{6.51}$$

$$J(t) = J_0 + J_1 e^{j\omega t} \tag{6.52}$$

where V_0 and J_0 are the dc components of voltage and current density, respectively, and V_1 and J_1 are the amplitudes of the small-signal sinusoidal components. The small-signal ac component of minority carrier hole density can be expressed as

$$\tilde{p}_n(x, t) = p_{n1}(x) e^{j\omega t} \tag{6.53}$$

The minority carrier concentration at the depletion-layer edge is given by

$$p_n = p_{n0} \exp\left[\frac{e(V_0 + V_1 e^{j\omega t})}{kT}\right] \tag{6.54}$$

For small-signal conditions, $V_1 \ll V_0$ and $\exp(x) \approx (1 + x)$. Thus we can rewrite Eqn (6.54) in the form

$$p_n \cong p_{n0} e^{\frac{eV_0}{kT}} \left(1 + \frac{eV_1}{kT} e^{j\omega t}\right)$$

yielding

$$p_n = p_{n0} e^{\frac{eV_0}{kT}} + \frac{p_{n0} e V_1}{kT} e^{eV_0/kT}\, e^{j\omega t} \tag{6.55}$$

The minority carrier electron density is given by a similar expression. The first term in Eqn (6.55) is the dc component of the hole density at the depletion-layer boundary. The second term is the small-signal ac component of the hole density. This ac component of the hole density is given by Eqn (6.53) for $\tilde{p}_n(x, t)$. Thus the second term in Eqn (6.55) is actually $\tilde{p}_n(x_n, t) = p_{n1}(x_n) e^{j\omega t}$.

The continuity equation, with the generation term equal to zero, for the small-signal ac component $\tilde{p}_n$ can be written as

$$j\omega \tilde{p}_n = \frac{-\tilde{p}_n}{\tau_p} + D_p \frac{\partial^2 \tilde{p}_n}{\partial x^2} \tag{6.56}$$

which can be rewritten in the form

$$\frac{\partial^2 \tilde{p}_n}{\partial x^2} - \frac{\tilde{p}_n}{D_p \tau_p / (1 + j\omega\tau_p)} = 0 \tag{6.57}$$

Defining

$$\tau_p^* = \frac{\tau_p}{1 + j\omega\tau_p} \tag{6.58}$$

We get from Eqn (6.57)

$$\frac{\partial^2 \tilde{p}_n}{\partial x^2} - \frac{\tilde{p}_n}{D_p \tau_p^*} = 0 \tag{6.59}$$

Equation (6.59) is similar to Eqn (6.11). Going through a procedure identical to the one used for obtaining Eqn (6.32), but with p_{n0} replaced by $[(p_{n0}eV_1)/(kT)]e^{j\omega t}$ in light of Eqn (6.55), we get

$$J(t) = \frac{J_0 + eV_1}{kT}\left[\frac{eD_p p_{n0}}{L_p/\sqrt{1 + j\omega\tau_p}} + \frac{eD_n n_{p0}}{L_n/\sqrt{1 + j\omega\tau_n}}\right] e^{(eV_0/kT)} e^{j\omega t} \tag{6.60}$$

Comparing Eqn (6.60) with Eqn (6.52), we get

$$J_1 = \frac{eV_1}{kT}\left[\frac{eD_p p_{n0}}{L_p/\sqrt{1 + j\omega\tau_p}} + \frac{eD_n n_{p0}}{L_n/\sqrt{1 + j\omega\tau_n}}\right] \exp\left(\frac{eV_0}{kT}\right) \tag{6.61}$$

The ac admittance Y is then given by

$$Y = \frac{J_1}{V_1} = g_d + j\omega c_d \tag{6.62}$$

For low frequencies, we have

$$\omega\tau_p,\ \omega\tau_n \ll 1 \tag{6.63}$$

Under the conditions indicated in Eqn (6.63), the equation for g_d becomes g_{d0}, given by

$$g_{d0} = \frac{e}{kT}\left[\frac{eD_p p_{n0}}{L_p} + \frac{eD_n n_{p0}}{L_n}\right] \exp^{eV_0/kT} \text{ mho/cm}^2 \tag{6.64}$$

Differentiating Eqn (6.32) for J with respect to V at $V = V_0$ results in

$$\frac{dJ}{dV} = \frac{e}{kT}\left[\frac{eD_p p_{n0}}{L_p} + \frac{eD_n n_{p0}}{L_n}\right] e^{eV_0/kT} \tag{6.65}$$

Equation (6.65) is identical to Eqn (6.64).

Similarly, the low-frequency diffusion capacitance C_{d0} is given by

$$C_{d0} = \frac{e}{kT}\left[\frac{eL_p p_{n0}}{2} + \frac{eL_n n_{p0}}{2}\right] e^{eV_0/kT} \text{ F/cm}^2 \tag{6.66}$$

Figure 6.9 is a schematic diagram showing the variation of normalized conductance and normalized diffusion capacitance as functions of normalized frequency $\omega\tau$.

The equivalent circuit of the ac admittance is shown in the inset. Figure 6.9 has been drawn assuming the dominance of one of the terms in Eqn 6.61. The diffusion capacitance can be seen to decrease with increasing frequency. For very high frequencies, C_d varies approximately as $\omega^{-1/2}$. Due to the dependence on exp (eV_0/kT), the diffusion capacitance increases with the direct current level. This makes C_d an important parameter at low frequencies and under forward-bias conditions.

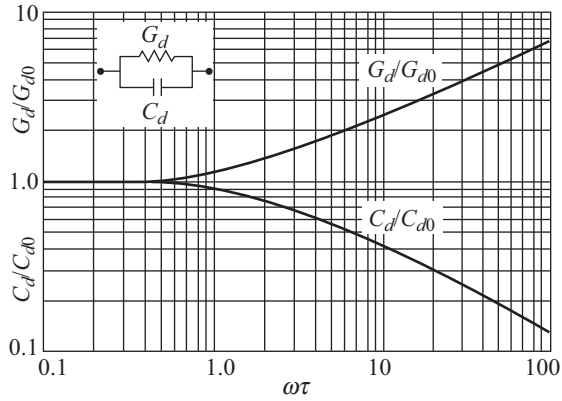


Fig. 6.9 Normalized diffusion conductance and diffusion capacitance versus normalized frequency

6.2.3 Equivalent Circuit

The small-signal equivalent circuit shown in the inset in Fig. 6.9 has been derived on the basis of the ideal current–voltage characteristics. A real *p-n* junction can have two other contributions to the equivalent circuit. These two are contributions due to junction capacitance and series resistance. Junction capacitance has already been discussed in detail in Chapter 5.

A finite series resistance of a *p-n* junction exists due to contributions from the neutral *p* and *n* regions and the contact regions. If V represents the applied voltage and r_s represents the total series resistance of the *p-n* junction, then we can write

$$V = V_J + Ir_s \tag{6.67}$$

where V_J represents the voltage across the junction. Figure 6.10 shows a plot of ln (I) versus V for a real *p-n* junction. We can see a deviation from the ideal current–voltage characteristic at higher applied voltages and currents.

One can see from Fig. 6.10 that a larger applied voltage is required to achieve the same current value when the series resistance of the *p-n* junction is included. A complete equivalent circuit of a *p-n* junction is shown in Fig. 6.11.

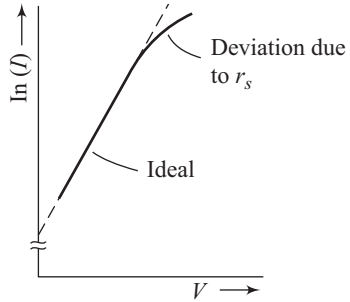


Fig. 6.10 Forward-biased *I-V* characteristics of a *p-n* junction diode showing the effect of series resistance

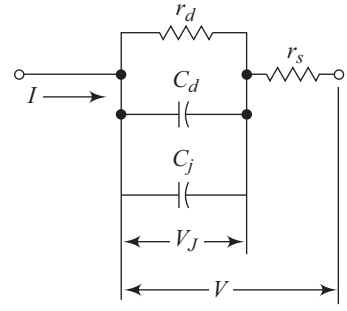


Fig. 6.11 Complete small-signal equivalent circuit of *p-n* junction

6.3 Generation–Recombination Currents

We have so far neglected the effect of generation–recombination currents within the space charge region in deriving the *I-V* characteristics of *p-n* junctions. In this section, we will make an attempt at including this contribution to the *I-V* characteristics.

You would recollect that the recombination rate, R, of excess electrons and holes, as given by the Shockley–Real–Hall theory, is

$$R = \frac{C_n C_p N_t (n_p - n_i^2)}{C_n (n + n_t) + C_p (p + p_t)} \tag{6.68}$$

where C_n and C_p are constants proportional to electron and hole capture cross sections, N_t is the total concentration of trapping centres, n and p are electron and hole concentrations, n_t is the electron concentration existing in the conduction band when the trap energy E_t coincides with the Fermi energy E_F, and p_t represents the corresponding hole concentration.

6.3.1 Reverse-bias Generation Current

Under reverse-bias conditions, the space charge region of a *p-n* junction does not contain mobile electrons and holes. Thus, in the space charge region, we have $n \approx p \approx 0$. The recombination rate R using Eqn (6.68) becomes

$$R = \frac{-C_n C_p N_t \, n_i^2}{C_n n_t + C_p p_t} \tag{6.69}$$

A negative recombination rate essentially implies a positive generation rate. Electrons and holes are being generated via the trap level to try to reestablish thermal equilibrium. You would recollect that the processes involved are electron and hole emission.

Reverse-bias generation current is shown schematically in Fig. 6.12.

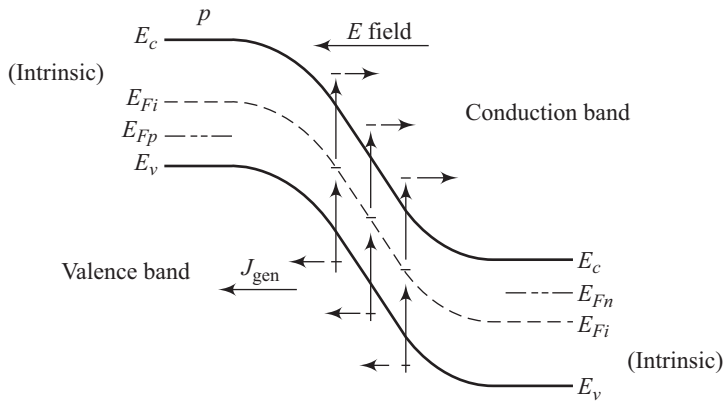


Fig. 6.12 Generation process in a reverse-biased *p-n* junction

The generated electrons and holes are swept away by the electric field in the space charge region. The direction of the flow of this component of current in the reverse-bias junction is the same as the normal reverse-bias current J_s. Thus the total current flowing through the reverse-bias junction is the sum of reverse-bias generation current and the ideal reverse-bias saturation current.

Let us make a simplifying assumption that the trap level is at the intrinsic Fermi level. Then

$$n_t = p_t = n_i \tag{6.70}$$

Using Eqn (6.70), Eqn (6.69) can be rewritten in the form

$$R = \frac{-n_i}{(1/N_t C_p) + (1/N_t C_n)} \tag{6.71}$$

Also, we have

$$\tau_{p0} = \frac{1}{C_p N_t}, \quad \tau_{n0} = \frac{1}{C_n N_t} \tag{6.72}$$

Equation (6.72) when substituted into Eqn (6.71) yields

$$R = -\frac{n_i}{\tau_{p0} + \tau_{n0}} \tag{6.73}$$

Let us at this stage define an average lifetime τ_m as

$$\tau_m = \frac{\tau_{p0} + \tau_{n0}}{2} \tag{6.74}$$

Using Eqn (6.74) in Eqn (6.73) results in

$$R = -\frac{n_i}{2\tau_m} = -G \tag{6.75}$$

From Eqn (6.75) it is clear that a negative recombination is in effect a generation rate. The generation current density, J_{gen}, is given by

$$J_{\text{gen}} = \int_0^W qG \, dx \tag{6.76}$$

where W is the width of the space charge region.

For a constant generation rate across the space charge region, we get

$$J_{\text{gen}} = \frac{qn_i W}{2\tau_m} \tag{6.77}$$

The generation current density indicated by Eqn (6.77) is in addition to the normal reverse saturation current density. Thus the total reverse-bias current density is given by

$$J_R = J_s + J_{\text{gen}} \tag{6.78}$$

Note that the reverse saturation current density, J_s, is bias-independent, whereas the generation current density is bias-dependent through the effective depletion region width, W. Thus, the total reverse-bias current density is not bias-independent.

Since the dependence of J_{gen} comes through the depletion region width, W, we must have

$$J_{gen} \propto (V_{bi} + V)^{1/2} \text{ (for abrupt junctions)} \tag{6.79}$$

$$J_{gen} \propto (V_{bi} + V)^{1/3} \text{ (for linearly-graded junction)} \tag{6.80}$$

6.3.2 Forward-bias Recombination Current

When a *p-n* junction is forward biased, electrons and holes are injected from opposite sides of the space charge region. Excess carriers moving through the space charge region can recombine and thus not contribute to the minority carrier distribution and hence the current.

Once again, for the recombination rate we can write

$$R = \frac{C_n C_p N_t \ (np - n_i^2)}{C_n (n + n_t) + C_p (p + p_t)} \tag{6.81}$$

which can be rewritten in the form

$$R = \frac{np - n_i^2}{\dfrac{(n + n_t)}{C_p N_t} + \dfrac{(p + p_t)}{C_n N_t}}$$

which on using Eqn (6.72) results in

$$R = \frac{np - n_i^2}{\tau_{p0}(n + n_t) + \tau_{n0}(p + p_t)} \tag{6.82}$$

Figure 6.13 shows the energy band diagram of a forward-biased *p-n* junction. Quasi-Fermi level E_{Fp} is shown in the figure alongwith the intrinsic Fermi level E_{Fi}.

We have already learnt that the electron and hole concentration, n and p, are given by

$$n = n_i \exp\left[\frac{E_{Fn} - E_{Fi}}{kT}\right] \tag{6.83}$$

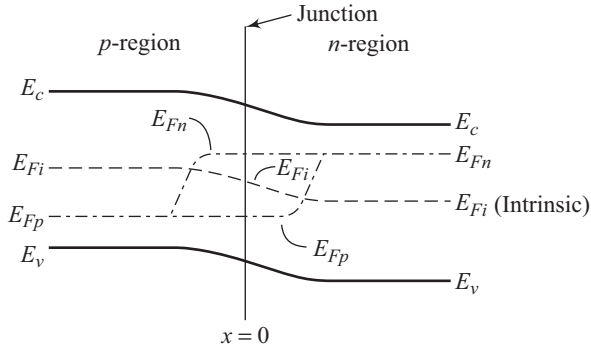


Fig. 6.13 Energy band diagram of a forward-biased *p-n* junction showing quasi-Fermi levels

and

$$p = n_i \exp\left[\frac{E_{Fi} - E_{Fp}}{kT}\right] \tag{6.84}$$

From Fig. 6.13, we can see that

$$(E_{Fn} - E_{Fi}) + (E_{Fi} - E_{Fp}) = eV \tag{6.85}$$

where V is the applied forward bias. We will again assume the trap level to be at the intrinsic Fermi level such that

$$n_t = p_t = n_i \tag{6.86}$$

The recombination rate can be shown to peak at the metallurgical junction ($x = 0$), where we have, from Eqn (6.85)

$$E_{Fn} - E_{Fi} = E_{Fi} - E_{Fp} = \frac{eV}{2} \tag{6.87}$$

Using Eqn (6.87) in Eqs (6.83) and (6.84), we get

$$n = n_i \exp\left(\frac{eV}{2kT}\right) \tag{6.88}$$

and

$$p = n_i \exp\left(\frac{eV}{2kT}\right) \tag{6.89}$$

Assuming the conditions in Eqn (6.86) and $\tau_{p0} = \tau_{n0} = \tau_0$, and using Eqs (6.88) and (6.89) in Eqn (6.82) leads to

$$R_m = \frac{n_i}{2\tau_0} \frac{[\exp(eV/kT) - 1]}{[\exp(eV/2kT) + 1]} \tag{6.90}$$

Assuming $V \gg kT/e$, Eqn (6.90) reduces to

$$R_m = \frac{n_i}{2\tau_0} \exp\left(\frac{eV}{2kT}\right) \tag{6.91}$$

The recombination current density, J_{rec}, is given by

$$J_{\text{rec}} = \int_0^W eR\, dx \tag{6.92}$$

Using Eqn (6.91) for R_m in Eqn (6.92) results in

$$J_{\text{rec}} = \frac{ex_m n_i}{2\tau_0} \exp\left(\frac{eV}{2kT}\right) \tag{6.93}$$

where x_m is the width of the space charge region where $R \approx R_m$. If $x_m \approx W$, then Eqn (6.93) can be rewritten as

$$J_{\text{rec}} \cong \frac{eWn_i}{2\tau_0} \exp\left(\frac{eV}{2kT}\right) = J_{r0} \exp\left(\frac{eV}{2kT}\right) \tag{6.94}$$

6.3.3 Net Forward-bias Current

When forward bias is applied to a *p-n* junction, holes are injected from the *p*-side and electrons are injected from the *n*-side of the junction. This results in a minority carrier distribution on either side of the junction. If some of the injected carriers recombine in the space charge region, an additional injection results from either side of the junction to compensate for the loss.

The total forward-bias current density is given by

$$J = J_{\text{rec}} + J_D \tag{6.95}$$

Neglecting the (–1) term and using Eqn (6.34), we have

$$J_D = J_S \exp\left(\frac{eV}{kT}\right) \tag{6.96}$$

Taking the natural log of Eqn (6.96) yields

$$\ln J_D = \ln J_S + \frac{eV}{kT} \tag{6.97}$$

Similarly, taking the natural log of Eqn (6.94) leads to

$$\ln J_{\text{rec}} = \ln J_{r0} + \frac{eV}{2kT} \tag{6.98}$$

The current–voltage characteristics of a *p-n* junction based on Eqs (6.97) and (6.98) are shown in Fig. 6.14.

Although the plots of $\ln J_D$ versus V and $\ln J_{\text{rec}}$ versus V are both straight lines, the slopes of these lines are different. The total current density is the sum of the current components. It is interesting to note that the recombination current density dominates at low-bias voltages whereas at high-bias voltages, the ideal diffusion current is the dominant factor.

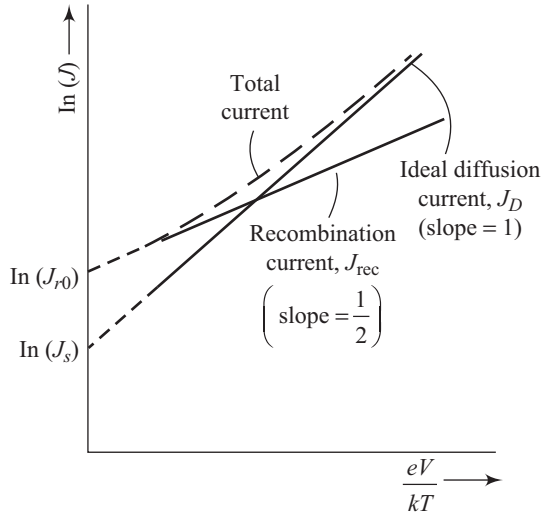


Fig. 6.14 Ideal diffusion, recombination, and total current in a forward-biased *p-n* junction

The general *p-n* junction current–voltage characteristics are given by

$$I = I_s \left[\exp\left(\frac{eV}{nkT} \right) - 1 \right] \tag{6.99}$$

The parameter n is called the ideality factor of the *p-n* junction and can have values between 1 and 2. One must, however, remember that at high current levels, an appreciable voltage drop takes place across the neutral semiconductor regions and the contact regions. This leads to a deviation from the ideal current–voltage characteristics in addition to the effects discussed in this section.

6.4 Junction Diode Switching Times

We have so far considered the response of a *p-n* junction to steady forward or reverse bias. Junction diodes are also used in switching applications where it is made to carry out a transition from one bias state to another. A simple circuit for studying the transition of a diode from initial forward bias to reverse bias is shown in Fig. 6.15.

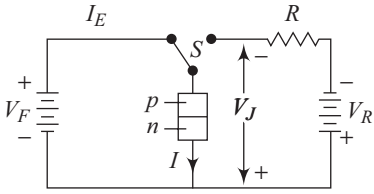


Fig. 6.15 Circuit for studying diode transient

A forward current, I_F, is initially made to flow through the *p-n* junction. At time $t = 0$, the switch S is used to ensure that a reverse current $I_R \approx V/R$ flows through the diode. The corresponding variation in the junction voltage as a function of time and the current transient are shown in Fig. 6.16.

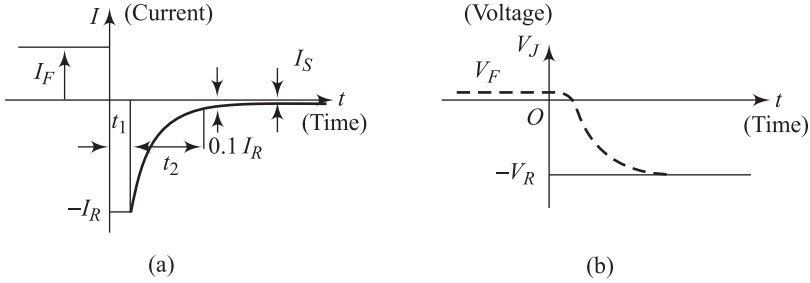


Fig. 6.16 (a) Junction current and (b) voltage variation with time for a *p-n* junction

From Fig. 6.16 (b) it can be seen that on reversing the bias, initially a high reverse current flows through the diode. This high reverse current is sustained for t_1 seconds. The reverse current then reduces to finally reach the reverse saturation current I_S. The time t_2 is the time required for the reverse current to reach a magnitude of $0.1I_R = 0.1V/R$. The time interval t_1 is also called the *storage time*.

The minority carrier density p_n for the *n*-side of the *p-n* junction as a function of time is given in Fig. 6.17.

From Fig. 6.17, we can see that the storage time t_1 is the time needed for the minority carrier concentration at the space charge edge to reach the thermal-equilibrium value.

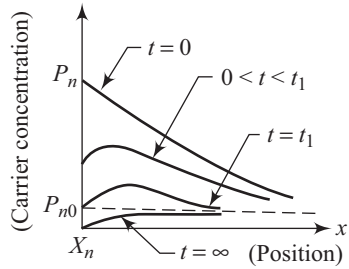


Fig. 6.17 Transient response of minority carrier density

For a one-sided p^+-n junction, equations for t_1 and t_2 can be obtained by solving the continuity equation

$$\frac{\partial p_n(x,t)}{\partial t} = D_p \frac{\partial^2 p_n(x,t)}{\partial x^2} - \left[\frac{p_n(x,t) - p_{n0}}{\tau_p}\right] \tag{6.100}$$

with suitable boundary conditions. These equations are

$$\text{erf}\sqrt{\frac{t_1}{\tau_p}} = \frac{1}{(1 + I_R/I_F)} \tag{6.101}$$

and

$$\text{erf}\sqrt{\frac{t_2}{\tau_p}} + \frac{\exp(-t_2/\tau_p)}{\sqrt{\pi(t_2/\tau_p)}} = 1 + 0.1\left(\frac{I_R}{I_F}\right) \tag{6.102}$$

Equations (6.101) and (6.102) are shown schematically in Fig. 6.18.

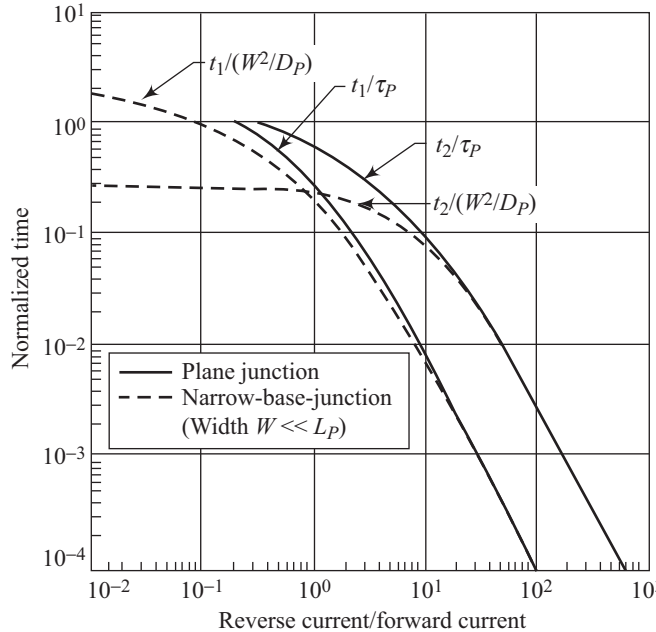


Fig. 6.18 Normalized time versus the ratio of reverse current to forward current

An approximation to Eqn (6.101) looks like

$$t_1 \approx \tau_p \ln\left[1 + \frac{I_F}{I_R}\right] \tag{6.103}$$

For large values of I_R/I_F and for $W \gg L_p$ (where W is the length of the n-type material), we get for the total transient time $(t_1 + t_2)$,

$$(t_1 + t_2) \simeq \frac{\tau_p}{2}\left(\frac{I_R}{I_F}\right)^{-1} \tag{6.104}$$

If I_R/I_F has large value but $W \ll L_p$, we have

$$t_1 + t_2 \simeq \frac{W^2}{2D_p}\left(\frac{I_R}{I_F}\right)^{-1} \tag{6.105}$$

Equation (6.101) is used as an experimental method for the determination of minority carrier lifetime. Deep level impurities such as gold in silicon can be used to reduce τ_p and hence result in fast switching diodes.

- At high current densities, minority carrier density is comparable to the majority carrier density and this condition is called high-injection condition. In such a situation, both drift and diffusion current components have to be considered. The finite resistivity of the quasi-neutral regions also plays a role under high injection condition.
- Diodes used as rectifiers have slow switching speeds. High-frequency applications, however, require high switching speeds.

Solved Problems

6.1 The n-side of a silicon p-n junction has a doping level of 5×10^{16} cm^{-3}. A forward bias of 0.55 V is applied on the p-n junction. Determine the minority carrier concentration at the edge of the space charge region. Assume that the p-n junction is at 300 K and $n_i = 1.5 \times 10^{10}$ cm^{-3}. Take $kT/e = 0.026$ V.

Solution

The minority carrier hole concentration, p_n, is given by

$$p_n = p_{n0} \exp\left(\frac{eV}{kT}\right) \tag{6.1.1}$$

Also,

$$p_{n0} = \frac{n_i^2}{N_d} \tag{6.1.2}$$

Putting the given values in Eqn (6.1.2) yields

$$p_{n0} = \frac{(1.5 \times 10^{10})^2}{5 \times 10^{16}} = 4.5 \times 10^3 \text{ cm}^{-3}$$

Substituting the value of p_{n0} in Eqn (6.1.1) results in

$$p_n = (4.5\times10^3)\exp\left(\frac{0.55}{0.026}\right)$$

which yields

$$p_n = 6.92 \times 10^{12}\ \text{cm}^{-3}.$$

The excess electron concentration in the *n*-region is also 6.92×10^{12} cm^{-3}. This value is much smaller than the thermal-equilibrium electron concentration, implying low injection condition.

6.2 A silicon *p-n* junction has the following parameters $N_a = 10^{16}$ cm^{-3}, $N_d = 5 \times 10^{16}$ cm^{-3}, $D_n = 25$ cm^2/s, $D_p = 10$ cm^2/s, $\tau_{p0} = 4 \times 10^{-7}$ s, and $\tau_{n0} = 2 \times 10^{-7}$ s.

Assuming $n_i = 1.5 \times 10^{10}$ cm^{-3} at 300 K, calculate the reverse saturation current density assuming the *p-n* junction to be ideal.

Solution

The reverse saturation current density for an ideal *p-n* junction is given by

$$J_s = \frac{eD_n n_{p0}}{L_n} + \frac{eD_p p_{n0}}{L_p} \tag{6.2.1}$$

Equation (6.2.1) can be rewritten in the form

$$J_s = en_i^2\left[\frac{1}{N_a}\sqrt{\frac{D_n}{\tau_{n0}}} + \frac{1}{N_d}\sqrt{\frac{D_p}{\tau_{p0}}}\right] \tag{6.2.2}$$

Putting the given values in Eqn (6.2.2) yields

$$J_s = (1.6\times10^{-19})(1.5\times10^{10})^2\left[\frac{1}{10^{16}}\sqrt{\frac{25}{2\times10^{-7}}} + \frac{1}{5\times10^{16}}\sqrt{\frac{10}{4\times10^{-7}}}\right]$$

This leads to

$$J_s = 4.39 \times 10^{-11}\ \text{A/cm}^2.$$

6.3 A silicon *p-n* junction diode consists of *p* and *n* regions with conductivities of 1000 Ω^{-1} m^{-1} and 20 Ω^{-1} m^{-1}, respectively. The minority carrier lifetimes in the two regions are 5 μs and 1 μs, respectively. Determine the ratio of hole current to electron current in the depletion layer, the reverse-bias saturation current density and the total current density flowing through the junction for a forward bias of 0.4 V. Assume T = 300 K, $n_i = 1.5 \times 10^{16}$ m^{-3}, $\mu_n = 0.13$ m^2/V s, and $\mu_p = 0.05$ m^2/V s.

Solution

The majority carrier densities are given by

$$p_{p0} \cong \frac{\sigma_p}{e\mu_p} = \frac{1000}{1.602\times10^{-19}\times0.05} = 1.248\times10^{23}\,\text{m}^{-3} \tag{6.3.1}$$

and

$$n_{n0} \cong \frac{\sigma_N}{e\mu_n} = \frac{20}{1.602 \times 10^{-19} \times 0.13} = 9.603 \times 10^{20}\,\text{m}^{-3} \tag{6.3.2}$$

Using mass-action law, the corresponding minority carrier densities are

$$n_{p0} = \frac{n_i^2}{p_{p0}} = \frac{(1.5 \times 10^{16})^2}{1.248 \times 10^{23}} = 1.80 \times 10^{9}\,\text{m}^{-3} \tag{6.3.3}$$

$$p_{n0} = \frac{n_i^2}{n_{n0}} = \frac{(1.5 \times 10^{16})^2}{9.603 \times 10^{20}} = 2.343 \times 10^{11}\,\text{m}^{-3} \tag{6.3.4}$$

Using the Einstein relation, we get

$$D_n = \mu_n \times kT = 0.13 \times 8.61 \times 10^{-5} \times 300 = 3.358 \times 10^{-3}\ \text{m}^2/\text{s} \tag{6.3.5}$$

and

$$D_p = \mu_p \times kT = 0.05 \times 8.61 \times 10^{-5} \times 300 = 1.291 \times 10^{-3}\ \text{m}^2/\text{s} \tag{6.3.6}$$

which yields

$$L_n = \sqrt{D_n \tau_n} = \sqrt{3.358 \times 10^{-3} \times 10^{-6}} = 5.79 \times 10^{-5}\,\text{m}$$

and

$$L_p = \sqrt{D_p \tau_p} = \sqrt{1.291 \times 10^{-3} \times 5 \times 10^{-6}} = 8.034 \times 10^{-5}\,\text{m}$$

The reverse saturation current density J_s is given by

$$J_s = \frac{eD_n n_{p0}}{L_n} + \frac{eD_p p_{n0}}{L_p} \tag{6.3.7}$$

Equation (6.3.7) can be rewritten as

$$J_s = \frac{en_{p0} L_n}{\tau_n} + \frac{ep_{n0}\, L_p}{\tau_p} \tag{6.3.8}$$

Putting the evaluated values in Eqn (6.3.8) yields

$$J_s = \frac{1.602 \times 10^{-19} \times 1.80 \times 10^{9} \times 5.79 \times 10^{-5}}{1 \times 10^{-6}} + \frac{1.602 \times 10^{-19} \times 2.343 \times 10^{11} \times 8.034 \times 10^{-5}}{5 \times 10^{-6}}$$

which results in

$$J_s = 0.016 \times 10^{-6} + 0.603 \times 10^{-6} = 0.619 \times 10^{-6}\ \text{A/m}^2 \tag{6.3.9}$$

The ratio of hole current to electron current is

$$\frac{0.603 \times 10^{-6}}{0.016 \times 10^{-6}} = 37.69$$

The total current density J is given by

$$J = J_s(e^{eV/kT} - 1)$$

which yields

$$J = 0.619 \times 10^{-6}(e^{0.4/0.026} - 1) = 2.97\ \text{A/m}^2$$

6.4 A *p-n* junction has specifications such that $N_a \gg N_d$, implying $p_{n0} \gg n_{p0}$. Assuming $T = 300$ K, $\tau_{p0} = 5 \times 10^{-7}$ s, and $I_{p0} = 5$ mA, calculate the diffusion capacitance. Take $kT/e = 0.026$ V for $T = 300$ K.

Solution

From Eqn (6.66) we can write

$$C_{d0} = \frac{e}{kT}\left[\frac{eL_p p_{n0}}{2} + \frac{eL_n n_{p0}}{2}\right] e^{eV_0/kT} \text{ F/cm}^2$$

For $p_{n0} \gg n_{p0}$,

$$C_{d0} \cong \frac{e}{kT}\left[\frac{eL_p p_{n0}}{2}\right] e^{eV_0/kT} \text{ F/cm}^2$$

Thus,

$$C'_{d0} = C_{d0}A = \frac{e}{2kT}[eL_p p_{n0}A]\, e^{eV_0/kT} \text{ F} \tag{6.4.1}$$

Equation (6.4.1) can be rewritten as

$$C'_{d0} = \frac{e}{2kT}\left[\frac{eL_p^2 p_{n0}A}{L_p}\right] e^{eV_0/kT} \tag{6.4.2}$$

We know

$$L_p^2 = D_p \tau_{p0} \tag{6.4.3}$$

Using Eqn (6.4.3) in Eqn (6.4.2), we get

$$C'_{d0} = \frac{e}{2kT}\left[\frac{ep_{n0}AD_p\tau_{p0}}{L_p}\right] e^{eV_0/kT}$$

which leads to

$$C'_{d0} = \frac{e}{2kT}[I_{p0}\tau_{p0}] \tag{6.4.4}$$

where $I_{p0} = \dfrac{AeD_p p_{n0}}{L_p} e^{eV_0/kT}$

Putting the given values in Eqn (6.4.4), we get

$$C'_{d0} = \frac{1}{2 \times (0.026)} \times 0.5 \times 10^{-3} \times 5 \times 10^{-7} \cong 4 \times 10^{-9} \text{ F}$$

6.5 Calculate the reverse-bias generation current density for the *p-n* junction indicated in Solved Problem 6.2. Assume $V_{bi} + V_R = 4$ V and $n_i = 1.5 \times 10^{10}$ cm^{-3}.

Solution

We know that

$$J_{gen} = \frac{en_i W}{2\tau_m} \tag{6.5.1}$$

where

$$\tau_m = \frac{\tau_{p0} + \tau_{n0}}{2} \tag{6.5.2}$$

Also

$$W = \left\{ \frac{2\varepsilon_s}{e} \left(\frac{N_a + N_d}{N_a N_d} \right) (V_{\text{bi}} + V_R) \right\}^{1/2} \tag{6.5.3}$$

Putting the given values in Eqn (6.5.3), we get

$$W = \left\{ \frac{2 \times 11.7 \times 8.85 \times 10^{-14}}{1.6 \times 10^{-19}} \left(\frac{10^{16} + 5 \times 10^{16}}{10^{16} \times 5 \times 10^{16}} \right) \times 4 \right\}^{1/2}$$

which yields

$$W = 0.79 \times 10^{-4} \text{ cm} = 0.79 \text{ μm}$$

Using Eqn (6.5.2), we can write

$$\tau_m = \frac{4 \times 10^{-7} + 2 \times 10^{-7}}{2} = 3 \times 10^{-7} \text{ s}$$

Putting the values of W and τ_m in Eqn (6.5.1) yields

$$J_{\text{gen}} = \frac{1.6 \times 10^{-19} \times 1.5 \times 10^{10} \times 0.79 \times 10^{-4}}{2 \times 3 \times 10^{-7}}$$
$$= 3.16 \times 10^{-7} \text{ A/cm}^2$$

It is important at this stage to compare this value of J_{gen} with the value of J_s equal to 4.39×10^{-11} A/cm^2.

6.6 Consider a silicon *p-n* junction diode at $T = 300$ K. Design the diode such that $J_n = 20$ A/cm^2 and $J_p = 5$ A/cm^2 at $V_a = 0.65$ V. Assume the values of the remaining parameters as given below:

$$D_n = 25 \text{ cm}^2/\text{s}, D_p = 10 \text{ cm}^2/\text{s}, \tau_{n0} = \tau_{p0} = 5 \times 10^{-7} \text{ s}, \varepsilon_r = 11.8$$

Solution

Using Eqn (6.29) in the text

$$J_p(x_n) = \frac{eD_p}{L_p} p_{n0} (e^{eV/kT} - 1)$$

Now:

$$L_p = \sqrt{D_p \tau_p}$$

$$p_{n0} = \frac{J_r(x_n) L_p}{eD_p [e^{eV/kT} - 1]}$$

$$= \frac{5 \times \sqrt{5 \times 10^{-7}}}{1.6 \times 10^{-19} \times \sqrt{10} \times [e^{0.65/0.0259} - 1]}$$

$$= 9.3 \times 10^4 \text{ cm}^{-3}$$

and using the law of mass action

$$p_{n0} = n_i^2 / N_d$$

$$\therefore \quad N_d = \frac{2.25 \times 10^{20}}{9.3 \times 10^4} = 2.42 \times 10^{15}\ \text{cm}^{-3}$$

Similarly, from Eqn (6.31) in the text

$$J_n(-x_p) = \frac{eD_n}{L_n} n_{p0} (e^{eV/kT} - 1)$$

Now

$$L_n = \sqrt{D_n \cdot \tau_n}$$

Again, using the law of mass action

$$n_{p0} = n_i^2 / N_a$$

we get

$$N_a = 1.01 \times 10^{15}\ \text{cm}^{-3}$$

Comment: In case, it is required, one can always vary the relative magnitude of the electron and hole current densities through a *p-n* junction structure by changing the doping concentrations in the device.

6.7 Consider a silicon *p-n* junction at $T = 300$ K, so that $n_i = 1.5 \times 10^{10}\ \text{cm}^{-3}$. Assume that the *n*-type doping is $1 \times 10^{16}\ \text{cm}^{-3}$ and that a forward bias of 0.6 V is applied to the *p-n* junction. Calculate the minority carrier hole concentration at the edge of the space charge region. Comment on the result obtained.

Solution

Using Eqn (6.9) in the text:

$$p_n = p_{n0}\, e^{eV/kT}$$

$$= \frac{2.25 \times 10^{20}}{1 \times 10^{16}} \times e^{0.6/0.0259}$$

$$= 2.59 \times 10^{14}\ \text{cm}^{-3}$$

Comment: Concentration of minority carriers at the edge of space charge region can be significantly high as compared to the equilibrium gradient, current contribution may also be large.

6.8 A silicon p^+-*n* junction has a donor doping of $5 \times 10^{16}\text{cm}^{-3}$ on the *n*-side and a cross-sectional area of $10^{-3}\ \text{cm}^2$. If $\tau_p = 1\ \mu\text{s}$ and $D_p = 10\ \text{cm}^2/\text{s}$, calculate the current with a forward bias of 0.5 V at 300 K.

Solution

As this is a p^+-*n* structure, the greater contribution to current will be from the holes present in the *n*-region. As a result, we can write the resulting diode equation as

$$I = eA \frac{D_p}{L_p} p_n e^{eV/kT}$$

where p_n is the hole concentration in the *n*-region and V is the applied forward bias.

We can find p_n using the law of mass action

$$p_n = n_i^2 / n$$

n_i at room temperature = 1.5×10^{10} cm^{-3}
Hence

$$p_n = 4.5 \times 10^3 \text{ cm}^{-3}$$
$$L_P = (D_p \tau_p)^{1/2}$$
$$= (10 \times 1 \times 10^{-6})^{1/2}$$
$$= 3.16 \times 10^{-3} \text{ cm}$$

Putting these values in the above expression for current we get

$$I = eA \frac{D_p}{L_p} p_n (e^{eV/kT})$$

$$= \frac{1.6 \times 10^{-19} \times 10^{-3} \times 10 \times 4.5 \times 10^3 \left(e^{\frac{0.5}{0.0259}} - 1 \right)}{3.16 \times 10^{-3}}$$

$$= 0.55 \text{ μA}$$

6.9 A silicon *p-n* junction diode is to be operated at $T = 300$ K and its relevant parameters are as follows:
$N_a = N_d = 10^{16}$ cm^{-3}, $D_n = 25$ cm^2/s, $D_p = 10$ cm^2/s, $\tau_{n0} = \tau_{p0} = 5 \times 10^{-7}$ s, $\varepsilon_r = 11.7$, mobility of electrons = 1350 cm^2/V s, mobility of holes = 450 cm^2/V s.
Calculate the electric field value deep in the *n*-region (far away from the space charge region) under the effect of an applied forward bias of 0.65 V.

Solution

In the text, we assumed (very strongly), for the derivation of equations, that no electric field exists in the neutral region. We will revisit this concept through this problem.
Given

$$N_a = N_d = 10^{16} \text{ cm}^{-3}$$
$$D_n = 25 \text{ cm}^2/\text{s}$$
$$D_p = 10 \text{ cm}^2/\text{s}$$
$$\tau_{n0} = \tau_{p0} = 5 \times 10^{-7} \text{ s}$$
$$\varepsilon_r = 11.7, \mu_n = 1350 \text{ cm}^2/\text{Vs}$$
$$\mu_p = 450 \text{ cm}^2/\text{Vs}$$

We know from Eqn (6.34) of the text

$$J = J_s \left[\exp\left(\frac{eV}{kT} \right) - 1 \right]$$

$$J_s = \left[\frac{eD_p p_{n0}}{L_p} + \frac{eD_n n_{p0}}{L_n} \right]$$

$$J_s = \left[\frac{eD_p n_i^2 / n_{n0}}{\sqrt{D_p \tau_p}} + \frac{eD_n n_i^2 / p_{p0}}{\sqrt{D_n \tau_n}} \right]$$

$$n_{n0} = p_{p0} = N_a = N_d = 10^{16}$$

$$\therefore \quad J_s = \left[\frac{1.6 \times 10^{-19} \times (10)^{\frac{1}{2}} \times 2.25 \times (10^{20}/10^{16})}{(5 \times 10^{-7})^{\frac{1}{2}}} + \frac{1.6 \times 10^{-19} \times (25)^{\frac{1}{2}} \times 2.25 \times (10^{20}/10^{16})}{(5 \times 10^{-7})^{\frac{1}{2}}} \right]$$

$$J_s = \frac{1.6 \times 10^{-19} \times 2.25 \times 10^{20}}{(5 \times 10^{-7})^{\frac{1}{2}} \times 10^{16}} [3.16 + 5]$$

$$= \frac{36}{7.071 \times 10^{-4}} (8.16)$$

$$J_s = 4.15 \times 10^{-11} \text{ A/cm}^2$$

Putting this obtained value of J_s in the equation for J we get

$$J = [4.15 \times 10^{-11}] \left[\exp\left(\frac{0.65}{0.0259} - 1 \right) \right]$$

$$J = 3.29 \text{ A/cm}^2$$

Now applying $J = \sigma E$

Far away from the space charge region, carrier population will be mainly dominated by electrons only

Hence

$$\sigma = eN_d \mu_n$$
$$= (1.6 \times 10^{-19}) \times (10^{16}) \times (1350)$$
$$= 2.16 \text{ mho cm}^{-1}$$

Now, we can find E as

$$E = \frac{J}{\sigma} = \frac{3.29}{2.16} = 1.52 \text{ V/cm}$$

Comment: This value of electric field is very small. Therefore, the assumption that we make for derivation of most equations, i.e., regions far away from the space charge region should be called *neutral*, is very much valid.

6.10 The reverse saturation current for a *p-n* junction diode is measured to be 2 μA at 300 K. Calculate the ac resistance of the diode at a forward bias of 100 mV. Assume ideality factor of the diode to be unity.

Solution

$$I = I_s \left[\exp\left(\frac{eV}{kT} \right) - 1 \right] \tag{6.10.1}$$

Also,

$$r_{ac} = \frac{dV}{dI} = \frac{1}{\frac{dI}{dV}} \tag{6.10.2}$$

Using Eqn (6.10.1), Eqn (6.10.2) leads to,

$$r_{ac} = \frac{1}{I_s \cdot \frac{e}{kT} \cdot \exp\left(\frac{eV}{kT}\right)} \tag{6.10.3}$$

We also know that $\frac{kT}{e}$ at $300\ \text{K} \approx 26\ \text{mV}$. Therefore putting values in Eqn (6.10.3) leads to

$$r_{ac} = \frac{1}{\frac{2\times10^{-6}}{26\times10^{-3}} \exp\left[\frac{0.1}{26\times10^{-3}}\right]}$$

yielding,

$$r_{ac} = \frac{26\times10^{-3}}{2\times10^{-6}\ \exp(3.85)} = \frac{26\times10^{-3}}{2\times10^{-6}\times47}$$

resulting in,

$$r_{ac} = 276.60\ \Omega$$

6.11 The reverse saturation current density of a germanium *p-n* junction diode is $200\ \text{mA/cm}^2$ at 300 K. Calculate the forward voltage needed to result in a forward current density of $10^5\ \text{A/m}^2$.

Solution

$$I = I_s\left[\exp\left(\frac{eV}{kT}\right) - 1\right] \tag{6.11.1}$$

or

$$J = J_s\left[\exp\left(\frac{eV}{kT}\right) - 1\right] \tag{6.11.2}$$

yielding,

$$\exp\left(\frac{eV}{kT}\right) - 1 = \frac{J}{J_s} = \frac{10^5}{200\times10^{-3}} = 5\times10^5$$

resulting in,

$$\exp\left(\frac{eV}{kT}\right) \simeq 5\times10^5$$

giving,

$$\frac{eV}{kT} = \log_e 5\times10^5 = 13.12$$

leading to,

$$\mathrm{V} = \frac{13.12 \times (1.38 \times 10^{-23}) \times 300}{1.6 \times 10^{-19}} = 0.34 \text{ V}$$

Recapitulation

- The Maxwell–Boltzmann approximation is mathematically expressed in the form

$$f_F(E) \approx \exp\left[-\frac{(E - E_F)}{kT}\right]$$

- When a *p-n* junction is biased,

$$n_p = n_{p0} \exp\left(\frac{eV}{kT}\right)$$

and

$$p_n = p_{n0} \exp\left(\frac{eV}{kT}\right)$$

- The diffusion length is related to the minority carrier lifetime through the expression

$$L_n^2 = D_n \tau_{n0}$$

and

$$L_p^2 = D_p \tau_{p0}$$

- The *p-n* junction *I-V* characteristic is given by

$$J = J_s\left[\exp\left(\frac{eV}{kT}\right) - 1\right]$$

where

$$J_s = \left[\frac{eD_p p_{n0}}{L_p} + \frac{eD_n n_{p0}}{L_n}\right]$$

- For a short diode

$$J_p(x) = \frac{eD_p p_{n0}}{W_n}\left[\exp\left(\frac{eV}{kT}\right) - 1\right]$$

- Diffusion conductance, g_d, is given as

$$g_d = \frac{dI_D}{dV}$$

and is evaluated at the point of interest.

- Low-frequency diffusion capacitance C_{d0} per unit area is given by

$$C_{d0} = \frac{e}{kT}\left[\frac{eL_p p_{n0}}{2} + \frac{eL_n n_{p0}}{2}\right] e^{eV_0/kT}$$

- Reverse-bias generation current density, J_{gen}, can be expressed in the form

$$J_{gen} = \frac{en_i W}{2\tau_m}$$

- Forward-bias recombination current density, J_{rec}, is approximately given by

$$J_{rec} \cong \frac{eWn_i}{2\tau_0} \exp\left(\frac{eV}{2kT}\right)$$

- Current–voltage characteristics of a *p-n* junction are generally given by

$$I = I_s \left[\exp\left(\frac{eV}{nkT}\right) - 1\right]$$

where n is called the ideality factor of the *p-n* junction.

- Storage time t_1 for the p^+-n junction is given by

$$\text{erf}\sqrt{\frac{t_1}{\tau_p}} = \frac{1}{(1 + I_R/I_F)}$$

- Time t_2 required for the reverse current to reach a magnitude of $0.1I_R$ is expressed as

$$\text{erf}\sqrt{\frac{t_2}{\tau_p}} + \frac{\exp(-t_2/\tau_p)}{\sqrt{\pi(t_2/\tau_p)}} = 1 + 0.1\left(\frac{I_R}{I_F}\right)$$

Exercises

Review Questions

6.1 Explain the main assumptions used to derive the Shockley equation.
6.2 Derive the relation between the minority carrier concentration on the *p*-side and the majority carrier concentration on the *n*-side of a *p-n* junction.
6.3 How is the relation derived in Review Question 6.2 modified when a *p-n* junction gets biased?
6.4 Explain long *p-n* junction approximation.
6.5 Derive the ideal current–voltage characteristics of a *p-n* junction.
6.6 How does the diffusion current expression for a short diode differ from that for a long diode?
6.7 Define diffusion conductance of a *p-n* junction and give its measurement units.
6.8 Derive expressions for diffusion capacitance and diffusion conductance for a *p-n* junction at low frequencies.
6.9 How does series resistance affect the current–voltage characteristics of a *p-n* junction?
6.10 Derive an expression for total reverse bias current for a *p-n* junction.
6.11 Derive an expression for recombination current for a forward-biased *p-n* junction.
6.12 Describe a technique to determine minority carrier lifetime using the transient response of a *p-n* junction. (This technique is also called the reverse recovery technique).

6.13 Express the total minority carrier concentration in terms of the thermal-equilibrium minority carrier concentration for a biased *p-n* junction.
6.14 What is long *p-n* junction approximation?
6.15 Show the *I-V* characteristic of a biased *p-n* junction graphically.
6.16 Define diffusion conductance, g_d, and illustrate the concept graphically.
6.17 What is the role of frequency in deciding the diffusion capacitance of a forward biased *p-n* junction?
6.18 Sketch the effect of finite resistance on the *I-V* characteristics of a *p-n* junction.
6.19 Draw the complete equivalent circuit of a *p-n* junction.
6.20 Derive the dependence of J_{gen} on applied bias for a reverse-biased abrupt *p-n* junction.
6.21 Derive an expression for J_{rec} for a forward-biased *p-n* junction.
6.22 Explain the concept of ideality factor of a *p-n* junction.
6.23 A *p-n* junction is taken to a reverse-bias condition from an initial forward-bias condition. What are the factors on which the switching time depends? Show the transient graphically.

Problems

6.1 A *p-n* junction consists of a *p*-region with a doping level of 1×10^{16} cm^{-3}. A forward bias of 0.5 V is applied across the junction. Calculate the minority carrier concentration at the edge of the space charge region. Take $n_i = 1.5 \times 10^{10}$ cm^{-3} at 300 K and assume $kT/e = 0.026$ V.

$$\left[Hint : n_p = n_{p0} \exp\left(\frac{eV}{kT}\right) \right]$$

Ans. 5.06×10^{12} cm^{-3}

6.2 The *n* side of a germanium *p-n* diode is doped to a level of 2×10^{16} cm^{-3}. A bias of 0.1 V is applied to the *p-n* junction. Calculate the minority carrier concentration at the edge of the space charge region. Assume the *p-n* junction temperature to be 300 K with $n_i = 2.4 \times 10^{13}$ cm^{-3}. Take the value of $kT/e = 0.026$ V.

Ans. 1.35×10^{12}cm^{-3}

6.3 Calculate the reverse saturation current density for the Si *p-n* junction with the following parameters:
$N_a = N_d = 5 \times 10^{16}$ cm^{-3}; $D_n = 25$ cm^2/s; $D_p = 10$ cm^2/s
$\tau_{p0} = \tau_{n0} = 3 \times 10^{-7}$ s
Take $n_i = 1.5 \times 10^{10}$ cm^{-3} at 300 K.

$$\left[Hint : J_s = en_i^2 \left[\frac{1}{N_a}\sqrt{\frac{D_n}{\tau_{n0}}} + \frac{1}{N_d}\sqrt{\frac{D_p}{\tau_{p0}}} \right] \right]$$

Ans. 1.07×10^{-11} A/cm^2

6.4 The *p* and *n* regions of a silicon *p-n* junction have resistivities of 1.0 Ω m and 0.2 Ω m, respectively. The minority carrier lifetimes in the two regions are 5 μs and 1 μs, respectively. Calculate the reverse-bias saturation current density of the *p-n* junction. Assume $T = 300$ K, $n_i = 1.5 \times 10^{16}$ m^{-3}, $\mu_n = 0.13$ m^2/Vs, and $\mu_p = 0.05$ m^2/Vs.

$$\left[Hint: J_s = \frac{en_{p0}L_n}{\tau_n} + \frac{ep_{n0}L_p}{\tau_p} \right]$$

Ans. 0.041×10^{-4} A/m^2

6.5 Determine the ratio of hole current to electron current for Problem 6.4. Suggest some methods to reduce the ratio to unity.

[*Hint:* $J_s = J_n + J_p$]

Ans. 1.41

6.6 A one-sided p^+-*n* silicon diode is made to go through a transient response with $I_F = 1.7$ mA and $I_R = 0.5$ mA. Calculate the storage time t_1, assuming minority carrier lifetime $\tau_p = 1 \times 10^{-7}$.

$$\left[Hint: \text{erf}\sqrt{\frac{t_1}{\tau_p}} = \frac{I_F}{I_F + I_R};\ \text{erf}(0.85) = 0.77 \right]$$

Ans. 0.72×10^{-7} s

6.7 A reverse recovery technique based set-up is being used to measure the minority carrier lifetime in a p^+-*n* junction. The set-up uses a forward-bias current $I_F =$ 2 mA and is switched to reverse bias with a total reverse-bias voltage of 2 V across a total series resistance of 4 KΩ. The storage time, t_1, is measured to be 1 μs. Calculate the minority carrier lifetime.

[*Hint:* erf (0.9) = 0.8]

Ans. 1.23×10^{-6} s

6.8 A silicon *p-n* junction diode has the following parameters at 300 K: $N_d = 10^{17}$ cm^{-3}; $N_a = 5 \times 10^{15}$ cm^{-3}; $D_n = 25$ cm^2/s, $D_p = 10$ cm^2/s, $\tau_{n0} = 5 \times 10^{-7}$ s, and $\tau_{p0} = 10^{-7}$ s. Calculate the diffusion capacitance of the *p-n* junction at a forward bias of 0.55 V.

$$\left[Hint: C_{d0} = \frac{e}{kT}\left[\frac{ep_{n0}\sqrt{D_p\tau_{p0}}}{2} + \frac{en_{p0}\sqrt{D_n\tau_{n0}}}{2} \right] \exp\left(\frac{eV_0}{kT} \right) \right]$$

Ans. 7.63×10^{-7} F/cm^2

6.9 A silicon *p-n* junction has the following parameters:

$N_a = 10^{16}$ cm^{-3}; $N_d = 5 \times 10^{16}$ cm^{-3}

$D_n = 25$ cm^2/s; $D_p = 10$ cm^2/s

$\tau_{p0} = 3 \times 10^{-7}$ s; $\tau_{n0} = 3 \times 10^{-7}$ s

Calculate the reverse-bias generation density for

$V_{\text{bi}} + V_R = 5$ V

$$\left[Hint: J_{\text{gen}} = \frac{en_i W}{2\tau_m} \right]$$

Ans. 3.52×10^{-7} A/cm^2

6.10 A device engineer applied a forward bias varying from 0.7 V to about 0.9 V to a *p-n* junction. He tried to draw the current–voltage characteristics of the *p-n* junction on a semi-log plot and obtained the variation of current as shown in Fig. 6.P10.1.

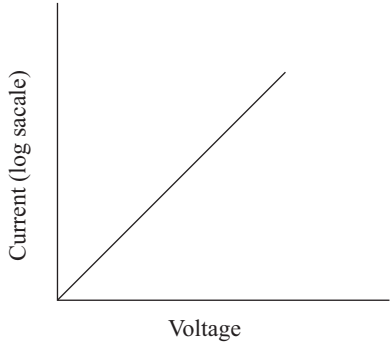


Fig. 6.P10.1 Current–voltage characteristics

Is his observation correct, taking into consideration all non-idealities of a *p-n* junction? Justify your answer with suitable explanation.

Ans. No. Discuss with your teacher

6.11 The reverse saturation current for a *p-n* junction diode is measured to be 3 μA at 300 K. Determine the ac resistance of the diode at a forward bias of 200 mV. The ideality factor of the device is unity.

$$\left[\text{Hint: } r_{\text{ac}} = \frac{1}{I_s \times \frac{e}{kT} \times \exp\left(\frac{eV}{kT}\right)} \right]$$

Ans. 3.95 Ω

6.12 The reverse saturation current density of a diode is measured to be 100 mA/cm^2 at 300 K. Determine the forward voltage needed that results in a forward current density of 10^5 A/m^2.

$$\left[\text{Hint: } \exp\left(\frac{eV}{kT}\right) - 1 = \frac{J}{J_s} \right]$$

Ans. 0.36 V

7 Metal–Semiconductor Junctions and Heterojunctions

- Metal–semiconductor contacts
- Band diagram using Schottky model
- Space charge width and junction capacitance
- *J-V* characteristic using emission model
- Schottky barrier lowering
- Tunnelling current
- Effect of surface states on metal–semiconductor contacts
- Metal–semiconductor ohmic contacts
- Specific contact resistance
- Energy band diagram of heterojunctions
- Two-dimensional (2D) electron gas in isotype heterojunctions

Learning Objectives

After going through this chapter the student will be able to

- draw energy band diagrams for heterojunctions using the Schottky model
- derive expressions for the space charge width and junction capacitance of metal–semiconductor junctions
- derive *J-V* characteristics of heterojunctions based on the emission model
- derive an expression for Schottky effect leading to barrier lowering
- understand the role of tunnelling current as the doping level of a semiconductor increases
- understand the effect of surface states and interface on the characteristics of metal–semiconductor contacts
- understand the behaviour of metal–semiconductor ohmic contacts
- define specific contact resistance
- draw energy band diagrams for common heterojunctions
- derive electron affinity rule
- understand the concept of 2D electron gas

Introduction

Semiconducting materials find application in a variety of devices. We have so far looked in detail at one such device, namely the *p*-*n* junction. Looking at what we have done so far, one would realize that the *p* and *n* regions of the junction are formed in the same semiconductor material. Such *p*-*n* junctions are called *homojunctions*. These are not the only type of junctions that are useful in semiconductor electronics. There are two interesting variations possible. The two sides of the *p*-*n* junction may be formed in different materials or one of the sides of a junction may actually be a metal. The first category of junctions is called a *heterojunction* and the other, metal–semiconductor junction, respectively. These devices are important and need to be understood well to fully exploit their potential. We will undertake the exercise of understanding these interesting junctions in this chapter. We will see that metal–semiconductor junctions can be either rectifying or ohmic in nature with appropriate consequences. Ohmic contacts, or more generally contacts, play a vital role in deciding the characteristics of semiconductor devices. The rectifying metal–semiconductor junctions are also called Schottky barriers.

7.1 Metal–Semiconductor Contacts

We will start our study of non-homojunctions with the metal–semiconductor contact that plays an important role in almost all junction devices.

7.1.1 Schottky Model

The energy band diagram for metal contacts to *n*-type semiconductors is shown schematically in Fig. 7.1.

Figures 7.1(a) and (b) represent the situation with $\phi_m > \phi_s$, where ϕ_m and ϕ_s are the work functions for the metal and semiconductor, respectively. The work function, as one would recollect, is the difference between the vacuum level and the Fermi level. Figure 7.1(a) shows the individual band diagram of the metal and the semiconductor when the two are maintained at a large distance from each other. From Fig. 7.1(b) it is clear that electrons in the *n*-type semiconductor face a barrier $(\phi_m - \phi_s)$ in trying to reach the metal, whereas the barrier to the reverse flow of electrons from the metal to the semiconductor is $(\phi_m - \chi_s)$, where χ_s is the electron affinity in the semiconductor. Figures 7.1(c) and (d) show the corresponding band diagrams for the *n*-type semiconductor with $\phi_m < \phi_s$.

Suppose we forward bias the junction with a voltage V. For the *n*-type semiconductor, this would amount to making the semiconductor negative with respect to the metal. The barrier for electrons in the *n*-type semiconductor is now modified to $e(V_{\text{bi}} - V)$, whereas the reverse barrier $(\phi_m - \chi_s)$ remains

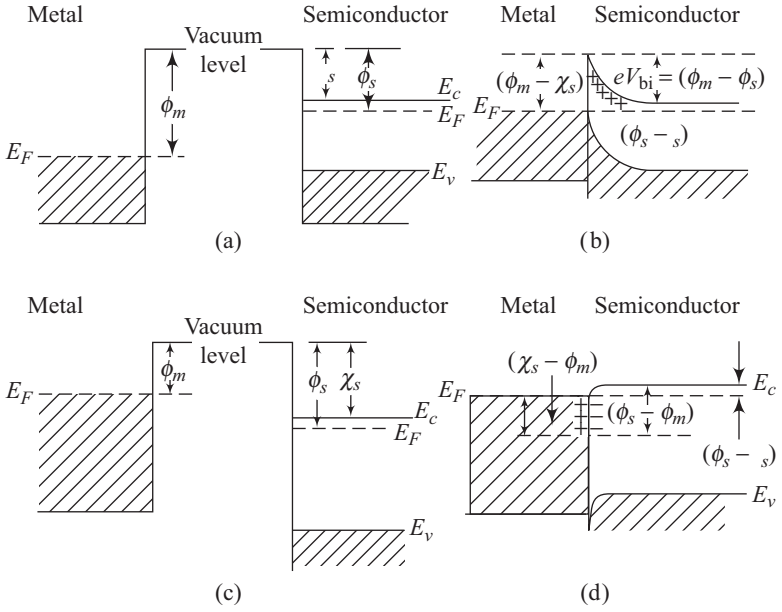


Fig. 7.1 Energy level diagrams of metal contacts to *n*-type semiconductors: (a) and (b) with $\phi_m > \phi_s$, (c) and (d) with $\phi_m < \phi_s$. Contact (b) acts as a rectifier, since a barrier $(\phi_m - \phi_s)$ exists in the conduction band of the semiconductor. Contact (d) is ohmic since virtually no barrier exists in the conduction band.

relatively unaffected by the applied voltage or even by the doping level of the semiconductor.

The barrier $(\phi_m - \chi_s)$ is called the *Schottky barrier* of the particular metal–semiconductor pair. Thus

$$\phi_{B0} = (\phi_m - \chi_s) \tag{7.1}$$

A comparison of Figs 7.1(b) and (d) reveals that the metal–semiconductor junction is rectifying for an *n*-type semiconductor if $\phi_m > \phi_s$. For $\phi_m < \phi_s$, the barrier is unaffected, and is thus called an *ohmic contact.*

The corresponding situation for a *p*-type semiconductor is shown in Fig. 7.2. For the *p*-type semiconductor, $\phi_m > \phi_s$ leads to an ohmic contact, whereas $\phi_m < \phi_s$ results in a rectifying contact.

Tables 7.1 and 7.2 indicate the electron affinities of some common semiconductors and the work functions of some common metals, respectively.

Table 7.1 Electron affinities of some common semiconductors

Element	Electron affinity, χ
Ge, germanium	4.13
Si, silicon	4.01
GaAs, gallium arsenide	4.07
AlAs, aluminium arsenide	3.5

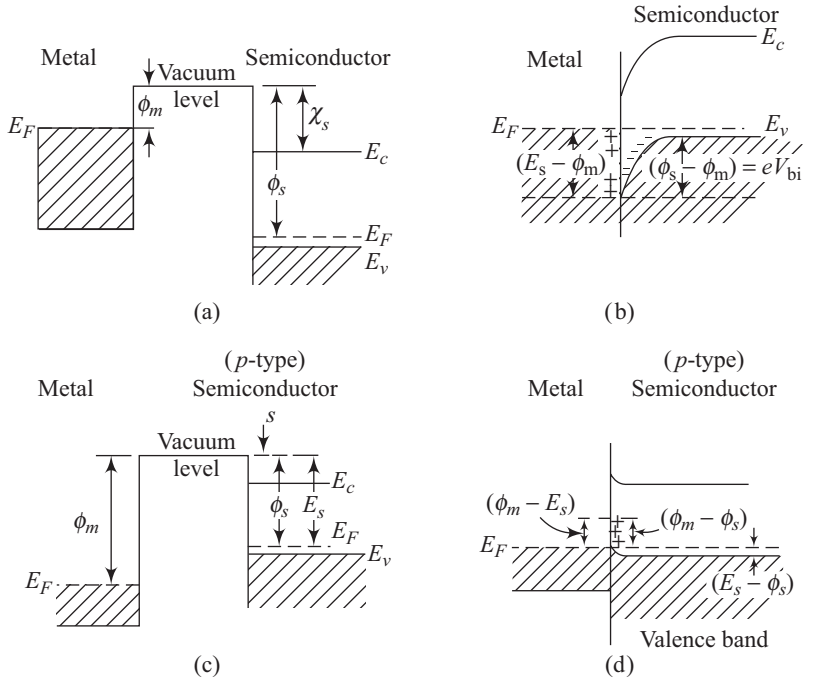


Fig. 7.2 Energy level diagrams of metal contacts to *p*-type semiconductors: (a) and (b) are for $\phi_m < \phi_s$. Contact (b) acts as a rectifier. In (c) and (d) the contact is ohmic.

Table 7.2 Work functions of some common elements

Element	Work function, ϕ_m
Ag, silver	4.26
Al, aluminium	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

7.1.2 Space Charge Width and Junction Capacitance

The ideal metal–semiconductor junction discussed so far has striking similarities with the *p-n* junction dealt with in earlier chapters. The electric field in the space charge region is determined using the Poisson equation, given by

$$\frac{dE}{dx} = \frac{\rho(x)}{\varepsilon_s} \tag{7.2}$$

where $\rho(x)$ is the space charge volume density, and the permittivity of the semiconductor is represented by the symbol ε_s. Assuming a uniformly doped semiconductor, we can write for the *n*-type using Eqn (7.2)

$$E = \int \frac{qN_d}{\varepsilon_S} dx = \frac{eN_d x}{\varepsilon_S} + C_n \tag{7.3}$$

where C_n is a constant of integration.

An important difference with respect to the *p*-*n* junction appears at this stage. A metal cannot sustain an electric field in its bulk and, therefore, *E*-field is zero inside the metal. Suppose $x = x_n$ represents the point where the electric field is zero at the space charge region edge in the semiconductor. Then, using Eqn (7.3) we have

$$E = 0 = \frac{eN_d x_n}{\varepsilon_S} + C_n$$

This leads to

$$C_n = -\frac{eN_d x_n}{\varepsilon_S} \tag{7.4}$$

Using Eqn (7.4) in Eqn (7.3) we have

$$E = -\frac{eN_d}{\varepsilon_s}(x_n - x) \tag{7.5}$$

Thus, the electric field is a linear function of distance and has its maximum value at the metal–semiconductor interface.

The space charge region width, *W*, can be expressed in a manner similar to the one used for the p^+-*n* one-side *p*-*n* junction. Thus

$$W = x_n = \left[\frac{2\varepsilon_s (V_{\text{bi}} + V)}{eN_d}\right]^{1/2} \tag{7.6}$$

where *V* represents the applied reverse bias. Similarly, the junction capacitance (per unit area), C_J, is given by

$$C_J = eN_d \frac{dx_n}{dV} = \left[\frac{e\varepsilon_s N_d}{2(V_{\text{bi}} + V)}\right]^{1/2} \tag{7.7}$$

Evaluating $(1/C_J)^2$ from Eqn (7.7) results in

$$\left(\frac{1}{C_J}\right)^2 = \frac{2(V_{\text{bi}} + V)}{e\varepsilon_s N_d} \tag{7.8}$$

Thus, the slope of the curve obtained using Eqn (7.8) yields the semiconductor doping N_d. The curve can also be used to obtain V_{bi}. The magnitude of V_{bi} can then yield the Schottky barrier ϕ_{B0} using the equation

$$eV_{\text{bi}} = \phi_{B0} - \phi_n \tag{7.9}$$

The value of ϕ_n is the difference between the Fermi level (E_F) and the conduction band of the semiconductor, and can be evaluated from the knowledge of N_d.

7.1.3 Characteristics Based on Emission Model

We will now attempt to model the current flowing in a metal–semiconductor junction assuming the carrier flow to be due to thermionic emission over the barrier. This essentially means that we are initially going to neglect tunnelling effects and image force barrier lowering. Figure 7.3 shows the barrier between a metal and an *n*-type semiconductor under different biasing conditions. It is clear from Fig. 7.3 that the zero-bias electron flux from the semiconductor into the metal is determined by the number of electrons having energy greater than or equal to qV_{bi}, which are directed per second towards the unit area of the relevant interface.

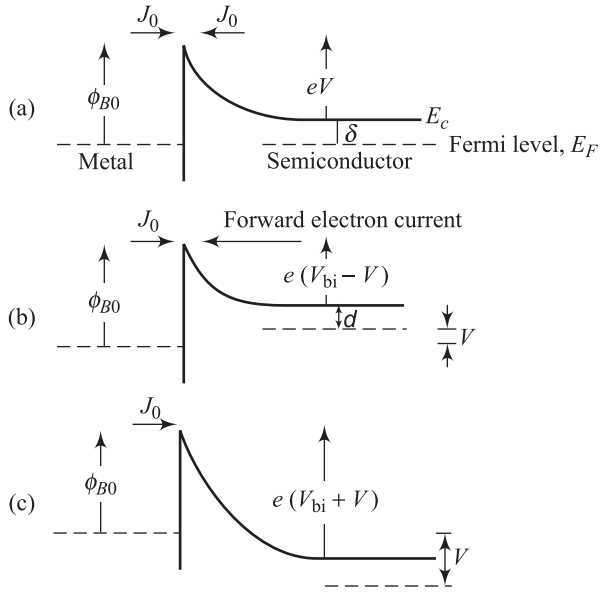


Fig. 7.3 Energy diagrams for a metal–semiconductor junction: (a) zero-bias voltage; (b) forward-bias voltage; (c) reverse-bias voltage

For a Maxwellian distribution of electrons we can write for J_0

$$J_0 = eN_d \left(\frac{kT}{2\pi m^*}\right)^{1/2} \exp\left(\frac{-eV_{bi}}{kT}\right) \tag{7.10}$$

where m^* represents the effective electron mass.

At zero bias, the net flow over the barrier must be zero. Thus, J_0 could also be written in terms of the electron flux entering from the metal into the semiconductor after going over the barrier ϕ_{B0}. Using Eqn (7.9), we can write

$$\phi_{B0} = eV_{bi} + \phi_n \tag{7.11}$$

Using simple semiconductor theory and assuming N_c to represent the effective density of states at E_c, we have

$$\exp\left(-\frac{\phi_n}{kT}\right) = \frac{N_d}{N_c} = \frac{N_d}{2(2\pi m^* kT/h^2)^{3/2}} \tag{7.12}$$

which yields

$$N_d = 2\left(\frac{2\pi m^* kT}{h^2}\right)^{3/2} \exp\left(\frac{-\phi_n}{kT}\right) \tag{7.13}$$

Putting Eqn (7.13) for N_d into Eqn (7.10) results in

$$J_0 = q(2)\left(\frac{2\pi m^* kT}{h^2}\right)^{3/2} \left(\frac{kT}{2\pi m^*}\right)^{1/2} \exp\left[\frac{-(\phi_n + eV_{\text{bi}})}{kT}\right]$$

which can be simplified to

$$J_0 = \frac{4\pi e m^* k^2 T^2}{h^3} \exp\left[\frac{-(\phi_n + eV_{\text{bi}})}{kT}\right] \tag{7.14}$$

Using Eqn (7.11) in Eqn (7.14) gives us

$$J_0 = \frac{4\pi}{h^3} e m^* k^2 T^2 \exp\left(\frac{-\phi_{B0}}{kT}\right) \tag{7.15}$$

which can be expressed in the form

$$J_0 = A^* T^2 \exp\left(\frac{-\phi_{B0}}{kT}\right) \tag{7.16}$$

Equation (7.16) is similar to the equation for current density due to the emission of electrons from a metal into vacuum over a barrier ϕ_{B0}. Such an equation would however involve free electron mass and A^* would be replaced by the Richardson constant A. The parameter A^* is, therefore, called the effective Richardson constant for thermionic emission.

Under an applied forward bias V, the effective barrier in the semiconductor becomes $e(V_{\text{bi}} - V)$ as shown in Fig. 7.3(b). The electron flow from the semiconductor into the metal thus gets boosted by a factor exp (eV/kT). The current–voltage characteristics of a metal–semiconductor junction can, therefore, be written as

$$J = J_0\left[\exp\left(\frac{eV}{kT}\right) - 1\right] \tag{7.17}$$

At this stage, it must be remembered that the simple emission model presented in this chapter assumes that the electrons emitted from the metal into the semiconductor face no difficulty in reaching the bulk of the semiconductor. Actual metal–semiconductor junctions, however, involve sufficiently thick depletion regions. This ensures that the electron flow is controlled by the field and diffusion equations. The analysis for realizing the current–voltage characteristics then becomes very complicated.

7.1.4 Schottky Effect

Let us first consider a metal–vacuum interface. Electrons in the metal encounter a barrier which is dependent upon the field strength at the metal surface. Figure 7.4(b) shows a schematic representation of electron energy $\phi(x)$ in units of e V near the metal.

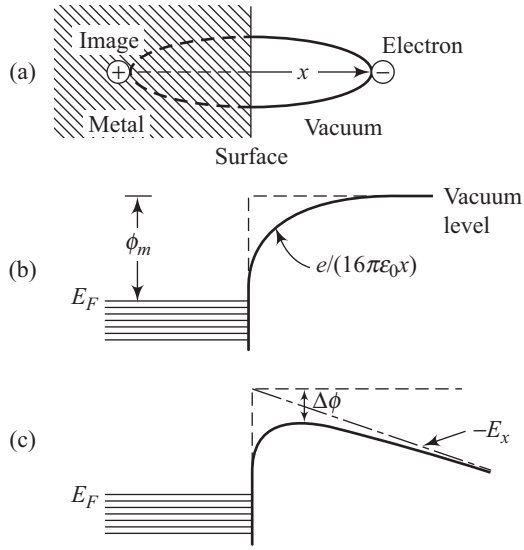


Fig. 7.4 Barrier at a metal–vacuum interface. (a) Electron in vacuum with image charge in the metal. (b) Electron energy barrier in the absence of applied field. (c) External applied field, E, reduces the barrier height by $\Delta\phi$.

An electron at a distance x from the metal surface experiences an attractive force towards the metal. Since the metal surface is a perfectly conducting sheet, the electric field lines must be perpendicular to the metal surface, at the interface, in accordance with Gauss's law. The field lines are, therefore, similar to the situation where the electron can be assumed to induce an image charge $+q$ at a distance $-x$ inside the metal surface [see Fig. 7.4(a)]. This method of analysing electrostatic problems is called *method of images*. The attractive force felt by the electron outside the metal is, therefore, given by

$$F = \frac{e^2}{4\pi\varepsilon_0 (2x)^2} \tag{7.18}$$

where ε_0 represents the dielectric constant of free space (8.85×10^{-14} F cm^{-1}). The electron energy $\phi(x)$ in units of e V can be obtained by integrating Eqn (7.18) from $x = \infty$ to x. The result is

$$\phi(x) = -\frac{e}{16\pi\varepsilon_0 x} \tag{7.19}$$

Another issue needs to be addressed at this stage. If we put $x = 0$, Eqn (7.19) leads to $\phi(x) = -\infty$ and not $-\phi_m$ as expected. However, if a small value like a few

angstroms is substituted in Eqn (7.19), the resultant value of $\phi(x)$ would be a few negative eV. For example, the magnitude of $\phi(4\ \text{Å})$ is still only –0.9 eV and Eqn (7.19) is generally used at distance of some hundreds of angstroms or more. Thus, although Eqn (7.19) does not yield the correct magnitude of potential at $x = 0$, it is still usable at finite distances from the metal surface.

Suppose an electric field E (V cm^{-1}) is applied near the metal–vacuum interface. The energy of an electron at a distance x from the metal surface then becomes

$$\phi(x) = -\left(\frac{e}{16\pi\varepsilon_0 x}\right) - Ex \tag{7.20}$$

Differentiating $\phi(x)$ with respect to x and equating it to zero results in $x = [e/(16\pi\varepsilon_0 E)]^{1/2}$, as the distance at which $\phi(x)$ is maximized. This maximum value of $\phi(x)$ can be evaluated by substituting $x = [e/(16\pi\varepsilon_0 E)]^{1/2}$ in Eqn (7.20), resulting in

$$\phi_{\max}(x) = -\left(\frac{eE}{4\pi\varepsilon_0}\right)^{1/2} \tag{7.21}$$

Thus, the barrier-lowering produced by the image force, $\Delta\phi$, in eV is given by

$$\Delta\phi = \left(\frac{eE}{4\pi\varepsilon_0}\right)^{1/2} \tag{7.22}$$

Having derived an expression for the Schottky barrier-lowering effect at a metal–vacuum interface, let us now focus again on the metal–semiconductor interface. The current density J_0 for a metal–semiconductor junction is therefore expected to follow the relation.

$$J_0 = A^* T^2 \exp\left[-\frac{(\phi_{B0} - \Delta\phi}{kT}\right] \tag{7.23}$$

with $\Delta\phi$ given by

$$\Delta\phi = \left(\frac{eE}{4\pi\varepsilon_s}\right)^{1/2} \tag{7.24}$$

The maximum electric field strength at the junction is given by

$$E_{\max} = \left[\frac{2eN_d(V_{\text{bi}} - V)}{\varepsilon_s}\right]^{1/2} \tag{7.25}$$

The electric field strength decreases linearly with distance from the junction. Since barrier-lowering effects are limited to small distances, Eqn (7.25) can be used as an approximate expression for electric field. Using Eqs (7.23)–(7.25), we can conclude that the reverse-bias current, I_R, for a metal–semiconductor junction varies as $\ln(I_R)$ proportional to $E^{1/2}$. Also

$$\ln(I_R) \propto (V_{\text{bi}} - V)^{1/4} \tag{7.26}$$

for a reverse-biased metal–semiconductor junction. Figure (7.5) shows typical reverse bias current–voltage characteristics of a Schottky barrier diode (metal–semiconductor junction). The increase in reverse bias current with increase in reverse-bias voltage is due to the barrier-lowering effect.

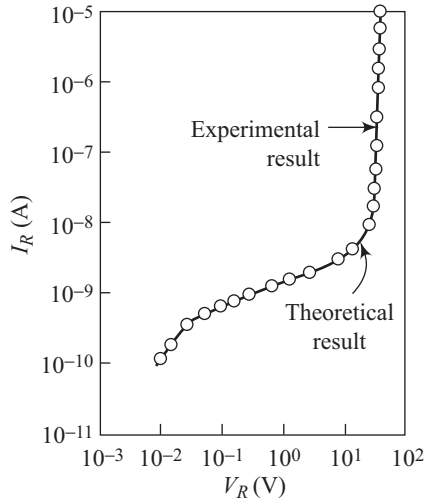


Fig. 7.5 Experimental and theoretical *I-V* characteristics for a Schottky diode

7.1.5 Tunnelling Current

From Eqn (7.6) it is clear that the space charge width, W, of a metal–semiconductor junction is inversely proportional to the square root of the semiconductor doping level. When the semiconductor doping level increases, the space charge width decreases. For a sufficiently heavily doped semiconductor, tunnelling current becomes the dominant carrier transport process. One must also remember that tunnelling is also likely to become the dominant transport mechanism at sufficiently low temperatures.

Tunnelling current J_t can be expressed as

$$J_t \propto \exp\left(-\frac{e\phi_{Bn}}{E_{00}}\right) \tag{7.27}$$

where the parameter E_{00} is given by

$$E_{00} = \frac{e\hbar}{2}\sqrt{\frac{N_d}{\varepsilon_s m^*}} \tag{7.28}$$

Thus, tunnelling current can be seen to increase exponentially with the semiconductor doping concentration.

The dominance of tunnelling current at high doping levels is of extremely high significance for semiconductor devices. Nearly all semiconductor devices require metal contacts and almost everywhere the need is to have as small a voltage drop across the contact region as possible. Putting metal contacts on heavily doped semiconductors then offers an important route to provide contacts for accessing the semiconductor device. The specific choice of the metal to be used also is more relaxed with such contacts.

7.2 Effect of Surface States and Interface

In our treatment of metal–semiconductor junctions, we have so far neglected the presence of any interfacial layer between the metal and the semiconductor. Another thing that we have neglected is the presence of surface states. These effects can substantially modify the characteristics of a metal–semiconductor junction. Surface states, as we know, are allowed electronic energy states within the energy band gap created due to the termination of the periodic potential at semiconductor surfaces. There are two types of surface states, namely *donor states* and *acceptor states*. Donor states are neutral if they contain an electron and become positively charged when no electron is present. On the other hand, acceptor states are neutral when they do not contain an electron, but become negatively charged when an electron occupies such a state.

Figure 7.6 shows the detailed band diagram of a metal–semiconductor (*n*-type) contact with an interfacial layer, δ, of the order of atomic distance.

In this figure, ϕ_{Bn} is the metal–semiconductor barrier height with image force lowering. Factors δ and Δ are the thickness of and the potential across the interfacial layer, respectively. Q_{sc} represents the space charge density in the semiconductor, whereas Q_{ss} and Q_M are surface-state density on the semiconductor and surface-charge density on the metal, respectively. The energy level at the surface is $e\phi_0$, the energy level coincided with the Fermi level before the formation of the metal–semiconductor contact. For charge neutrality on the surface, all states below $e\phi_0$ should have been filled.

Let us assume that the semiconductor has acceptor surface states with a density D_s states per cm^2 per eV. Let us also assume that the density D_s is constant over the energy range $e\phi_0$ to E_F. Without going into the details, we will simply write a

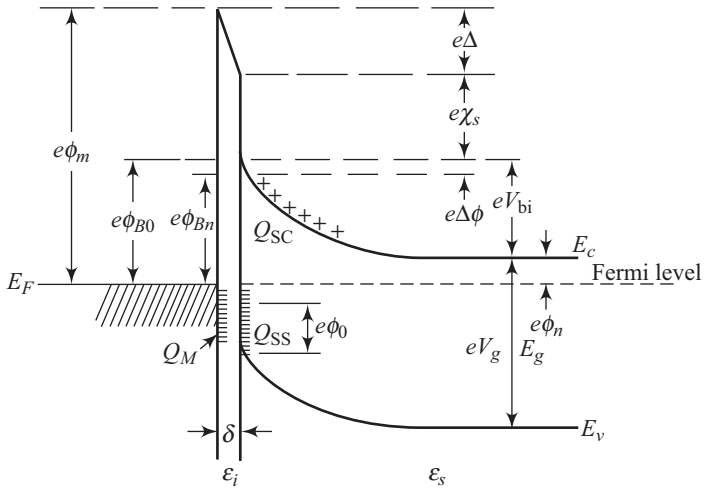


Fig. 7.6 Energy band diagram for metal–(*n*-type) semiconductor junction with interfacial layer

final expression relating the various terms and then look closely at two limiting cases. This final expression is

$$E_g - e\phi_0 - e\phi_{Bn} = \frac{1}{eD_s}\sqrt{2e\varepsilon_s N_d(\phi_{Bn} - \phi_n)} - \frac{\varepsilon_i}{eD_s\delta}\left[\phi_m - (\chi_s + \phi_{Bn})\right] \tag{7.29}$$

The derviation of Eqn (7.29) assumes that the interfacial layer has a thickness of a few angstroms and is essentially transparent to electrons.

We will now consider two interesting limiting cases.

Case 1 $D_s \to \infty$

Equation (7.29) leads to

$$E_g - e\phi_0 - e\phi_{Bn} = 0$$

which implies

$$\phi_{Bn} = \frac{1}{e}(E_g - e\phi_0) \tag{7.30}$$

Equation (7.30) can be used to conclude the following:

(i) The barrier height is independent of material parameters such as the work function and semiconductor electron affinity.
(ii) Barrier height is dependent upon the semiconductor band gap and surface energy $e\phi_0$.
(iii) Fermi level at the interface is pinned at the value $e\phi_0$ above the valence band, by the surface states.

Case 2 $D_s\delta \to 0$

Equation (7.29) reduces to

$$\phi_{Bn} = \phi_m - \chi_s \tag{7.31}$$

which is identical to Eqn (7.1) as expected.

7.3 Metal–Semiconductor Ohmic Contacts

Ohmic contacts refer to low-resistance, bias-independent metal–semiconductor contacts. The tunnelling dominated metal–semiconductor contacts discussed earlier in this chapter are one type of ohmic contacts. Such ohmic contacts are quite common and very useful.

Figures 7.1(d) and 7.2(d) show the band diagrams when the metal–semiconductor contact behaves like an ohmic contact even if the semiconductor is not heavily doped. While Fig. 7.1(d) shows a contact between a metal and an *n*-type semiconductor, Fig. 7.2(d) is for a metal and a *p*-type semiconductor. Thus for $\phi_m < \phi_s$, a metal–semiconductor contact behaves like an ohmic contact. From Fig. 7.1(d), it is clear that when a positive voltage is applied to the metal, the electrons in the semiconductor encounter no barrier in flowing from the

semiconductor to the metal. On the other hand, if a positive voltage is applied to the semiconductor, the effective barrier for electrons flowing from the metal into the semiconductor is approximately $\phi_{Bn} = \phi_n$. One would appreciate that ϕ_n has a fairly small magnitude for moderate to heavily doped semiconductors. Thus, *electrons can conveniently flow between the metal and the semiconductor for* $\phi_m < \phi_s$*, for a metal–(n-type) semiconductor junction*. A similar situation prevails for a metal–(*p*-type) semiconductor junction, for $\phi_m > \phi_s$. One must, however, remember that Figs 7.1 and 7.2 were drawn assuming the total absence of any surface states and are, therefore, too ideal to be true. In practice, ohmic contacts are obtained by fabricating Schottky contacts on heavily doped n- and p-type semiconductors. The width of the depletion region formed in a Schottky contact is dependent upon the doping concentration on the semiconductor side as expressed by Eqn. (7.6). For doping concentrations exceeding 10^{19} cm^{-3}, the depletion width becomes several atomic layers thick and direct tunnelling of carriers through such thin layers become possible. This tunnelling dominant current is present in practical ohmic constants.

7.3.1 Specific Contact Resistance

Specific contact resistance, ρ_c, is usually taken as a figure of merit for the quality of ohmic contacts. The quantity is defined as the reciprocal of the derivative of current density with respect to voltage, evaluated at zero bias. Mathematically, this can be written in the form

$$\rho_c = \left(\frac{\partial J}{\partial V} \right)^{-1} \Bigg|_{V=0} \tag{7.32}$$

and has units of Ω cm^2. The objective of a device engineer is to realize as small a value of ρ_c as possible (< 10^{-6} Ω cm^2 is considered good).

We have already seen that the current–voltage characteristics for the metal–semiconductor junction, according to the emission model (conduction is due to thermionic emission), are given by

$$J_n = A^* T^2 \exp\left(-\frac{e\phi_{Bn}}{kT} \right) \left[\exp\left(\frac{eV}{kT} \right) - 1 \right] \tag{7.33}$$

Using Eqs (7.32) and (7.33), we get

$$\rho_c = \frac{(kT/e) \exp[e\phi_{Bn}/(kT)]}{A^* T^2} \tag{7.34}$$

Thus, the specific contact resistance decreases very fast with decreasing barrier height. Tunnelling current, on the other hand, is governed by Eqs (7.27) and (7.28). For the specific contact resistance ρ_c, we can now write

$$\rho_c \propto \exp\left[\frac{2\sqrt{\varepsilon_s m_n^*}}{\hbar} \frac{\phi_{Bn}}{\sqrt{N_d}} \right] \tag{7.35}$$

Thus, for a tunnelling contact, the specific contact resistance is a strong function of the semiconductor doping level.

- The spectral response measurement of metal–semiconductor junctions can be used to determine barrier heights. This technique uses the fact that the photon energy exceeding the barrier height but less than semiconductor band gap, can lead to photoemission of electrons from the metal into the semiconductor.
- Another method of determining the barrier height depends upon the reverse bias variation of space charge capacitance.
- It is always more difficult to obtain good ohmic contacts on wide-band gap materials. Usually tunnelling contacts are most feasible for such semiconductors.

7.4 Heterojunctions

Metal–semiconductor junction discussed in this chapter is a deviation from the simple homojunction we learnt earlier in this book. We will now make an attempt to understand another type of junction called heterojunction. A heterojunction is formed between two different semiconductors. Heterojunctions find widespread applications in devices such as photodiodes, phototransistors, lasers, and so on.

7.4.1 Energy Band Diagram

Let us begin by considering the energy band diagram for the GaAs–Ge heterojunction system. We will use this system to understand certain basic procedures to obtain the energy band diagram of heterojunctions. One of the first things to notice in the GaAs–Ge system is the nearly identical lattice constants at 300 K. In fact, for GaAs and Ge, the lattice constants are identical to within 0.08 per cent (5.654 Å for GaAs and 5.658 Å for Ge). The coefficients of linear expansion with temperature are also very similar for these materials.

The procedure followed for drawing the energy band diagram is shown in Fig. 7.7. Figure 7.7(a) shows the energy diagrams for the two materials separately. The diagram has been drawn by assuming a constant vacuum level of energy. These band diagrams are relevant for the individual materials before they are brought into contact with each other.

From Fig. 7.7(a), it is clear that the Fermi energy level E_{Fp} is lower in electron energy than E_{Fn}. The two Fermi levels must come together when the materials are brought into contact with each other. To ensure this, a small number of electrons pass from GaAs to Ge. This movement of electrons at the interface between the two semiconductors bends up the E_c energy level in GaAs as shown in Fig. 7.7 (b). The band bending is represented by the symbols V_{bin} and V_{bip}. The band bending must accommodate the original difference in the Fermi energy levels $E_{Fp} - E_{Fn}$.

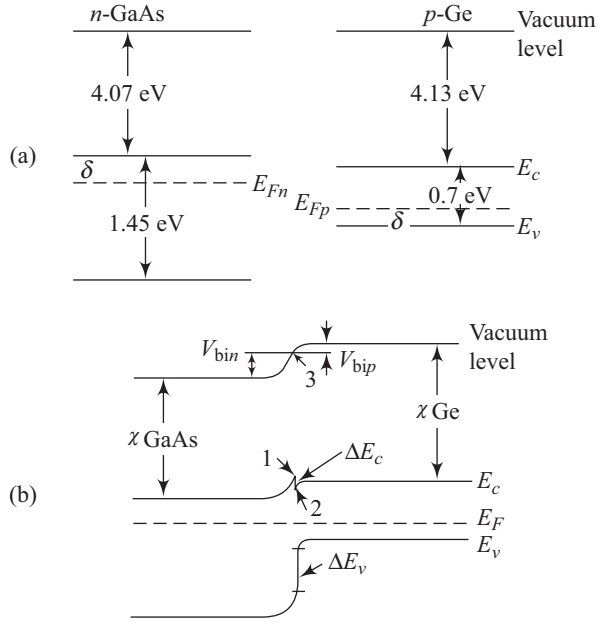


Fig. 7.7 Energy band diagram for *n-p* GaAs–Ge

Thus, we have

$$E_{Fp} - E_{Fn} = (\chi_{Ge} + E_{g(Ge)} - \delta_{Ge}) - (\chi_{GaAs} + \delta_{GaAs}) = V_{bin} + V_{bip} \tag{7.36}$$

Let us assume that the transition regions are completely depleted over distance x_n and x_p such that

$$\frac{x_n}{x_p} = \frac{N_a}{N_d} \tag{7.37}$$

The application of Poisson's equation then leads to the relations

$$V_{bin} = \frac{N_d x_n^2}{2\varepsilon_{GaAs}} \tag{7.38}$$

and

$$V_{bip} = \frac{N_a x_p^2}{2\varepsilon_{Ge}} \tag{7.39}$$

Combining Eqs (7.38) and (7.39) yields

$$\frac{V_{bin}}{V_{bip}} = \frac{N_a \varepsilon_{Ge}}{N_d \varepsilon_{GaAs}} \tag{7.40}$$

The conduction band energy step ΔE_c can be seen from Fig. 7.7(b) to be

$$\Delta E_c = \delta_{GaAs} + V_{bin} - (E_{g(Ge)} - \delta_{Ge}) + V_{bip} \tag{7.41}$$

Using Eqn (7.36) in Eqn (7.41) leads to

$$\Delta E_c = \chi_{Ge} - \chi_{GaAs} \tag{7.42}$$

The valence band energy step ΔE_v using geometrical considerations in Fig. 7.7(b) can be seen to be given by

$$\Delta E_v = [E_{g(\text{GaAs})} - E_{g(\text{Ge})}] - (\chi_{\text{Ge}} - \chi_{\text{GaAs}}) \tag{7.43}$$

Combining Eqs (7.42) and (7.43) results in

$$\Delta E_c + \Delta E_v = E_{g(\text{GaAs})} - E_{g(\text{Ge})} \tag{7.44}$$

The ideal situation represented by Eqn (7.44) is called the *electron affinity rule*. It serves as a good starting point for any study of heterojunctions.

The simple conclusions drawn with respect to ΔE_c, ΔE_v, and their sum are important as first-order approximations for heterojunctions of all doping levels. Figure 7.8 presents the band diagram for a (*p*-GaAs)–(*n*-Ge) heterojunction.

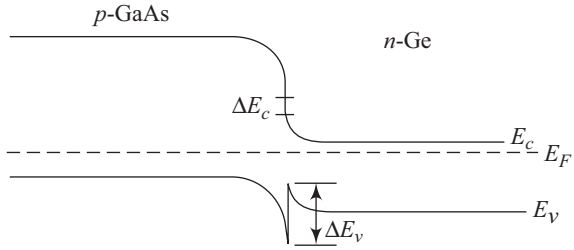


Fig. 7.8 Energy band diagram for the *p-n* GaAs–Ge heterojunction

The simple model of the heterojunction presented in this section assumes no charge to be present at the interface between the two semiconductors. Such charges would exist, for example, in situations where energy states are present at the interface. Such interface states would arise in heterojunctions between materials having lattice constants differing by more than around 1 per cent. If the materials have a large difference in their coefficients of expansion, a strain disorder can develop at the interface when the junction cools down from its growth temperature.

Heterojunctions in which the dopant type changes between the two sides of a junction are called *anisotype*, whereas heterojunctions where there is no change in the dopant type as one crosses the junction are called *isotype* heterojunctions.

7.4.2 Two-dimensional Electron Gas

We will now try to understand an interesting effect demonstrated by isotype heterojunctions. The thermal-equilibrium energy band diagram of a GaAs–AlGaAs (*n-n*) heterojunction is shown in Fig. 7.9.

To differentiate between the two semiconductors, the symbol n_1 has been used for smaller band gap material GaAs, whereas the symbol n_2 has been used for the wider band gap material AlGaAs. To explain the special properties of this type of heterojunction, we will assume that GaAs is lightly doped, whereas AlGaAs is moderately to heavily doped. To maintain equilibrium, electrons flow from the wide-band gap material into the low-band gap GaAs. This leads to the formation

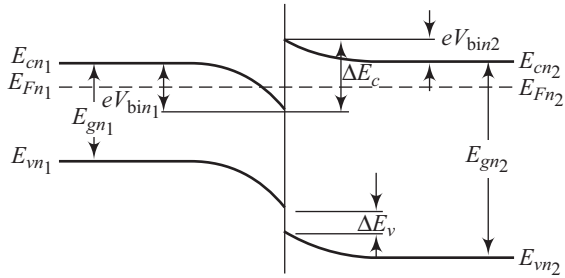


Fig. 7.9 Energy band diagram of an *n*-*n* heterojunction in thermal equilibrium

of an accumulation layer of electrons in the potential well around the interface. Electrons contained in this potential well follow the basic quantum mechanical principle that their energies are quantized. Thus, electron energy levels are *quantized* in a direction perpendicular to the interface, whereas the electrons are free to move in the other two spatial directions. Such a behaviour of electrons is referred to as *two-dimensional electron gas.*

The electrons comprising the two-dimensional electron gas lie in the region of low impurity doping, thus minimizing impurity-scattering effects. Electron mobility values realizable in this region are much larger than the values for electrons present along with ionized donors. The electrons in the potential well can be further isolated from the ionized impurities by sandwiching an intrinsic layer of graded Al_x $Ga_{1-x}As$ (the mole fraction x varies with distance) between *n*-type AlGaAs and intrinsic GaAs. Figure 7.10 shows the conduction band edges in thermal-equilibrium conditions for such a heterojunction.

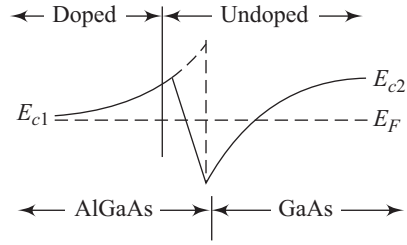


Fig. 7.10 Conduction band edge for graded heterojunction

7.4.3 Quantum Confinement of Carriers

Heterojunctions can be used to construct quantum well structures. In its simplest form a quantum well structure consists of a thin layer of narrow gap material such as GaAs sandwiched between two thick layers of wider band gap material like GaAlAs. The thin narrow band gap material then forms a quantum well in one dimension. In the growth direction (z-direction), the electron and hole levels are bound states of the well and are known as *subbands*. There are three sets of subbands, namely electron subbands, light-hole subbands, and heavy-hole subbands.

Each subband has its own set of quantum numbers. Motion in the x-y plane is however unrestricted. Thus we have a two-dimensional carrier system in the well. The structures have been made possible due to the development of advanced

growth techniques like Molecular Beam Epitaxy (MBE) and Metal-Organic Chemical Vapour Deposition (MOCVD). These techniques allow semiconductor scientists to grow extremely thin (~ 10 Å) with excellent control of composition. These structures have found applications in areas like lasers and LEDs.

- A multilayered heterojunction arrangement with the typical layer thickness of around 100 Å results in a superlattice structure. Alternate layers of GaAs and $Al_xGa_{1-x}As$ for example could result in such a superlattice. Such structures require the use of sophisticated processes such as molecular beam epitaxy.
- Heterojunctions are also used in photonic devices such as semiconductor lasers, photodetectors, and photovoltaic cells.

Solved Problems

7.1 A metal–semiconductor contact is formed between gold and *n*-type silicon doped to a level $N_d = 5 \times 10^{16}$ cm^{-3} at 300 K. Calculate the ideal Schottky barrier height, built-in potential barrier, space charge width at zero bias, and maximum electric field for zero bias. Assume the work function of gold to be 5.1 eV, $\chi_s = 4.01$ eV, and the effective density of states function $N_c = 2.8 \times 10^{19}$ cm^{-3}.

Solution

The ideal Schottky barrier height, ϕ_{B0}, is given by

$$\phi_{B0} = \phi_m - \chi_s \tag{7.1.1}$$

Putting the given values in Eqn (7.1.1) results in

$$\phi_{B0} = 5.1 - 4.01 = 1.09 \text{ eV}$$

Also,

$$\phi_n = \frac{kT}{e} \ln\left(\frac{N_c}{N_d}\right)$$

Putting the given values in this equation results in

$$\phi_n = 0.026 \ln\left(\frac{2.8 \times 10^{19}}{5 \times 10^{16}}\right) = 0.165 \text{ V}$$

In units of eV, $\phi_n = 0.165$ eV

Built-in potential barrier, V_{bi}, is given by

$$V_{\text{bi}} = \frac{\phi_{B0} - \phi_n}{e} \tag{7.1.2}$$

where both ϕ_{B0} and ϕ_n have to be substituted in units of eV. Thus, we get

$$V_{\text{bi}} = \frac{1.09 - 0.165}{1} = 0.925 \text{ V} \tag{7.1.3}$$

Space charge width x_n at zero bias is expressed in the form

$$x_n = \left[\frac{2\varepsilon_s V_{\text{bi}}}{eN_d}\right]^{1/2} \tag{7.1.4}$$

Substituting the calculated value of V_{bi} from Eqn (7.1.3) and other known values in Eqn (7.1.4) leads to

$$x_n = \left[\frac{2 \times 11.7 \times (8.85 \times 10^{-14)}\,(0.925)}{(1.6 \times 10^{-19})\,(5 \times 10^{16})}\right]^{1/2}$$

$$= 0.155 \times 10^{-4} \text{ cm}$$

The maximum electric field $|E_{\max}|$ at zero bias is given by

$$|E_{\max}| = \frac{eN_d x_n}{\varepsilon_s}$$

$$= \frac{(1.6 \times 10^{-19})\,(5 \times 10^{16})\,(0.155 \times 10^{-4})}{(11.7)\,(8.85 \times 10^{-14})}$$

$$= 11.98 \times 10^4 \text{ V cm}^{-1}$$

7.2 A plot of $(1/C_J)^2$ versus reverse bias, V, for a particular metal–silicon (n-type) contact is a straight line. The intercept of the line is at a magnitude of 0.45 V and the slope of the curve is $\sim 6 \times 10^{13}$. Evaluate the actual barrier height.

Solution

The slope of the $(1/C_J)^2$ versus V curve, is given by $2/(e\varepsilon_s N_d)$. Thus

$$N_d = \frac{2}{e\varepsilon_s\,(\text{slope})} \tag{7.2.1}$$

$$= \frac{2}{(1.6 \times 10^{-19})\,(11.7)\,(8.85 \times 10^{-14}) \times 6 \times 10^{13}}$$

$$= 2.01 \times 10^{17} \text{ cm}^{-3}$$

Also

$$\phi_n = \frac{kT}{e} \ln\left(\frac{N_c}{N_d}\right) \tag{7.2.2}$$

$$= (0.026) \ln\left(\frac{2.8 \times 10^{19}}{2.01 \times 10^{17}}\right)$$

$$= 0.128 \text{ V} \tag{7.2.3}$$

With ϕ_n and ϕ_{Bn} expressed in V, the barrier height ϕ_{Bn} is given by

$$\phi_{Bn} = V_{\text{bi}} + \phi_n \tag{7.2.4}$$

where $V_{\text{bi}} = 0.45$ = the magnitude of the intercept on the voltage axis.
Thus, using Eqs (7.2.3) and (7.2.4), we get

$$\phi_{Bn} = 0.45 + 0.128 = 0.578 \text{ V}$$

7.3 Calculate the Schottky barrier-lowering for an Si–metal contact, assuming an electric field $E = 10^4$ V cm $^{-1}$. Also calculate the position of the maximum barrier height.

Solution

The Schottky barrier-lowering, $\Delta\phi$, is given by

$$\Delta\phi = \sqrt{\frac{eE}{4\pi\varepsilon_s}} \tag{7.3.1}$$

Putting the given values in Eqn (7.3.1) leads to

$$\Delta\phi = \sqrt{\frac{(1.6\times10^{-19})(10^4)}{(4\pi)\,(11.7)\,(8.85\times10^{-14})}}$$
$$= 0.011 \text{ V}$$

The position x_m, where the barrier height is maximized, is given by

$$x_m = \sqrt{\frac{e}{16\pi\varepsilon_s E}} \tag{7.3.2}$$
$$= \sqrt{\frac{1.6\times10^{-19}}{16\times\pi\times(11.7)\,(8.85\times10^{-14})\,(10^4)}}$$
$$= 5.54\times10^{-7} \text{ cm} = 55.4 \text{ Å}$$

7.4 Calculate the reverse-saturation current density for a Au–(*n*-type) Si Schottky diode. Assume $e\phi_{Bn}$ for the barrier to be 0.82 eV and the effective Richardson constant equal to 114 A/K^{2}cm^2. Take $T = 300$ K. Neglect the Schottky barrier-lowering effect.

Solution

Reverse-saturation current density, J_0, is given by

$$J_0 = A^* T^2 \exp\left(-\frac{e\phi_{Bn}}{kT}\right) \tag{7.4.1}$$

Putting the given values in Eqn (7.4.1), we get

$$J_0 = (114)(300)^2 \exp\left(-\frac{0.82}{0.026}\right)$$
$$= 2.06\times10^{-7} \text{ A/cm}^2$$

7.5 An *n-p* heterojunction is formed between GaAs and Ge. Calculate conduction band and valence band steps ΔE_c and ΔE_v, and thereby check the validity of the electron affinity rule. Assume $E_{g(\text{GaAs})} = 1.45$ eV, $E_{g(\text{Ge})} = 0.7$ eV, $\chi_{\text{GaAs}} = 4.07$ eV and $\chi_{\text{Ge}} = 4.13$ eV. Assume the absence of interface states.

Solution

The conduction band step, ΔE_c, is given by

$$\Delta E_c = \chi_{\text{Ge}} - \chi_{\text{GaAs}} \tag{7.5.1}$$

Putting the given values in Eqn (7.5.1) results in

$$\Delta E_C = 4.13 - 4.07 = 0.06 \text{ eV} \tag{7.5.2}$$

The valence band step, ΔE_V, can be evaluated using the expression

$$\Delta E_v = [E_{g(\text{GaAs})} - E_{g(\text{Ge})}] - (\chi_{\text{Ge}} - \chi_{\text{GaAs}})$$

leading to

$$\Delta E_v = (1.45 - 0.7) - 0.06 = 0.69 \text{ eV} \tag{7.5.3}$$

Using Eqs (7.5.2) and (7.5.3) results in

$$\Delta E_c + \Delta E_v = 0.69 + 0.06 = 0.75 = E_{g(\text{GaAs})} - E_{g(\text{Ge})}$$

7.6 Consider an ideal chromium to *n*-type silicon Schottky diode at $T = 300$ K. Assume the semiconductor is doped at a concentration of $N_d = 3 \times 10^{15}$ cm^{-3}. Determine the following.

(a) Ideal Schottky barrier height
(b) Peak electric field with an applied reverse bias voltage of 5 V
(c) Depletion layer capacitance per unit area for a reverse bias voltage of 5 V

(Given: work function for chromium = 4.5 e V, electron affinity for Si = 4.01 e V, $\varepsilon_0 = 8.854 \times 10^{-14}$ F/cm, $\varepsilon_{\text{Si}} = 11.7$)

Solution

(a) Schottky barrier height, $\phi_B = \phi_m - \chi$ as per Eqn (7.31) in the text

$\phi_m = 4.5$ e V (given)

$\chi_{\text{Si}} = 4.01$ e V (given)

$\therefore \quad \phi_B = 4.5 - 4.01 = 0.49$ e V

(b) From Eqn (7.3) in the text, we find

$$|E_{\max}| = \frac{eN_d x_n}{\varepsilon_s}$$

where

$$x_n = \left[\frac{2\varepsilon_s (V_{\text{bi}} + V_R)}{eN_d}\right]^{1/2}$$

Assuming the built-in potential, V_{bi}, to be very small as compared to the applied reverse bias of 5 V, and hence ignoring it, we get

$$x_n = \left[\frac{2(11.7)(8.854 \times 10^{-14})(5)}{(1.6 \times 10^{-19})(3 \times 10^{15})}\right]^{1/2}$$

$$= 1.469 \times 10^{-4} \text{ cm}$$

$\therefore \quad E_{\max}$ can be found as

$$= \frac{1.6 \times 10^{-19} \times 3 \times 10^{15} \times 1.51 \times 10^{-4}}{11.7 \times 8.854 \times 10^{-14}}$$

$$\approx 6.81 \times 10^4 \text{ V/cm}$$

(c) From Eqn (7.7) in the text we get

$$C_J = \left[\frac{e\varepsilon_s N_d}{2(V_{\text{bi}} + V)}\right]^{1/2}$$

$$= \left[\frac{1.6 \times 10^{-19} \times 11.7 \times 8.854 \times 10^{-14} \times 3 \times 10^{15}}{2 \times 5}\right]^{1/2}$$

$$= 6.88 \times 10^{-9} \text{ F/cm}^2$$

7.7 Figure 7.7.1 depicts the energy band diagram for a typical metal–*n*-type semiconductor contact. Take a careful look at the diagram and answer the following questions.

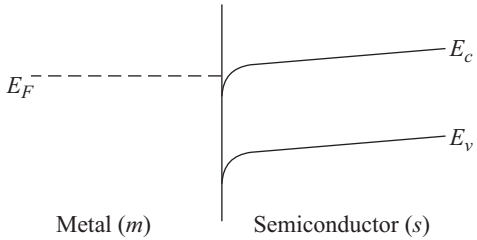


Fig. 7.7.1 Energy band diagram of a metal–*n*-type semiconductor junction

(a) What kind of contact exists between the metal and the semiconductor?

(b) Which is greater φ_m or φ_s, where φ has its usual meaning?

(c) Has any bias been applied to the structure? If no, then justify your answer. If yes, then write all applicable possibilities with justifications, which might be true for the above-mentioned structure out of the four options given below:

(i) Positive bias to the metal
(ii) Negative bias to the metal
(iii) Positive bias to the semiconductor
(iv) Negative bias to the semiconductor

Solution

(a) As there is almost free flow of electrons from the semiconductor to the metal (as is obvious from the energy band diagram), hence this is an ohmic contact.

(b) ϕ stands for the workfunction. Semiconductor workfunction (ϕ_s) is greater as tilting of the valence and conduction bands has been shown to be towards the Fermi level. It means that for equilibrium to be established a positive charge is being developed on the metal and a negative charge on the semiconductor. This can happen only when the semiconductor workfunction (ϕ_s) is greater.

(c) Yes, bias has been applied as a tilting is shown on the semiconductor side.
Options (i) and (iv) are valid for possible application of bias.
This is due to the fact that, if one applies positive bias to the metal, it will result in an electric field towards the semiconductor. This is because electrons move in the direction opposite to that of the electric field. Hence, there is a downhill movement of electrons. This justifies the tilting of energy bands on the semiconductor side.

Similarly, the tilting can also be explained for option (iv), since a negative bias on the semiconductor would produce the same effect as option (i).

7.8 Which of the metals listed below will you choose so as to make an ohmic contact to *n*-type InAs? (Tick all those applicable). Justify your choice of answer with a brief explanation.

(a) Gold (b) Silver (c) Copper (d) Aluminium

Solution

We can choose any of the above metals so as to form an ohmic contact with *n*-type InAs. This is essentially due to the phenomenon of Fermi level pinning and for *n*-type InAs, Fermi level (E_F) at the interface is pinned above the conduction band edge as shown in Fig. 7.8.1.

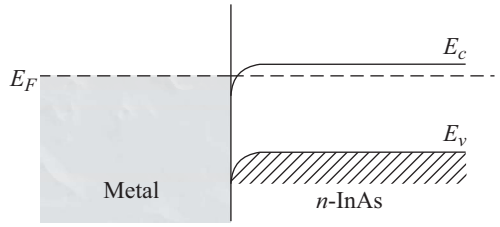


Fig. 7.8.1 Fermi level pinning

7.9 A contact is formed between a metal having a workfunction of 4.3 eV and an Si substrate doped with 10^{17} indium atoms (electron affinity = 4 eV).

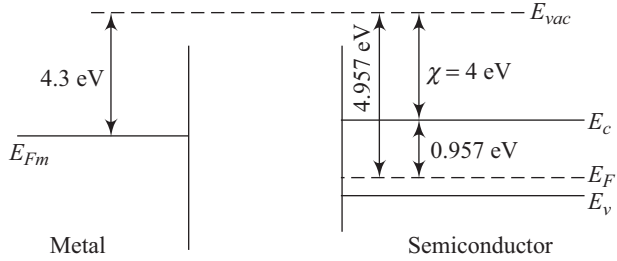
(a) Draw an equilibrium band diagram, showing a numerical value for eV_0.
(b) Draw the band diagram with a 0.3 V forward bias. Repeat for a reverse bias of 2 V.

Solution

$\phi_m = 4.3$ eV [Given]
$\chi_{Si} = 4$ eV
$p_0 = N_a = 10^{17}$ cm^{-3}
$\therefore$ from Eqn (3.43) in the text

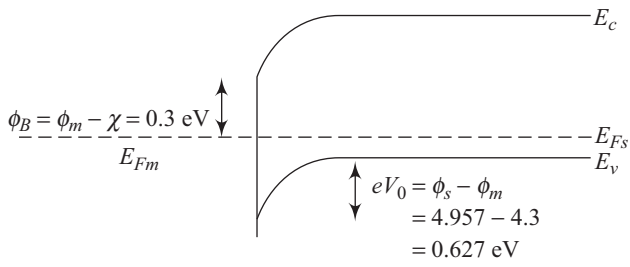
$$E_i - E_F = kT \ln \frac{p_0}{n_i} = 0.0259 \ln \frac{10^{17}}{1.5 \times 10^{10}} = 0.407 \text{ e V}$$

(a) Before contact
$\phi_s = 4 + 0.55 + 0.407 = 4.957$ eV



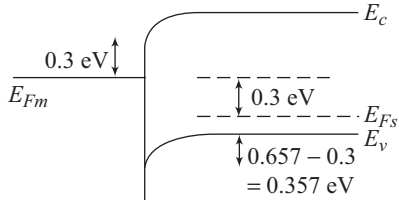
After contact

(b) Applied voltage =

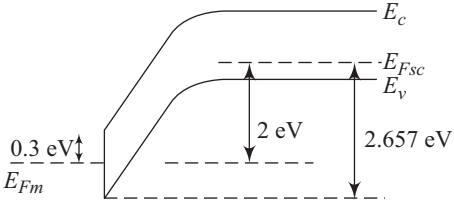


Forward bias = 0.3 V

(c) Applied voltage =



Reverse bias = 2 V



7.10 A silicon sample with the Fermi level 0.0259 e V above the intrinsic level is shown in Fig. 7.10.1. A metallic contact of gold is made on the *A* side. Consider that the sample is operated at 300 K.

Fig. 7.10.1 Silicon sample

(a) Sketch and label the energy band diagram for the metal–semiconductor contact developed across *A*.

(b) Now, this sample is doped with indium impurities on the *B* side as shown in the shaded region in the figure with a concentration of 9.9×10^{14} cm^{-3} and then the same metal contact is made on the *B* side as well. Sketch and label the energy band diagram for the metal–semiconductor contact developed across *B*.

(c) Comment on the behaviour of the metal–semiconductor contacts on both sides, *A* and *B*.

(d) Sketch qualitatively, the band diagram for part (a), if the semiconductor sample temperature is increased to 400 K. How is the energy band diagram different from the one drawn for part (a)? Explain briefly.

Solution

Using Eqn (3.42) in the text

$$n_0 = n_i\, e^{(E_F - E_i)/kT}$$

We can find the electron doping concentration in the given silicon sample as

$$n_0 = 1.5 \times 10^{10}\, e^{0.0259/0.259} = 3.3 \times 10^{14}\ \text{cm}^{-3}$$

We know that the electron affinity for silicon (χ) = 4.05 e V

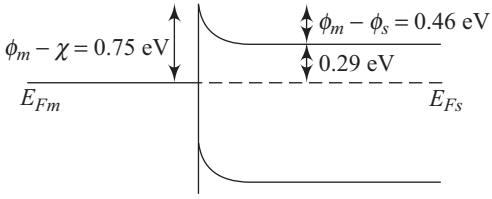
$\therefore$ ϕ_s (workfunction of the semiconductor) = $\chi + (E_c - E_F)$

$= 4.05 + 0.29$

$= 4.34$ e V

Workfunction of gold = 4.8 e V

Therefore, we can draw the energy band diagram for the metal–semiconductor contact across *A* as



(b) On the *B* side, *p*-type impurities in the form of indium atoms are added, which gives us a case of compensated semiconductor.

Hence, net concentration of *p*-type on the *B* side = $(9.9 \times 10^{14}) - (3.3 \times 10^{14})$

$= 6.6 \times 10^{14}\ \text{cm}^{-3}$

$\therefore$ $E_i - E_F$ on the *B* side $= kT \ln \dfrac{6.6 \times 10^{14}}{1.5 \times 10^{10}}$

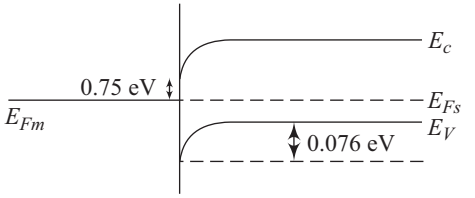
$= 0.276$ e V

$\therefore$ $\phi_s = 4.05 + 0.55 + 0.276$

$= 4.876$ e V

Now, on the *B* side, $\phi_m < \phi_s$ for *p*-type.

Hence, the resulting energy band diagram will be:



(c) On both sides, *A* and *B*, while making a metal contact there exists a barrier for flow of carriers from the semiconductor to the metal. Hence, for both cases, a rectifying contact is developed.

(d) If the operating temperature is increased to 400 K, n_i will increase to $\approx 8 \times 10^{12}\ \text{cm}^{-3}$ and hence ϕ_s will increase even further.

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

$$\Rightarrow E_F - E_i = kT \ln \frac{n_0}{n_i} = (kT)_{400\,\text{K}} \left[\ln \frac{3.3 \times 10^{14}}{8 \times 10^{12}} \right]$$

$$= 0.126 \text{ eV}$$

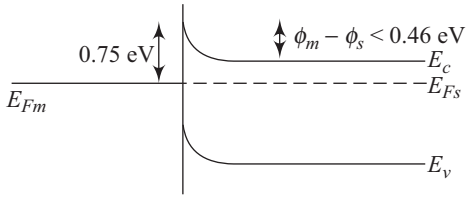
$\therefore$ New value of $\phi_s = 4.05 + (0.55 - 0.126)$

$$= 4.474$$

Still $\phi_m > \phi_s$

Hence, again a rectifying contact is made, but this time with a lower barrier $(\phi_m - \phi_s)$ for the carrier flowing from the semiconductor to the metal.

The resulting energy band diagram is shown below:



Recapitulation

- A metal–(*n*-type) semiconductor is rectifying for $\phi_m > \phi_s$ and ohmic for $\phi_m < \phi_s$.

  Space charge width, W, for a metal–(*n*-type) semiconductor junction is given by

  $$W = \left[\frac{2\varepsilon_s (V_{\text{bi}} + V)}{eN_d} \right]^{1/2}$$

  and the junction capacitance, C_J, is expressed as

  $$C_J = \left[\frac{e\varepsilon_s N_d}{2(V_{\text{bi}} + V)} \right]^{1/2}$$

- The reverse-saturation current, J_0, for a metal–semiconductor contact is given by

  $$J_0 = A^* T^2 \exp\left(-\frac{\phi_{B0}}{kT} \right)$$

  and the *I*-*V* characteristic is expressed in the form

  $$J = J_0 \left[\exp\left(\frac{eV}{kT} \right) - 1 \right]$$

- Schottky barrier-lowering $\Delta\phi$ can be mathematically written as

  $$\Delta\phi = \left(\frac{eE}{4\pi\varepsilon_s} \right)^{1/2}$$

- Tunnelling current density J_t is proportional to exp $(-e\phi_{Bn}/E_{00})$ with

  $$E_{00} = (e\hbar/2)\sqrt{N_d/(\varepsilon_s m^*)}$$

  where m^* is the effective mass.
- When surface state density $D_s \to \infty$

  $$\phi_{Bn} = \frac{1}{e}(E_g - q\phi_0)$$

- Specific contact resistance, ρ_c, is given by

  $$\rho_c = \left(\frac{\partial J}{\partial V}\right)^{-1}\Bigg|_{V=0}$$

  and has units of Ω cm^2.
- For a *p*-*n* heterojunction between *p*-Ge and *n*-GaAs

  $$\frac{V_{bin}}{V_{bip}} = \frac{N_a \varepsilon_{Ge}}{N_d \varepsilon_{GaAs}}$$

  $$\Delta E_c = \chi_{Ge} - \chi_{GaAs}$$

  and

  $$\Delta E_c + \Delta E_v = E_{g(GaAs)} - E_{g(Ge)}$$

- Isotype heterojunctions demonstrate the phenomenon of two-dimensional electron gas, wherein the electron energy levels are quantized in a direction perpendicular to the interface.
- Practical Ohmic contacts on heavily doped ν and π-type semiconductors.
- Heterojunctions can be used to construct quantum well structure.

Exercises

Review Questions

7.1 Draw the energy band diagram for the metal–(*p*-type) semiconductor, assuming no surface states exist.
7.2 Under what conditions does a metal–(*n*-type) semiconductor have non-rectifying properties? Sketch the relevant band diagram.
7.3 Derive an expression for the junction capacitance of a metal–semiconductor junction.
7.4 Derive the current–voltage characteristics for a Schottky contact using the emission model.
7.5 What is the Schottky barrier-lowering effect?
7.6 When is tunnelling the dominant conduction mechanism in a metal–semiconductor junction?
7.7 How do surface states affect the metal–semiconductor barrier?
7.8 Define specific contact resistance and give its SI units.
7.9 What are isotype and anisotype heterojunctions?
7.10 What is the importance of lattice constants in the fabrication of heterojunctions?

7.11 Draw the energy band diagram of an *n-p* GaAs–Ge heterojunction, assuming no interface states.
7.12 Demonstrate the validity of the electron affinity rule for the GaAs–Ge heterojunction.
7.13 Explain the concept of two-dimensional electron gas.
7.14 How are practical Ohmic contacts formed?
7.15 What are the three set of subbands?
7.16 Explain the terms, MBE and MOCVD.
7.17 How is a quantum well structure formed using GaAs and GaAl As?
7.18 Draw energy band diagrams for the following:
(a) metal–(*n*-type) semiconductor junction
(b) metal–(*p*-type) semiconductor junction
7.19 Differentiate between rectifying and ohmic metal–semiconductor contacts.
7.20 Derive expressions for the space charge region width and junction capacitance for metal–(*n*-type) semiconductor for contacts.
7.21 Explain the *J-V* characteristics for a metal–semiconductor contact using emission model.
7.22 Derive an expression for Schottky barrier-lowering.
7.23 Why does tunnelling current dominate in metal–semiconductor junctions using heavily doped semiconductors?
7.24 Draw a band diagram for a metal–(*n*-type) semiconductor incorporating the interfacial layer.
7.25 Why is specific contact resistance, and not contact resistance, used to qualify a contact?
7.26 Prove the electron affinity rule for a heterojunction.
7.27 Explain the formation of two-dimensional electron gas using suitable energy band diagram.

Problems

7.1 A Schottky contact is made out of Au and (*n*-Ge) with a doping level $N_d = 10^{16}$ cm^{-3} at 300 K. Evaluate the ideal Schottky barrier height, built-in potential, space charge region width at zero bias, and maximum magnitude of electric field at zero bias. Assume $\phi_{Au} = 5.1$ eV, $\chi_s = 4.13$ eV, $N_c = 1.04 \times 10^{19}$ cm^{-3}.

$$\left[Hint: V_{bi} = \frac{\phi_{B0} - \phi_n}{e}; |E_{max}| = \frac{eN_d x_n}{\varepsilon_s} \right]$$

Ans. 0.97 eV, 0.79 V, 0.37×10^{-4} cm, 4.18×10^4 Vcm^{-1}

7.2 Consider a Ti–(*n*-type GaAs) Schottky contact with doping level $N_d = 10^{15}$ cm^{-3}. Calculate the ideal barrier height, built-in potential barrier, space charge region width at zero bias, and magnitude of maximum electric field at zero bias. Assume, $\phi_{Ti} = 4.33$ eV, $\chi_{GaAs} = 4.07$ eV, $N_c = 4.7 \times 10^{17}$ cm^{-3}, $\varepsilon_{GaAs} = 13.1$.

Ans. 0.26 V, 0.1 V, 0.38×10^{-4} cm, 5.24×10^3 V cm^{-1}

7.3 Graph of $(1/C_J)^2$ versus reverse bias V is revealed to be a straight line for a particular metal–Ge (n-type) Schottky contact. The intercept of the straight line is at a magnitude of 0.3 V and the slope is $\sim 2 \times 10^{13}$. Calculate the actual barrier height. Assume $\varepsilon_{Ge} = 10$ and $N_c = 1.04 \times 10^{19}$ cm^{-3}.

$$\left[\textit{Hint: } N_d = \frac{2}{e\varepsilon_s (\text{slope})} \text{ and } \phi_{Bn} = V_{\text{bi}} + \phi_n \right]$$

Ans. 0.38 V

7.4 The plot of $(1/C_J)$ versus reverse bias, V, for a metal-(n-type) GaAs contact is a straight line. The doping level of GaAs is 10^{15} cm^{-3}. What is the slope of the straight line?

$$\left[\textit{Hint: } \text{Slope} = \frac{2}{e\varepsilon_s N_d} \right]$$

Ans. 1.08×10^{16}

7.5 A Schottky contact is formed using a metal with Ge. An electric field of magnitude $E = 5 \times 10^4$ V cm^{-1} exists. Calculate the Schottky barrier-lowering and the position of the maximum barrier height. Assume $\varepsilon_{Ge} = 16$.

$$\left[\textit{Hint: } \Delta\phi = \sqrt{\frac{eE}{4\pi\varepsilon_s}} \text{ and } x_m = \sqrt{\frac{e}{16\pi\varepsilon_s E}} \right]$$

Ans. 0.021 V, 2.12×10^{-7} cm

7.6 A metal–GaAs Schottky contact results in a Schottky barrier-lowering of 0.02 V. Determine the magnitude of the electric field in the semiconductor. Assume $\varepsilon_{GaAs} = 13.1$.

$$\left[\textit{Hint: } E = \frac{(\Delta\phi)^2 \, 4\pi\varepsilon_s}{e} \right]$$

Ans. 3.64×10^4 V cm^{-1}

7.7 Calculate the reverse-saturation current density for a tungsten–silicon Schottky contact. Assume a barrier height of 0.7 V and take the value of the effective Richardson constant A^* to be 114 A/K^2 cm^2.

$$\left[\textit{Hint: } J_0 = A^* T^2 \exp\left(-\frac{e\phi_{Bn}}{kT} \right) \right]$$

Ans. 2.08×10^{-5} A/cm^2

7.8 For the tungsten–silicon Schottky contact indicated in Problem 7.7, calculate the forward bias required to generate a forward current density of 5 Acm^{-2}.

$$\left[\textit{Hint: } V \simeq \frac{kT}{e} \ln\left(\frac{J}{J_0} \right) \right]$$

Ans. 0.322 V

7.9 A heterojunction is formed between n^+-$Al_{0.3}$ $Ga_{0.7}$ As and n-GaAs. The conduction band step ΔE_c is found to be 0.28 eV. Assume E_g ($Al_{0.3}$ $Ga_{0.7}$ As) = 1.85 eV and $E_{g\text{GaAs}}$ = 1.43 eV. Calculate the valence band step ΔE_v.

[*Hint:* $\Delta E_c + \Delta E_v = E_{g(Al_{0.3}Ga_{0.7}As)} - E_{g(GaAs)}$]

Ans. 0.14 eV

7.10 An n-p heterojunction is fabricated between GaAs and Ge. The donor concentration N_d of GaAs is 10^{16} cm^{-3} and the acceptor concentration N_a of Ge is 5×10^{16} cm^{-3}. Evaluate the ratio V_{bin}/V_{bip} for the heterojunction. Assume ε_{Ge} = 16 and ε_{GaAs} = 13.1.

$$\left[\textit{Hint:} \frac{V_{bin}}{V_{bip}} = \frac{N_a \varepsilon_{Ge}}{N_d \varepsilon_{GaAS}} \right]$$

Ans. 6.11

7.11 A GaAs sample ($\chi = 4.07$ e V), as shown in Fig 7.P11.1, is kept at $T = 300$ K. It has interface states 0.7 e V below the conduction band level and is doped with 10^{13} tellurium impurity atoms per cm^3. It is brought in contact with gold on the A side:

Fig. 7.P11.1 GaAs sample

(a) Sketch and label the energy band diagram for the metal–semiconductor contact developed across A.

(b) Now, this sample is doped with cadmium impurities on the B side (as shown in the shaded region in the figure) with a concentration of 4×10^{18} cm^{-3} and then the same metal contact is made on the B side after passivating the semiconductor surface to remove almost all interface states. Sketch and label the energy band diagram for the metal semiconductor contact developed across B.

(c) Sketch and label the band diagram for the junction made between A and B, if any. Otherwise, justify your answer.

(d) Comment on the behaviour of the metal–semiconductor contacts on both sides A and B.

[*Hint:* Refer to Solved Problem 7.10]

7.12 Reverse saturation current densities of a Schottky barrier diode and an ordinary p-n junction diode are given as 5.98×10^{-5} A/cm^2 and 3.66×10^{-11} A/cm^2 respectively. Calculate the forward bias voltage required to generate a forward bias current density of 10 A/cm^2 for both these diodes. Comment on the result obtained.

Ans. For Schottky diode: 0.312 V
for p-n junctrion diode:0.682 V

7.13 In Solved Problem 7.7, the original *n*-type semiconductor is replaced by a *p*-type semiconductor. Assuming that everything else remains the same, draw the resulting energy band diagram. Also state the type of a contact that would be developed.

7.14 The ideal Schottky barrier height of a metal *n*-type semiconductor contact is 0.15V at 300 K. If the doping level in the semiconductor is $N_d = 8 \times 10^{16}$ cm^{-3} at 300 K, calculate the effective density of states function N_c.

Ans. 2.56×10^{19} cm^{-3}

7.15 A plot of $(1/C_J)^2$ versus reverse bias, V, for a particular metal–silicon (*n*-type) contact is a straight line. The doping concentration, N_d, is 5×10^{17} cm^{-3}. Calculate the slope of the straight line.

Ans. 2.42×10^{13}

8 Bipolar Junction Transistors

- Fundamental processes involved in bipolar junction transistor (BJT)
- Current components and their relations
- Notations and configurations used
- Different current gains for transistors
- Minority carrier distributions in different regions of a transistor
- Different models for bipolar transistors
- Frequency limitation for BJT
- High-frequency transistors
- Switching characteristic of BJT
- Schottky-clamped transistor

Learning Objectives

After going through this chapter the student will be able to

- understand the physical processes involved in a bipolar transistor
- understand different current components in transistors and the relations between them
- understand common notations and configurations
- derive expressions for common current gains
- derive expressions for current gains in different regions of a transistor
- solve numericals based on output characteristics of a transistor
- solve numericals based on excess carrier concentrations in different regions of a transistor
- understand Ebers–Moll model, Gummel–Poon model, hybrid-pi model, and h-parameter equivalent circuit model
- derive expressions for A_{ve} and A_{vb}
- understand frequency limitations and their causes for a BJT
- understand the issues involved in the design of high-frequency transistors
- understand the switching characteristics of a bipolar junction transistor
- understand the functioning of a Schottky transistor
- solve numericals based on the expression of collector–emitter voltage at saturation
- solve numericals based on A_{ve} and A_{vb}
- solve numericals based on the expression for cut-off frequency

Introduction

The *p-n* junction studied so far is an important milestone in the understanding of solid state electronics. The phenomenal growth of this field is, however, due to the invention of transistor, which led to the large-scale replacement of the then-dominating vacuum tubes. We will start our study of the transistor with the bipolar junction transistor in this chapter. In simple terms, a bipolar junction transistor (BJT) consists of three differently doped semiconductor regions. Two of these regions are doped with one type dopant (acceptor or donor) and the third with the another type. A BJT essentially consists of two *p-n* junctions placed back to back. The word 'bipolar', is used to indicate the role of both the types of charge carriers, namely electrons and holes. We will see in this chapter how such a simple structure can result in a host of effects such as current gain, power gain, and so on. Some important models that have been used to explain the characteristics of a BJT will also be briefly discussed in this chapter. Some important limitations of this device would be taken care of in other types of transistors, which will be discussed in the subsequent chapters.

8.1 Fundamentals of Bipolar Junction Transistors

As already mentioned a bipolar junction transistor consists of two *p-n* junctions placed back to back. The two *p-n* junctions are coupled to each other by a semiconducting region, which is common to both the junctions. This common region is known as the *base* region of the transistor. The other two regions of the same conductivity type are called the *emitter* and the *collector*. Thus the junction between the emitter and the base is one of the *p-n* junctions comprising the transistor. This junction is referred to as the *emitter junction*, whereas the junction between the collector and the base is called the collector junction. Under no external applied bias and in thermal equilibrium the Fermi energy levels in all the three regions must lie along the same line. Figure 8.1 shows the diagram of an *n-p-n* transistor alongwith its zero-bias energy band diagram.

In Fig. 8.1 the emitter and collector regions are of *n*-type, whereas the base region is of *p*-type, such a transistor is called an *n-p-n* transistor. The other type of transistor is called a *p-n-p* transistor. In the *p-n-p* transistor, the emitter and collector regions are *p*-type and the base region is *n*-type.

At zero bias the net current density is zero throughout the transistor. The net electron-current density, due to diffusion and drift, is equal to zero. Similarly, the net hole current density is also zero. To understand the basic control action of a transistor, let us assume the emitter junction to be forward biased and the collector junction to be reverse biased. It so happens that this also is the most common state of operation of a transistor. This biasing arrangement alongwith the energy band diagram is shown in Fig. 8.2.

Electrons are injected from the emitter region, where they happen to be majority carriers, into the base region, where they are minority carriers. In the base region,

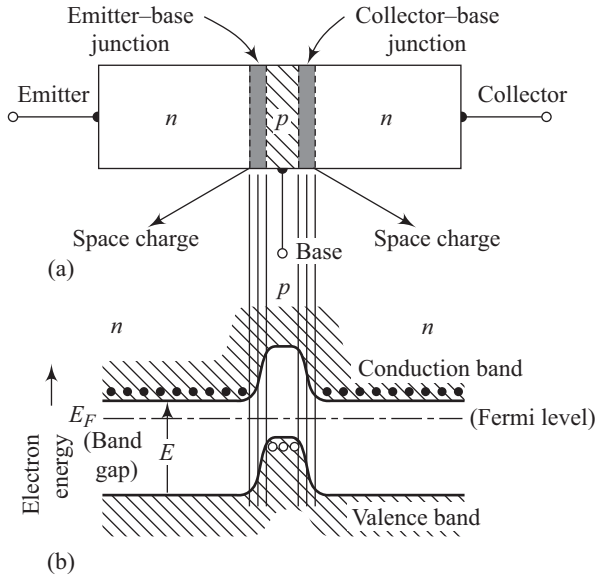


Fig. 8.1 (a) Schematic diagram of an *n-p-n* transistor, (b) zero-bias energy band diagram

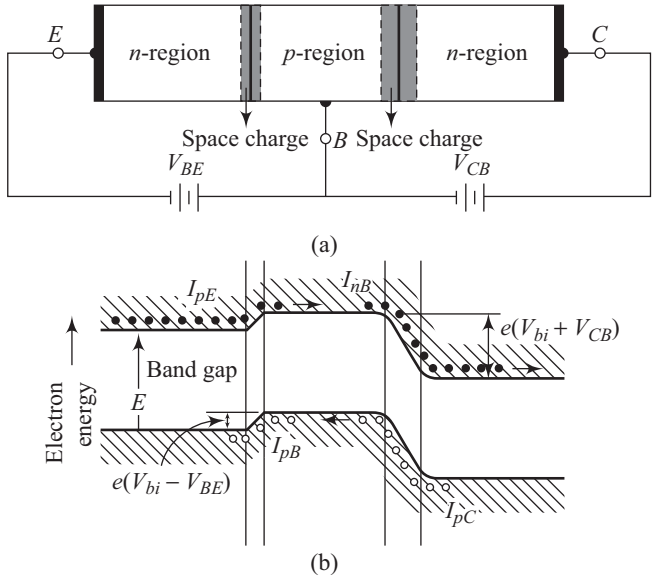


Fig. 8.2 Biasing arrangement and energy band diagram of an *n-p-n* transistor

a small fraction of these electrons recombines with the majority carrier holes. The remaining electrons diffuse across the thin base region and reach the space charge region of the collector junction. The *n*-type collector region is positively biased, due to which the electrons are collected by the collector. The collected electrons dominate the collector current. Thus, *the collector current is controlled*

by the emitter-to-base voltage. The power gain resulting due to the transistor action is possible because the base-to-emitter voltage and the base current are very small.

Figure 8.3 shows the symbolic representations of *n-p-n* and *p-n-p* transistors.

The emitter terminal in the symbolic representation has an arrow which points in the direction of emitter current (opposite to the direction of electron flow).

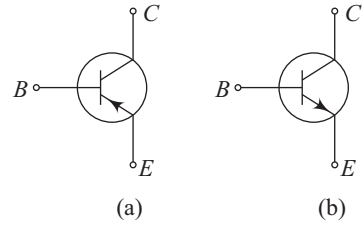


Fig. 8.3 Symbolic representation of (a) *p-n-p* transistor and (b) *n-p-n* transisor

To understand the physical mechanism of a bipolar junction transistor, let us go back to the energy band diagram depicted in Fig. 8.2(b). Due to the forward bias across the emitter junction, the potential barrier across the emitter–base junction gets lowered to $e(V_{\text{bi}} - V_{BE})$, where V_{BE} represents the forward bias across the emitter–base junction. This lowering of the potential barrier results in the diffusion of electrons from the *n*-type emitter to the *p*-type base. A flow of holes from the base region to the emitter region also takes place but can be treated as negligible since the emitter is usually doped much more heavily than the base. The diffusion of electrons into the base region results in an electron profile in the base that has a higher density at the emitter end of the base than at the collector end. The minority carrier electrons in the base region traverse through this region by diffusion. The collector–base junction is reverse biased and thus the potential barrier across this junction increases to $e(V_{\text{bi}} + V_{CB})$ where V_{CB} represents the applied bias across the collector–base junction. The electrons that arrive at the beginning of the collector junction space charge layer on the base side then slide down the potential hill, as

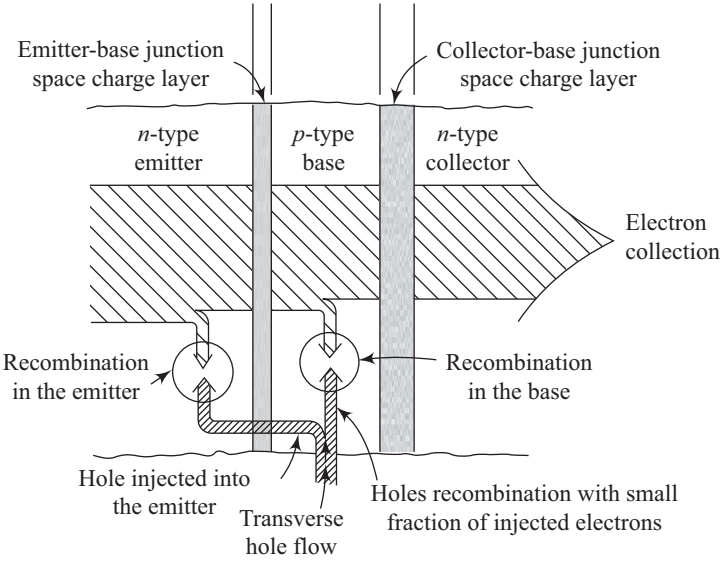


Fig. 8.4 Internal current flow in an *n-p-n* transistor

shown in Fig. 8.2(b), and are finally collected by the collector terminal. As already discussed, a small percentage of electrons injected into the base recombine with the holes present in the base. Under thermal equilibrium conditions, these lost electrons are replenished by the base current.

The base current is kept low by ensuring a thin base having high carrier lifetimes. The internal currents flowing in an *n-p-n* transistor are depicted schematically in Fig. 8.4.

8.2 Current Components and Relations

We have so far considered the base–emitter junction to be forward biased and the base–collector junction to be reverse biased. This biasing arrangement is also called the *forward-active* operating mode. Figure 8.5 shows the complete biasing arrangement for a transistor in the forward-active operating mode.

We must remember that positive conventional current is opposite to the direction of electron current. The directions indicated for I_E, I_B, and I_C should be viewed in this context. The emitter current I_E consists of two components. I_{nE} is due to electrons crossing from the emitter into the base and I_{pB} is due to holes crossing from the base into the emitter. The component I_{pB} consists of holes which are moving away from the collector and therefore are not collected by the collector. The objective of a device engineer is to keep the ratio I_{nE}/I_{pB} as high as possible by ensuring that the emitter is much more heavily doped than the base. The directions of I_E and I_C in Fig. 8.5 have been indicated for conventional currents. This is the reason for I_E and I_C to have directions that are opposite to the direction of the flow of electrons. Some electrons entering the base region are lost due to recombination with holes in this region and the rest are collected by the collector. This collector current is represented by I_{nB} in Fig. 8.5. The electron current $(I_{nE} - I_{nB})$ constitutes the base current. The collector junction is reverse biased. In addition to I_{nB}, therefore, there exists a collector reverse saturation current I_{CS}. The collector saturation current consists of holes moving from the collector region towards the base region and electrons moving from the base region to the collector region. Referring to Fig. 8.5, we can write

$$I_C = I_{CS} - I_{nB} \tag{8.1}$$

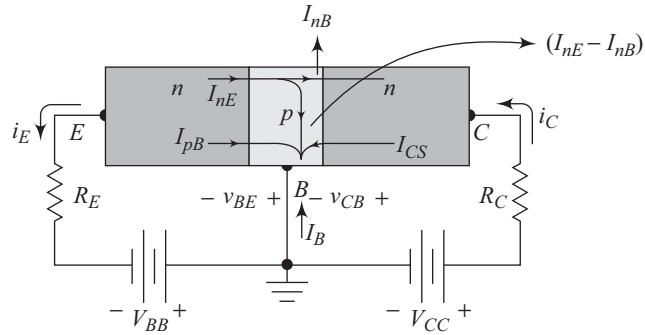


Fig. 8.5 Biasing arrangement for an *n-p-n* transistor in the forward-active mode

8.3 Important Notations and Configurations

One must remember at this stage that although the biases applied on a transistor are dc, the transistor generally has ac inputs and outputs. It is therefore important to develop a set of notations before developing any further understanding of the transistor. Figure 8.6 shows a schematic diagram of the various important variations for collector current. Although collector current has been taken as a representative example, the notations are very general.

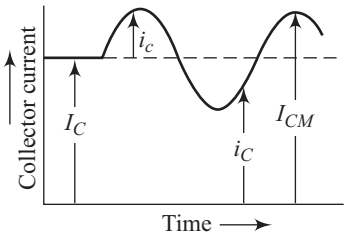


Fig. 8.6 Important notations for current

The same notation is indicated in a tabular form in Table 8.1.

Table 8.1 Popular notations for transistors

Parameter	Letter	Subscript	Example
Total instantaneous quantity	Lowercase	Uppercase	i_C
Instantaneous varying component	Lowercase	Lowercase	i_c
Dc quantity	Uppercase	Uppercase	I_C
Ac rms quantity	Uppercase	Lowercase	I_c

A transistor is a three-terminal device with two junctions. Each of these junctions requires a biasing arrangement. One of the best ways of biasing the two junctions is to keep one of three terminals common. This results in the following three configurations, namely common–emitter (CE), common–base (CB), and common–collector (CC). These configurations are shown schematically in Fig. 8.7.

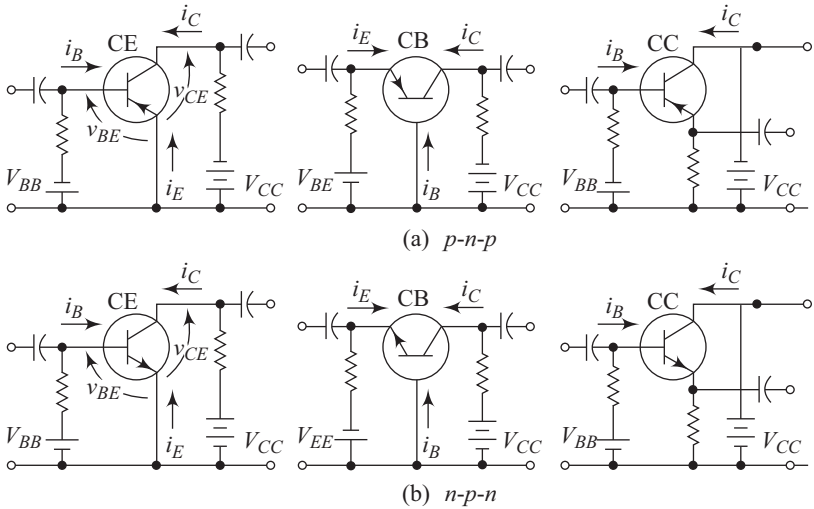


Fig. 8.7 Important configurations for transistors

8.4 BJT Characteristics

A bipolar junction transistor (BJT) can be operated in four possible modes. The modes are characterized by the biasing configurations of the emitter-base and base-collector junctions. In the *active* or *normal mode* the emitter-base junction is forward biased whereas the base-collector junction is reverse biased. Both the junctions are reverse biased in *cut-off mode* and both the junctions are forward biased in the *saturation mode*. In the *inverse active* mode, the emitter-base junction is reverse biased whereas the base-collector junction is forward biased.

The I-V characteristics are dependent upon the configuration used for the BJT. Figure 8.8 shows the I-V characteristics for the common-base configuration of a *p-n-p* transistor.

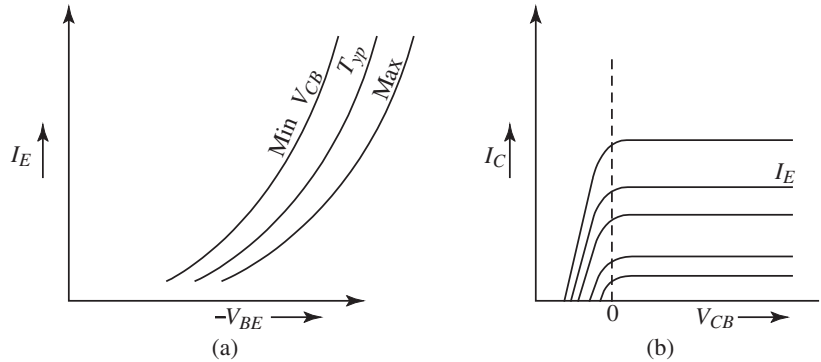


Fig. 8.8 I-V characteristics for common-base configuration (a) Input (b) output characteristics

The input characteristics shown in Fig. 8.8(a) can be expressed as,

$$V_{EB} = f_1(I_E, V_{EB}) \tag{8.2}$$

Thus the emitter-base voltage is a function of the emitter current and collector-base voltage. The three curves indicated in Fig. 8.8(a) are for minimum, typical, a and maximum collector-base voltage. The negative sign has been used before V_{BE} because V_{BE} is negative for *p-n-p* transistors and the graph depicts positive voltage values. The input characteristics simply represent the emitter-base junction diode. The output characteristics are expressed in the form,

$$I_C = f_2(I_E, V_{CB}) \tag{8.3}$$

Fig. 8.8(b) shows the output characteristics for different I_E values with the magnitude of I_E going up with increasing I_C values. In the active region of the characteristics, the emitter-base junction is forward biased whereas the base-collector junction is reverse biased. Both the junctions are forward biased in saturation region. The cut-off region is characterized by the fact that both the junctions are reverse biased.

The input and output characteristics for the common-emitter configuration are shown schematically in Fig. 8.9 for a *p-n-p* transistor.

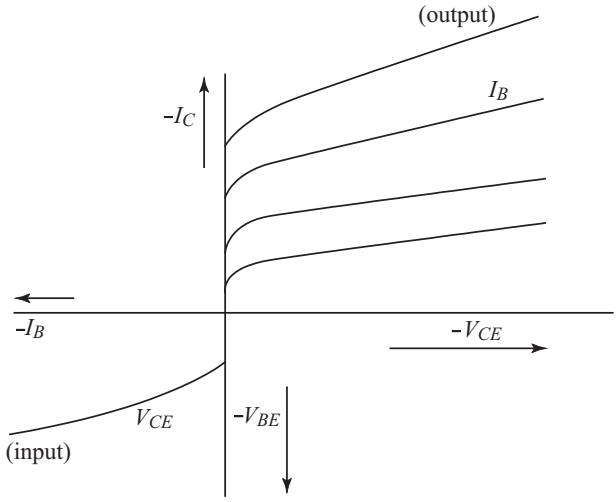


Fig. 8.9 Input and output characteristics for common-emitter configuration for a *p-n-p* transistor

Input characteristics can be expressed as,

$$V_{BE} = f_1(I_B, V_{CE}) \tag{8.4}$$

and the output characteristics are given by,

$$Ic = f_2(I_B, V_{CE}) \tag{8.5}$$

The input characteristics are the forward characteristics of the emitter-base diode. The output characteristics show the dependence of I_C on V_{CE} for different I_B values.

The collector terminal is common to the input and output circuits in the common-collector configuration. This configuration is also called the emitter follower. This is because in this configuration the input signal and the output voltage can be nearly equal for small V_{BE} or large gain. The input resistance is high and the output resistance is low. The configuration is therefore very suitable as a buffer, i.e. to couple to a load having large current demands.

8.5 Current Gains for Transistor

As we have seen, the emitter current i_E consists primarily of a large diffusion of electrons from the emitter to the base, in the case of an *n-p-n* transistor. Some of these electrons are lost due to recombination within the base, and the rest are collected by the collector. Suppose a parameter α is used to represent the modification in the emitter current due to the base region recombination.

Taking into account the collector junction reverse saturation current I_{CS}, we can write

$$\alpha_F i_E = + i_C - i_{CS} \tag{8.6}$$

Thus,

$$\alpha_F = \frac{i_C - I_{CS}}{i_E} \tag{8.7}$$

Under usual operating conditions, the quantity I_{CS} can be neglected, leading to

$$\alpha_F = \frac{i_C}{i_E} \tag{8.8}$$

The quantity α is called the common-base current gain. Since $i_C < i_E$, $\alpha < 1$. The objective of a device engineer is to bring α as close to unity as possible.

The ratio of collector current to base current is also an important quantity and is called the common-emitter current gain. The symbol β is used to represent it; thus

$$\beta = \frac{i_C}{i_B} \tag{8.9}$$

Since the base current i_B is generally very small in comparison to i_C, the common-emitter current gain is much larger than unity (values of 100 or more are common).

The base-collector junction of a BJT is reverse biased. The depletion region of a reverse biased junction is dependent upon the applied bias as given by Eqn. (5.53). For a lightly doped base region, the depletion layer width can be a significant part of the total base region width. The effective base width is then reduced with increase in the applied reverse bias in the base-collector junction. This is called *Early effect* after it was first interpreted by J.M. Early. The effect is also referred to as *base narrowing* or base-width modulation. Early effect leads to an increase in β with applied base-collector reverse bias. The overall effect is that the collector current starts increasing with applied collector voltage rather than maintaining a constant value as expected from simple treatment. At a large-enough collector voltage the depletion region can fill the entire base region. This is called the *punch-through condition*. When such a condition happens, for example, in a p^+-n-p^+ transistor, holes are directly swept from the emitter region to the collector region and the transistor action is completely lost. This condition is therefore also referred to as the breakdown condition of a transistor.

8.6 Minority Carrier Distribution

We have already seen the dominant role played by diffusion in deciding the characteristics of a transistor. Diffusion, you would recollect, is, in turn, determined by the carrier gradient. The minority carrier distribution is thus a very important characteristic in deciding the transistor characteristics. As a demonstrative example, we will consider an n-p-n transistor and concentrate primarily on the active mode of the transistor (the BE junction is forward biased and the BC junction is reverse biased). The doping concentrations in the emitter, base, and collector regions are assumed to be N_E, N_B, and N_C, respectively. The corresponding widths of the three regions are W_E, W_B, and W_C.

8.6.1 Base Region

To make the mathematics simpler, we will treat each region separately. A schematic diagram of the base region alongwith the space charge regions on the two sides is shown in Fig. 8.10.

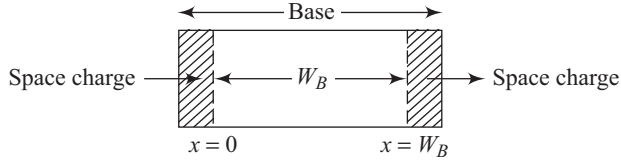


Fig. 8.10 Schematic diagram of the base showing space charge regions

The transport equation in the base region, neglecting electric field in the neutral base region, is given by

$$D_n = \frac{\partial^2[\delta n(x)]}{\partial x} - \frac{\delta n(x)}{\tau_{n0}} = 0 \tag{8.10}$$

leading to

$$\frac{\partial^2[\delta n(x)]}{\partial x^2} - \frac{\delta n(x)}{L_n^2} = 0 \tag{8.11}$$

where D_n and τ_{n0} are the minority carrier electron diffusion coefficient and lifetime in the base region, respectively. The excess electron concentration δn is defined through the relation,

$$\delta n(x) = n_p(x) - n_{p0} \tag{8.12}$$

The general solution of Eqn (8.11) is

$$\delta n(x) = Ae^{(x/L_n)} + Be^{-(x/L_n)} \tag{8.13}$$

where L_n is the minority carrier diffusion length in the base region.
We will first evaluate $\delta n(x)$ at the two edges of the neutral base region, i.e., at $x = 0$ and $x = W_B$, leading to

$$\delta n\,(x = 0) = A + B \tag{8.14}$$

and

$$\delta n(x = W_B) = Ae^{W_B/L_n} + Be^{-W_B/L_n} \tag{8.15}$$

Since the junction between the base and the emitter is forward biased, the boundary condition at $x = 0$ (see Fig. 8.10) is

$$\delta n(x = 0) = n_p\,(x = 0) - n_{p0} = n_{p0}[e^{eV_{BE}/kT} - 1] \tag{8.16}$$

The junction between the base and the collector is reverse biased, leading to a boundary condition at $x = W_B$, which can be expressed as

$$\delta n\,(x = W_B) = n_p\,(x = W_B) - n_{p0} = 0 - n_{p0} \tag{8.17}$$

leading to

$$\delta n\,(x = W_B) = -n_{p0} \tag{8.18}$$

Using Eqs (8.14) and (8.16), we get

$$A + B = n_{p0}[e^{eV_{BE}/kT} - 1] \tag{8.19}$$

Also, using Eqs (8.15) and (8.18), we get

$$Ae^{W_B/L_n} + Be^{-W_B/L_n} = -n_{p0} \tag{8.20}$$

Using Eqs (8.19) and (8.20) and solving for A and B yields

$$A = \frac{n_{p0}e^{-W_B/L_n}[e^{eV_{BE}/kT} - 1] + n_{p0}}{e^{-W_B/L_n} - e^{W_B/L_n}} \tag{8.21}$$

and

$$B = \frac{-n_{p0}e^{W_B/L_n}[e^{eV_{BE}/kT} - 1] - n_{p0}}{e^{-W_B/L_n} - e^{W_B/L_n}} \tag{8.22}$$

You would recollect that

$$\sinh z = \frac{e^Z - e^{-Z}}{2} \tag{8.23}$$

Using Eqn (8.23) in Eqs (8.21) and (8.22) results in

$$A = \frac{-n_{p0}e^{-W_B/L_n}[e^{eV_{BE}/kT} - 1] - n_{p0}}{2\sinh\left(\dfrac{W_B}{L_n}\right)} \tag{8.24}$$

and

$$B = \frac{n_{p0}e^{W_B/L_n}[e^{eV_{BE}/kT} - 1] + n_{p0}}{2\sinh\left(\dfrac{W_B}{L_n}\right)} \tag{8.25}$$

Substituting the values of A and B from Eqs (8.24) and (8.25) into Eqn (8.13) leads to

$$\delta n(x) = \frac{n_{p0}\left\{[e^{eV_{BE}/kT} - 1]\sinh\left(\dfrac{W_B - x}{L_n}\right) - \sinh\left(\dfrac{x}{L_n}\right)\right\}}{\sinh\left(\dfrac{W_B}{L_n}\right)} \tag{8.26}$$

If $W_B < L_n$, then we can use the approximation

$$\sinh(x) \approx x \text{ for } x \ll 1 \tag{8.27}$$

Using approximation (8.27) in Eqn (8.26) results in

$$\delta n(x) \approx \frac{n_{p0}}{W_B}\{[e^{eV_{BE}/kT} - 1]\,(W_B - x) - x\} \tag{8.28}$$

8.6.2 Emitter Region

A schematic diagram of the emitter region is shown in Fig. 8.11.

Since the emitter region has only one space charge region, the edge of the space charge region has been defined to be $x = 0$ and the edge of the emitter region to be $x = W_E$. Thus W_E is the width of the neutral emitter region. In Fig. 8.11, the positive x-direction has been defined from right to left, unlike the positive x direction in Fig. 8.10. The redefinition of the x direction and magnitude, we must remember, does not take away the importance of some general conclusions regarding the minority carrier concentration profile.

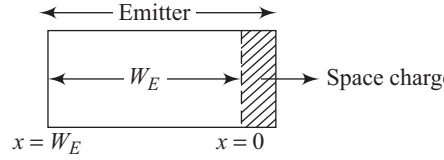


Fig. 8.11 Schematic diagram of an emitter showing space charge region

In the absence of any electric field in the neutral emitter region, the transport equation in the emitter region is given by

$$D_p \frac{\partial^2[\delta p(x)]}{\partial x^2} - \frac{\delta p(x)}{\tau_{p0}} = 0 \tag{8.29}$$

where D_p and τ_{p0} are the minority carrier hole diffusion length and minority carrier lifetime, respectively. Equation (8.29) can be rewritten in the form

$$\frac{\partial^2[\delta p(x)]}{\partial x^2} - \frac{\delta p(x)}{L_p^2} = 0 \tag{8.30}$$

where the excess hole concentration, δp, is given by

$$\delta p(x) = p_n(x) - p_{n0} \tag{8.31}$$

Equation (8.30) has a general solution of the form

$$\delta p(x) = A_1 e^{x/L_p} + B_1 e^{-x/L_p} \tag{8.32}$$

Excess minority carrier hole concentrations at the two points $x = 0$ and $x = W_E$ are given by

$$\delta p(0) = A_1 + B_1 \tag{8.33}$$

and

$$\delta p(W_E) = A_1 e^{W_E/L_p} + B_1 e^{-W_E/L_p} \tag{8.34}$$

Since the emitter–base junction is forward biased, we must have

$$\delta p(0) = p_{n0}[e^{eV_{BE}/kT} - 1] \tag{8.35}$$

Assuming an infinite surface recombination velocity at $x = W_E$ leads to

$$\delta p\,(W_E) = 0 \tag{8.36}$$

Solving for A_1 and B_1 using Eqs (8.33)–(8.36) and substituting into Eqn (8.32), following a procedure that is similar to the one followed in the preceding section, yields

$$\delta p(x) = \frac{p_{n0}[e^{eV_{BE}/kT} - 1] \sinh\left(\frac{W_E - x}{L_p}\right)}{\sinh(W_E/L_p)} \tag{8.37}$$

Assuming a thin emitter region, i.e, W_E to have a small value , Eqn (8.37) can be rewritten in the form

$$\delta p(x) \approx \frac{p_{n0}}{W_E}[e^{eV_{BE}/kT} - 1](W_E - x) \tag{8.38}$$

If, on the other hand, W_E is comparable to L_p, $\delta p(x)$ displays an exponential dependence on x.

8.6.3 Collector Region

Figure 8.12 shows a schematic diagram of the collector region.

In the collector region, the space charge associated with the base-collector junction is towards the left edge and the right edge is the end of the collector region. The transport equation for minority carrier holes in the collector region has the form

$$D_p \frac{\partial^2[\delta p(x)]}{\partial x^2} - \frac{\delta p(x)}{\tau_{p0}} = 0 \tag{8.39}$$

The diffusion coefficient, D_p, and the minority carrier lifetime, τ_{p0}, used in Eqn (8.39) are valid in the collector region and are in general different in magnitude with respect to the magnitudes applicable in the emitter region. Equation (8.38) can be rewritten in the form

$$\frac{\partial^2[\delta p(x)]}{\partial x^2} - \frac{\delta p(x)}{L_p^2} = 0 \tag{8.40}$$

where

$$L_p^2 = D_p \tau_{p0} \tag{8.41}$$

The excess carrier concentration $\delta p(x)$ can be expressed as

$$\delta p(x) = p(x) - p_{n0} \tag{8.42}$$

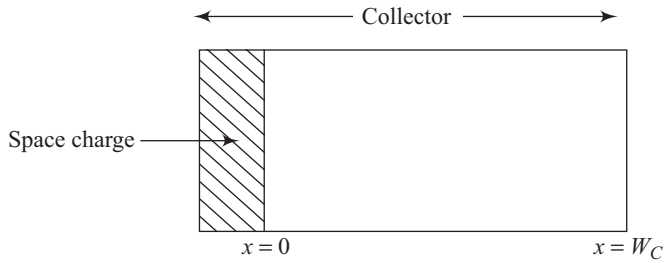


Fig. 8.12 Schematic diagram of the collector region

Equation (8.40) has solutions of the form

$$\delta p(x) = A_2 e^{x/L_p} + B_2 e^{-x/L_p} \tag{8.43}$$

For a very long collector, the excess carrier concentration $\delta p(x = W_C)$ cannot remain finite unless $A_2 = 0$. Thus Eqn (8.43) can be rewritten in the form

$$\delta p(x) = B_2 e^{-x/L_p} \tag{8.44}$$

The boundary condition at the edge of the space charge region is given by

$$\delta p\ (x = 0) = p(x = 0) - p_{n0} = 0 - p_{n0} = -p_{n0} \tag{8.45}$$

Using Eqn (8.45) in Eqn (8.44) leads to

$$B_2 = -p_{n0} \tag{8.46}$$

Using Eqn (8.46) in Eqn (8.44) results in

$$\delta p(x) = -p_{n0} e^{-x/L_p} \tag{8.47}$$

- The number of impurities per unit area in the base is called the Gummel number. The Gummel number for silicon bipolar transistors is between 10^{12} and 10^{13} cm^{-2}.
- The ratio of collector current and base current decreases with an increase in collector current. This effect is called Webster effect.

8.7 Models for Bipolar Junction Transistors

A transistor is a complicated device involving two junctions. It is therefore quite natural to develop some equivalent circuit models which can replace transistors in a circuit. We will now consider, in some detail, three important models used for bipolar junction transistors.

8.7.1 Ebers–Moll Model

A bipolar junction transistor consists of two interacting junction diodes and the Ebers–Moll equivalent circuit uses this interaction as its basic principle. Figure 8.13 is a schematic diagram of the biasing arrangement and the current directions commonly used in understanding this model of the transistor. Notice that both emitter-base and base-collector junctions have been assumed to be forward biased. One must state at this stage that such a biasing arrangement does not take away the generality of this model. We have so far considered the active mode of operation of a transistor, in which the emitter-base junction is forward biased and the base-collector junction is reverse biased. A transistor can, however, operate in two other modes, namely saturation mode and cut-off mode. In the saturation mode, both the junctions of a transistor are forward biased, whereas in the cut-off mode both the junctions are reverse biased.

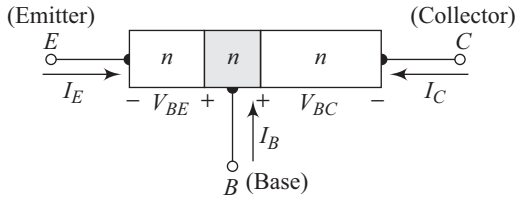


Fig. 8.13 Biasing arrangement and current directions for transistor

From Fig 8.13, it is clear that

$$I_E + I_B + I_C = 0 \tag{8.48}$$

Notice that the direction of I_E used in Fig. 8.13 is opposite to the one used in Fig. 8.5. In general, the collector current arises due to two contributions, namely, the electrons reaching the collector due to emission from the emitter and the current contribution due to the base-collector junction. If I_E represents the emitter current and α_F represents the common–base current gain, then, for the collector current I_C, we can write

$$I_C = \alpha_F I_E - I_R \tag{8.49}$$

where I_R represents the base-collector contribution. In the active mode, the base–collector junction is reverse biased, and Eqn (8.49) can be rewritten in the form

$$I_C = \alpha_F I_E + I_{CS} \tag{8.50}$$

where I_{CS} is the BC junction reverse saturation current. Current I_E is, in turn, given by

$$I_E = I_{ES}[e^{eV_{BE}/kT} - 1] \tag{8.51}$$

In the saturation mode, the BC junction is forward biased and therefore, for I_R in Eqn (8.49), we can write

$$I_R = I_{CS}[e^{eV_{BC}/kT} - 1] \tag{8.52}$$

Putting Eqn (8.48) in Eqn (8.49) leads to

$$I_C = \alpha_F I_{ES}[e^{eV_{BE}/kT} - 1] - I_{CS}[e^{eV_{BC}/kT} - 1] \tag{8.53}$$

One must at this stage appreciate that nothing prevents us from interchanging the roles of the emitter junction and the collector junction. In such a situation, we can write for emitter current I_E,

$$I_E = \alpha_R I_R - I_F \tag{8.54}$$

where α_R is the common-base current gain in the inverse mode and I_F is the emitter-base junction current. Equation (8.54) can be written as

$$I_E = \alpha_R I_{CS}[e^{eV_{BC}/kT} - 1] - I_{ES}[e^{eV_{BE}/kT} - 1] \tag{8.55}$$

Equations (8.53) and (8.55) represent the Ebers–Moll equations, and the corresponding equivalent circuit is shown in Fig. 8.14.

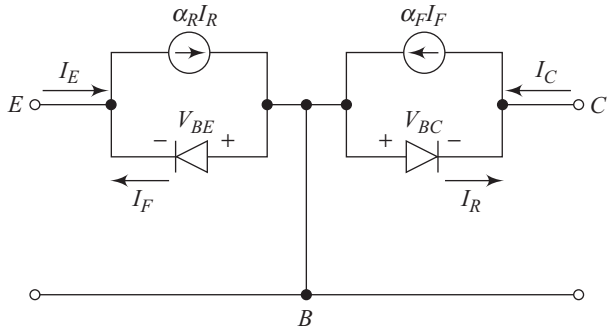


Fig. 8.14 Ebers–Moll equivalent circuit

The equivalent circuit consists of two current sources and two ideal junctions. Each current source is dependent upon the voltages across the other junction. A quantitative analysis reveals that the parameters α_F, α_R, I_{CS}, and I_{ES} are not independent quantities, but are inter-related through the relation

$$\alpha_R I_{CS} = \alpha_F I_{ES} \tag{8.56}$$

Information about the parameters α_R, I_{CS}, α_F, and I_{ES} is generally provided in data sheets by the manufacturer.

The Ebers–Moll equivalent circuit model is valid for any mode of operation of a transistor. For example, in the saturation mode, both base-emitter and base-collector junctions are forward biased therefore V_{BE} and V_{BC} to be used in Eqs (8.53) and (8.55) are positive quantities. In a lot of applications involving transistors, the collector–emitter voltage at saturation, $V_{CE(\text{sat})}$ is an important quantity. This quantity is given by

$$V_{CE(\text{Sat})} = V_{BE} - V_{BC} \tag{8.57}$$

Substituting the value of I_E from Eqn (8.55) into Eqn (8.48) leads to

$$-(I_B + I_C) = \alpha_R I_{CS}[e^{eV_{BC}/kT} - 1] - I_{ES}[e^{eV_{BE}/kT} - 1] \tag{8.58}$$

Using Eqs (8.58) and (8.53) and solving for V_{BE} we get

$$V_{BE} = \frac{kT}{e} \ln\left[\frac{I_C(1-\alpha_R) + I_B + I_{ES}(1-\alpha_F\alpha_R)}{I_{ES}(1-\alpha_F\alpha_R)}\right] \tag{8.59}$$

Equations (8.53) and (8.58) can also evaluate V_{BC} by substituting for [exp (eV_{BE}/kT)–1] from Eqn (8.53) into Eqn (8.58). We get

$$V_{BC} = \frac{kT}{e} \ln\left[\frac{\alpha_F I_B - (1-\alpha_F) I_C + I_{CS}(1-\alpha_F\alpha_R)}{I_{CS}(1-\alpha_F\alpha_R)}\right] \tag{8.60}$$

Substituting for V_{BE} and V_{BC} from Eqs (8.59) and (8.60) into Eqn (8.57) leads to

$$V_{CE(\text{sat})} = \frac{kT}{e} \ln\left[\frac{I_C(1-\alpha_R) + I_B}{\alpha_F I_B - (1-\alpha_F) I_C} \times \frac{I_{CS}}{I_{ES}}\right] \tag{8.61}$$

Using Eqn (8.56) in Eqn (8.61) results in

$$V_{CE(\text{sat})} = \frac{kT}{e} \ln \left[\frac{I_C(1-\alpha_R) + I_B}{\alpha_F I_B - (1-\alpha_F) I_C} \times \frac{\alpha_F}{\alpha_R} \right] \tag{8.62}$$

Diagrams like Fig. 8.1.1 can be constructed from Eqn (8.62) using known values of transistor parameters.

8.7.2 Gummel–Poon Model

Ebers–Moll model discussed in the preceding section is effective in managing many transistor applications. The model, however, does not provide an in-depth picture of the physical effects going on in a transistor. This also makes the Ebers–Moll model quite ineffective in handling situations such as non-uniform doping in the base region of transistors. The Gummel–Poon model takes into account the physical processes taking place within a bipolar junction transistor, and is therefore quite capable of handling non-ideal situations.

Let us consider an *n-p-n* transistor. The electron current density in the base region is given by

$$J_n = e\mu_n n(x) E + qD_n \frac{dn(x)}{dx} \tag{8.63}$$

One would recollect that an electric field E would exist whenever the doping in the base region is non-uniform. The electric field E in such a situation is given by

$$E = \frac{kT}{e} \frac{1}{p(x)} \frac{dp(x)}{dx} \tag{8.64}$$

Substituting Eqn (8.64) into Eqn (8.63) yields

$$J_n = e\mu_n n(x) \left(\frac{kT}{e} \right) \left(\frac{1}{p(x)} \right) \frac{dp(x)}{dx} + qD_n \frac{dn(x)}{dx} \tag{8.65}$$

Einstein's relation is given by

$$D_n = \frac{kT}{e} \mu_n \tag{8.66}$$

Using Eqn (8.66) in Eqn (8.65) results in

$$\begin{aligned} J_n &= \frac{eD_n}{p(x)} \left[n(x) \frac{dp(x)}{dx} + p(x) \frac{dn(x)}{dx} \right] \\ &= \frac{eD_n}{p(x)} \frac{d}{dx}(pn) \end{aligned} \tag{8.67}$$

This can be rewritten in the form

$$\frac{J_n}{eD_n} \frac{p(x)}{} = \frac{d}{dx}(pn)$$

Integrating this equation across the base region leads to

$$\frac{J_n}{eD_n} \int_0^{W_B} p(x)dx = \int_0^{W_B} \frac{d}{dx}(pn)dx \tag{8.68}$$

Assuming the electron current density to be constant across the base region and a constant diffusion coefficient results in

$$\frac{J_n}{eD_n}\int_0^{W_B} p(x)dx = \int_0^{W_B} n(x)\frac{dp(x)}{dx}dx$$

$$= n(W_B)\,p(W_B) - n(0)p(0) \tag{8.69}$$

In the active mode, the BE junction is forward biased, whereas the BC junction is reverse biased. We, therefore, have

$$\mathrm{x}(0) = n_{p0}e^{eV_{BE}/kT} \tag{8.70}$$

and

$$n(W_B) = 0 \tag{8.71}$$

Using Eqs (8.70) and (8.71) in Eqn (8.69) yields

$$\frac{J_n}{eD_n}\int_0^{W_B} p(x)\,dx = -n_{p0}e^{eV_{BE}/kT}\,p(0) \tag{8.72}$$

We know

$$n_{p0}\,p(0) = n_i^2 \tag{8.73}$$

Equations (8.73) and (8.72) can be combined to give

$$\frac{J_n}{qD_n}\int_0^{W_B} p(x)dx = -n_i^2 e^{eV_{BE}/kT}$$

which can be simplified to

$$J_n = \frac{-qD_n n_i^2 e^{eV_{BE}/kT}}{\int_0^{W_B} p(x)dx} \tag{8.74}$$

$\int_0^{W_B} P(x)dx$ is called the base Gummel number, Q_B, and is a measure of total majority carrier charge in the base.

A similar analysis in the emitter region results in a hole current density given by

$$J_p = \frac{-eD_p n_i^2 e^{qV_{BE}/kT}}{\int_0^{W_E} n(x)dx} \tag{8.75}$$

$\int_0^{W_E} n(x)dx$ is called the emitter Gummel number, Q_E.

8.7.3 Hybrid-pi Model

The Ebers–Moll model and the Gummel–Poon model discussed in the preceeding sections do not take into account the junction capacitances associated with bipolar junction transistors. These models therefore tend to fail when an analysis is done of transistors at high frequencies. Figure 8.15 is a schematic representation of the hybrid-pi model or Giacoletto model of a transistor in the common-emitter configuration. Since all capacitive elements have been included, the model is obviously valid at high frequencies.

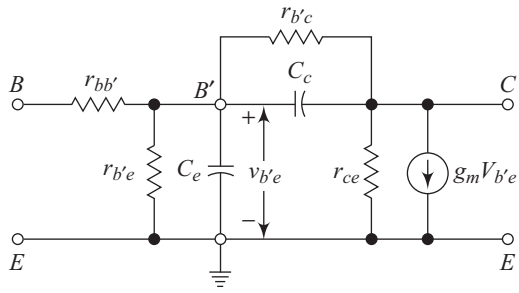


Fig. 8.15 Hybrid–pi model for CE configuration

In Fig. 8.15, B represents the external base terminal and B' is the active internal part of the base region. The hybrid-pi model makes the following assumptions.

(i) Resistances and capacitances used in the model are frequency independent.
(ii) Under a given biasing condition, resistances and capacitances of the model do not change due to small-signal variations.

We will now discuss the various parameters used in Fig. 8.15 in brief.

Resistance $r_{bb'}$ An ohmic base spreading resistance exists between the external base terminal B and the internal active base region B'.

g_m $V_{b'e}$ (current source) The parameter g_m is called the mutual conductance and is given by

$$g_m = \left.\frac{\partial I_C}{\partial V_{B'E}}\right|_{CE} \tag{8.76}$$

Any small change, $v_{b'e}$, in the emitter–base junction produces a corresponding change in the small-signal collector current. This results in the current source indicated by $g_m V_{b'e}$ in Fig. 8.15.

Resistance $r_{b'e}$ Any small change in the base–emitter forward–bias voltage produces a corresponding change in the emitter current. This, in turn, produces a change in the minority carrier base recombination current. This effect is included in the hybrid-pi model as resistor $r_{b'e}$.

Capacitance C_e The excess minority carrier storage in the base region results in the diffusion capacitance C_e between the base region B' and emitter E.

Resistance $r_{b'c}$ Any change in the base-collector reverse-bias voltage produces a corresponding change in the depletion width of the base-collector junction. This results in a change in the effective base width. The phenomenon that results in a dependence of the effective base width on the collector voltage is called *Early effect* and is referred to as *base width modulation.* Any change in the effective base width produces a corresponding change in the emitter and collector currents. This is because a change in the base width results in a change in the slope of the minority carrier distribution in the base. This feedback effect between base (input) and collector (output) is represented by resistance $r_{b'c}$.

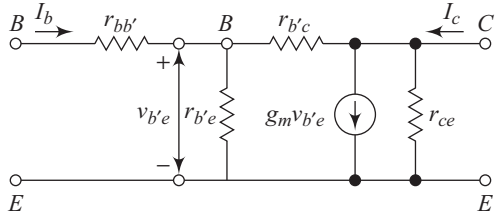


Fig. 8.16 Hybrid-pi model at low frequency

Resistance r_{ce} The resistance existing between the collector and the emitter is represented by r_{ce}.

Capacitance C_c The collector junction barrier capacitance is represented by C_c in the hybrid-pi model.

At low frequencies, all capacitances are negligible, and the hybrid-pi model for the bipolar junction transistor reduces to the one shown in Fig 8.16.

8.7.4 *h*-parameter Equivalent Circuit Model

The *h*-parameter equivalent circuit model considers a transistor to be represented as a two-port network. A general two-port network is shown in Fig. 8.17.

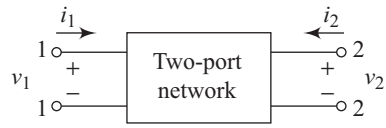


Fig. 8.17 General two-port network

In the figure, v_1, i_1, v_2, and i_2 are the terminal quantities representing voltages and currents. Any pairs of the terminal variables may arbitrarily be considered to be independent. Let us assume i_1 and v_2 to be independent variables. For the other terminal quantities, we can write

$$v_1 = f_1(i_1, v_2) \tag{8.77}$$

$$i_2 = f_2(i_1, v_2) \tag{8.78}$$

Taking total differentials of Eqs (8.77) and (8.78), we get

$$dv_1 = \frac{\partial v_1}{\partial i_1} di_1 + \frac{\partial v_1}{\partial v_2} dv_2 \tag{8.79}$$

and

$$di_2 = \frac{\partial i_2}{\partial i_1} di_1 + \frac{\partial i_2}{dv_2} dv_2 \tag{8.80}$$

If the two-port network shown in Fig 8.15 represents a device working over the linear region of operation with constant slopes, the partial derivatives become constants. Let us at this stage introduce four parameters (h_i, h_r, h_f, and h_o) such that Eqs (8.79) and (8.80) can be rewritten in the forms

$$V_1 = h_i I_1 + h_r V_2 \tag{8.81}$$

$$I_2 = h_f I_1 + h_o V_2 \tag{8.82}$$

where V_1, V_2, I_1, and I_2 represent sinusoidal changes. Out of the four h parameters used in Eqs (8.81) and (8.82), two can be defined by currents and voltages in the form

$$h_i = \frac{V_1}{I_1} = \text{short-circuit input impedance; } V_2 = 0 \tag{8.83}$$

and

$$h_f = \frac{I_2}{I_1} \text{ short-circuit forward current gain, } V_2 = 0 \tag{8.84}$$

If an open circuit is maintained between terminals 1 and 2, we must have $I_1 = 0$. Equations (8.81) and (8.82) then lead to

$$h_r = \frac{V_1}{V_2} = \text{open-circuit reverse voltage gain; } I_1 = 0 \tag{8.85}$$

and

$$h_o = \frac{I_2}{V_2} = \text{open-circuit output admittance; } I_1 = 0 \tag{8.86}$$

The h coefficients are also known as *hybrid parameters*.

A bipolar junction transistor has three elements, namely an emitter, a base, and a collector. To represent a BJT as an h-parameter circuit, therefore, requires one of the basic transistor elements to be treated as common. As already discussed, this gives rise to three configurations, namely common-emitter (CE), common-base (CB), and common-collector (CC). The corresponding h parameters use a second subscript to indicate the particular configuration. For example, the parameter h_f is represented by h_{fe}, h_{fb}, and h_{fc} in the common-emitter, common-base, and common-collector configurations, respectively.

Figure 8.18 shows the common-emitter configuration of a transistor, with its corresponding parameter equivalent circuit.

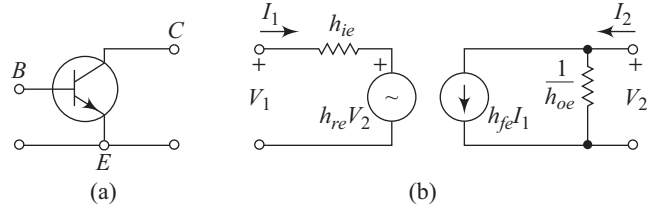


Fig. 8.18 (a) Common-emitter configuration, (b) equivalent circuit

8.8 Important BJT Configurations

In this section, we will use the h-parameter model to derive some important relations valid for specific configurations of transistors. One important application of bipolar junction transistors is as amplifiers. We will use typical amplifier circuits to derive these important relations.

8.8.1 Common-emitter Amplifier

Figure 8.19 shows the diagram of a common-emitter amplifier using a *p-n-p* transistor alongwith the relevant *h*-parameter equivalent circuit.

One must, however, remember that we must use the equivalent circuit for small input signal levels. The dc bias sources have been removed in the ac equivalent circuit since ac voltage cannot appear across dc bias sources. The output blocking capacitor C_B offers negligible reactance and therefore negligible signal voltage appears across it. This is why the blocking capacitor has been removed from the *h*-parameter equivalent circuit.

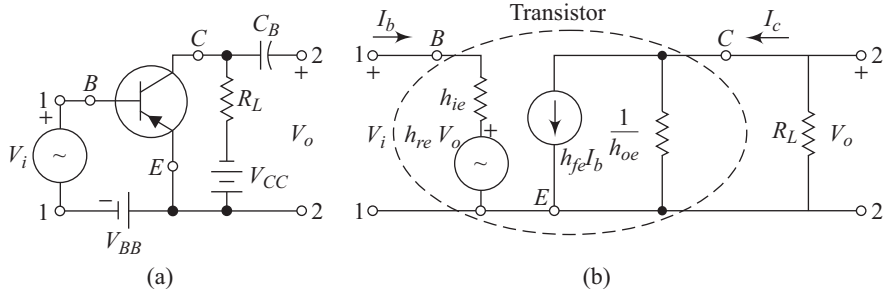


Fig. 8.19 (a) Common-emitter amplifier (b) *h*-parameter equivalent circuit

Let us use V_i and V_o for the rms input and output voltages, respectively. From Fig. 8.19(b), we can write

$$V_i = h_{ie}\, I_b + h_{re}\, V_o \tag{8.87}$$

and

$$I_C = h_{fe}\, I_b + h_{oe}\, V_o \tag{8.88}$$

The operating frequency is considered low enough so that internal capacitive reactances are negligible. With the direction assigned to I_C in Fig. 8.19, we can write

$$I_C = -\frac{V_o}{R_L} \tag{8.89}$$

where R_L is the load resistor. Using Eqn (8.89) in Eqn (8.88) leads to

$$\frac{-V_o}{R_L} = h_{fe} I_b + h_{oe} V_o$$

yielding

$$-V_o - h_{oe}\, R_L\, V_o = h_{fe}\, I_b\, R_L$$

which implies

$$-I_b = \frac{(h_{oe} R_L + 1)\, V_o}{h_{fe} R_L} \tag{8.90}$$

Substituting the expression for I_b from Eqn (8.90) into Eqn (8.87) yields

$$V_i = -\frac{h_{ie}V_o(h_{oe}R_L + 1)}{h_{fe}R_L} + h_{re}V_o$$

$$= V_o \frac{[-h_{ie}(h_{oe}R_L + 1) + h_{re}]}{(h_{fe}R_L)}$$

This gives

$$\frac{V_o}{V_i} = A_{ve} = \frac{-1}{\left[\frac{h_{ie}(h_{oe}R_L + 1)}{h_{fe}R_L}\right] - h_{re}} \tag{8.91}$$

Usually, the product $h_{oe}R_L$ can be neglected in comparison to unity in Eqn (8.91). This implies

$$A_{ve} = \frac{V_o}{V_i} = \frac{-1}{[h_{ie}/(h_{fe}R_L)] - h_{re}} \tag{8.92}$$

The generator $h_{re}V_0$ in Fig. 8.19 represents a feedback effect, which makes input quantities slightly dependent upon the properties of the output circuit. Usually this effect is small so that we can assume $h_{re} = 0$. Thus Eqn (8.92) further reduces to

$$A_{ve} = \frac{-h_{fe}R_L}{h_{ie}} \tag{8.93}$$

A simplified *h*-parameter equivalent circuit for the common-emitter amplifier is shown in Fig. 8.20.

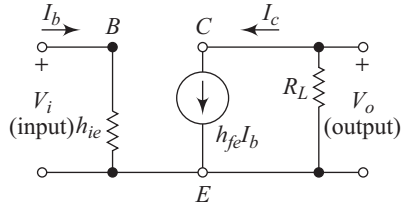


Fig. 8.20 Simplified *h*-parameter *CE* amplifier circuit

The ratio h_{fe}/h_{ie} can be further written in the form

$$\frac{h_{fe}}{h_{ie}} = \frac{|I_c/I_b|}{|V_{be}/I_b|} = \frac{I_c}{V_{be}} \tag{8.94}$$

The ratio h_{fe}/h_{ie}, therefore, relates the current in the output circuit to the voltage in the input. This ratio is called *transconductance* and the symbol g_m is used for it. Thus, Eqn (8.94) can also be rewritten in the form

$$A_{ve} = -g_m R_L \tag{8.95}$$

8.8.2 Common-base Amplifier

A common-base amplifier along with its *h*-parameter equivalent circuit is shown in Fig. 8.21.

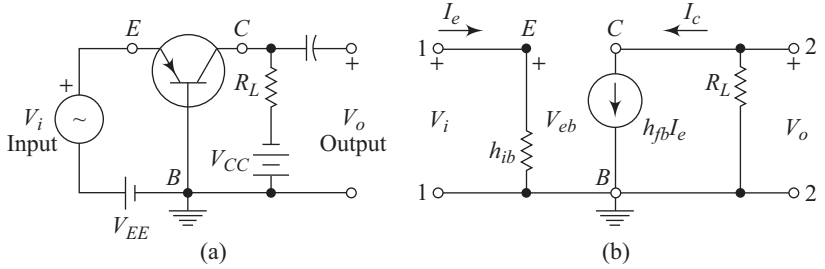


Fig. 8.21 (a) Common-base amplifier circuit (b) *h*-parameter equivalent circuit

Comparing Figs 8.20 and 8.21, we can see that, for the common-base configuration, the input current is I_e instead of I_b in the common-emitter circuit. Thus we can write

$$h_{ib} I_e = h_{ie} I_b \tag{8.96}$$

Also, we can express I_b in the form

$$I_b = \frac{I_e}{1 + h_{fe}} \tag{8.97}$$

Using Eqs (8.92) and (8.93), we can get

$$h_{ib} = \frac{h_{ie}}{1 + h_{fe}} \tag{8.98}$$

By definition,

$$I_C = -\alpha_f I_e \tag{8.99}$$

where α_f is the small-signal collector–emitter current gain. Equation (8.95) can be rewritten in the form

$$I_C = -\frac{h_{fe}}{1 + h_{fe}} I_e \tag{8.100}$$

From Fig. 8.19, we can write

$$V_{eb} = h_{ib}\, I_e = \frac{h_{ie}}{1 + h_{fe}} I_e \tag{8.101}$$

The corresponding transconductance, g_m, becomes

$$g_m = \frac{I_C}{V_{eb}} \tag{8.102}$$

Using Eqs (8.100) and (8.101) in Eqn (8.102) leads to

$$g_m = -\frac{\left[h_{fe}/(1 + h_{fe})\right] I_e}{\left[h_{ie}/(1 + h_{fe})\right] I_e} = \frac{-h_{fe}}{h_{ie}} \tag{8.103}$$

Output voltage V_o can be written as

$$V_o = I_C R_L = g_m V_{eb} R_L \tag{8.104}$$

The voltage gain A_v becomes

$$A_{vb} = \frac{V_o}{V_i} = \frac{V_o}{V_{eb}}$$

Using V_o from Eqn (8.104) yields

$$A_{vb} = g_m R_L \tag{8.105}$$

8.8.3 Common-collector Amplifier

A common-collector amplifier with its equivalent circuit incorporating g_m is shown in Fig. 8.22.

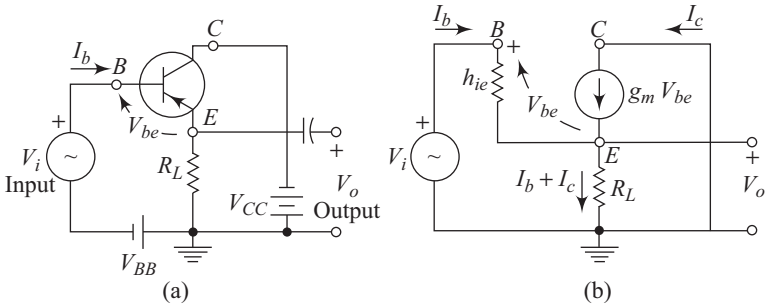


Fig. 8.22 (a) Common-collector amplifier, (b) equivalent circuit

From Fig. 8.22 (a), the voltage equation for the input loop becomes

$$V_i - V_{be} - V_o = 0 \tag{8.106}$$

When the input signal V_{be} is small or the gain is large, Eqn. (8.106) simplifies to

$$V_i \cong V_o \tag{8.107}$$

Thus the output voltage equals or *follows* the input voltage. This is why the circuit is called an *emitter follower*.

The voltage gain A_{vc} for this amplifier configuration is given by

$$A_{vc} = \frac{h_{fe} R_L}{h_{ie}(1 + g_m R_L)} = \frac{g_m R_L}{(1 + g_m R_L)} \tag{8.108}$$

8.9 Thermal Runaway

The collector current I_C decides the power dissipated in a transistor. At large I_C values considerable power gets dissipated within a transistor leading to heating of the device. The power dissipated is given by the product of I_C and the

base-collector voltage V_{BC}. The dissipated power is due to the scattering collisions encountered by the carriers as they are swept across the base-collector junction depletion region. Power transistors that operate at high currents often have heat sinks attached to the device externally to carry away the heat. Increase in device temperature can lead to changes in the material parameters. In Si and Ge the carrier lifetime τ_p increases with increase in temperature since thermal re-excitation from recombination centres becomes effective as the temperature rises. The reduced i_B leads to an increase in β values. The mobility however changes as $T^{-3/2}$ in the lattice-scattering range. From the Einstein relation Eqn (3.107) it can then be seen that D_p decreases with increase in temperature. The increased transit time leads to a fall in β. Of these two competing processes, the effect due to increase in lifetime is more dominating; therefore, the net effect of increase in temperature is an increase in β value. The increased β results in a higher I_C for a given base current, thereby raising the temperature of the device further. The cascading effect is called thermal runaway and can cause a permanent failure of the device.

8.10 Kirk Effect

Another interesting effect takes place at high collector currents. At high currents, a large number of mobile carriers can build up in the base and the collector regions. For a *p-n-p* transistor these mobile carriers are holes injected from the heavily doped emitter region. A schematic representation of the consequence of presence of these holes is shown schematically in Fig. 8.23. Figure 8.23(i) shows the situation at low and moderate collector currents, whereas Fig. 8.23(ii) depicts the situation at high collector currents.

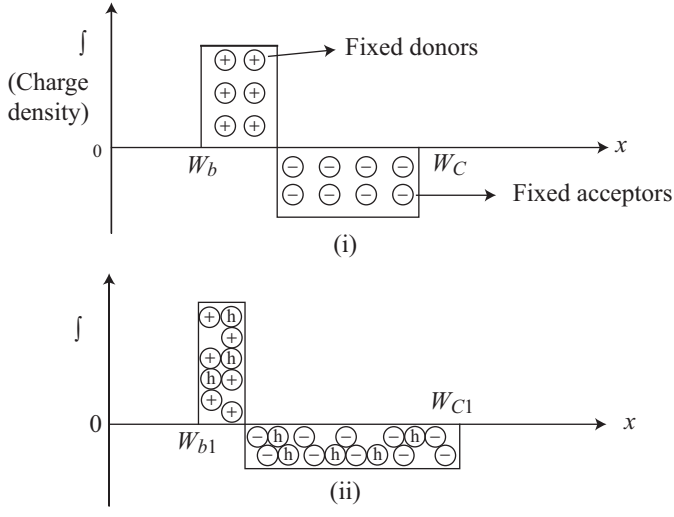


Fig. 8.23 Space charge distribution for a *p-n-p* transistor at (i) moderate currents (ii) high currents

The polarity of mobile charges in the base region adds to the charge due to fixed donor charges, whereas charge due to mobile charges gets subtracted from the charge due to the fixed acceptors in the collector region's depletion region. To maintain the reverse bias across the base–collector junction, a smaller depletion width is thus required in the base region and a larger depletion width in the collector region as shown in Fig. 8.23(ii). The effective widening of the neutral base region is called the *Kirk effect*. The Kirk effect leads to a drop in the current gain and an increase in the base transit time.

8.11 Frequency Limitation for Transistor

Transit time and junction charging times play important roles in the operation of a bipolar junction transistor. Any change in the base–emitter voltage V_{BE} leads to a corresponding change in the carriers injected from the emitter into the base region. For example, if V_{BE} needs to increase, additional carrier need to be injected from the emitter into the base region. These excess carriers then diffuse through the base region, suffer some recombinations there, and finally get collected in the collector region.

The total delay time τ_D between the emitter and the collector consists of the following four time constants:

$$\tau_D = \tau_e + \tau_b + \tau_d + \tau_c \tag{8.109}$$

where τ_e is the charging time of emitter–base junction capacitance, τ_b is the transit time in the base region, τ_d is the transit time in the collector depletion region, and τ_c is the charging time of the collector capacitance. Charging time, τ_e, of the emitter–base junction capacitance is given by

$$\tau_e = R_e(C_{je} + C_p) \tag{8.110}$$

where R_e is the diffusion resistance of the emitter junction, C_{je} is the emitter junction capacitance, and C_p represents all the parasitic capacitances between the base and the emitter.

The transit time in the base region, τ_b, is expressed in the form

$$\tau_b = \frac{W_B^2}{2D_n} \tag{8.111}$$

where W_B is the width of the base region. The collector depletion region transit time, τ_d is given by

$$\tau_d = \frac{W_{\text{dc}}}{v_S} \tag{8.112}$$

where W_{dc} is the BC junction space charge width and v_S is the electron saturation velocity. The last time constant, τ_c, representing the charging time of the collector capacitance is given by

$$\tau_c = R_c\,(C_{BC} + C_S) \tag{8.113}$$

where C_{BC} is the BC junction capacitance, R_C is the collector region series resistance, and C_S is the collector-to-substrate capacitance. Generally R_C is very small for the epitaxial-layer based transistor and thus τ_c is negligible.

Let us now try to physically understand the effect of increasing the frequency on the junction transistor performance. As the frequency of the input signal applied to a transistor increases, the transit time of the carriers from the emitter to the collector, becomes comparable to the time period of the input signal. When this happens, the output is no more in phase with the input signal. This leads to a corresponding decrease in the magnitude of the current gain.

The cut-off frequency of a transistor, f_T, is given as

$$f_T = \frac{1}{2\pi\tau_D} \tag{8.114}$$

When the signal frequency equals the cut-off frequency, the magnitude of the common–emitter current gain equals unity.

8.12 Webster Effect

At low injection levels, the transit time in the base region is given by Eqn. (8.111). At high injection levels, τ_b gets reduced by up to a factor of 2. As the emitter injects a high minority carrier concentration into the base region, the majority carrier concentration also increases significantly above the equilibrium value. The minority carrier concentration decreases as we go from the base-emitter region to the base-collector junction. The majority carrier concentration also changes accordingly. This results in diffusion of majority carriers from the emitter to base. This diffusion of majority carriers has the potential to upset the drift-diffusion balance that maintains a quasi-equilibrium distribution in the base region. To cancel this effect, a built-in electric field develops in the base region that creates an opposing majority carrier drift current component. This built-in field assists the minority carrier transport from the emitter towards the collector region, thereby resulting in the reduction of the transit time τ_b. This causes the Webster Effect.

8.13 High-frequency Transistors

Transistors used in high-frequency circuitry are designed keeping in view the following basic guidelines:

(i) The physical size of the device should be kept small.
(ii) The base width should be kept low to reduce transit time.
(iii) The emitter and collector areas should be kept small to reduce junction capacitance.

As is often the case in technology, there are trade-offs involved. If transistors have to handle high power, they need to be large. This obviously contradicts the design guidelines for the high-frequency transistor. Interdigitation is often resorted to

increase the useful emitter edge while keeping the overall limitation on the emitter area. Figure 8.24 shows the typical interdigitated transistor geometry used in microwave transistors.

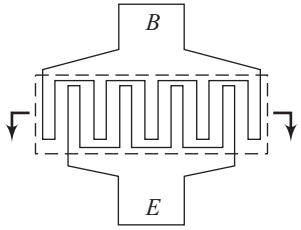


Fig. 8.24 Interdigitated transistor geometry

The operating frequency of a transistor is dependent upon the RC time constants associated with the emitter, base, and collector regions. The resistances associated with each of these regions are to be minimized to reduce the corresponding RC time constants. Metallization schemes and patterns are chosen to reduce the series resistance. The series base resistance associated with an *n-p-n* device can be lowered by incorporating a p^+ diffusion region between the surface contact area and the active base region. It can be further reduced by increasing the doping level and using heterojunction by sandwiching the systems such as AlGaAs/GaAs. Such a transistor is called a *heterojunction bipolar transistor* (HBT). The energy band diagram of a typical HBT is shown in Fig. 8.25. The barrier for electron injection ($e\phi_n$) is seen to be much smaller than the barrier for hole injection ($e\phi_p$). This is achieved by ensuring that the emitter is made of a wider band gap material. The reason behind boosting electron injection and suppressing hole injection is the higher mobility of electrons in comparison with that of holes. It is for the same reason that Si *n-p-n* transistors are generally preferred to *p-n-p* transistors. This is because the mobility and diffusion coefficients of electrons are higher than those of holes.

Packaging also plays a vital role in the design of high-frequency transistors. An optimum package reduces the parasitic effect of resistance, inductance, and capacitance at higher frequency.

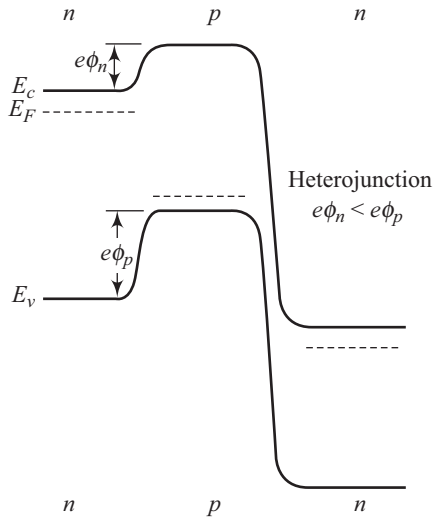


Fig. 8.25 Energy band diagram of HBT

8.14 Switching Characteristics of BJT

A transistor can go through a switching operation in some applications. Typically a switching operation consists of taking a transistor from the cut-off mode to the saturation mode and back. You would recollect that in the cut-off mode the BE and BC junctions are both generally reverse biased. In the saturation mode, on the other hand, both BE and BC junctions are forward biased. A typical common-emitter configuration for an *n-p-n* transistor is shown schematically in Fig. 8.26.

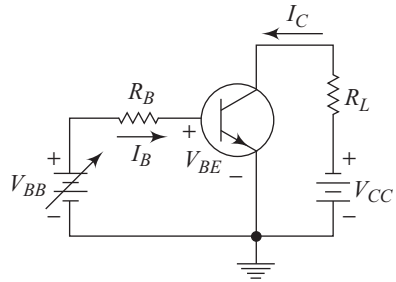


Fig. 8.26 Common-emitter configuration for *n-p*-n transistor

Using Kirchhoff's loop law in the CE loop, we get

$$V_{CE} = V_{CC} - I_C R_L \tag{8.115}$$

This linear relationship between the collector current and collector–emitter voltage is often referred to as the *load line*. Figure 8.27 shows the output characteristics along with the load line for a typical bipolar transistor.

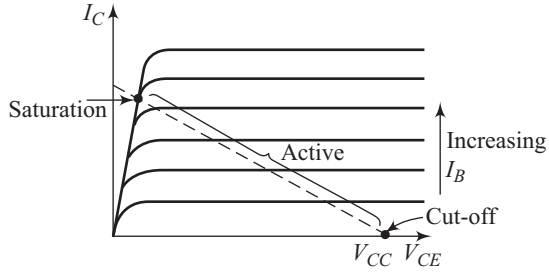


Fig. 8.27 Output characteristics of transistor with load line

The three most important modes of the operation of a transistor, namely *saturation mode*, *active mode*, and *cut-off mode*, are also shown in Fig. 8.27. The variable nature of V_{BB} in Fig. 8.26 indicates a base power supply that can be altered. The input drive voltage required to carry out a typical switching operation is shown in Fig. 8.28.

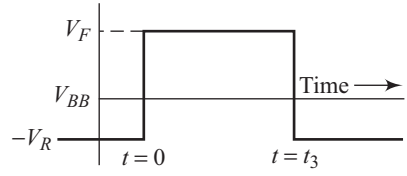


Fig. 8.28 Input drive voltage for switching a transistor

Due to the applied input drive, the *n-p-n* transistor shown in Fig. 8.26 is initially in the cut-off mode. At $t = 0$, the voltage V_{BB} switches to V_F and takes the transistor eventually to the saturation mode. Figure 8.29 shows the dependence of collector current on time. Different important time intervals are also indicated on this figure. We will discuss these time intervals one by one.

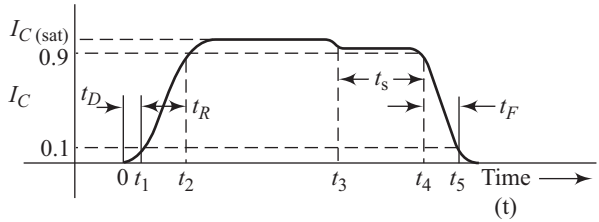


Fig. 8.29 Collector current as a function of time lapsed

$\mathbf{0 \le t \le t_1}$ In this time period, the base current supplies the charge to take the base–emitter junction from reverse bias to slight forward bias. The collector current rises to 10 per cent of its final value during this time period. This is also referred to as the delay time t_D.

$\mathbf{t_1 \le t \le t_2}$ The base current during this period supplies charge to drive the base–emitter junction from a near-cut-off condition to a near-saturation condition. The collector current increases from 10–90 per cent of its value. The time period is therefore referred to as the rise time t_R.

$\mathbf{t > t_2}$ The base current drives the transistor into saturation and establishes the final minority carrier distribution throughout the transistor.

$\mathbf{t_3 \le t \le t_4}$ The base current reverses its direction at the beginning of this time interval. The reverse base current pulls the excess stored charge carriers from the emitter and base regions. In the saturation mode, both base–emitter and the base–collector junctions are forward biased. The time taken for the complete removal of the stored charge in the base is called *storage time* t_s. During the storage time, the collector current reduces to 90 per cent of its maximum saturation value.

$\mathbf{t_4 \le t \le t_5}$ During this time interval, the collector current drops from 90 per cent to 10 per cent of its saturation value. The base–collector junction is reverse biased, but the stored excess charge carriers in the base are still being removed. This time period is called the fall time t_F.

8.14.1 Schottky Transistor

In switching applications, all attempts are to be made to reduce the storage time t_s of the bipolar junction transistor. One method of achieving this is by incorporating a Schottky diode between the base and collector terminals of a normal *n-p-n* transistor. A diagram of such a Schottky-clamped transistor (sometimes referred to simply as Schottky transistor) is shown in Fig. 8.30 (a). Figure 8.30(b) shows the corresponding symbolic representation.

In the active mode, the base–collector junction is reverse biased. This also ensures the Schottky diode is also reverse biased and hence effectively disengages from the circuit. When the bipolar transistor is driven to saturation, the base–collector junction gets forward biased and so does the Schottky diode. The effective turn-on voltage of a Schottky diode is around half of that for a *p-n* junction. Thus, the excess base current gets shunted through the Schottky diode. This leads to a reduction in the excess stored charge in the base and collector regions of the bipolar

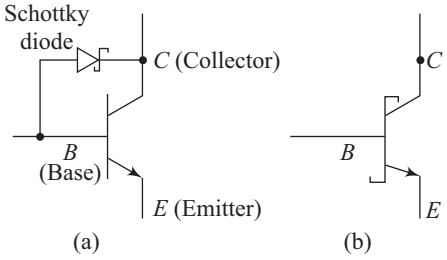


Fig. 8.30 (a) Schottky-clamped transistor; (b) Symbolic representation

transistor. This is due to decrease in the base–collector junction voltage, which is now decided by the turn-on voltage of the Schottky diode. The reduced excess stored charge also results in a reduction in the value of storage time t_s. Values in the region of a few nanoseconds are common for such transistors.

- Gummel–Poon model is very accurate for modelling but requires around 25 parameters.
- Microwave transistors are characterized by using scattering parameters such as S_{11}, S_{22}, S_{12} and S_{21}. S_{11} and S_{22} are input and output reflection coefficients, respectively. S_{21} and S_{12} are forward and reverse transmission gains, respectively.

Solved Problems

8.1 The output characteristics of a particular transistor are shown is Fig. 8.1.1. Calculate the common-base current gain at point Q for the transistor.

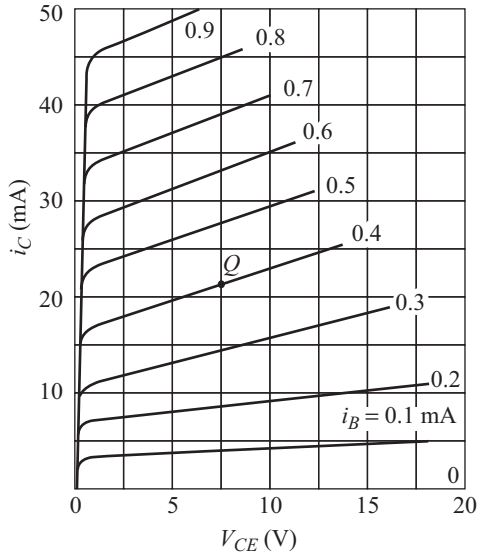


Fig. 8.1.1 Output characteristics of a transistor

Solution

As can be seen from Fig. 8.1.1, the output characteristics consist of plots of collector current i_C as a function of collector–emitter voltage V_{CE} for different base currents i_B. Common-base current gain α is given by

$$\alpha = \frac{i_C}{i_E} \tag{8.1.1}$$

Putting the given values in Eqn (8.1.1), we get

$$\alpha = \frac{21.0}{21.0 + 0.4} = \frac{21.0}{21.4} = 0.98$$

8.2 A silicon *n-p-n* transistor consists of the emitter and base regions uniformly doped to 5×10^{18} cm^{-3} and 5×10^{16} cm^{-3}, respectively. The base-emitter forward bias $V_{BE} = 0.6$ V. The neutral base width $W_B = 3$ μm and the minority carrier electron diffusion length in the base region is $L_n = 15$ μm. Calculate the excess minority carrier concentration in the base region at (a) $x = W_B/3$ and (b) $x = 0$.

Solution

(a) $$n_{p0} = \frac{n_i^2}{5 \times 10^{16}} = \frac{(1.5 \times 10^{10})^2}{5 \times 10^{16}} = 4500 \text{ cm}^{-3} \tag{8.2.1}$$

The excess minority carrier concentration, $\delta n(x)$, is given by

$$\delta n(x) \simeq \frac{n_{p0}}{W_B}\{(e^{eV_{BE}/kT} - 1)\,(W_B - x) - x\} \tag{8.2.2}$$

At $x = W_B/3$, Eqn (8.2.2) yields

$$\delta n\left(x = \frac{W_B}{3}\right) = \frac{4.5 \times 10^3}{3 \times 10^{-4}}\left\{(e^{0.6/0.026} - 1)\left(W_B - \frac{W_B}{3}\right) - \frac{W_B}{3}\right\}$$

This can be simplified to

$$\delta n\left(x = \frac{W_B}{3}\right) = \frac{4.5 \times 10^3}{3 \times 10^{-4}}\left\{(e^{0.6/0.026} - 1)\left(\frac{2}{3}W_B\right) - \frac{W_B}{3}\right\}$$

Substituting the given value of W_B results in

$$\delta n\left(x = \frac{W_B}{3}\right) = 1.5 \times 10^7\left\{(e^{0.6/0.026} - 1)\left(\frac{2}{3} \times 3 \times 10^{-4}\right) - 1 \times 10^{-4}\right\}$$

$$= 3.15 \times 10^{13} \text{ cm}^{-3}$$

(b) At $x = 0$, we get

$$\delta n\,(x = 0) = 1.5 \times 10^7\{(1.05 \times 10^{10})\,(3 \times 10^{-4})\}$$
$$= 4.73 \times 10^{13} \text{ cm}^{-3}$$

8.3 A bipolar transistor has the following significant parameters at $T = 300$ K.

$$\alpha_F = 0.98, \quad \alpha_R = 0.18, \quad I_C = 2 \text{ mA}, \quad I_B = 60 \text{ μA}$$

Calculate the collector–emitter voltage at saturation.

Solution

The collector–emitter voltage at saturation $V_{CE(sat)}$ is given by

$$V_{CE}(\text{sat}) = \frac{kT}{e} \ln \left[\frac{I_C(1-\alpha_R) + I_B}{\alpha_F I_B - (1-\alpha_F) I_C} \left(\frac{\alpha_F}{\alpha_R} \right) \right] \tag{8.3.1}$$

Putting the given values in Eqn (8.3.1) yields

$$V_{CE(\text{sat})} = 0.026 \ln \left[\frac{2(1-0.18) + 0.06}{(0.98)(0.06) - (1-0.98)2} \left(\frac{0.98}{0.18} \right) \right]$$

$$V_{CE(\text{sat})} = 0.16 \text{ V}$$

8.4 A common-emitter amplifier circuit uses a load resistor $R_L = 3$ kΩ. The bipolar transistor used has the following parameters: $h_{ie} = 1$ kΩ, $h_{fe} = 25$, $h_{re} = 2 \times 10^{-4}$ and $h_{oe} = 15 \times 10^{-6}$ mho.

Evaluate transconductance g_m and voltage gain A_{ve}.

Solution

Transconductance g_m is given by

$$g_m = \frac{h_{fe}}{h_{ie}} \tag{8.4.1}$$

Putting the given values in Eqn (8.4.1)

$$g_m = \frac{25}{1 \times 10^3} = 0.025 \text{ mho}$$

The voltage gain, A_{ve}, can be evaluated using the Eqn

$$A_{ve} = - g_m R_L \tag{8.4.2}$$

Substitution of proper values into Eqn (8.4.2) results in

$$A_{ve} = - 0.025 \times 3 \times 10^3 = - 75$$

8.5 A particular *n-p-n* transistor has the following parameters at 300 K:

$I_E = 1.5$ mA $\quad C_{je} = 1.2$ pF
$W_B = 0.4$ μm $\quad D_n = 25$ cm^2/s
$W_{dc} = 2.5$ μm $\quad R_C = 25$ Ω
$C_{BC} = 0.15$ pF $\quad C_S = 0.12$ pF

Calculate the total emitter-to-collector delay time, τ_D, and the cut-of frequency of the transistor. Assume saturation velocity $v_s = 10^7$ cm/s.

Solution

The emitter–base junction charging time, τ_e, is given by

$$\tau_e = R_e C_{je} \tag{8.5.1}$$

where the resistance R_e is expressed as

$$R_e = \frac{kT}{e} \frac{1}{I_E} = \frac{0.026}{1.5 \times 10^{-3}} = 17.3 \, \Omega \tag{8.5.2}$$

Using the value of R_e from Eqn (8.5.2) in Eqn (8.5.1) results in

$$\tau_e = 17.3 \times 1.2 \times 10^{-12} = 20.8 \text{ ps} \tag{8.5.3}$$

The transit time τ_b in the base is given by

$$\tau_b = \frac{W_B^2}{2D_n} \tag{8.5.4}$$

Putting the given values in Eqn (8.5.4) yields

$$\tau_b = \frac{(0.4 \times 10^{-4})^2}{2 \times 25} = 32 \text{ ps}$$

The collector depletion region transit time, τ_d, is given by

$$\tau_b = \frac{W_{\text{dc}}}{v_s} \tag{8.5.5}$$

Putting the relevant values in Eqn (8.5.5) results in

$$\tau_d = \frac{2.5 \times 10^{-4}}{10^7} = 2.5 \times 10^{-11} \text{ s} = 25 \text{ ps}$$

The collector capacitance charging time, τ_c, is expressed as

$$\tau_c = R_c\,(C_{BC} + C_S) \tag{8.5.6}$$

Substituting the relevant values, we get

$$\tau_c = 25(0.15 \times 10^{-12} + 0.12 \times 10^{-12}) = 6.75 \text{ ps} \tag{8.7.5}$$

Finally, the total emitter-to-collector delay time is given by

$$\tau_D = 20.8 + 32 + 25 + 6.75 = 84.6 \text{ ps} \tag{8.5.8}$$

The cut-off frequency f_T is expressed as

$$f_T = \frac{1}{2\pi\tau_D} = \frac{1}{2\pi(84.6 \times 10^{-12})} = 1.88 \text{ GHz}$$

8.6 For the *p-n-p* transistor biased as shown in Fig. 8.6.1 draw and explain the *I-V* characteristic when *V* is swept from nominal negative values to nominal positive values. (Assume equal reverse saturation currents for both the junctions).

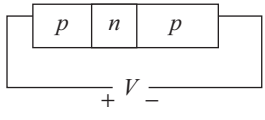
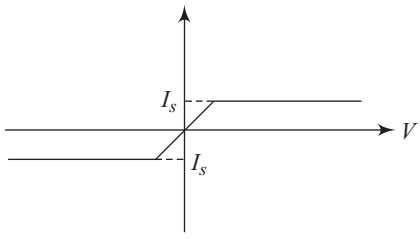


Fig. 8.6.1 *p-n-p* transistor biasing

Solution

Here, I_s is the reverse saturation current. When voltage V is swept from small negative values to small positive values, the only current which flows in the device is a small reverse saturation current. Beyond some particular values of positive and negative voltages, breakdown may occur, but we will not be taking that into consideration for this particular problem. With suitable selection of doping levels for the *p*, *n*, and *p* regions, we can arrive at another useful electronic device, i.e., a diac.



8.7 (a) For an Si *p-n-p* transistor biased in the active region with $\gamma = 1$, find the approximate upper frequency limit, if the frequency response is dominated by transit time delay. (Width of the base region = 0.5 μm, $D_p = 15\ \text{cm}^2/\text{s}$)

(b) What changes do you suggest in the device mentioned above, so as to increase its upper frequency limit?

(c) With the suggested changes, will any compromises be made in the operation of this device? Explain briefly.

Solution

(a) It can be derived from Eqn (8.111) that, for a *p-n-p* transistor, transit time in the base region, τ_n, would be given by

$$\tau_n = \frac{W_b^2}{2D_p}$$

Substituting the values for W_b and D_p in the above equation, we get

$$\tau_n = \frac{(0.5\,\mu\text{m})^2}{2(15\ \text{cm}^2/\text{s})}$$

$$= \frac{0.5 \times 0.5 \times 10^{-6} \times 10^{-6}}{2 \times 15 \times 10^{-4}}$$

$$= \frac{0.25}{30} \times 10^{-8}\ \text{s}$$

$$= 8.33 \times 10^{-11}\ \text{s} = 83.3\ \text{ps}$$

Now, ignoring the delay caused by movement of carriers in the emitter and the collector, i.e., τ and τ_C, and taking into account only τ_B to calculate the upper frequency limit,
we get

$$f_T = \frac{1}{2\pi\tau_B} \qquad \text{[From Eqn (8.114)]}$$

$$= \frac{1}{2 \times 3.14 \times 83.3 \times 10^{-12}}$$

$$= \frac{10^{12}}{2 \times 3.14 \times 83.3} = \frac{1000}{523.12} \times 10^9$$

$$= 1.91\ \text{GHz}$$

(b) To increase the upper frequency limit, the following must be adopted:

(i) Physical size of the device should be kept small.

(ii) Base width should be kept small to reduce transit time.

(iii) Base, emitter, and collector areas should be kept small to reduce junction capacitance.

(c) If one reduces the physical size of the device, then its power handling capacity will also be affected. So, generally a careful optimization must be made by the device engineer between power handling capacity and device size.

8.8 Doping profiles of two transistors are shown in Fig. 8.8.1

(a) Which profile will you select for designing a prototype transistor. Why?

(b) As a designer, your objective is to design a BJT which can operate up to a high frequency. Out of these two profiles, which one will you use for your design and why?

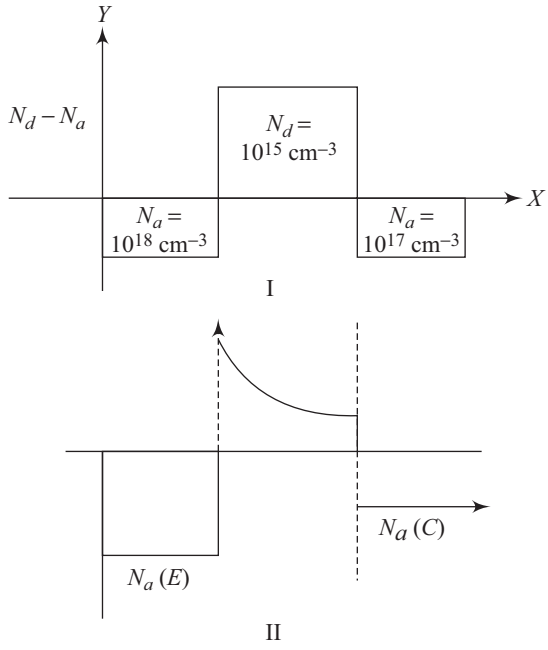


Fig. 8.8.1 Doping profiles of two transistors

Solution

As can be clearly observed, the second doping profile shows non-uniform doping in the base region.

(a) Profile I should be selected, as it is uniformly doped in the base region.

(b) Profile II should be selected for designing a BJT to be used at high frequencies because of lesser transit time and an increased value of β.

Recapitulation

- In a bipolar junction transistor, the collector current is controlled by the base–emitter voltage.
- CE, CB, and CC are three important configurations used in circuit involving transistors.
- Common-base current gain, α, is given by

$$\alpha = \frac{i_C}{i_E}$$

- Common-emitter current gain, β, is expressed as

$$\beta = \frac{i_C}{i_B}$$

- Minority carrier concentration, $\delta n(x)$, in the base region, if we assume $W_B < L_n$, is given by

$$\delta n(x) = \frac{n_{p0}}{W_B}\left\{[e^{eV_{BE}/kT} - 1]\,(W_B - x) - x\right\}$$

- Minority carrier concentration, $\delta p(x)$, in the emitter region for a thin emitter is given by

$$\delta p(x) = \frac{p_{n0}}{W_E}[e^{eV_{BE}/kT} - 1]\,(W_E - x)$$

- Minority carrier concentration, $\delta p(x)$, in the collector region is expressed in the form

$$\delta n(x) = -p_{n0}\, e^{-x/L_p}$$

- Using the Ebers–Moll equivalent circuit, the collector–emitter voltage at saturation is given by

$$V_{CE(\text{sat})} = \frac{kT}{e}\ln\left[\frac{I_C(1-\alpha_R) + I_B}{\alpha_F I_B - (1-\alpha_F)I_C}\left(\frac{\alpha_F}{\alpha_R}\right)\right]$$

- According to the Gummel–Poon model, the electron current density, J_n, in the base region is given by

$$J_n = \frac{-qD_n n_i^2 e^{eV_{BE}/kT}}{\int_0^{W_B} p(x)\,dx}$$

- Voltage gian, A_{ve}, in the common-emitter configuration is given by

$$A_{ve} = \frac{-h_{fe}R_L}{h_{ie}} = -g_m R_L$$

- Voltage gain, A_{vb}, in the common-base configuration is expressed as

$$A_{vb} = g_m R_L$$

- Total delay time, τ_D, from the emitter to the collector is given by

$$\tau_D = \tau_e + \tau_b + \tau_d + \tau_c$$

where $\tau_e = R_e\,(C_{je} + C_p)$, $\tau_b = W^2{}_B / 2D_n$, $\tau_d = W_{dc}/v_S$, and $\tau_c = R_c\,(C_{BC} + C_S)$.

- Cut-off frequency f_T of a transistor is obtained using the expression

$$f_T = \frac{1}{2\pi\tau_D}$$

- Early effect leads to an increase in β with applied base-collector reverse bias.

- For CB configuration,

  $$V_{EB} = f_1(I_E, V_{CB})$$

  and

  $$I_C = f_2(I_E, V_{CB})$$

- For CE configuration,

  $$V_{BE} = f_1(I_B, V_{CE})$$

  and

  $$I_C = f_2(I_B, V_{CE})$$

- Kirk effect leads to a drop in the current gain and an increase in the base transit time.

Exercises

Review Questions

8.1 Give the symbolic representations for *n-p-n* and *p-n-p* transistors.
8.2 Discuss the motion of electrons from the emitter to the collector in detail.
8.3 Describe common-emitter, common-base and common-collector configurations.
8.4 Define common-base current gain.
8.5 Derive expressions for the minority carrier distributions in the base region of an *n-p-n* transistor.
8.6 Draw the Ebers–Moll equivalent circuit for a bipolar junction transistor.
8.7 Use the Gummel–Poon model to derive an expression for the electron current density in the base region of a transistor.
8.8 Draw the hybrid-model equivalent circuit for the common–emitter configuration.
8.9 Draw and discuss the *h*-parameter equivalent circuit for the common–emitter configuration of a bipolar transistor.
8.10 Differentiate between common–emitter, common–base, and common–collector configurations.
8.11 Define cut-off frequency of a transistor.
8.12 Plot the typical dependence of collector current on time for a bipolar transistor and explain its important regions.
8.13 How does a Schottky transistor reduce the total storage time?
8.14 Draw a schematic diagram showing different current components in a bipolar transistor.
8.15 Show the three important configurations of a bipolar transistor.
8.16 Define common-base current gain and common-emitter current gain. (Give their measurement units.
8.17 Derive expressions for minority carrier distribution in the emitter and collector regions of a bipolar transistor.

8.18 Explain the important parameters used in the Ebers–Moll equivalent circuit of BJT.
8.19 Why is the Gummel–Poon model a more accurate model for a transistor? What is the significance of the Gummel number?
8.20 Explain the different components of the hybrid-pi model for CE configuration of a BJT.
8.21 Define h_i, h_r, h_f, and h_o for a general two-port circuit.
8.22 Derive an expression for the cut-off frequency of a bipolar transistor.
8.23 What is a load line of a transistor?
8.24 Define the following terms: rise time, storage time, fall time.
8.25 Draw the excess minority carrier concentration profile in all the three regions of an *n-p-n* transistor for the following:
(a) Forward active region
(b) Reverse active region
(c) Saturation region
(d) Cut-off region
8.26 How does choosing a heterojunction at the emitter–base end of a BJT ensure the enhancement of emitter injection efficiency? Explain with the help of a diagram.
8.27 What is Early effect?
8.28 Explain punch-through condition.
8.29 Explain active or normal mode of a BJT.
8.30 What is the difference between cut-off mode and saturation mode of a BJT.
8.31 Give a graphical representation of the input characteristics of BJT in the common-base configuration.
8.32 Sketch the Input and Output characteristics of a pnp transistor in common-Emitter configuration.
8.33 What is the other name for common-collector configuration?
8.34 Explain the process of thermal runaway.
8.35 Explain Kirk effect.
8.36 Describe Webster effect.

Problems

8.1 In the output characteristics of a transistor, $i_B = 25\ \mu A$ and $i_c = 1.27$ mA. Calculate the common-base current gain for the particular values of i_B and i_C.

$$\left[Hint: \alpha = \frac{i_C}{i_E} \right]$$

Ans. 0.98.

8.2 A germanium transistor has the Q point corresponding to $i_B = 30\ \mu A$ and $i_c = 1.3$ mA. Calculate the common-emitter current gain for the given Q point.

$$\left[Hint: \beta = \frac{i_C}{i_B} \right]$$

Ans. 43.33

8.3 The emitter and base regions of an *n-p-n* silicon transistor are doped uniformly to levels of 10^{19} cm^{-3} and 5×10^{16} cm^{-3}, respectively. The base-emitter forward bias $V_{BE} = 0.55$ V and the neutral base width $W_B = 2.5$ μm. The minority carrier electron diffusion length, L_n, in the base is 16 μm. Determine the excess minority carrier concentration in the base region at $x = W_B/2$.

$$\left[\textit{Hint: } \delta n(x) \approx \frac{n_{p0}}{W_B} \{ [e^{eV_{BE}/kT} - 1] (W_B - x) - x \} \right]$$

Ans. 3.46×10^{12} cm^{-3}

8.4 A silicon *n-p-n* bipolar transistor has the emitter and base regions uniformly doped to 5×10^{18} cm^{-3} and 5×10^{16} cm^{-3}, respectively. A forward bias of 0.6 V is applied to the base-emitter junction. The neutral emitter region width, W_E, is 3 μm. Calculate the excess minority carrier concentration at the emitter edge of the emitter-base space charge region.

$$\left[\textit{Hint: } \delta p(x) = \frac{p_{n0}}{W_E} \left[\exp\left(\frac{eV_{BE}}{kT} \right) - 1 \right] (W_E - x) \right]$$

Ans. 4.74×10^{11} cm^{-3}

8.5 An *n-p-n* bipolar transistor is biased such that it is in the active region. Calculate the minority carrier concentration with respect to the thermal equilibrium value at $x = 2L_p$ in the collector region.

$$\left[\textit{Hint: } \delta p(x) = - p_{n0} e^{-x/L_p} \right]$$

Ans. 86% of thermal equilibrium value

8.6 At what point in the collector region of an *n-p-n* transistor does the minority carrier concentration reach 90% of its thermal equilibrium value. Assume that the transistor is operating in the active mode.

Ans. $x = 2.3L_p$

8.7 A particular bipolar transistor has the following important parameters at $T = 300$ K. Obtain the collector–emitter voltage under saturation conditions.
$\alpha_F = 0.99$, $\alpha_R = 0.17$, $I_C = 3$ mA, $I_B = 70$ μA.

$$\left[\textit{Hint: } V_{CE(\text{sat})} = \frac{kT}{e} \ln \left[\frac{I_C (1 - \alpha_R) + I_B}{\alpha_F (I_B) - (1 - \alpha_F) I_C} \left(\frac{\alpha_F}{\alpha_R} \right) \right] \right]$$

Ans. 0.154 V

8.8 A load resistor R_L of 2.5 kΩ is used in a common-emitter amplifier circuit. The transistor used in the circuit has the following parameters:
$h_{ie} = 1.2$ kΩ, $h_{fe} = 26$,
$h_{re} = 3 \times 10^{-4}$, $h_{oe} = 14 \times 10^{-6}$ mho
Calculate the transconductance g_m and the voltage gain A_{ve} for the circuit.

$$\left[\textit{Hint: } g_m = \frac{h_{fe}}{h_{ie}}; A_{ve} = - g_m R_L \right]$$

Ans. 0.022 mho, −55

8.9 The transistor mentioned in Problem 8.8 is to be used in a common-base amplifier circuit. The load resistor R_L used is of 2 kΩ. Calculate the voltage gain, A_{vb}.
[*Hint:* $A_{vb} = g_m R_L$]

Ans. 44

8.10 An *n-p-n* transistor has the following parameters at 300 K:
$I_E = 2$ mA, $C_{je} = 1.5$ pF
$W_B = 0.5$ μm $D_n = 25$ cm^2/s
$W_{dc} = 2.6$ μm $R_C = 30$ Ω
$C_{BC} = 0.17$ pF $C_S = 0.13$ pF
Calculate the cut-off frequency of the transistor assuming saturation velocity $v_s = 10^7$ cm/s.

$$\left[\textit{Hint: } f_T = \frac{1}{2\pi\tau_D} \right]$$

Ans. 1.52 GHz

8.11 Consider a *p-n-p* bipolar transistor with an emitter doping of 10^{18} cm^{-3} and a base doping of 10^{17} cm^{-3}. The quasi-neutral region width in the emitter is 1 μm and in the base 0.2 μm. Use $\mu_n = 1000$ cm^2/Vs and $\mu_p = 300$ cm^2/Vs. The minority carrier lifetime in the base is 10 ns. Calculate the emitter efficiency, the base transport factor, and the current gain of the transistor biased in the forward active mode. Assume that there is no recombination in the depletion region. Calculate the saturation voltage of a bipolar transistor biased with a base current of 1 mA and a collector current of 10 mA. Use $\alpha_R = 0.993$ and $\alpha_F = 0.2$.

Ans. $\gamma = 0.94$, $\beta = 11.5$

8.12 Consider a typical *n-p-n* transistor. Transit time for the electrons across the base of this transistor is 100 ps, the width of the depletion region towards the collector is 1 μm, and electrons cross this region with their scattering limited velocity. Due to the capacitance action prevailing at emitter-base junction, the total charging time of this capacitance is around 30 ps. If the collector capacitance and resistance values are given to be 0.1 pF and 10 Ω, respectively, find the cut-off frequency of this BJT.

Ans. 1.1 GHz

8.13 The common-base current gain for a particular operating condition is 0.97. If the collector current is 20 mA, calculate the base current.

Ans. 0.62 mA

8.14 The transconductance of a bipolar transistor, g_m is 0.03 *mho*. If $h_{fe} = 28$, calculate h_{ie}.

Ans. 933.3 Ω

9 Field-effect Transistor

- Junction-field-effect transistor
- Metal–semiconductor field-effect transistor
- Basic MOS structure
- Capacitance–voltage characteristics of MOS
- Interface traps and oxide charge
- MOSFET characteristics
- Types of MOSFET

Learning Objectives

After going through this chapter the student will be able to

- understand the operating principles of the JFET and the MESFET
- understand the *I-V* characteristics of the JFET
- derive expressions for transconductance of the MESFET in different regions of its *I-V* characteristics
- differentiate between normally off and normally on MESFETs
- understand the operation of the HEMT
- solve numericals based on the *I-V* characteristics of the JFET and the MESFET
- understand the important characteristics of a basic MOS capacitor
- derive the expressions for the work-function difference and the maximum space charge width of the MOS structure
- derive the expressions for C/C_{ox} and C_{min} for an MOS capacitor
- understand the effect of interface traps and oxide charge on *C-V* characteristics of the MOS structure
- understand the operating principle of the MOSFET
- differentiate between enhancement type and depletion type MOSFETs
- solve numericals based on the characteristics of an MOS capacitor
- solve numericals based on the *I-V* characteristics of the MOSFET

Introduction

The previous chapter dealt with bipolar junction transistors (BJT). Another popular transistor is the field-effect transistor. Several variations of the field-effect transistor (FET) are available. Some of these are metal-oxide–semiconductor (MOSFET), junction (JFET), and metal–semiconductor (MESFET) FET. If the oxide layer is replaced by a general insulator, the corresponding FET is called MISFET (metal–insulator–semiconductor FET). Unlike the BJT, an FET is a majority carrier device having three terminals. Voltage is the controlling factor in an FET. Technology enables the growth of a native oxide layer on the surface of silicon; this has proved to be a milestone in the development of the MOSFET. This functional MOSFET was first demonstrated by Kahng and Atalla in 1960. The stupendous growth of integrated circuits can, to a large extent, be attributed to the Si MOSFET. FET will be the subject of discussion in this chapter. Apart from the fundamentals of the common types of FETs, the chapter will offer an insight into the capacitance–voltage characteristics of the MOS structure.

9.1 Junction-field-effect Transistor

A junction-field-effect transistor (JFET) consists of a conducting channel with two ohmic contacts on either end. The conducting channel can either be an *n*-type or a *p*-type semiconductor. Accordingly, the JFET is called either an *n*-channel or a *p*-channel JFET. One of the ohmic contacts is called the source and the other is called the drain. A schematic diagram of an *n*-channel JFET is shown in Fig. 9.1.

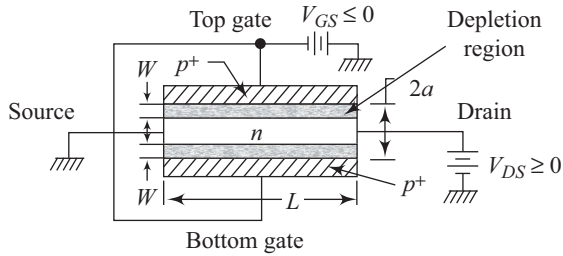


Fig. 9.1 Schematic diagram of an *n*-channel JFET

Two p^+ doped regions on opposite sides of the *n*-channel complete the JFET structure. In actual operation, the two *p*-*n* junctions are reverse biased. *The magnitude of reverse bias applied is used to modulate the cross-sectional area used for current flow*. We must also remember that only one type of carrier is used in this device, thus making it unipolar.

Figure 9.2 shows the symbolic representation of an *n*-channel JFET with appropriate applied bias.

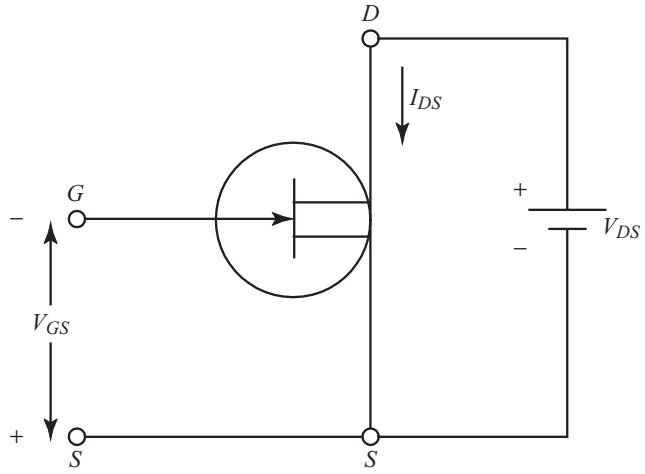


Fig. 9.2 Symbolic representation of an *n*-channel JFET with appropriate applied bias

9.1.1 Operating Principle

Let us consider a JFET with channel length L, channel width W, and channel depth $2d$. A schematic representation of such a JFET is shown in Fig. 9.3.

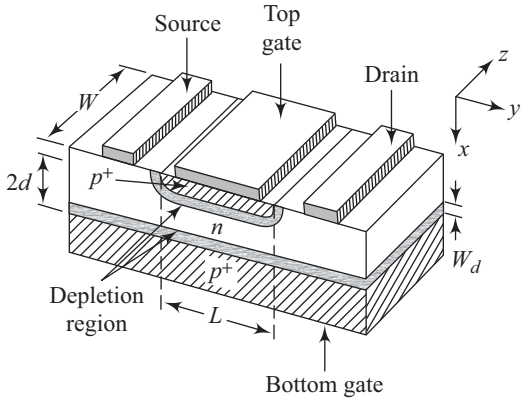


Fig. 9.3 Detailed layout of the JFET

Now let us assume that the top and bottom gates are connected together. The source terminal is kept grounded and the gate voltage, V_{GS}, and drain voltage, V_{DS}, are both measured with respect to the source. Under normal operating conditions we have

$$V_{GS} \leq 0 \quad \text{and} \quad V_{DS} \geq 0 \tag{9.1}$$

Resistance R of the channel is given by

$$R = \rho \frac{L}{A} \tag{9.2}$$

where ρ is the resistivity of the channel which, in this case, is of n-type. Substituting the expression for conductivity σ in Eqn (9.2) leads to

$$R = \frac{L}{e\mu_n N_d A} \tag{9.3}$$

where N_d is the donor concentration of the channel. If W_d represents the width of the depletion region due to the two reverse biased junctions, then the area A of the channel is given by

$$A = 2W(d - W_d) \tag{9.4}$$

Substituting Eqn (9.4) in Eqn (9.3) leads to

$$R = \frac{L}{2e\mu_n N_d W(d - W_d)} \tag{9.5}$$

Figure 9.4 shows the change in output characteristics with an increasing V_{DS} (I_D versus V_{DS}).

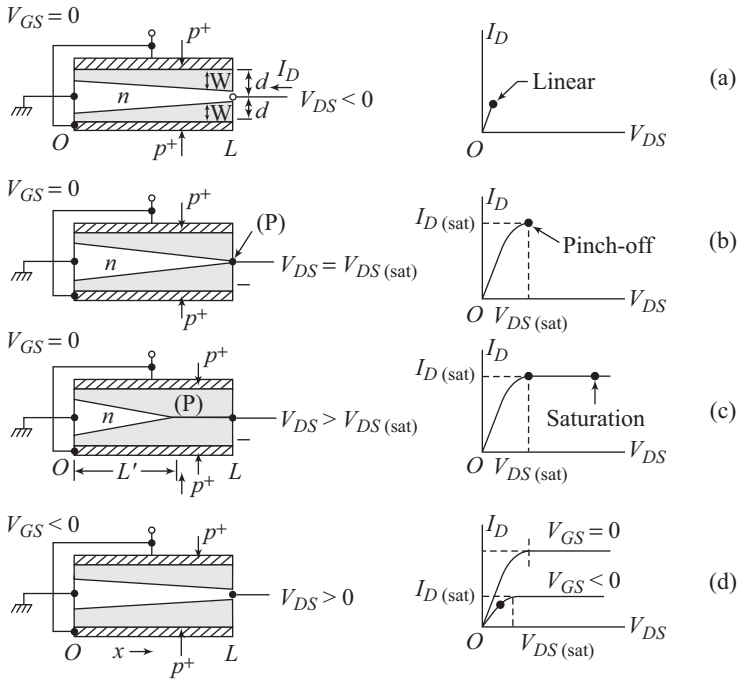


Fig. 9.4 Output characteristics of the JFET for changing V_{DS} values

The following conclusions can be drawn from Fig. 9.4:

- Figure 9.4(a) shows the I-V characteristics for $V_{GS} = 0$ and small V_{DS} values. One must understand at the outset that, for non-zero values of V_{DS}, the reverse bias across the p^+-n junction is not constant but increases progressively as one goes from source to drain. Therefore the depletion region width increases, reducing the effective channel width, as we go from source to drain. As the

voltage V_{DS} increases, the effective channel cross-sectional area decreases, thus leading to an increasing channel resistance, R. This slows down the rise in drain current with V_{DS}. This portion of the output curve is called the *linear region* [Fig. 9.4(a)].

- The increase in depletion region width with an increasing V_{DS} ultimately reaches a level where the two depletion regions touch each other at the drain terminal. At $V_{DS} = V_{DS(\text{sat})}$, the depletion region width, W_d, becomes equal to d at the drain. From our understanding of the p^+-n junction this results in

$$V_{DS(sat)} = \frac{eN_d d^2}{2\varepsilon_s} - V_{\text{bi}} \quad \text{for } V_{GS} = 0 \tag{9.6}$$

where V_{bi} is the built-in potential for the gate-channel p-n junction. This is shown as the pinch-off point in Fig. 9.4(b). A drain current, designated as $I_{D(\text{sat})}$, flows through the depletion region into the drain terminal. This current is analogous to the collector-base junction current in the case of a bipolar junction transistor as discussed in the preceding chapter. In practice, however, the channel width does not become zero but has a small minimum value δ. This is because a zero channel width would lead to infinite current density. The drift velocity v_d is given by $v_d = \mu_n \varepsilon$, where ε is the electric field. At extremely small channel widths, the electric field attains very high values and at such high fields μ_n is inversely proportional to the electric field. The drift velocity therefore remains constant as a function of applied voltage and the current saturates.
- As V_{DS} is increased further, the pinch-off point, P, shifts towards the source as shown in Fig. 9.4(c). The voltage at the pinch-off point remains constant at $V_{DS(\text{sat})}$. The potential drop in the channel, between the source and the pinch-off point remains unchanged. This constant potential drop ensures that the number of electrons per unit time travelling from the source terminal to the pinch-off point, P, remains constant. This results in a constant current beyond the pinch-off point as shown in Fig. 9.4(c).
- Figure 9.4(d) shows a comparison of the output characteristics for $V_{GS} = 0$ and $V_{GS} < 0$. When $V_{GS} < 0$, the depletion region width, W_d, is more than the corresponding magnitude with $V_{GS} = 0$, for any value of V_{DS}. This results in a comparatively lower cross-sectional area of the channel available for conduction. This ensures a larger value of resistance and thus a lower value of drain current for any given value of V_{DS}. This effect is demonstrated in Fig. 9.4(d). The corresponding $V_{DS(\text{sat})}$ is given by

$$V_{DS\,(\text{sat})} \quad \frac{eN_d d^2}{2\varepsilon_s} - V_{\text{bi}} - V_{GS} \tag{9.7}$$

One must remember that the absolute value of V_{GS} must be used in Eqn (9.7) for an n-channel JFET. An immediate consequence of Eqn (9.7) is a reduction in the magnitude of V_{DS} at which the onset of pinch-off takes place. This reduction in V_{DS} is, in fact, equal to the magnitude of V_{GS}.

9.1.2 Current–Voltage Characteristics

Figure 9.5 is a schematic diagram of an *n*-channel JFET in the pre-pinch-off state.

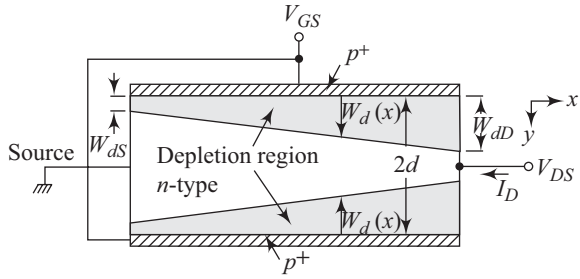


Fig. 9.5 An *n*-channel JFET in the pre-pinch-off condition

Let us consider an elemental section of the channel of length *dx*. The voltage drop, *dV*, across this section is given by

$$dV = I_D dR = \frac{I_D dx}{2e\mu_n N_d W[d - W_d(x)]} \tag{9.8}$$

The depletion layer width at a distance *x* from the source is given by

$$W_d(x) = \sqrt{\frac{2\varepsilon_s[V(x) + V_{GS} + V_{\text{bi}}]}{eN_d}} \tag{9.9}$$

From Eqn (9.8) the product of a constant I_D and *dx* is expressible as

$$I_D dx = 2e\mu_n N_d W[d - W_d(x)]\, dV \tag{9.10}$$

Equation (9.9) leads to

$$[W_d(x)]^2 = \frac{2\varepsilon_s[V(x) + V_{GS} + V_{\text{bi}}]}{eN_d}$$

$$\therefore \quad eN_d[W_d(x)]^2 = 2\varepsilon_s[V(x) + V_{GS} + V_{\text{bi}}] \tag{9.11}$$

Differentiating Eqn (9.11) results in

$$dV = \frac{eN_d}{\varepsilon_s} W_d d(W_d) \tag{9.12}$$

Substituting Eqn (9.12) into Eqn (9.10) gives

$$I_D dx = 2e\mu_n N_d W[d - W_d(x)] \frac{eN_d}{\varepsilon_s} W_d d(W_d) \tag{9.13}$$

Integrating Eqn (9.13) over the entire channel length, *L*, yields

$$I_D = \frac{1}{L} \int_{W_{dS}}^{W_{dD}} 2e\mu_n N_d W[d - W_d(x)] \frac{eN_d}{\varepsilon_s} W_d d(W_d) \tag{9.14}$$

$$I_D = \frac{W\mu_n e^2 N_d^2}{\varepsilon_s L}\left[d\left(W_{dD}^2 - W_{dS}^2\right) - \frac{2}{3}\left(W_{dD}^3 - W_{dS}^3\right)\right] \tag{9.15}$$

Using Eqn (9.9) we can write

$$W_{dD}\sqrt{\frac{2\varepsilon_s[V_{DS} + V_{GS} + V_{\text{bi}}]}{eN_d}} \tag{9.16}$$

And

$$W_{dS} = \sqrt{\frac{2\varepsilon_s[V_{GS} + V_{\text{bi}}]}{eN_d}} \tag{9.17}$$

Substituting Eqs (9.16) and (9.17) into Eqn (9.15) yields

$$I_D = I_p\left[\frac{V_{DS}}{V_P} - \frac{2}{3}\left(\frac{V_{DS} + V_{GS} + V_{\text{bi}}}{V_p}\right)^{3/2} + \frac{2}{3}\left(\frac{V_{GS} + V_{\text{bi}}}{V_p}\right)^{3/2}\right] \tag{9.18}$$

where

$$I_p = \frac{W\mu_n e^2 N_d^2 d^3}{\varepsilon_s L} \tag{9.19}$$

and

$$V_p = \frac{eN_d d^2}{2\varepsilon_s} \tag{9.20}$$

Voltage V_P is called the *pinch-off voltage*. V_P is the total voltage at which $W_{dD} = d$, i.e., $(V_{DS} + V_{GS} + V_{\text{bi}})$.

Figure 9.6 is a plot of ideal current–voltage characteristics of a JFET with $V_P = 3.5$ V at $V_{GS} = 0$. For negative values of V_{GS}, V_P will shift towards lower values as already discussed.

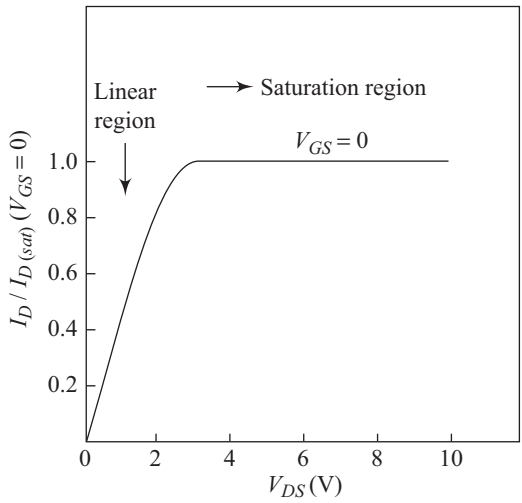


Fig. 9.6 Ideal *I-V* characteristics of JFET for $V_{GS} = 0$

Two regions of the *I-V* curve deserve special mention.

Linear region

For $V_{DS} \ll V_{GS} + V_{\text{bi}}$ Eqn (9.18) reduces to

$$I_D \simeq \frac{I_P}{V_P}\left[1 - \sqrt{\frac{V_{GS} + V_{\text{bi}}}{V_P}}\right] V_{DS} \tag{9.21}$$

Channel or drain conductance, g_D, is defined as

$$g_D = \left.\frac{\partial I_D}{\partial V_{DS}}\right|_{V_{GS}=\text{const}} \tag{9.22}$$

Using Eqn (9.21) for I_D in Eqn (9.22) gives us

$$g_D = \frac{I_P}{V_P}\left[1 - \sqrt{\frac{V_{GS} + V_{\text{bi}}}{V_P}}\right] \tag{9.23}$$

Transconductance, g_m, of a JFET is defined as

$$g_D = \left.\frac{\partial I_D}{\partial V_{GS}}\right|_{V_{DS}=\text{const}} \tag{9.24}$$

In the linear region we get

$$g_m = \frac{I_P}{2V_P^2}\sqrt{\frac{V_P}{V_{GS} + V_{\text{bi}}}}\, V_{DS} \tag{9.25}$$

Saturation region

In this region

$$V_P = V_{DS} + V_{GS} + V_{\text{bi}} \tag{9.26}$$

Thus, the drain current, $I_{D(\text{sat})}$, becomes

$$I_{D(\text{sat})} = I_p\left[\frac{1}{3} - \left[\frac{V_{GS} + V_{\text{bi}}}{V_P}\right] + \frac{2}{3}\left[\frac{V_{GS} + V_{\text{bi}}}{V_P}\right]^{3/2}\right] \tag{9.27}$$

and the saturation voltage, $V_D(\text{sat})$, is given by

$$V_{D(\text{sat})} = V_P - V_{GS} - V_{\text{bi}} \tag{9.28}$$

From Eqn (9.27) it is clear that $I_{D(\text{sat})}$ is not a function of V_{DS} in the saturation region. Thus the channel conductance is zero in the saturation region. The corresponding transconductance g_m in the saturation region is given by

$$g_m = \frac{I_P}{V_P}\left[1 - \sqrt{\frac{V_{GS} + V_{\text{bi}}}{V_P}}\right] \tag{9.29}$$

which can also be written as

$$g_m = \frac{2W\mu_n eN_d d}{L}\left[1 - \sqrt{\frac{V_{GS} + V_{\text{bi}}}{V_P}}\right] \tag{9.30}$$

9.2 Metal–Semiconductor Field-effect Transistor

A metal–semiconductor field–effect transistor (MESFET) is very similar to the JFET in operational characteristics. The *p-n* junction at the gate electrode of the JFET is replaced with a rectifying metal–semiconductor contact in a MESFET. MESFETs are generally fabricated using *n*-type III–V compound semiconductors, such as gallium arsenide, as epitaxial layers on semi-insulating substrates. Semi-insulating substrates help in reducing parasitic capacitance and high electron mobilities help in minimizing series resistance. The relatively high saturation velocities also result in high cut-off frequencies.

A schematic view of a typical MESFET is shown in Fig. 9.7. The metal–semiconductor junction depletion layer that controls the current in the channel between the source and the drain is also shown in the figure.

9.2.1 Normally-off and Normally-on MESFETs

Some FETs have a conductive channel at $V_{GS} = 0$. These devices are called 'normally-on' FETs. Another kind do not have any conductive channel at $V_{GS} = 0$ and are known as 'normally-off' FETs.

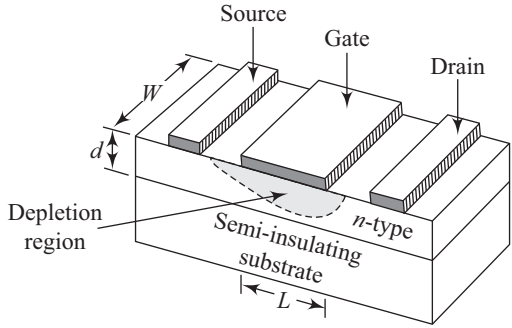


Fig. 9.7 Schematic view of a MESFET

In a normally-off MESFET, the built-in potential (V_{bi}) of the metal–semiconductor junction at the gate is sufficient to deplete the entire channel region. Hence a conductive channel is absent at $V_{GS} = 0$. Such devices are constructed by incorporating a very thin epitaxial layer on semi-insulating substrate, using semiconductors like GaAs. A positive bias must be applied at the gate electrode (Fig. 9.7) for any channel current to flow. *The minimum positive voltage required to be applied to the gate electrode to initialize a channel current is called* the *threshold voltage* V_T. Threshold voltage, V_T, is thus given by

$$V_T = V_{bi} - V_P \tag{9.31}$$

where V_P is the pinch-off voltage given by Eqn (9.20).
From Eqn (9.31) we get

$$V_{bi} = V_T + V_P \tag{9.32}$$

Substituting the expression for V_{bi} from Eqn (9.32) into Eqn (9.27) results in

$$I_{D(\text{sat})} = \frac{I_P}{2}\left\{\frac{1}{3} - \left[1 + \left(\frac{V_{GS} + V_T}{V_P}\right)\right] + \frac{2}{3}\left[1 + \left(\frac{V_{GS} + V_T}{V_P}\right)\right]^{3/2}\right\} \tag{9.33}$$

where the current I_P has been divided by 2 to account for the fact that the MESFET shown in Fig. 9.7 does not have a lower channel. The polarity of V_{GS} for MESFET is opposite to that of a JFET. Incorporating this in Eqn (9.33) leads to

$$I_{D(\text{sat})}\frac{I_P}{2}\left\{\frac{1}{3} - \left[1 - \left(\frac{V_{GS} - V_T}{V_P}\right)\right] + \frac{2}{3}\left[1 - \frac{(V_{GS} - V_T)}{V_p}\right]^{3/2}\right\} \tag{9.34}$$

Assuming $[(V_{GS} - V_T)/V_P] \ll 1$ and using Taylor series expansion in Eqn (9.34) results in

$$I_{D(\text{sat})} \approx \frac{I_P}{2}\left(\frac{1}{4}\right)\frac{(V_G - V_T)^2}{V_P^2} \tag{9.35}$$

Substituting for I_P from Eqn (9.19) and V_P from Eqn (9.20) in Eqn (9.35) yields

$$I_{D(\text{sat})} \approx \frac{W\mu_n\varepsilon_s}{2dL}(V_{GS} - V_T)^2 \tag{9.36}$$

Typical characteristics of the normally-on and normally-off modes of the MESFET's operation are shown in Fig. 9.8.

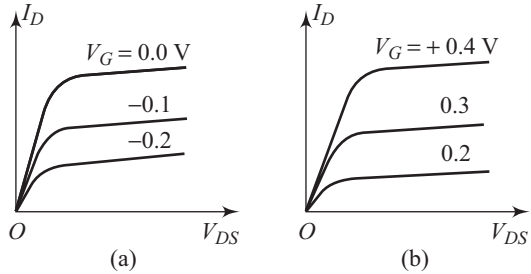


Fig. 9.8 Output characteristic of (a) normally-on and (b) normally-off MESFETS

The basic shape of the *I-V* curve is similar for the normally-on and normally-off modes. A normally-off MESFET requires positive gate voltages to allow reasonable gate currents. The normally-on mode, on the other hand, will allow drain currents to flow even for negative V_{GS} values.

9.2.2 High-electron-mobility Transistor

A high transconductance in a MESFET requires a high channel conductivity. To an extent, the channel conductivity can be increased by increasing the channel

carrier concentration. Increased doping in the channel, however, also leads to increased scattering of charge carriers by the ionized impurities. This, in turn, results in lowering of the effective carrier mobility. Band gap engineering approaches can be used to increase the channel carrier concentration without a corresponding increase in the doping level. One such method is to use a channel consisting of a thin undoped well (e.g., GaAs) bounded on either side by wider band gap, doped barrier semiconductors (e.g., AlGaAs). Such a configuration is called *modulation doping* and is shown schematically in Fig. 9.9.

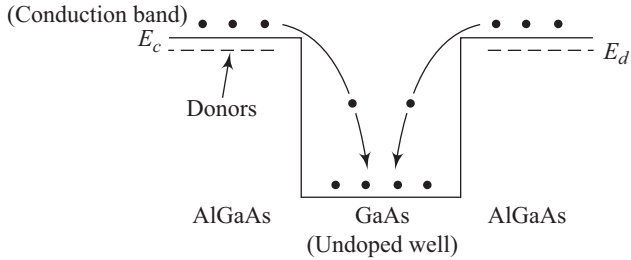


Fig. 9.9 Trapping of electrons in modulation doped structures

As shown in Fig. 9.9, electrons from the conduction band of the doped AlGaAs barrier layer, fall into the well and get trapped there. Since the source of the electrons is the AlGaAs barrier material on either side, there is no impurity scattering of the electrons in the well. This marked reduction in scattering results in higher mobility which, in turn, gets reflected in a higher channel conductivity. This effect is even more pronounced at low temperatures because of low lattice (phonon) scattering. Such a transistor is called a high-electron-mobility transistor (HEMT) or a modulation doped field-effect transistor (MODFET).

Figure 9.9 is a simplified diagram that does not consider the band bending at AlGaAs/GaAs interface. The band bending and its effect on the energy band diagram at the AlGaAs/GaAs interface in shown in Fig. 9.10.

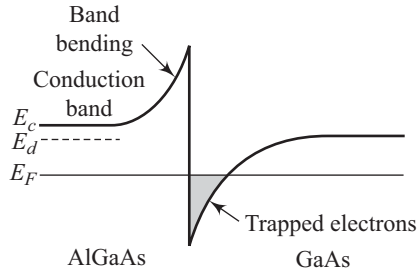


Fig. 9.10 Energy band diagram for an AlGaAs/GaAs system including band bending

Due to the band-bending effect, electrons accumulate at the corners of the well. From Fig. 9.10 it is clear that only one heterojunction is sufficient to trap the electrons. This arrangement can result in extremely high channel electron

concentrations with high mobility since the electrons are spatially separated from the ionized impurities that are the source of these electrons. The potential barrier at the AlGaAs/GaAs interface prevents the electrons from diffusing into the smaller band gap GaAs layer. The trapped electrons in the approximately triangular well are also referred to as a two-dimensional electron gas (a term used to emphasize the fact that the conduction occurs in a thin sheet of charge), or simply as 2-DEG. Mobilities in the region of 300,000 $cm^2/V\,s$ at 77 K and 2,000,000 $cm^2/V\,s$ at 4 K are achievable. These characteristics translate into extremely high cut-off frequencies and fast response times.

We must end this section by pointing out that the AlGaAs/GaAs heterojunction is not the only system used in HEMTs. Other combinations like InGaAsP/InP are equally promising. Because HEMTs very often use extremely thin layers, small lattice mismatches can be tolerated. Such thin semiconductor layers having small lattice mismatches are called *pseudomorphic* and the corresponding HEMTs are also called *pseudomorphic* HEMTs.

- MESFETs encounter short-channel effects when channel lengths of less than 1 μm are involved. There are two main consequences:
  (a) Saturated drain current occurs due to velocity saturation and does not require true pinch-off.
  (b) An effective channel-length reduction is observed after pinch-off. This occurs due to a significant intrusion of the depletion region into the channel length.
- Low noise and reasonable gains have been obtained at operating frequencies around 35 GHz using HEMTs. Cut-off frequencies around 100 GHz are possible using HEMTs with 0.25 μm channel length.
- Cut-off frequency f_T for a JFET is given by

$$f_T = \frac{2\mu_n e N_d d^2}{\pi \varepsilon_s L^2}$$

9.3 Basic MOS Structure

The third type of field-effect transistor that we shall discuss in detail is the MOSFET or the metal-oxide-semiconductor field-effect transistor. This device is extremely important and involves an understanding of the basic MOS structure. In its most fundamental form, the MOS structure (also called MOS capacitor) consists of a semiconductor substrate on which an oxide (silicon dioxide is very common) is deposited. A metal contact on the top of the oxide completes the structure. The top metal layer is also sometimes replaced with a high-conductivity polycrystalline semiconductor. Silicon is a natural choice for such structures due to the ease with which the oxide layer can be obtained. Figure 9.11 shows the basic MOS structure.

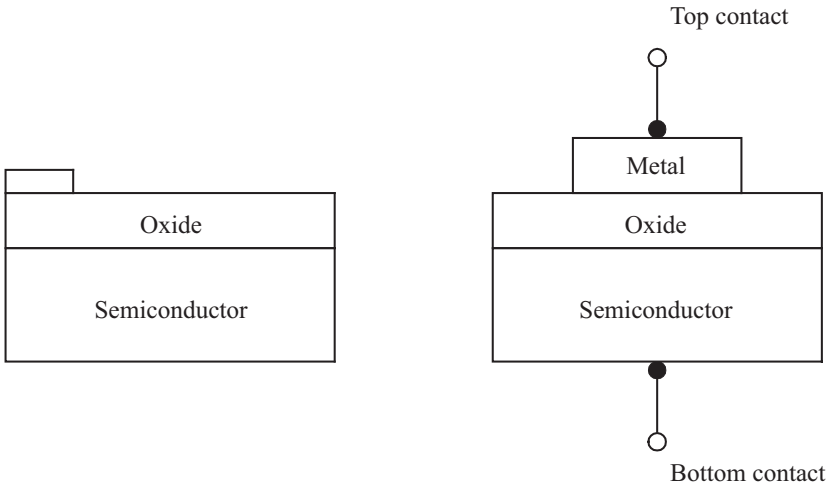


Fig. 9.11 Basic MOS structure

Figure 9.12 shows an MOS capacitor formed using a *p*-type semiconductor. The top metal electrode is given a negative voltage with respect to the semiconductor substrate.

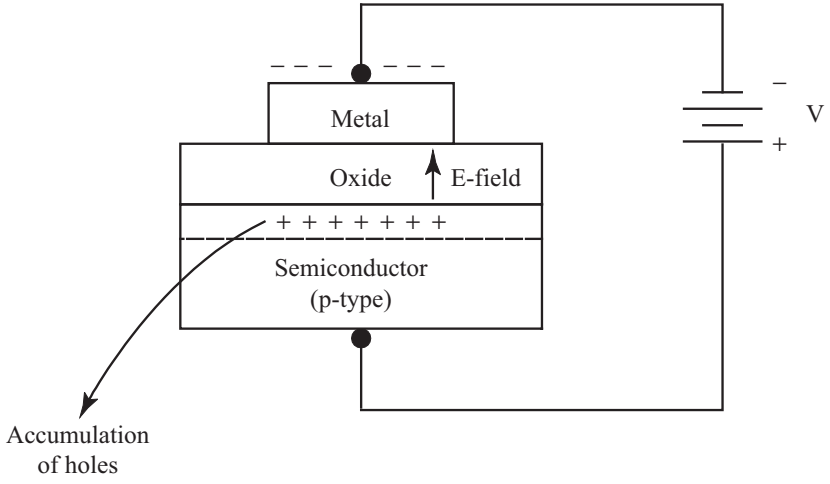


Fig. 9.12 An MOS capacitor with negative voltage on the top electrode

Due to the applied potential, a negative charge exists on the top metal plate. The resulting induced electric field exerts a force on the majority carrier holes and they move towards the oxide–semiconductor interface. An *accumulation layer* of holes develops at the oxide–semiconductor junction. This accumulation layer provides the positive charge for the equilibrium electric-field distribution.

Figure 9.13 shows the MOS capacitor formed on a *p*-type semiconductor but with voltage applied on the top metal plate.

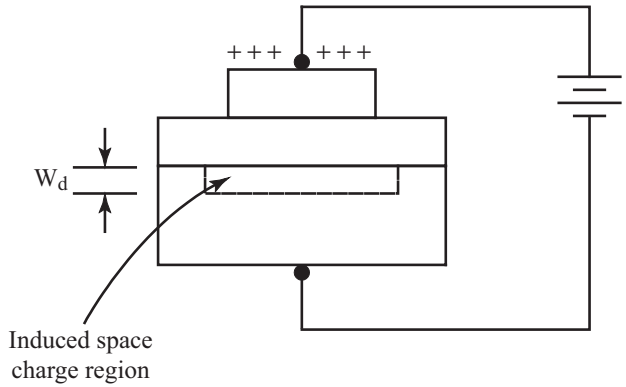


Fig. 9.13 An MOS capacitor using *p*-type semiconductor, with positive voltage applied on the top metal plate

As a result of the applied bias, a positive charge exists on the top metal plate. Due to the resultant electric field, the majority carrier holes experience a force away from the semiconductor–oxide interface. A space charge region consisting of negatively charged, fixed, ionized acceptor atoms is formed in the semiconductor. The resulting energy band diagram is shown in Fig. 9.14(b), for comparison, the band diagram for negative bias on the top plate is shown in Fig. 9.14(a).

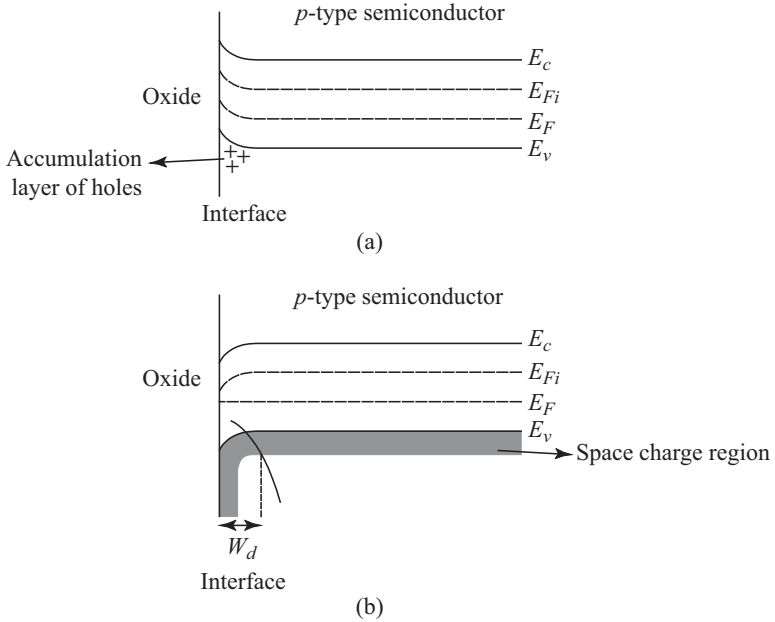


Fig. 9.14 Energy band diagram of MOS structures on a *p*-type semiconductor (a) negative gate voltage (b) positive gate voltage

In Fig. 9.14(b) the conduction and valence band edges bend somewhat like those in a *p*-*n* junction and a space charge region develops. The conduction band and the intrinsic Fermi level bend to move closer to the Fermi energy level. On the other hand, when a negative bias is applied to the metal gate, the conduction band and the intrinsic Fermi level bend away from the Fermi energy level as shown in Fig. 9.14(a). The valence band comes closer to the Fermi level at the semiconductor–oxide interface. This leads to an accumulation of holes as shown in Fig. 9.14(a). The *p*-type majority carrier concentration in the accumulation layer is far higher than that in the bulk of the semiconductor.

The diagram presented in Fig. 9.14(b) is valid for moderate positive voltages applied to the top metal electrode of the MOS capacitor. As the positive voltage applied on the top gate increases, the corresponding positive and negative charges on the MOS capacitor also increase. The greater negative charge in the MOS capacitor results in a larger space charge region accompanied by greater band bending. Ultimately a stage is reached when the intrinsic Fermi level at the surface falls below the Fermi energy level as shown in Fig. 9.15. When this happens, the conduction band becomes closer to the Fermi level than the valence band and the semiconductor region immediately adjacent to the oxide–semiconductor interface effectively becomes *n*-type. The surface of the semiconductor thus gets inverted from a *p*-type to an *n*-type semiconductor.

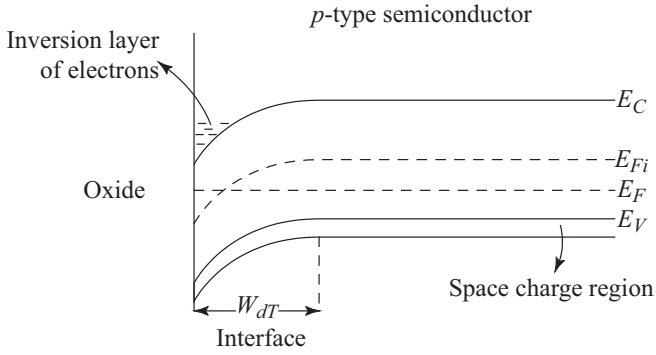


Fig. 9.15 Band diagram showing inversion layer formation

The inversion layer so formed has some interesting consequences which will be discussed later in this chapter.

So far we have assumed the semiconductor to be *p*-type. Similar results are obtained for *n*-type semiconductors with applied voltages of opposite polarity. Figure 9.16 shows the energy band diagram of an MOS formed using an *n*-type semiconductor substrate for different bias polarities.

Due to the positive voltage applied to the metal gate, an accumulation of electrons occurs in the semiconductor. A moderate negative voltage applied to the metal gate results in a positive space charge region. At high negative voltages, the

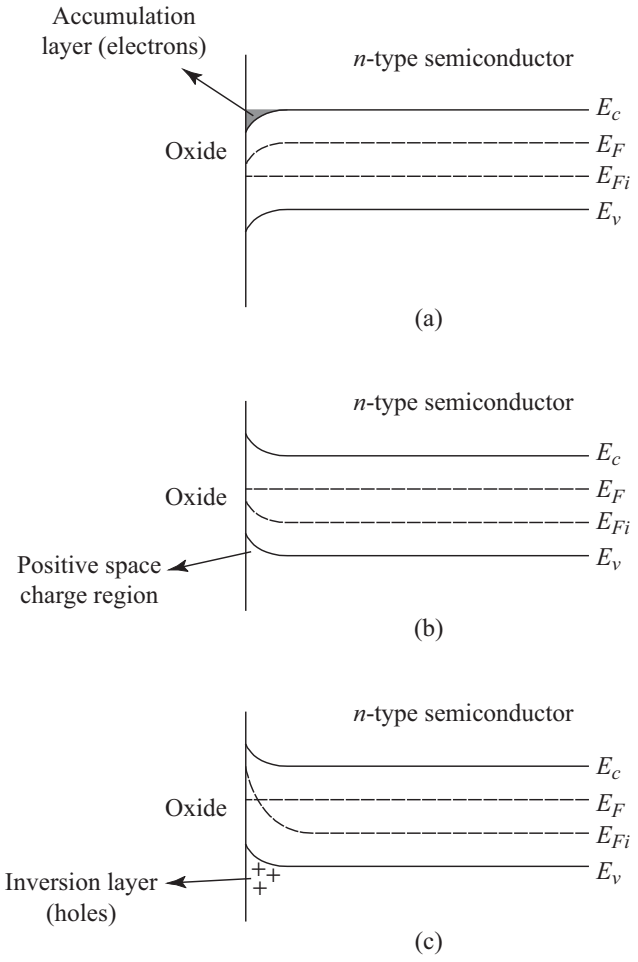


Fig.9. 16 Energy band diagram for the MOS structure on *n*-type semiconductor for different bias conditions

bands bend so much that the intrinsic Fermi level goes above the Fermi energy level, thereby producing an inversion layer. In this state, the valence band is closer to the Fermi level than the conduction band. Thus the semiconductor surface close to the oxide is inverted to *p*-type characteristics from its original *n*-type.

9.3.1 Depletion Layer Thickness

We have seen that a space charge region develops in the semiconductor if the metal gate is given a positive voltage. Figure 9.17 shows a detailed energy band diagram of the MOS structure using a *p*-type semiconductor with the space charge region.

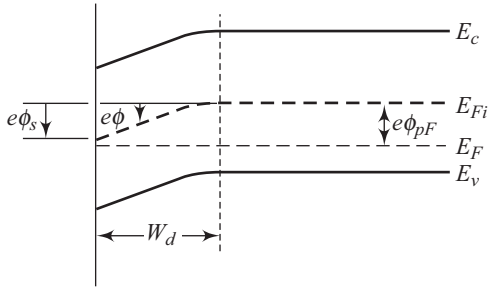


Fig. 9.17 Energy band diagram of MOS structure on *p*-type semiconductor

Let us define a potential, ϕ, as the difference between the intrinsic Fermi Energy, E_{Fi}, in the bulk and its value around the interface. ϕ is zero in the bulk of the semiconductor. As can be seen from Fig. 9.17, surface potential, ϕ_s, is the difference (in volts) between E_{Fi} at the surface of the semiconductor and in its bulk. Potential ϕ_{pF} is the difference (in volts) between the intrinsic Fermi level, E_{Fi}, and the Fermi energy level, E_F. ϕ_{pF} is given by the equation

$$\phi_{pF} = \frac{kT}{e} \ln\left(\frac{N_a}{n_i}\right) \tag{9.37}$$

where N_a is the acceptor doping concentration in the *p*-type semiconductor and n_i is the intrinsic carrier concentration.

The space charge width, W_d, can be written using an expression similar to that for a one-sided abrupt *p-n* junction by replacing V_{bi} with ϕ_S. Thus we get

$$W_d = \left(\frac{2\varepsilon_s \phi_s}{eN_a}\right)^{1/2} \tag{9.38}$$

where ε_S denotes the permittivity of the semiconductor.

An interesting situation arises when $\phi_S = 2\phi_{pF}$ as shown in Fig. 9.18. As we can see the Fermi level at the surface is above the intrinsic Fermi level whereas that in the bulk is below the intrinsic Fermi energy level.

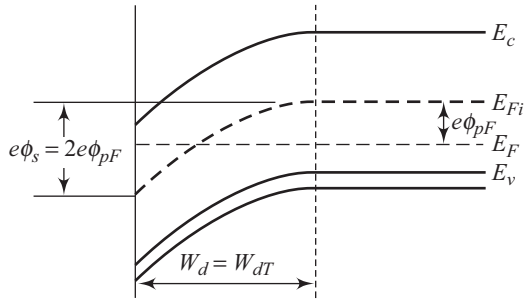


Fig. 9.18 Energy band diagram showing the threshold inversion point

Furthermore, when $\phi_S = 2\phi_{pF}$, the difference between the intrinsic Fermi energy and the Fermi energy is the same at the surface and in the bulk. Thus the minority

carrier electron concentration at the surface becomes equal to the majority carrier hole concentration in the bulk. This is the *threshold inversion point*. A little more voltage applied at the gate will take the MOS capacitor into the inversion mode. This voltage at which $\phi_s = 2\phi_{pF}$ is, therefore, called the *threshold voltage*. The space charge region reaches its maximum width at this point. We can obtain this maximum space charge region width by replacing ϕ_S by $2\phi_{pF}$ in Eqn (9.38). Thus the maximum space charge width W_{dT} is given by

$$W_{dT} = \left(\frac{4\varepsilon_s \phi_{pF}}{eN_a} \right)^{1/2} \tag{9.39}$$

A similar result can also be obtained for *n*-type semiconductors. Figure 9.19 shows the energy band diagram at threshold voltage for an MOS capacitor formed on an *n*-type semiconductor substrate.
Then, we have

$$\phi_{nF} = \frac{kT}{e} \ln \left(\frac{N_d}{n_i} \right) \tag{9.40}$$

where N_d is the donor concentration in the *n*-type semiconductor.

The maximum space charge width, W_{dT}, is given by

$$W_{dT} = \left(\frac{4\varepsilon_s \phi_{nF}}{eN_d} \right)^{1/2} \tag{9.41}$$

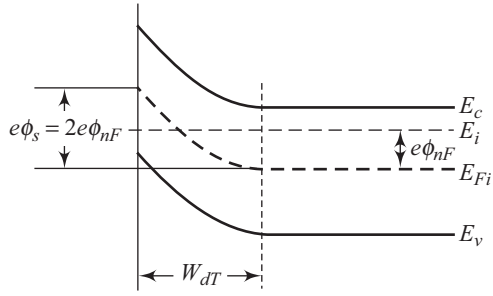


Fig. 9.19 Energy band diagram showing the threshold condition of an MOS capacitor on *n*-type semiconductor

9.3.2 Work-function Difference

Figure 9.20 shows the energy band diagram of an MOS system formed using a *p*-type semiconductor with the three components separated and as a composite system. In the figure, ϕ_m is the work-function and ϕ'_m is the modified work-function of the metal. The modified metal work-function gives the energy required to inject an electron from the metal into the conduction band of the oxide. Similarly χ and χ' are the electron affinity and modified electron affinity, respectively, of the semiconductor. The potential drop across the oxide for zero

gate voltage is designated as $V_{\text{oxide}(0)}$. ϕ_{s0} represents the surface potential for zero applied gate voltage. Equating the difference between the oxide conduction band and the Fermi energy on the metal and semiconductor sides from Fig. 9.20 we get

$$e\phi'_m + eV_{\text{oxide}(0)} = e\chi' + \frac{E_g}{2} - e\phi_{so} + e\phi_{pF} \tag{9.42}$$

Equation (9.42) can be rewritten in the form

$$V_{\text{oxide}} + \phi_{s0} = \left[\phi'_m - \left(\chi' + \frac{E_g}{2e} + \phi_{pF}\right)\right] \tag{9.43}$$

The quantity $[\phi'_m - (\chi' + E_g/2e + \phi_{pF})$ is called the metal–semiconductor work-function difference and is denoted by the symbol ϕ_{ms}.

For an ideal MOS diode formed on a *p*-type semiconductor at $V = 0$, $\phi_{mS} = 0$. Energy band diagram for an ideal situation is shown in Fig. 9.21.

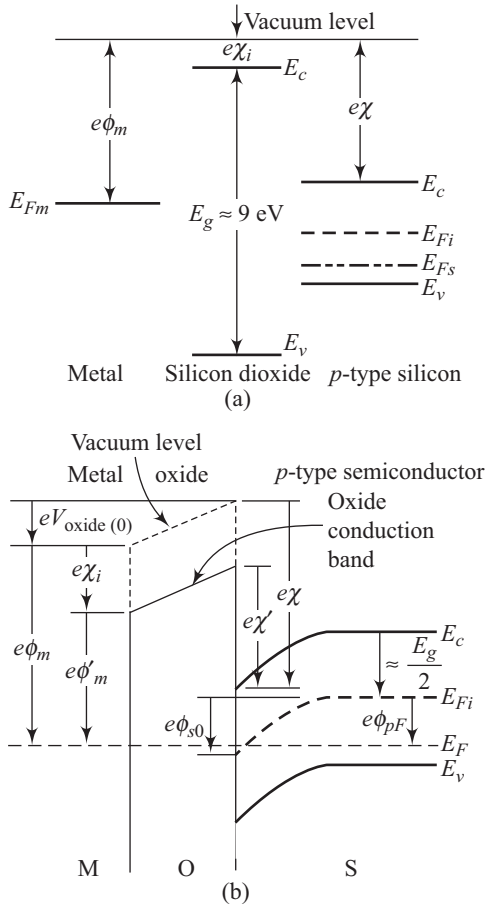


Fig. 9.20 Energy band diagram of MOS formed on *p*-type semiconductor. (a) separated components (b) composite band diagram

Under these conditions, the energy band is flat under no applied bias. Such an ideal situation is valid provided that

- the only charges that exist under any biasing conditions are the ones in the semiconductor and equal but opposite charges on the metal surface adjacent to the oxide.
- the resistivity of the oxide is infinite, thereby ensuring that no carrier transport takes place across the oxide under any dc-biasing conditions.

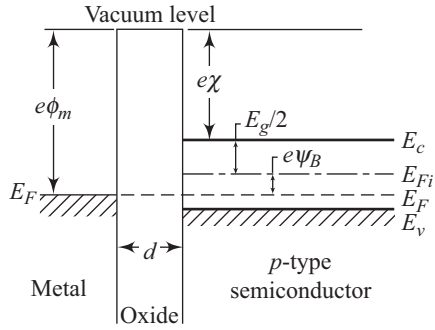


Fig. 9.21 Energy band diagram of an ideal MOS diode on *p*-type semiconductor

9.4 Capacitance–Voltage Characteristics of MOS Capacitor

Figure 9.22(a) shows the band diagram of an ideal MOS structure. The corresponding charge distribution is shown in Fig. 9.22(b).

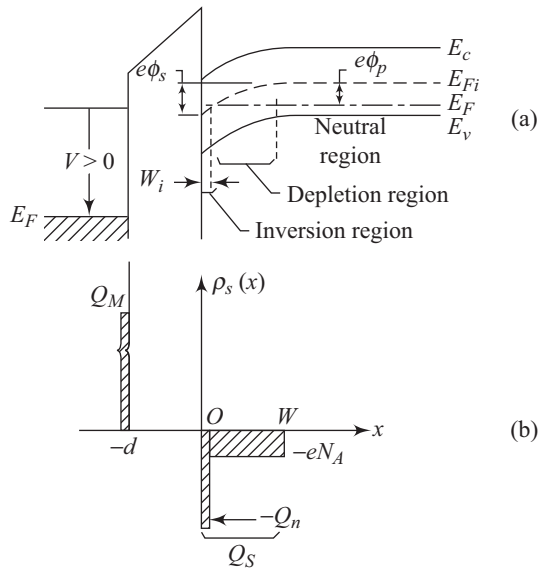


Fig. 9.22 (a) Energy band diagram of ideal MOS structure (b) Charge distribution

In the ideal situation, there is no work-function difference; therefore, the applied voltage, V, appears partly across the oxide layer (V_{ox}) and partly across the semiconductor (ϕ_s). Thus,

$$V = V_{ox} + \phi_s \tag{9.44}$$

Figure 9.23(a) shows the electric-field distribution across the ideal MOS structure and Fig. 9.23(b) shows the corresponding potential distribution. E_{ox} represents the field in the oxide and Q_s is the charge per unit area in the semiconductor.

The potential V_{ox} across the oxide is given by

$$V_{ox} = E_{ox}\, t_{ox} \tag{9.45}$$

where t_{ox} represents the oxide thickness.

In terms of the charge Q_S, Eqn (9.45) can be rewritten in the form

$$V_{ox} = \frac{|Q_s|\, t_{ox}}{\varepsilon_{ox}} = \frac{|Q_s|}{C_{ox}} \tag{9.46}$$

where C_{ox} is the oxide capacitance per unit area.

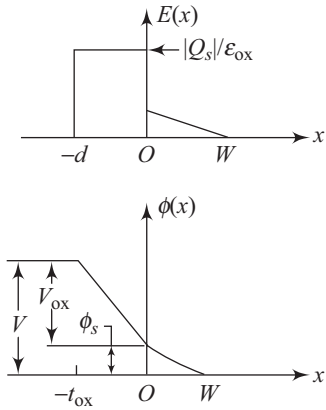


Fig. 9.23 (a) Electric field distribution across ideal MOS structure
(b) Potential distribution

The total capacitance, C, in the MOS structure arises due to a series combination of the oxide capacitance and the semiconductor depletion-layer capacitance C_J. Thus,

$$C = \frac{C_{ox} C_J}{C_{ox} + C_J} \text{ F/cm}^2 \tag{9.47}$$

The depletion region capacitance C_J is, in turn, represented by

$$C_J = \frac{\varepsilon_s}{W_d} \tag{9.48}$$

From Eqn (9.38) we get

$$\phi_s = \frac{eN_a W_d^2}{2\varepsilon_s} \tag{9.49}$$

Using Eqs (9.49), (9.44), (9.46), and (9.47) we get

$$\frac{C}{C_{\text{ox}}} = \frac{1}{\sqrt{1 + \dfrac{2\varepsilon_{\text{ox}}^2 V}{eN_a \varepsilon_s t_{\text{ox}}^2}}} \tag{9.50}$$

From Eqn (9.50) it is clear that the capacitance of the MOS structure with *p*-type semiconductor and positive bias decreases with increasing gate bias. The width of the depletion region continues to increase with increasing positive voltage. Ultimately at high positive voltage, there occurs an inversion of the semiconductor surface. At this point, the width of the depletion region becomes constant and does not increase with increase in applied voltage. As we have seen, this happens at an applied voltage that results in $\phi_S = 2\phi_{pF}$. Thus, using Eqn (9.37), ϕ_S at inversion is given as

$$\phi_s = \frac{2kT}{e} \ln\left(\frac{N_a}{n_i}\right) \tag{9.51}$$

The applied gate voltage at this stage, called the threshold voltage, V_T, can be written from Eqn (9.44) as

$$V_T = V_{\text{ox}} + 2\phi_B = \frac{|Q_s|}{C_{\text{ox}}} + 2\phi_B \tag{9.52}$$

The charge Q_s is given by qN_aW_{dT}, where W_{dT}, the maximum space charge width, is given by Eqn (9.39). Using this expression for W_{dT} and substituting the resultant expression for Q_S in Eqn (9.52) yields

$$V_T = \frac{\sqrt{2\varepsilon_s eN_a(2\phi_B)}}{C_{\text{ox}}} + \phi_B \tag{9.53}$$

At strong inversion the capacitance remains constant at a magnitude C_{min}, that can be obtained from Eqn (9.47) by substituting $C_J = \varepsilon_s / W_{dT}$. Thus

$$C_{\text{min}} = \frac{\varepsilon_{\text{ox}}}{t_{\text{ox}} + \left(\dfrac{\varepsilon_{\text{ox}}}{\varepsilon_s}\right) W_{dT}} \tag{9.54}$$

When the applied gate voltage is negative for an MOS structure using a *p*-type semiconductor, an accumulation of holes takes place at the semiconductor surface. The total capacitance in the accumulation region becomes close to the oxide capacitance given by $\varepsilon_{\text{ox}}/t_{\text{ox}}$. Figure 9.24 shows the typical capacitance–voltage characteristics of an ideal MOS diode. The dashed curve has been obtained using Eqs (9.50), (9.53), and (9.54) (also known as depletion

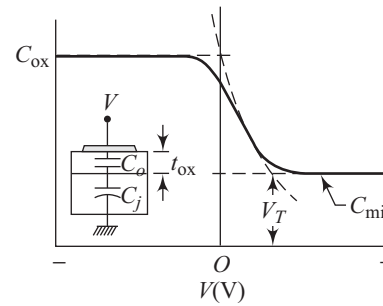


Fig. 9.24 Typical *C-V* characteristic for an ideal MOS structure using a *p*-type semiconductor

approximation), whereas the solid curve is valid when using exact calculations. Both the curves can be seen to match quite satisfactorily. We must remember that an MOS structure based on an *n*-type semiconductor would display similar *C-V* characteristics but with a reversal of the gate bias.

The capacitance–voltage curve shown in Fig. 9.24 is valid at high frequencies so that the changes in the gate voltage result in a corresponding increase in the charges at the edge of the depletion region. If the measurement frequencies are low, on the other hand, then the minority carrier electron concentrations can follow the variations in the ac gate signal. Thus a charge exchange with the inversion layer at a rate decided by the measurement signal is possible. When this happens, the capacitance magnitude at strong inversion matches that due to the oxide layer, C_{ox}. Figure 9.25 shows the ideal low-frequency *C-V* characteristics for an MOS capacitor formed on an *n*-type silicon substrate.

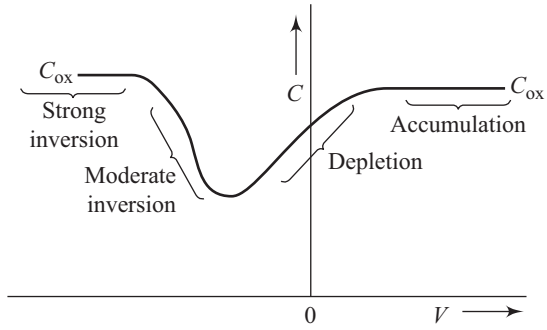


Fig. 9.25 Ideal, low-frequency *C-V* curve for an MOS capacitor using an *n*-type silicon substrate

Typically, high frequency corresponds to values of the order of 1 MHz, whereas low frequency effects dominate in the frequency range of 5 to 100 Hz.

9.4.1 Interface Traps and Oxide Charge

The equilibrium of an MOS structure is, in general, decided by other factors apart from the work-function difference discussed earlier in this chapter. Charges in the oxide and traps at the Si–SiO_2 interface are also deciding factors for the equilibrium of the MOS structure. In general there are four entities included in this category and these are (i) interface-trapped charge, (ii) fixed-oxide charge, (iii) oxide-trapped charge, and (iv) mobile ionic charge. The charges and traps along with their locations are shown schematically in Fig. 9.26.

Interface-trapped charges arises due to the Si–SiO_2 interface characteristics and depend upon the chemical composition of the Si–SiO_2 interface. The interface-trap density (interface traps per unit area) depends upon the orientation of the silicon substrate, with ⟨111⟩ orientation resulting in an interface trap density that is an order of magnitude higher than in the ⟨100⟩ orientation. Low temperature (~450 °C) hydrogen annealling can be used to neutralize interface-trapped

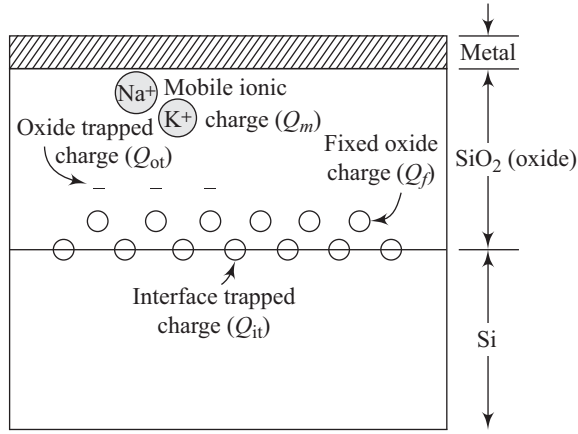


Fig. 9.26 Charges and traps in oxide and interface

charges. A symbol Q_{it} is generally used for interface-trapped charge and, using annealling procedures, it is possible to realize values as low as 10^{10} cm^{-2} for Q_{it} in ⟨100⟩ oriented silicon wafers.

Fixed-oxide charge is represented by Q_f and is generally positive. Oxidation and annealling conditions as well as orientation decide the nature and magnitude of fixed-oxide charge. The source of this positive fixed charge is the ionic silicon left near the interface at the oxidation completion stage and the uncompleted silicon bonds such as Si–Si or Si–O. For the ⟨100⟩ and the ⟨111⟩ surfaces typical values of fixed oxide charge are in the region of 10^{10} cm^{-2} and 5×10^{10} cm^{-2}, respectively.

Defects in silicon dioxide result in the oxide-trapped charges Q_{ot}. Certain specialized processes like *X*-ray radiation exposure or high-energy electron bombardment can also result in oxide-trapped charges. Low-temperature annealling can be used to remove process related oxide-trapped charges.

Mobile ionic charges result due to sodium or other alkali ions. Represented by Q_m, these charges are mobile at high temperatures and under high-voltage conditions.

9.4.2 Effect of Oxide Charge on *C-V* Characteristics

The oxide charges discussed in the previous section have a profound effect on the observed capacitance–voltage characteristics of an MOS capacitor. Suppose a positive sheet of charge Q_o per unit area exists within the oxide as shown in Fig. 9.27(a). This positive sheet charge induces negative charges partly in the metal and partly in the semiconductor.

The resulting electric-field distribution is shown in Fig. 9.27(a). The electric-field distribution assumes the work-function difference to be zero. As we have seen, the flat-band condition implies that no charges are induced in the

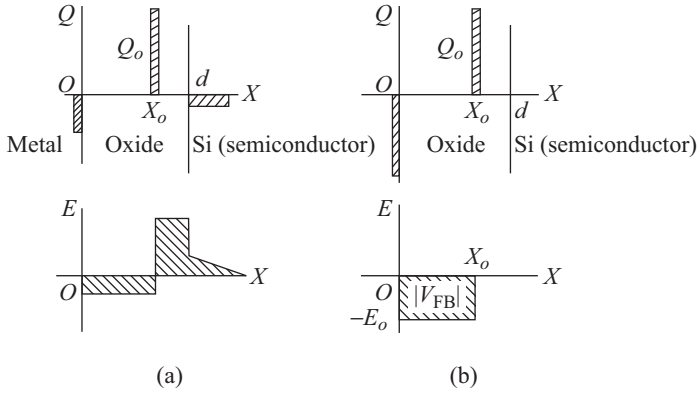


Fig. 9.27 (a) Effect of sheet charge on electric field for $V_{GS} = 0$; (b) Charge and electric field for flat-band condition

semiconductor. To realize this state, a negative voltage needs to be applied to the metal, resulting in more negative charges accumulating on the metal. The electric field distribution shifts downward and eventually vanishes at the semiconductor surface. This situation is depicted in Fig. 9.27(b). The flat-band voltage, V_{FB}, is thus given by

$$V_{FB} = -E_O X_O = \frac{Q_o}{\varepsilon_{ox}} X_O = -\frac{Q_o}{C_{ox}} \frac{X_o}{X_{ox}} \tag{9.55}$$

From Eqn (9.55) it can be concluded that the flat-band voltage depends on the density of the sheet charge, Q_o, and its location, X_o, within the oxide. Two cases deserve special mention. These are

Case (i) $X_o = 0$
In this case the sheet charge is located very close to the metal and induces no charges in the semiconductor. $V_{FB} = 0$ for this case.

Case (ii) $X_o = X_{ox}$
The sheet charge is located close to the semiconductor (one example is Q_f). The resulting flat-band voltage is maximum and is given by

$$V_{FB} = -\frac{Q_o}{C_{ox}} \tag{9.56}$$

For a non-zero work-function difference $q\phi_{ms}$ and non-zero Q_f, Q_m, and Q_{ot}, the actual experimental capacitance–voltage curve is shifted from the ideal theoretical curve by an amount given by

$$V_{FB} = \phi_{ms} - \left(\frac{Q_f + Q_m + Q_{ot}}{C_{ox}} \right) \tag{9.57}$$

Figure 9.28 shows *C-V* characteristics of the MOS structure under three different situations. Curve (i) is the ideal situation. The parallel shift in the *C-V* curve, governed by Eqn (9.57), is shown by curve (ii) in Fig. 9.28. In the presence

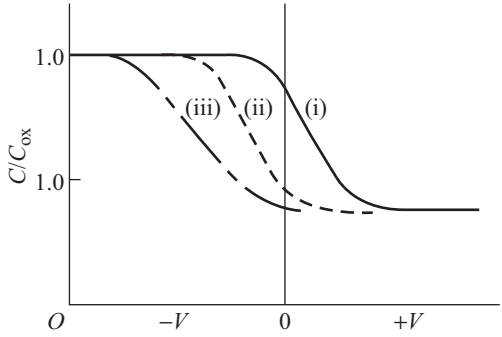


Fig. 9.28 *C-V* characteristics in the presence of fixed oxide charge and interface traps for MOS structure

of significant interface-trapped charge, the *C-V* curve gets displaced with an amount that is surface potential dependent as shown in curve (iii).

9.5 MOS Field-effect Transistor

The MOSFET is the most important device in the development of very large-scale integrated circuits (VLSI) which will be discussed in detail in a later chapter. A schematic view of a typical MOSFET is shown in Fig. 9.29.

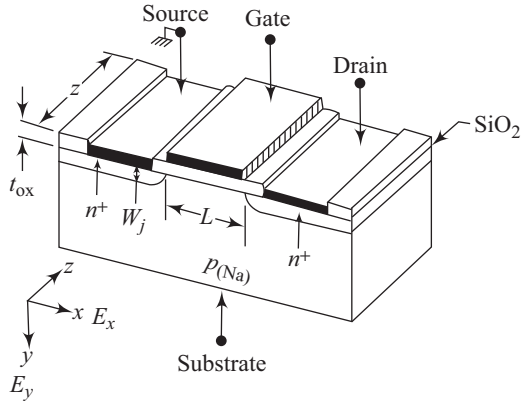


Fig. 9.29 A typical MOSFET

MOSFET is a four-terminal device including the *p*-type semiconductor substrate. The other three terminals are the source, the gate, and the drain. As is true of other FETs, the gate terminal can consist of metal, heavily doped polysilicon or silicide. In Fig. 9.29, W_j represents the junction depth and N_a is the acceptor concentration.

Suppose some voltage is applied to the gate resulting in an inversion of the semiconductor surface. The situation is shown in Fig. 9.30(a). A small applied

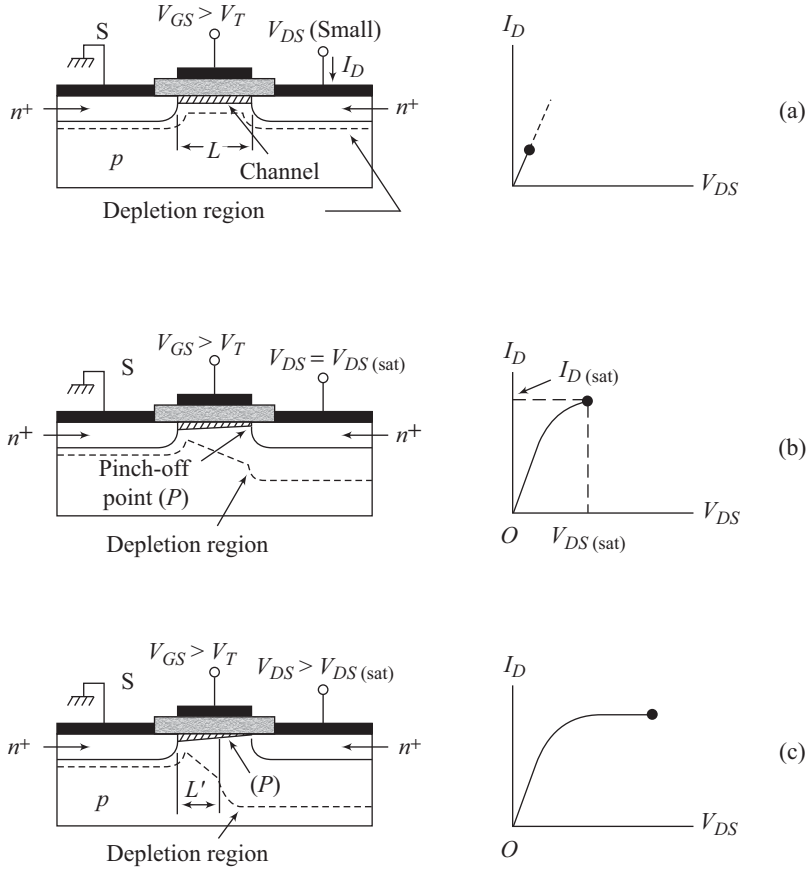


Fig. 9.30 MOSFET output *I-V* characteristics

drain voltage will cause electrons to flow from the source to the drain through the conducting channel. The channel acts as a resistive layer and the drain current, I_{DS}, varies linearly with drain voltage, V_{DS}. The corresponding curve is also shown in Fig. 9.30(a).

As the drain voltage increases, the voltage drop within the oxide near the drain terminal reduces. This leads to a corresponding decrease in the induced inversion charge density near the drain. Finally a stage is reached when the inversion layer width is reduced to zero at $x = L$. This is called the pinch-off point [Fig. 9.30(b)]. Once this state is achieved, the drain current remains constant. This is because for $V_{DS} > V_{DS(sat)}$, the voltage at the pinch-off point, P, remains fixed at $V_{DS(sat)}$. This ensures that the number of charge carriers arriving at point P from the source remains constant, thereby fixing the current between the drain and the source. All that happens with increasing V_{DS} is a movement of the point P away from the drain terminal. The channel length thus gets effectively reduced to L' as shown in Fig. 9.30(c). The complete *I-V* curve is also shown in Fig. 9.30(c).

9.5.1 MOSFET Characteristics

We will now derive the basic MOSFET characteristics under ideal conditions by assuming the following.

- The gate structure is an ideal MOS structure without interface traps, fixed-oxide charges, and work-function difference.
- Only drift currents are significant.
- Carrier mobility is constant in the inversion layer.
- The channel is uniformly doped.
- The reverse-leakage current is negligible.
- The transverse field (E_y in the y-direction) is much larger than the longitudinal field.

A schematic representation of a MOSFET biased in the linear region is shown in Fig. 9.31.

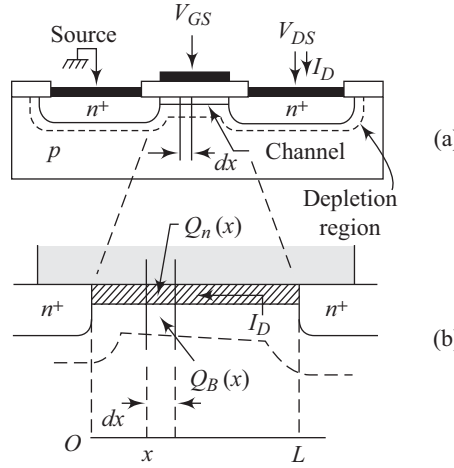


Fig. 9.31 A MOSFET biased in the linear region

The total charge, Q_s, induced in the semiconductor per unit area at a distance x can be obtained using Eqs (9.44) and (9.46). Thus we get

$$Q_s(x) = -[V_{GS} - \phi_S(x)]C_{ox} \tag{9.58}$$

where $Q_s(x)$ is the surface potential at the location x and C_{ox} is the gate oxide capacitance given by ε_{ox}/t_{ox}.

Charge $Q_s(x)$ can also be expressed as

$$Q_s(x) = Q_n(x) + Q_{SC}(x) \tag{9.59}$$

where $Q_n(x)$ is the inversion layer charge at position x and $Q_{SC}(x)$ is the corresponding depletion region space charge. From Eqn (9.59) we can write

$$Q_n(x) = Q_S(x) - Q_{SC}(x) \tag{9.60}$$

Using Eqn (9.58) in Eqn (9.60) we get

$$Q_n(x) = -[V_{GS} - \phi_S(x)]C_{ox} - Q_{SC}(x) \tag{9.61}$$

At inversion, the surface potential $\phi_s(x)$ can be written in the form

$$\phi_s = 2\phi_B + V(x) \tag{9.62}$$

where $V(x)$ is the reverse bias between the point x and the grounded source electrode in Fig. 9.31(a). Also, the space charge $Q_{SC}(x)$ is given by

$$Q_{SC}(x) = -eN_a W_{dT} \tag{9.63}$$

On using Eqn (9.39) this results in

$$Q_{SC}(x) = \sqrt{2\varepsilon_s eN_a[V(x) + 2\phi_B]} \tag{9.64}$$

Substituting Eqn (9.64) in Eqn (9.61) yields

$$Q_n(x) = -[V_{GS} - V(x) - 2\phi_B]C_{ox} + \sqrt{2\varepsilon_s eN_a[V(x) + 2\phi_B]} \tag{9.65}$$

Channel conductivity at a position x is given approximately by (Fig. 9.31)

$$\sigma(y) = q\, n(y)\, \mu_n(y) \tag{9.66}$$

If the mobility in the channel is assumed constant, the channel conductance can be expressed in the form

$$g = \frac{Z}{L}\int_0^{W_i} \sigma(y)dy \tag{9.67}$$

Using Eqn (9.66) in Eqn (9.67) we get

$$g = \frac{Z}{L}\mu_n \int_0^{W_i} qn(y)dy \tag{9.68}$$

The term $\int_0^{W_i} qn(y)dy$ is the total charge per unit area in the inversion layer. Thus

$$g = \frac{Z\mu_n}{L}|Q_n| \tag{9.69}$$

Channel resistance dR of an elemental section dx in Fig. 9.28(b) is given by

$$dR = \frac{dx}{Z\mu_n |Q_n(x)|} \tag{9.70}$$

The voltage drop dV across this elemental section is

$$dV = I_D dR = \frac{I_D dx}{Z\mu_n |Q_n(x)|} \tag{9.71}$$

where I_D is the position-independent drain current.
Substituting for $Q_n(x)$ from Eqn (9.65) and integrating from the source ($x = 0$, $V = 0$) to the drain ($x = L$, $V = V_{DS}$) results in

$$I_D = \frac{Z}{L}\mu_n C_{ox}\left\{\left[V_{GS} - 2\phi_B - \frac{V_{DS}}{2}\right]V_{DS} - \frac{2}{3}\frac{\sqrt{2\varepsilon_s eN_a}}{C_{ox}}\left[(V_{DS} + 2\phi_B)^{3/2} - (2\phi_B)^{3/2}\right]\right\} \tag{9.72}$$

Figure 9.32 shows the current–voltage characteristics of an ideal MOSFET as governed by Eqn (9.72).

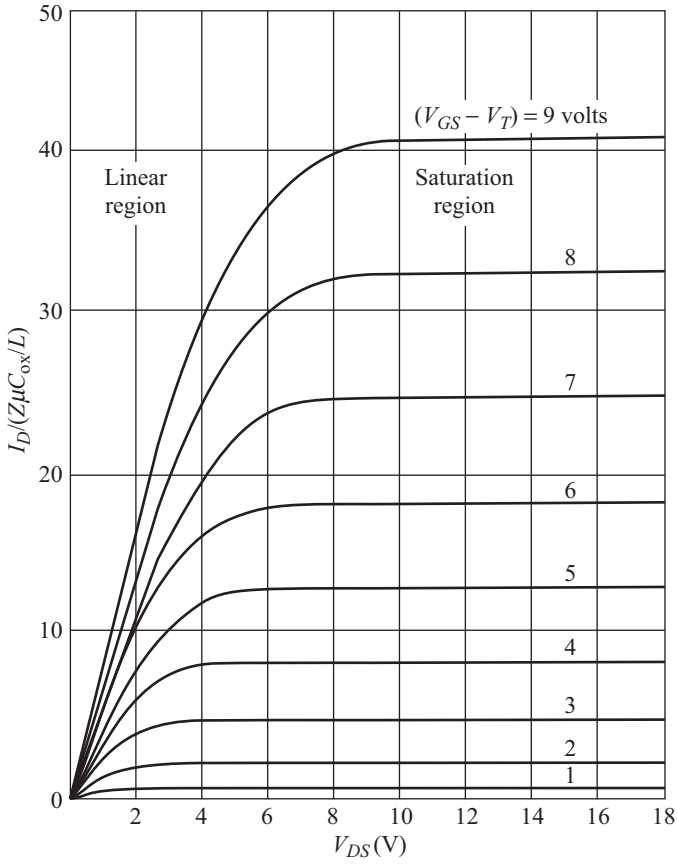


Fig. 9.32 Current–voltage characteristics of an ideal MOSFET

For any particular value of V_{GS}, the drain current initially increases linearly with drain voltage. The increase in the drain current then slows down to eventually approach a saturation value.

Linear region

For small values of V_{DS} [$V_{DS} \ll (V_{GS} - V_T)$] Eqn (9.72) can be rewritten in the form

$$I_D \simeq \frac{Z}{L}\mu_n C_{ox}(V_{GS} - V_T)V_{DS} \tag{9.73}$$

where the threshold voltage, V_T, is given by

$$V_T = \frac{\sqrt{2\varepsilon_s e N_a (2\phi_B)}}{C_{ox}} + 2\phi_B \tag{9.74}$$

Thus a plot of I_D versus V_{GS} for a particular value of V_{DS} can be used to find the threshold voltage from the extrapolated value on the V_{GS} axis. The channel conductance, g_D, and the transconductance, g_m, can be deduced using

$$g_D \equiv \left.\frac{\partial I_D}{\partial V_{DS}}\right|_{V_{GS}=\text{const}} = \frac{Z}{L}\mu_n C_{\text{ox}}(V_{GS} - V_T) \tag{9.75}$$

and

$$g_m \equiv \left.\frac{\partial I_D}{\partial V_{GS}}\right|_{V_{DS}=\text{const}} = \frac{Z}{L}\mu_n C_{\text{ox}} V_{DS} \tag{9.76}$$

Saturation region

As the drain voltage is increased a point is reached when the charge in the inversion layer $Q_n(x)$ at $x = L$ becomes zero and the number of mobile electrons at the drain gets reduced drastically. This point is called pinch-off point analogous to the pinch-off point in the JFET. The corresponding drain voltage and drain current are designated $V_{DS(\text{sat})}$ and $I_{D(\text{sat})}$, respectively. The saturation region exists for drain voltages larger that $V_{DS(\text{sat})}$. Putting $Q_n(L) = 0$ in Eqn (9.65) results in

$$V_{DS(\text{sat})} = V_{GS} - 2\phi_B + K_1^2\left[1 - \sqrt{1 + 2V_{GS}/K_1^2}\right] \tag{9.77}$$

where $K_1 = \sqrt{\varepsilon_s e N_a}\,/C_{\text{ox}}$

Substituting Eqn (9.77) in Eqn (9.72) gives us

$$I_{D(\text{sat})} \cong \frac{Z\mu_n \varepsilon_{\text{ox}}}{2t_{\text{ox}}L}(V_{GS} - V_T)^2 \tag{9.78}$$

Equation (9.78) can be seen to be identical to Eqn (9.36) for the MESFET with t_{ox} replacing channel depth, d, and ε_{ox} replacing ε_s.

For low substrate doping and thin oxide layers, V_T in the saturation region is still given by Eqn (9.74). In the saturation region the channel conductance is zero for an ideal MOSFET. Transconductance, g_m, in the saturation region can be derived from Eqn (9.78)

$$g_m \equiv \left.\frac{\partial I_D}{\partial V_{GS}}\right|_{V_{DS}=\text{const}} = \frac{Z\mu_n \varepsilon_{\text{ox}}}{t_{\text{ox}}L}(V_{GS} - V_T) \tag{9.79}$$

9.5.2 Short Channel Effect

As the channel length reduces, the portion of the channel where the carriers travel with a saturation velocity starts becoming substantial. At extremely short channel lengths, carriers can be practically assumed to travel with saturation velocity. Under this situation, the drain current is given by the product of channel charge per unit area, width of the channel, and the saturation velocity. The equation for $I_{D\,(\text{sat})}$ then becomes,

$$I_{D(\text{sat})} \approx ZC_i(V_{GS} - V_T)\nu_S \tag{9.80}$$

Comparing Eqs (9.80) and (9.78) we can see that due to the short channel effect, the drain current shows a linear dependence on $(V_{GS} - V_T)$ instead of the quadratic dependence at higher channel lengths.

9.5.3 Control of Threshold Voltage

The threshold voltage decides the voltage at which a given MOS device gets switched on. A circuit designer needs a variety of choices in the value of threshold voltage for different applications. Obviously a MOS with a threshold voltage greater than the source voltage is not acceptable. Some special applications also require extremely low threshold values. Methods to control the threshold voltage are therefore important. Using Eqs (9.52) and (9.57), the equation for threshold voltage including the voltage required to achieve a flat band is given by

$$V_T = \phi_{ms} - \left[\frac{Q_f + Q_m + Q_{\text{ox}}}{C_{\text{ox}}}\right] + \frac{|Q_S|}{C_{\text{ox}}} + 2\phi_\beta \tag{9.81}$$

Equation (9.81) can be used in a variety of ways to control the threshold voltage. The term ϕ_{ms} can be modified by varying the choice of the gate conductor material. Substrate doping variations can be used to modify ϕ_B. $|Q_S|$ can be controlled by changing substrate doping. C_{ox} is controllable by adjusting the thickness and dielectric constant of the insulator layer. Fixed-oxide charge Q_f is dependent upon the annealing conditions and choice of orientation.

9.5.4 Substrate Bias Effect

In Fig. 9.30, no bias has been applied between the source *S* and the substrate. It is possible to apply a bias between the source and the substrate. If a reverse bias is applied between the substrate and the source (for *n*-channel device V_B is negative), the depletion region gets widened. Due to the additional charge in the depletion region, the threshold voltage required to get an inversion layer increases. For a *p*-channel device on the other hand, the threshold voltage becomes more negative with a reverse bias applied between the source and the substrate. This technique is often employed to increase the threshold voltage of MOSFET devices.

9.5.5 Subthreshold Characteristics

Equation (9.78) leads to the conclusion that $I_{D\,(\text{sat})}$ becomes zero as V_{GS} is reduced to V_T. In real devices, some drain current flows even below the threshold voltage. This current flow is called the subthreshold current. The current is due to the prevailing weak inversion in the channel region that exists between flat band and threshold. This leads to a diffusion current flowing from the source to the drain.

In the subthreshold region, the drain current, I_D depends exponentially on the gate bias V_{Gs}. A plot of $\ln I_D$ as a function of V_{Gs} is therefore linear. The slope, *S*, of this straight line is called the subthreshold slope. State-of-the art MOSFET devices have typical *S* values of ~ 70 mv/decade at room temperature. Smaller values of S imply a better suitability for the MOSFET to act as a switch. The parameter *S* is a measure of the ability of the gate potential to modulate the drain current I_D. Value of S is high for MOSFETs having heavy channel doping or if the silicon–silicon dioxide interface has substantial number of fast interface states. At very small gate voltage values, the subthreshold current is reduced to the leakage current

due to the source/drain junctions. This decides the off-state leakage current and therefore the standby power dissipation for many CMOS (complementary MOS) circuits that involve both *n*-channel and *p*-channel MOSFET devices.

9.5.6 Equivalent Circuits for MOSFET

Figure 9.33 shows an equivalent circuit of the MOSFET device. The equivalent circuit contains several important components. There exists an overlap between the gate and drain and the gate and source regions. Both these overlaps result in capacitances.

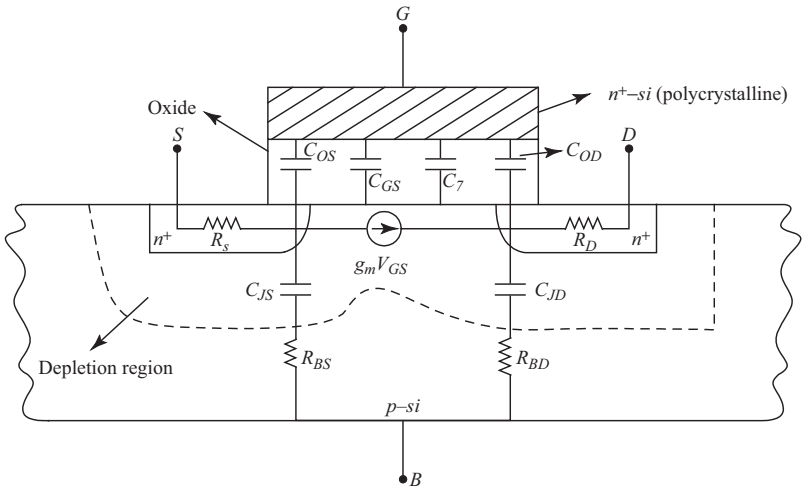


Fig. 9.33 Equivalent circuit of MOSFET

The overlap between the gate and drain regions results in the capacitance C_{OD}. The capacitance C_{OD} is also called the *Miller overlap capacitance*. The capacitance C_{OS} results from the overlap of the gate and the source regions. The overlap capacitances can be minimized by using a *self-aligned gate*. In this process, the gate itself is used to mask the source/drain implantations. A lateral straggle during implantation however does not allow the complete elimination of the overlap capacitances. In the equivalent circuit, $(R_S + R_D)$ represents the total source/drain series resistance R_{SD}. This series resistance leads to a *wastage* of a part of the applied voltage due to the ohmic drop. C_{JS} and C_{JD} represent the depletion capacitances associated with the source and drain regions. The drain current is represented by a (gate) voltage controlled constant-current source. R_{BS} and R_{BD} represent resistances in the substrate between the bulk contact and the source and drain.

9.5.7 MOSFET Scaling and Hot Electron Effects

Advances in integrated circuit technology have been made possible due to shrinking or scaling down of devices. This scaling has resulted in several advantages like

improved packing density, speed, and power dissipation. Pioneering work in this area was done by Dennard at IBM. According to him various structural parameters of the MOSFET have to be scaled together for any device to function properly. What that means is that lateral dimensions like channel width and length as well as vertical dimension like source/drain junction depths and gate insulator thickness have to be reduced together to maintain effectiveness of the device. Power supply voltages have also to be simultaneously reduced. This however is not always possible due to other system-related constraints. Due to this, the longitudinal electric fields in the pinch-off region and the transverse electric fields across the gate oxide increase as the MOSFETs are scaled. The resultant effects are collectively called the *hot electron effects* and *short channel effects*.

The hot electron effects result in some special characteristics. Some of these are listed below.

(i) In the pinch-off region, the electrons gain kinetic energy at the expense of electrostatic potential energy. They move higher up in the conduction band and some electrons can surmount the 3.1 eV potential barrier between the Si channel and the gate oxide. These electrons constitute an additional gate current and thus the input impedance of the MOSFET gets reduced.

(ii) Some electrons that surmount the barrier between the Si channel and gate oxide, get trapped in the gate oxide as fixed oxide charges. This leads to an increase in the flat band voltage and therefore increases the threshold voltage.

(iii) Hot electrons can also rupture Si–H bonds that exist at $Si–SiO_2$ interface leading to the creation of fast interface states. These states create stress that degrade MOSFET parameters like transconductance and subthreshold slope.

(iv) Problems mentioned in (ii) and (iii) can be controlled by using lightly doped drain (LDD). Use of a lightly doped drain leads to an increase in the depletion width associated with the drain-channel junction and a consequent reduction of electric field.

(v) Hot carrier effects are less cause of concern for *p*-channel MOSFETs. This is due to two reasons. First, channel mobility of holes is around half that of electrons; thus for the same electric field, there are few hot holes in comparison to hot electrons. Second, the barrier for hole injection in the valence band between Si and SiO_2 is higher (5 eV) than the 3.1 eV barrier encountered by electrons in the conduction band. This also means that LDD is often not used for *p*-channel device.

(vi) Hot electrons also create secondary electron-hole pairs by impact ionization. This leads to an increase in drain current in saturation region with drain bias at high voltages. This ultimately results in a decrease in output impedance.

9.5.8 Drain-induced Barrier Lowering

As long as the channel length of a MOSFET is long, there is no electrostatic interaction between the source and the drain. As the channel length becomes

small and either the source/drain junctions are too deep or channel region doping is low, the source and the drain start interacting electrostatically. The drain bias then starts affecting the source-to-channel potential barrier. In fact the source-to-channel potential barrier gets lowered. This effect is called the Drain-induced Barrier Lowering (DIBL). DIBL results in punch-through leakage or breakdown between the source and the drain. The result is that the gate control is lost. To solve the problem, the source/drain junctions should be made sufficiently shallow in step with the reduction in channel length. Also, the channel doping needs to be made sufficiently high so that the drain bias is unable to control the source junction. One way of achieving this is to do an *anti-punch-through* implantation in the channel region.

9.5.9 Short Channel and Narrow Width Effect

The threshold voltage depends upon the depletion charge under the gate. A portion of the depletion region charges are shared with the source and the drain. For long channel devices, the portion of depletion charge under the gate that is shared by source and drain is extremely small. As the channel length reduces, the shared charge becomes a sizable portion of the total depletion charge. The threshold voltage V_T then starts decreasing as a function of channel length. This leads to difficulties in controlling the threshold voltage. This phenomenon is called the short channel effect (SCE). An opposite effect called the reverse short channel effect also exists. The source/drain implants result in the creation of Si point defects. For n-channel MOSFETs, a pile up of B takes place during B doping of the channel region. These two causes lead to an increase in threshold voltage with decreasing channel length before it finally decreases due to SCE. This effect is therefore aptly called reverse short channel effect (RSCE).

As the channel width gets reduced, the depletion charge belonging to the gate increases. This leads to an increase in threshold voltage with a decrease of width of MOSFET device. This effect is called the *narrow width effect*.

9.5.10 Gate-induced Drain Leakage

As the gate becomes more negatively biased, a depletion region forms in the n-type drain. For doping in the region of $\sim 10^{18}$ cm^{-3}, the depletion widths are narrow enough to permit band-to-band tunnelling to take place. This creates electron–hole pairs with the electrons being collected by the drain. This is known as gate-induced drain leakage (GIDL). If the doping in the drain region is lower, then the depletion widths are too wide to permit tunnelling whereas too high doping results in most of the applied voltage dropping across the gate oxide. GIDL is an important factor that decides the off-state leakage current in MOSFET devices.

9.5.11 Comparison of BJT with MOSFET

The BJT utilizes the injection of minority carriers across a forward biased junction whereas a MOSFET mainly utilizes the control of the current flowing through

two terminals by a voltage applied on a third terminal. A FET is a majority carrier device and is therefore often referred to as a unipolar transistor. A BJT on the other hand is a minority carrier device. Since a BJT utilizes both electrons and holes, it is called bipolar transistor. Surface defects and surface states play a very crucial role in MOSFET in comparison to the BJT device. Unlike BJTs, FET devices are characterized by a high input impedance since the control voltage is applied across a reverse-biased junction or Schottky barrier, or across an insulator. This make FET device ideal for applications like controlled switching between a conducting state and a non-conducting stage. The input impedance of BJT devices is not as high as that of FET devices. The high input impedance also makes FET device an ideal choice for digital circuits and integrated circuits involving semiconductor memory devices and microprocessors.

9.5.12 Types of MOSFET

There are two basic types of MOSFETs, namely enhancement and depletion type MOSFETs.

Enhancement type MOSFET

Suppose the channel conductance of an *n*-channel MOSFET is very low at zero gate bias. A positive voltage must then be applied to the gate to form the *n*-channel. This type of *n*-channel MOSFET is called an enhancement or normally-off *n*-channel MOSFET. Similarly a normally-off *p*-channel MOSFET is called a *p*-channel enhancement MOSFET. Typical schematic diagrams, *I-V* characteristics, and symbolic representations of *n*-channel and *p*-channel enhancement type MOSFETs are shown in Fig. 9.34.

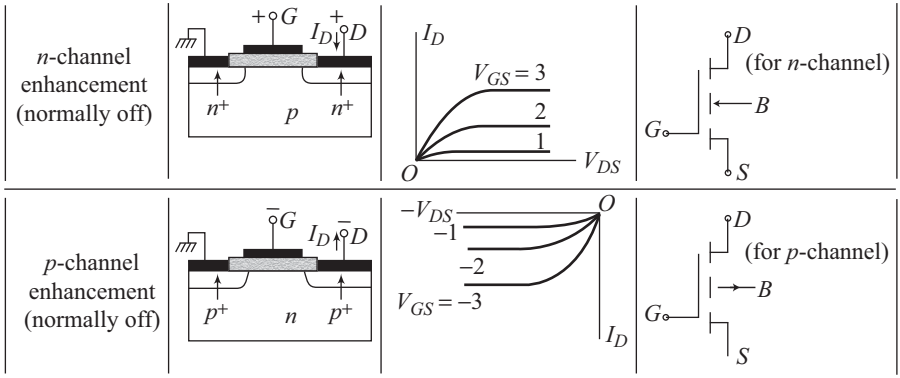


Fig. 9.34 Schematic diagram and *I-V* characteristics of enhancement MOSFETs

Depletion Type MOSFET

If an *n*-channel exists at zero bias, a negative voltage has to be applied on the gate electrode to reduce the channel conductance. Such a MOSFET is then called a depletion or normally-on *n*-channel MOSFET. The corresponding device using

a *p*-channel is called a *p*-channel depletion MOSFET. Figure 9.35 shows the *n*-channel and *p*-channel depletion MOSFETs along with their *I*-*V* characteristics and symbolic representations.

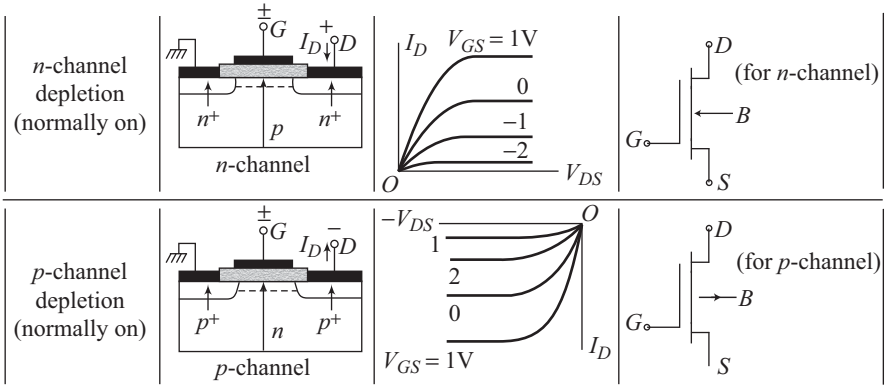


Fig. 9.35 Depletion MOSFETs and their *I*-*V* characteristics

- The maximum operating frequency, f_T, is defined as the frequency at which the output current is equal to the input current. In other words, the MOSFET can no longer function as an amplifier. The frequency, f_T, is given by

$$f_T = \frac{\mu_n V_{DS}}{2\pi L^2} \text{ for } V_{DS} \leq V_{DS(\text{sat})}$$

  where μ_n represents the carrier mobility and L, the channel length.
- In the presence of fixed-oxide charge and work-function difference, the threshold voltage V_T becomes

$$V_T = V_{\text{FB}} + 2\phi_B + \frac{\sqrt{2\varepsilon_s eNa(2\phi_B)}}{C_{ox}}$$

Solved Problems

9.1 An *n*-channel silicon JFET is formed with $N_d = 5 \times 10^{16}$ cm^{-3}, $N_a = 10^{19}$ cm^{-3}, $d = 1.2$ μm, $L = 18$ μm, $W = 80$ μm, and $\mu_n = 1350$ cm^2/V s. Calculate the pinch-off voltage, pinch-off current, and the drain current at pinch-off assuming $V_{GS} = 0$.

Solution

Pinch-off voltage V_P is given by

$$V_P = \frac{eN_d d^2}{2\varepsilon_s} \tag{9.1.1}$$

Putting the given values in Eqn (9.1.1) yields

$$V_P = \frac{(1.6\times10^{-19})\times(5\times10^{16})\times(1.2\times10^{-4})^2}{2\times11.7\times8.85\times10^{-14}}$$

$$= 55.6 \text{ V}$$

Pinch-off current, I_P, is given by

$$I_P = \frac{W\mu_n e^2 N_d^2 d^3}{\varepsilon_s L} \tag{9.1.2}$$

Putting values in Eqn (9.1.2) we get

$$I_P = \frac{80\times10^{-4}\times1350\times(1.6\times10^{-19})^2\times(5\times10^{16})^2\times(1.2\times10^{-4})^3}{11.7\times8.85\times10^{-14}\times18\times10^{-4}}$$

$$= 640.8\times10^{-3} \text{ A}$$

Built-in voltage, V_{bi}, can be obtained using the equation

$$V_{\text{bi}} = \frac{kT}{e}\ln\left(\frac{N_a N_d}{n_i^2}\right) \tag{9.1.3}$$

yielding

$$V_{\text{bi}} = 0.026\ln\left(\frac{5\times10^{16}\times10^{19}}{(1.5\times10^{10})^2}\right)$$

$$= 0.92 \text{ V}$$

Finally, drain current at pinch-off, $I_{D(\text{sat})}$, is given by

$$I_{D(\text{sat})} = I_p\left[\frac{1}{3} - \left(\frac{V_{GS}+V_{\text{bi}}}{V_P}\right) + \frac{2}{3}\left(\frac{V_{GS}+V_{\text{bi}}}{V_P}\right)^{3/2}\right] \tag{9.1.4}$$

Putting the values of V_P, I_P, and V_{bi} and substituting $V_{GS} = 0$ yields

$$I_{D(\text{sat})} = 641\times10^{-3}\left[\frac{1}{3} - \frac{0.92}{55.6} + \frac{2}{3}\left(\frac{0.92}{55.6}\right)^{3/2}\right]$$

$$= 204\times10^{-3} \text{ A}$$

9.2 An *n*-channel MESFET is to be designed using a GaAs-gold Schottky barrier contact. The barrier height obtained is ϕ_{Bn} = 0.9 V and the *n*-channel doping concentration is $N_d = 3\times10^{15}$ cm^{-3}. If the threshold voltage desired is 0.3 V, calculate the required channel thickness. Assume $N_c = 4.7\times10^{17}$ cm^{-3} where N_c is the effective density-of-states function in the conduction band. Let T be equal to 300 K take ε_s to be 13.1.

Solution

For a Schottky contact

$$\phi_n = \frac{kT}{e}\ln\left(\frac{N_c}{N_d}\right) \tag{9.2.1}$$

Putting the given values in Eqn (9.2.1) we get

$$\phi_n = 0.026 \ln\left(\frac{4.7 \times 10^{17}}{3 \times 10^{15}}\right)$$

$$= 0.131 \text{ V}$$

Built-in voltage, V_{bi}, is expressed in the form

$$V_{bi} = \phi_{Bn} - \phi_n \tag{9.2.2}$$

Putting the calculated value of ϕ_n and the given value of ϕ_{Bn} in Eqn (9.2.2) results in

$$V_{bi} = 0.9 - 0.131 = 0.769 \text{ V}$$

Threshold voltage, V_T, is related to pinch-off voltage, V_P, through the equation

$$V_T = V_{bi} - V_P$$

which implies

$$V_P = V_{bi} - V_T = 0.769 - 0.3 = 0.469 \text{ V}$$

The pinch-off voltage, V_P, is given by

$$V_P = \frac{ed^2 N_d}{2\varepsilon_s}$$

yielding

$$d^2 = \frac{2\varepsilon_s V_P}{eN_d}$$

resulting in

$$d = \sqrt{\frac{2\varepsilon_s V_P}{eN_d}} \tag{9.2.3}$$

Putting the known values of different parameters in Eqn (9.2.3) yields

$$d = \sqrt{\frac{2 \times 13.1 \times 8.85 \times 10^{-14} \times 0.469}{1.6 \times 10^{-19} \times 3 \times 10^{15}}}$$

$$= 0.476 \text{ μm}$$

9.3 An MOS structure is formed using a *p*-type silicon wafer with $N_a = 5 \times 10^{16}$ cm^{-3}. Calculate the maximum space charge width. Assume $T = 300$ K and intrinsic carrier concentration, $n_i = 1.5 \times 10^{10}$ cm^{-3}.

Solution

The maximum space charge width, W_{dT}, is given by

$$W_{dT} = \left(\frac{4\varepsilon_s \phi_{pF}}{eN_a}\right)^{1/2} \tag{9.3.1}$$

where

$$\phi_{pF} = \frac{kT}{e} \ln\left(\frac{N_a}{n_i}\right) \tag{9.3.2}$$

Putting the given values in Eqn (9.3.2) leads to

$$\phi_{pF} = 0.026 \ln\left(\frac{5 \times 10^{16}}{1.5 \times 10^{10}}\right)$$
$$= 0.39 \text{ V}$$

Substituting the value of ϕ_{pF} in Eqn (9.3.1) results in

$$W_{dT} = \left[\frac{4 \times 11.7 \times (8.85 \times 10^{-14})(0.39)}{(1.6 \times 10^{-19})(5 \times 10^{16})}\right]^{1/2}$$
$$= 0.142 \text{ μm}$$

9.4 Consider the aluminium–silicon dioxide–silicon, MOS system at T = 300 K. Assume $\phi'_m = 3.20$ V, $\chi' = 3.25$, and $E_g = 1.11$ eV. Suppose p-type doping concentration is $N_a = 10^{15}$ cm^{-3}. Calculate the work-function difference for the system. Take $n_i = 1.5 \times 10^{10}$ cm^{-3}.

Solution

The work-function difference, ϕ_{ms}, is given by

$$\phi_{ms} = \left[\phi'_m - \left(\chi' + \frac{E_g}{2e} + \phi_{pF}\right)\right] \tag{9.4.1}$$

From Eqn (9.3.2) we know

$$\phi_{pF} = \frac{kT}{e} \ln\left(\frac{\text{Na}}{n_i}\right) \tag{9.4.2}$$

Putting the given values in Eqn (9.4.2) results in

$$\phi_{pF} = 0.026 \ln\left(\frac{10^{15}}{1.5 \times 10^{10}}\right)$$
$$= 0.289 \text{ V}$$

Substituting the value of ϕ_{pF} and other parameters in Eqn (9.4.1) leads to

$$\phi_{ms} = [3.20 - (3.25 + 0.555 + 0.289)]$$
$$= -0.894 \text{ V}$$

9.5 An ideal n-channel MOSFET has the following parameters:

$L = 1.3$ μm; $\mu_n = 660$ cm^2/Vs;

$C_{ox} = 7 \times 10^{-8}$ F/cm^2; $V_T = 0.66$ V

What should be the channel width, Z, such that $I_{D(\text{sat})} = 5$ mA for $V_{GS} = 5$ V?

Solution

Saturation drain current I_D(sat) is given by

$$I_{D(\text{sat})} \cong \frac{Z\mu_n \varepsilon_{\text{ox}}}{2t_{\text{ox}}L}(V_{GS} - V_T)^2$$

which implies

$$I_{D(\text{sat})} \cong \frac{Z\mu_n C_{\text{ox}}}{2L}(V_{GS} - V_T)^2 \tag{9.5.1}$$

Putting the given values in Eqn (9.5.1) results in

$$5\times10^{-3} = \frac{Z(660)\,(7\times10^{-8})}{2(1.3\times10^{-4})}(5-0.66)^2$$

or

$$Z = \frac{5\times10^{-3}\times2\times1.3\times10^{-4}}{660\times7\times10^{-8}\times(4.34)^2}$$

$$= 0.00149 \text{ cm} = 14.9 \text{ μm}$$

9.6 Calculate V_T of an n-channel Si MOSFET for an n^+-polysilicon gate with gate oxide thickness = 100 Å, $N_a = 10^{18}$ cm^{-3}, and a fixed oxide charge of 5×10^{10} qC/cm^2. Repeat this calculation for a substrate bias of –2.5 V. Take $\phi_{ms} = -1.5$ V.

Solution

Taking into account Eqn (9.52) given in the text and including all the effects of interface trap charges, etc., the modified expression for V_T can be written as

$$V_T = V_{\text{FB}} + 2\phi_F - \frac{Q_d}{C_i}$$

$$V_{\text{FB}} = 2\phi_F - \frac{Q_i}{C_i}$$

$$C_i = \frac{\varepsilon_i}{d} = \frac{8.85\times10^{-14}\times3.9}{100\times10^{-8}} = 3.452\times10^{-7} \text{ F/cm}^2$$

$$V_{\text{FB}} = \phi_{\text{ms}} - \frac{Q_i}{C_i} = -1.5 - \frac{5\times10^{10}\times1.6\times10^{-19}}{3.452\times10^{-7}} = -1.523\text{V}$$

$$\phi_F = \frac{kT}{e}\ln\frac{N_a}{n_i} = 0.0259\ln\left(\frac{10^{18}}{1.5\times10^{10}}\right) = 0.467 \text{ V}$$

$$W = \sqrt{\frac{2\varepsilon_s(2\phi_F)}{eN_a}} = \sqrt{\frac{2(11.8)(8.854\times10^{-14})(2\times0.467)}{1.6\times10^{-19}\times10^{18}}}$$

$$= 3.49\times10^{-6} \text{ cm}$$

$$Q_d = -eN_aW_m$$

$$V_T = V_{FB} + 2\phi_F - \frac{Q_d}{C_i}$$

$$= -1.523 + 2(0.467) + \frac{1.6 \times 10^{-19} \times 10^{18} \times 3.49 \times 10^{-6}}{3.452 \times 10^{-7}}$$

$$= 1.03 \text{ V}$$

Now, with $V_B = -2.5$ V, depletion charge increases. So, instead of band bending of $2\phi_F$, now we have band bending of $2\phi_F + V_B$.

$\therefore$ the new width will be

$$W_m = \sqrt{\frac{2\varepsilon_s(2\phi_F + V_B)}{eN_a}}$$

$$= 6.695 \times 10^{-6} \text{ cm}$$

$$Q_d = -eN_aW_m$$

$$= -1.071 \times 10^{-6} \text{ C}$$

$$-\frac{Q_d}{C_i} = \frac{1.071 \times 10^{-6}}{3.452 \times 10^{-7}} = 3.103 \text{ V}$$

$\therefore$ $V_T = -1.523 + 0.934 + 3.103 = 2.514$ V

9.7 For a typical MOS structure with an *n*-substrate (doping concentration = 5×10^{17}cm^{-3}), n^+-poly, gate oxide thickness of 80 Å, interface charge = 10^{11} qC/cm^2, and $\phi_{ms} = -0.15$ V

(a) Calculate V_T.
(b) Draw the band diagram for this MOS structure.
(c) State whether this MOS structure will be in accumulation, depletion, or inversion near the Si–oxide interface under equilibrium conditions. Justify your answer.

Solution

(a) Taking into account Eqn (9.52) and including the effect of interface trapped charges, the modified expression for V_T can be written as

$$V_T = \phi_{ms} - \frac{Q_i}{C_i} - \frac{Q_d}{C_i} + 2\phi_F$$

where ϕ_{ms} is the difference in work functions of poly and substrate, Q_i represents the charges at the interface, Q_d represents the depletion charges, and $\phi_F = \frac{kT}{e} \ln(N_d/n_i)$.

Calculating all individual terms in the above expression we get

$$\phi_F = \frac{kT}{e} \ln\left(\frac{N_d}{n_i}\right)$$

$$= 0.0259 \ln\left(\frac{5 \times 10^{17}}{1.5 \times 10^{10}}\right) = -0.448 \text{ V}$$

$$W_m = 2\sqrt{\frac{\varepsilon_s|\phi_F|}{eN_d}} = 2\sqrt{\frac{11.8\times 8.854\times 10^{-14}\times 0.448}{1.6\times 10^{-19}\times 5\times 10^{+17}}}$$

$$= 4.83\times 10^{-6}\ \text{cm}$$

$$Q_d = eN_dW_m = 1.6\times 10^{-19}\times 5\times 10^{17}\times 4.83\times 10^{-6}$$

$$= 3.87\times 10^{-7}\ \text{C/cm}^2$$

$$C_i = \frac{\varepsilon_i}{d} = \frac{3.9\times 8.854\times 10^{-14}}{80\times 10^{-8}} = 4.31\times 10^{-7}\ \text{F/cm}^2$$

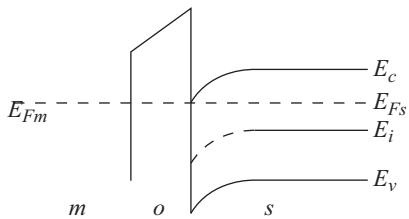
Putting all these values in the expression for V_T

$$V_T = -0.15 - \frac{10^{11}\times 1.6\times 10^{-19}}{4.31\times 10^{-7}} - \frac{3.87\times 10^{-7}}{4.31\times 10^{-7}} - (2\times 0.448)$$

$$= -0.15 - 0.037 - 0.897 - 0.896$$

$$= -1.98\ \text{V}$$

(b)



(c) Since band bending at the Si–SiO_2 interface will be downward, the material at the interface will be more *n*-type. Hence, the structure will be in accummulation mode under thermal equilibrium.

9.8 During a laboratory experiment a device engineer tried to find the value of D_n of electrons in the channel of a typical Si *n*-MOS structure and found this value to be very close to 35 cm^2/s at 300 K. But the person working on this experiment was not at all happy with the results obtained and suspected some error in the experiment.

(a) Is the engineer correct in suspecting this error?
(b) If yes, then state whether he should have obtained a higher or lower value of D_n than what he actually obtained? Justify your answer. Else, suggest a suitable experimental set-up to determine the required value of D_n.

(Given: mobility of electrons in silicon at 300 K = 1350 cm^2/Vs)

Solution

(a) Yes, the engineer is correct in suspecting the error.
(b) He should have obtained a lower value of D_n as

$$D_n = \frac{kT}{e}\mu_n$$

and μ_n, i.e., mobility of electrons in the channel will be lower than the given value, i.e., 1350 cm^2/Vs.

Justification: Mobility of electrons in the channel will always be lower than that in the bulk.

Recapitulation

- Saturation drain–source voltage, $V_{DS(\text{sat})}$, for an ν-channel JFET is given by

$$V_{DS(\text{sat})} = \frac{eN_d d^2}{2\varepsilon_s} - V_{\text{bi}} - V_{GS}$$

- Pinch-off voltage, V_P, and current, I_D, for an *n*-channel JFET can be obtained using the equations

$$V_P = \frac{eN_d d^2}{2\varepsilon_s} \text{ and } I_P = \frac{W\mu_n e^2 N_d^2 d^3}{\varepsilon_s L}$$

- In the linear region, the transconductance of a JFET, g_m, takes the form

$$g_m = \frac{I_P}{2V_P^2}\sqrt{\frac{V_P}{V_{GS}+V_{\text{bi}}}}\,V_{DS}$$

whereas, in the saturation region, it is given by

$$g_m = \frac{I_P}{V_P}\left[1 - \sqrt{\frac{V_{GS}+V_{\text{bi}}}{V_P}}\right]$$

- Saturation drain current $I_{D(\text{sat})}$ for a MESFET is expressible as

$$I_{D(\text{sat})} \approx \frac{W\mu_n\varepsilon_s}{2dL}(V_{GS}-V_T)^2$$

- Space charge width, W_d, for an MOS structure using *p*-type semiconductor is given by

$$W_d = \left(\frac{2\varepsilon_s\phi_s}{eN_a}\right)^{1/2}$$

where ϕ_s is the surface potential.

The maximum space charge width, W_{dT}, can be obtained using

$$W_{dT} = \left(\frac{4\varepsilon_s\phi_{pF}}{eN_a}\right)^{1/2}$$

- The quantity $\left[\phi_m' - \left(\chi' + \frac{E_g}{2e} + \phi_{pF}\right)\right]$ is called the metal–semiconductor work-function difference, ϕ_{ms}.
- For an MOS structure using a *p*-type semiconductor, the *C*-*V* characteristics in the depletion region are given by

$$\frac{C}{C_{\text{ox}}} = \frac{1}{\sqrt{1+\frac{2\varepsilon_{\text{ox}}^2 V}{eN_a\varepsilon_s t_{\text{ox}}^2}}}$$

- At strong inversion, the minimum capacitance, C_{min}, can be determined using the expression

$$C_{\min} \frac{\varepsilon_{ox}}{t_{ox} + \left(\frac{\varepsilon_{ox}}{\varepsilon_s}\right) W_{dT}}$$

- The flat-band voltage is given by

$$V_{FB} = \phi_{mS} - \left(\frac{Q_f + Q_m + Q_{ot}}{C_{ox}}\right)$$

- *I-V* characteristics for an ideal MOSFET are expressible in the form

$$I_D = \frac{Z}{L} \mu_n C_{ox} \left\{ \left[V_{GS} - 2\phi_B - \frac{V_{DS}}{2} \right] V_{DS} \right.$$

$$\left. - \frac{2}{3} \frac{\sqrt{2\varepsilon_s e N_a}}{C_{ox}} \left[(V_{DS} + 2\phi_B)^{3/2} - (2\phi_B)^{3/2} \right] \right\}$$

- For a MOSFET the saturation drain current can be obtained using

$$I_{D(sat)} \simeq \frac{Z \mu_n \varepsilon_{ox}}{2 t_{ox} L} (V_{GS} - V_T)^2$$

- With short channel effect, the expression for $I_{D(sat)}$ is,

$$I_{D(sat)} \approx z\, c_i\, (V_{GS} - V_T)\, V_S$$

- Expression for threshold voltage including the voltage required to achieve flat band is given by

$$V_T = \phi_{ms} - \left[\frac{Q_f + Q_m + Q_{ox}}{C_{ox}} \right] + \frac{|Q_s|}{C_{ox}} + 2\phi_B$$

- Overlap capacitances can be minimized by using a self-aligned gate.
- Hot electron effect leads to an increase in the flat-band voltage.
- GIDL decides the leakage current in MOSFET devices.

Exercises

Review Questions

9.1 Explain the operating principle of the JFET using a suitable sketch.
9.2 Derive an expression for the channel resistance of a JFET.
9.3 Draw the output characteristics of a typical JFET and explain the different regions.
9.4 Derive an expression for drain current for a JFET and hence define pinch-off current and voltage.
9.5 Give some differences in the operation of the JFET and the MESFET.
9.6 What are normally-off MESFETs? Explain using a suitable schematic diagram.

9.7 Describe the working of a high-electron-mobility transistor.
9.8 Show the formation of accumulation layer, positive space charge region, and inversion layer for the MOS structure using an *n*-type semiconductor, with the help of a suitable band diagram.
9.9 Derive an expression for the space charge width for an MOS structure on a *p*-type semiconductor.
9.10 Define threshold voltage for a MOS structure.
9.11 What does the term work-function difference signify for an MOS structure?
9.12 Draw the *C*-*V* characteristic curve of an ideal MOS structure and explain the important regions.
9.13 Explain the terms:
(a) Interface-trapped charge
(b) Fixed-oxide charge
(c) Oxide-trapped charge
(d) Mobile ionic charge
9.14 How do the *C*-*V* characteristics of an MOS structure get modified in the presence of oxide charge?
9.15 Explain the *I*-*V* characteristics of a typical MOSFET.
9.16 Derive an expression for drain current as a function of drain voltage for an ideal MOSFET.
9.17 Give the salient features of the linear region and the saturation region of the output curve of an ideal MOSFET.
9.18 Give the operating principle of an enhancement type MOSFET.
9.19 Explain the *I*-*V* characteristics of a JFET using suitable schematics.
9.20 Derive expressions for the transconductance, g_m, for a JFET in the linear and saturation regions.
9.21 Derive the expression for $I_{D(\text{sat})}$ for a MESFET.
9.22 Derive the operating principle of a HEMT.
9.23 Show the formation of accumulation layer, depletion layer, and inversion layer in a MOS structure formed on *p*-type semiconductor using suitable energy band diagrams.
9.24 Derive an expression for the maximum space charge width, W_{dT}, for an MOS structure.
9.25 Explain the term work-function difference with the help of a suitable diagram.
9.26 Derive an expression for C/C_{ox} for an MOS structure in depletion. How does oxide-charge affect the *C*-*V* characteristics of an MOS structure?
9.27 Use a suitable diagram to show the important region of I-V characteristics of a MOSFET.
9.28 Derive an expression for $I_{D(\text{sat})}$ for a MOSFET.
9.29 Differentiate between enhancement type and depletion type MOSFETs.
9.30 What is the consequence if the channel width never actually become zero?
9.31 Give the symbolic representation of *n*-channel JJET.
9.32 Give the symbolic representation of *n*-channel depletion MOSFET.
9.33 What is short channel effect?
9.34 Give some methods of controlling threshold voltage.
9.35 What is substrate bias effect?

9.36 Describe the subthreshold characteristics of a MOSFET.
9.37 Sketch the equivalent circuit of a MOSFET.
9.38 What is a self-aligned gate?
9.39 What are hot electron effects?
9.40 Describe the process of LDD.
9.41 What is Drain-induced Barrier Lowering (DIBL)?
9.42 Describe the short channel effect.
9.43 What is reverse short channel effect?
9.44 Describe narrow width effect.
9.45 What is Gate-Induced Drain Leakage (GIDL)?
9.46 Present a comparision of BJT and MOSFET devices.

Problems

9.1 An *n*-channel JFET is fabricated using silicon with important parameter values of $N_d = 10^{16}$ cm^{-3}; $N_a = 5 \times 10^{18}$ cm^{-3}, $d = 1$ μm, $L = 20$ μm, $W = 70$ μm, and $\mu_n = 1350$ cm^2/V s. Evaluate the pinch-off voltage, pinch-off current, and the drain-current at pinch-off assuming $V_{GS} = 0$. Take $n_i = 1.5 \times 10^{10}$ cm^{-3} at $T = 300$ K.

$$\left[\textit{Hint: } V_P = \frac{eN_d d^2}{2\varepsilon_s}, I_P = \frac{W\mu_n e^2 N_d^2 d^3}{\varepsilon_s L} \right.$$

$$\left. \text{and } I_{D(\text{sat})} = I_P \left\{ \frac{1}{3} - \left(\frac{V_{GS} + V_{\text{bi}}}{V_P} \right) + \frac{2}{3} \left(\frac{V_{GS} + V_{\text{bi}}}{V_P} \right)^{3/2} \right\} \right]$$

Ans. 7.7 V, 11.68×10^{-3} A; 0.0029 A

9.2 Consider a silicon *n*-channel JFET at $T = 300$ K. The p^+n junction has a uniform doping concentration of $N_a = 5 \times 10^{18}$ cm^{-3} and $N_d = 10^{16}$ cm^{-3}. Assume the metallurgical channel thickness, d, to be 0.8 μm. Calculate the internal pinch-off voltage, V_P, and the actual gate-to-source voltage to be applied to achieve pinch-off, V_P'.
[*Hint:* $V_P' = V_{\text{bi}} - V_P$]

Ans. –4.09 V

9.3 A silicon *p*-channel *p-n* JFET at 300 K has a gate doping concentration of $N_d = 5 \times 10^{18}$ cm^{-3}. Determine the channel doping concentration and channel thickness if the actual gate-to-source voltage to be applied to achieve pinch-off, V_P', is to be 2.5 V. Is the answer unique?

$$\left[\textit{Hint:} \text{ Assume } N_a \text{ and use } V_{\text{bi}} = \frac{kT}{e} \ln \left[\frac{N_a N_d}{n_i^2} \right] \right]$$

and $V_P = V_{bi} + V_P'$]

Ans. 5×10^{16} cm^{-3}; 0.297 μm; solution is not unique and depends on the choice of N_a.

9.4 A GaAs–gold Schottky barrier contact is being used to design an *n*-channel MESFET. The barrier height obtained is ϕ_{Bn} = 0.9 V and the doping concentration of the *n*-channel is $N_d = 5 \times 10^{15}$ cm^{-3}. The desired threshold voltage is 0.25 V. Calculate the required channel thickness. Assume $N_c = 4.7 \times 10^{17}$ cm^{-3}, where N_c is the effective density-of-states function in the conduction band. Assume $T = 300$ K and take ε_s to be 13.1.

$$\left[\textit{Hint: } \phi_n = \frac{kT}{e} \ln\left(\frac{N_c}{N_d}\right) \text{ and } V_{\text{bi}} = \phi_{Bn} - \phi_n . \text{Also, } V_P = \frac{qe^2 N_d}{2\varepsilon_s} \right]$$

Ans. 0.393 μm

9.5 An *n*-channel MESFET is to be designed using materials such that $\phi_{Bn} = 0.85$ V, $N_c = 4.7 \times 10^{17}$ cm^{-3}, $V_T = 0.28$ V, $\varepsilon_s = 13.1$, and $V_P = 0.4$ V. Calculate the doping concentration of the required *n*-channel.

[*Hint:* $V_{\text{bi}} = V_T + V_P$ and $V_{\text{bi}} = \phi_{Bn} - \phi_n$]

Ans. 6.8×10^{14} cm^{-3}

9.6. An MOS structure is fabricated using a *p*-type semiconductor substrate doped to $N_a = 5 \times 10^{16}$ cm^{-3}. The silicon dioxide insulator thickness is $t_{\text{ox}} = 600$ Å. The gate material is n^+ polysilicon with $\phi_{ms} = -1.1$ V. Assuming $Q_{\text{ot}} = 10^{11}$ electron charges per cm^3, calculate the flat-band voltage. Take $\varepsilon_{\text{ox}} = 3.9$.

$$\left[\textit{Hint: } C_{\text{ox}} = \frac{\varepsilon_{\text{ox}}}{t_{\text{ox}}} \text{ and } V_{\text{FB}} = \phi_{ms} - \frac{Q_{\text{ot}}}{C_{\text{ox}}} \right]$$

Ans. –1.38 V

9.7 An MOS structure on *p*-type semiconductor shows a flat-band voltage of –1.3 V. The semiconductor substrate has a doping level of $N_a = 10^{16}$ cm^{-3} and the oxide thickness $t_{\text{ox}} = 600$ Å. The metal–semiconductor work-function difference is $\phi_{ms} = -1.0$ V. Assume $\varepsilon_{0\text{x}} = 3.9$. Determine the oxide-trapped charge, Q_{ot}.

$$\left[\textit{Hint: } V_{\text{FB}} = \phi_{ms} - \frac{Q_{\text{ot}}}{C_{\text{ox}}} \right]$$

Ans. 1.07×10^{11} electronic charge per cm^2

9.8 An MOS structure is formed using an *n*-type silicon wafer with $N_d = 10^{16}$ cm^{-3}. Determine the maximum space charge width assuming $T = 300$ K and $n_i = 1.5 \times 10^{10}$ cm^{-3}.

$$\left[\textit{Hint: } \phi_{nF} = \frac{kT}{e} \ln\left(\frac{N_d}{n_i}\right); W_{dT} = \left(\frac{4\varepsilon_s \phi_{nF}}{eN_d}\right)^{1/2} \right]$$

Ans. 0.301 μm

9.9 An MOS capacitor is based on an aluminium–silicon dioxide–silicon system at $T = 300$ K. Assume $\phi_m' = 3.20$ V. $\chi' = 3.25$, and $E_g = 1.11$ eV. The *p*-type doping concentration is $N_a = 5 \times 10^{15}$ cm^{-3}. Determine the work-function difference for the system, assuming $n_i = 1.5 \times 10^{10}$ cm^{-3}.

$$\left[Hint: \phi_{ms} = \left\{ \phi_m' - \left(\chi' + \frac{E_g}{2e} + \phi_{pF} \right) \right\} \right]$$

Ans. –0.935 V

9.10 An ideal *n*-channel MOSFET has the following parameters:

$L = 1.2$ μm; $\mu_n = 660$ cm^2/V s;

$C_{ox} = 5 \times 10^{-8}$ F/cm^2; $V_T = 0.6$ V; $Z = 12$ μm.

Calculate $I_{D(sat)}$ for $V_{GS} = 4$ V.

$$\left[Hint: I_{D(sat)} \approx \frac{Z \mu_n \varepsilon_{ox}}{2 t_{ox} L} (V_{GS} - V_T)^2 \right]$$

Ans. 1.9 mA

9.11 (a) A JFET structure has been prepared with a channel of width less than the total width of the two depletion regions at zero bias. Discuss how this device can be used as an FET.

(b) Why, as a designer, are you interested in scaling the MOS devices and what problems will you face during scaling?

(c) Variation of space charge density in the semiconductor as a function of surface potential for *p*-type silicon with $N_a = 10^{15}$ cm^{-3} at room temperature is shown in Fig. 9.P11.1.

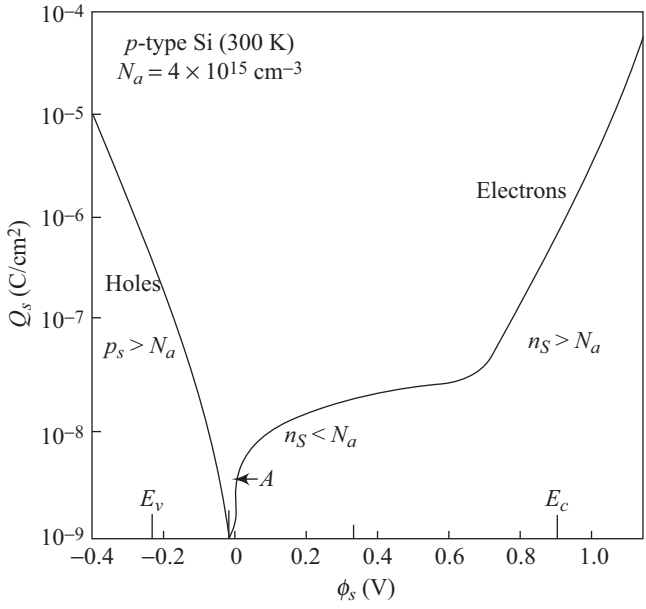


Fig. 9.P11.1 Variation of space charge density with surface potential

Is the claim that point *A* corresponds to flat-band condition true? Justify with physical explanation.

(d) For an ideal MOS capacitor on *p*-type Si with $N_a = 10^{16}$ cm^{-3}, the *C-V* curve is shown in Fig. 9.P11.2. Assume the thickness of SiO_2 to be 100 Å. Find the voltage/capacitances at the points shown on the curve.

(e) Explain the effect of drain induced barrier lowering on threshold voltage. Why is the DIBL effect more dominant in short channel devices? What is the effect of DIBL on the performance of the device? As a designer, how will you reduce this effect?

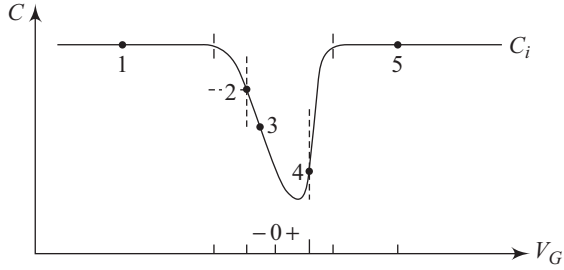


Fig. 9.P11.2 *C-V* curve for an ideal MOS capacitor

9.12 (a) C_{GD}, the gate–drain overlap capacitance is also known as _____

(b) GIDL affects the _____ current in short channel MOSFETs.

a. linear region b. on-state

c. off-state leakage d. substrate leakage

(c) MOS devices are fabricated on _____ surface of Si since it has the least interface charge density.

(i) 101 (ii) 100 (iii) 111 (iv) 110

(d) For flat band condition in an ideal MOS capacitor ϕ_s = _____

(e) The quantities that can be determined using an MOS *C-V* curve

(a) C_i (b) Substrate doping

(c) Threshold voltage (d) Only (a) and (b)

(e) (a), (b), and (c)

9.13 Energy levels in three separated materials that form an MOS system are shown in Fig. 9.P13.1(a) and energy band diagram for the MOS system composed of the materials indicated in Fig. 9.P13.1(a) is shown in Fig. 9.P13.1(b)

Assuming that the oxide and Si–SiO_2 interface are free of charges,

(a) How is charge transfer through the oxide possible to attain equilibrium?

(b) Find the thickness of the silicon dioxide (t_{ox}) as shown in Fig. 9.P13.1(b).

(c) What is the flat-band voltage?

Ans. (b) 114 nm, (c) –0.8 V

9.14 Assuming an ideal case, draw the energy band diagram of an MOS capacitor with *n*-type of substrate, clearly indicating the surface potential and bulk potential,

(a) for positive voltage from the metal to the semiconductor

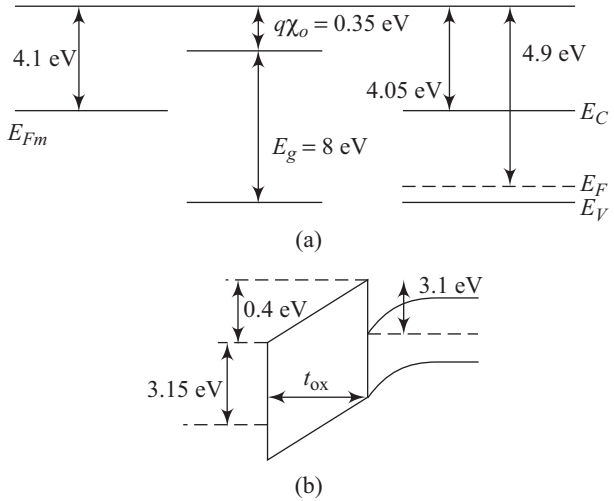


Fig. 9.P13.1 (a) Energy levels in materials (b) Energy band diagram for an MOS system

(b) for large negative voltage from the metal to the semiconductor

(c) no voltage applied from the metal to the semiconductor

(d) if the metal in this MOS structure is replaced by n^+-polysilicon, so that $\phi_{ms} \ll 0$, then draw the equilibrium band diagram.

9.15 (a) What are the advantages of a 2DEG FET over a silicon MOSFET?

(b) In a p^+-n junction, the n-doping N_d is doubled. How do the following change, if everything else remains unchanged (Indicate only increase or decrease)?

(i) Built-in potential

(ii) Breakdown voltage

(iii) Junction capacitance

(c) Write the significant differences between MOSFET and MESFET.

9.16 An n-channel silicon FET has a pinch-off voltage, V_p of 50 V. If $N_d = 7 \times 10^{16}$ cm^{-3} for the device, calculate the channel depth.

Ans. 1.92 μm

9.17 An n-channel MESFET uses a GaAs-gold Schottky barrier contact. The difference between the Fermi level (E_F) and the conduction band of the semiconductor is given by, $\phi_n = 0.12$V. Calculate the n-channel doping concentration N_d. Assume, $N_c = 4.7 \times 10^{17}$ cm^{-3}.

Ans. 4.65×10^{15} cm^{-3}

10 Opto-electronic Devices

- Optical absorption
- Photovoltaic cells
- Photodetectors
- Light-emitting diodes
- Laser diodes

Learning Objectives

After going through this chapter the student will be able to

- understand the mechanisms involved in optical absorption in semiconductors
- derive the expression for flux dependence on depth within a semiconductor
- understand the *I-V* characteristics of solar cells
- derive the expression for maximum power delivered by solar cells
- define conversion efficiency of solar cells
- understand the effect of series resistance on power delivered by solar cells
- understand the operation of heterojunction solar cells and amorphous silicon solar cells
- solve numericals based on optical absorption in semiconductors
- solve numericals based on the I-V characteristics of solar cells
- derive the expression for the gain of a photoconductor
- derive the expression for the total photocurrent density of a photodiode
- understand the operation of PIN and avalanche photodiodes
- understand the operation of a phototransistor
- understand the mechanism of operation of the LED and the loss mechanisms involved
- understand the role of critical angle in LED devices
- understand the operation of laser diodes
- solve numericals based on photoconductor gain
- solve numericals based on photodiodes
- solve numericals based on LED devices

Introduction

Semiconductor devices play important roles in the designing of systems using optical stimulants. A variety of semiconductor devices can be used to detect, generate, or convert optical energy to electrical energy. We will study some of these devices in this chapter. Optical absorption can be suitably exploited through the use of photovoltaic cells to generate electricity from the sun. Another very important area where semiconductor devices can play a pivotal role is in light wave communication and guidance. Light emitting diodes and laser diodes serve as important sources of light, whereas photodetectors of different types are used to detect light signals for further processing. This chapter will thus lay the foundations of some important application areas of semiconductor devices. We will limit our discussion to the essentials of opto-electronic systems.

10.1 Optical Absorption

Light can manifest itself as wave and particle. The particular nature of light that is responsible for producing a specific interaction is highly dependent upon the result of this interaction. When treated as particles, light is assumed to consist of photons with energy E which is related to the frequency ν through the equation

$$E = h\nu \tag{10.1}$$

where h is the Planck's constant. The frequency ν is, in turn, related to the wavelength λ through the relation

$$\nu = \frac{c}{\lambda} \tag{10.2}$$

where c is the velocity of light. Another useful mathematical expression is

$$\lambda = \frac{c}{\nu} = \frac{hc}{E} = \frac{1.24}{E}\ \mu\text{m} \tag{10.3}$$

where the value of E to be used in Eqn (10.3) should be in eV.

Photons incident on a semiconductor can enter the semiconductor and interact with it in a variety of ways. They can interact with the lattice atoms and produce heat. Photons can also interact with donor, acceptors, or defects. They can exchange their energy with valence electrons and these excited valence electrons can then jump from the valence band to the conduction band. Thus the photon can be responsible for the generation of an electron–hole pair. In fact, it is this process of generation of electron–hole pairs that will be of much interest during our discussions in this chapter.

10.1.1 Absorption Coefficient

The interaction between photons and a semiconductor depends upon the frequency of the photons and the energy band gap of the semiconductor. If the photon energy, as given by Eqn (10.1), is greater than the energy band gap, E_g,

the photon gets absorbed in the semiconductor material and the photon energy is used by a valence electron to raise itself to the conduction band. If, on the other hand, the photon energy is less than the energy band gap, E_g, the photon travels through the semiconductor and the semiconductor appears to be transparent to the incident light.

Figure 10.1 shows the basic absorption processes taking place for different values of $h\nu$.

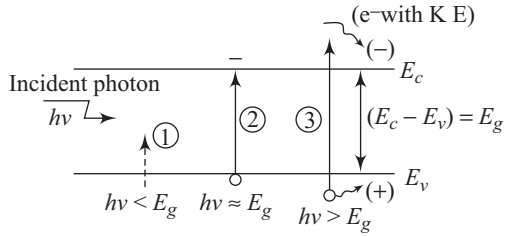


Fig. 10.1 Basic absorption processes

For $h\nu < E_g$, the photon is not absorbed. For $h\nu = E_g$, the electron has just sufficient energy to transfer itself from the valence band to the conduction band. For $h\nu > E_g$, the electron reaches the conduction band with excess kinetic energy which can then get dissipated as heat in the semiconductor.

Suppose that the intensity of the photon flux incident on a semiconductor element of thickness dx at a depth x is $I_\nu(x)$. Also let us assume the photon flux coming out of the element of thickness dx is given by $I_\nu(x + dx)$ as shown in Fig. 10.2.

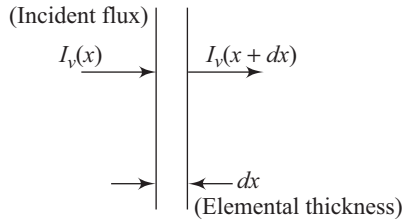


Fig. 10.2 Photon flux passing through elemental thickness

We can define a parameter called the absorption coefficient as the relative number of photons absorbed per unit distance. A symbol α is used to denote this parameter. It has units of cm^{-1}. Energy absorbed per unit time in depth dx can be written in the form

$$\alpha I_\nu(x)\, dx \tag{10.4}$$

From Fig. 10.2 we get

$$I_\nu(x + dx) - I_\nu(x) = \frac{d}{dx} I_\nu(x) dx = -\alpha I_\nu(x)\, dx \tag{10.5}$$

The negative sign emphasizes the fact that the intensity goes down with increasing depth inside the semiconductor. Equation (10.5) can be rewritten in the form

$$\frac{dI_v(x)}{dx} = -\alpha I_v(x) \tag{10.6}$$

Let us assume that the initial incident flux is I_{v0}, i.e.,

$$I_v(0) = I_{v0} \tag{10.7}$$

Equation (10.6) has solution of the form

$$I_v(x) = I_{v0} e^{-\alpha x} \tag{10.8}$$

Thus, the intensity of photon flux decreases exponentially with depth inside the semiconductor. It is also clear that for $x = \frac{1}{\alpha}$, Eqn (10.8) yields

$$I_v\left(x = \frac{1}{\alpha}\right) = \frac{I_{v0}}{e} \tag{10.9}$$

Thus at the depth given by the reciprocal of the absorption coefficient, the intensity of photon flux falls to $1/e$ of its initial value. A large value of absorption coefficient leads to absorption of photons within a small depth, whereas a small value of absorption coefficient means that the photons are absorbed within a large depth.

The absorption coefficient of a semiconductor is a very strong function of band gap energy and photon energy. Figure 10.3 shows the dependence of absorption coefficient on wavelength for some important semiconductors.

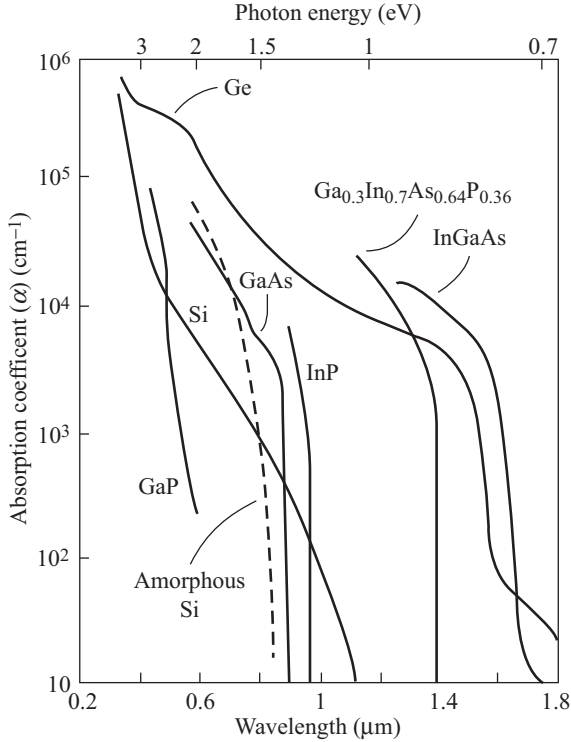


Fig. 10.3 Absorption coefficient dependence on wavelength of some semiconductors

A rapid increase in the value of α can be seen for $h\nu > E_g$ or for $\lambda < 1.24/E_g$. This is expected because it is only under these circumstances that a significant absorption of photons takes place as discussed earlier. Figure 10.4 compares the light spectrum with the band gap energies of some common semiconductors. It is a useful figure since the application of a photonic device will be based on a relative comparison of the band gap of the semiconductor with the photon wavelength. Silicon and gallium arsenide , for example, absorb the visible part of the spectrum, whereas semiconductors like GaP and CdS are transparent to the red part of the spectrum.

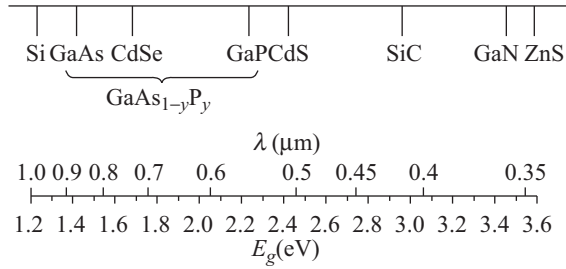


Fig 10.4 Comparison of energy band gaps with light spectrum

10.1.2 Excess Carrier Generation Rate

Photons with $h\nu > E_g$ get absorbed in the semiconductor and create electron–hole pairs as excess carriers. The product $\alpha I_v(x)$ gives the rate of absorption of energy per unit volume and has units of energy/cm^3 s. If we assume that each absorbed photon with energy $\geq h\nu$ creates one EHP (electron–hole pair), then the generation rate of EHP, g', due to photons can be written in the form

$$g' = \frac{\alpha I_v(x)}{h\nu} \tag{10.10}$$

The units of g' are cm^{-3} s^{-1}. The ratio $I_v(x)/h\nu$ also gives the photon flux.

10.2 Photovoltaic Cells

A photovoltaic cell is a semiconductor device that converts photon energy into electrical energy. Since the most abundantly available natural source of photon energy is the sun, a photovoltaic cell is also referred to as a solar cell.

10.2.1 *p-n* Junction Solar Cells

A semiconductor *p-n* junction was the first device to be used as a solar cell. When a *p-n* junction is created, majority carrier holes go from the *p*-region towards the *n*-region and majority carrier electrons go from the *n*-region towards the *p*-region. At equilibrium, a depletion region develops that has a built-in electric field. If incident photon radiations can create additional electron–hole pairs then

the built-in electric field can be used to sweep them across the junction to generate current. Holes generated within the space charge region and within a diffusion length from the edge of the space charge region, in the *n*-region and the electrons created across the space charge region, as also those created within a diffusion length in the *p*-region will be the chief contributors.

Figure 10.5 shows an illuminated *p-n* junction with a resistive load R_L. If I_L represents the generated photocurrent, the potential drop across the load produces a forward bias on the *p-n* junction. The corresponding forward bias current, I_F, is in a direction opposite to the photocurrent I_L. The resultant current is decided by the relative magnitudes of the two component currents. The net *p-n* junction current is then given by

$$I = I_L - I_F = I_L - I_S\left[\exp\left(\frac{eV}{kT}\right) - 1\right] \tag{10.11a}$$

where I_s is the reverse saturation current of the *p-n* junction.

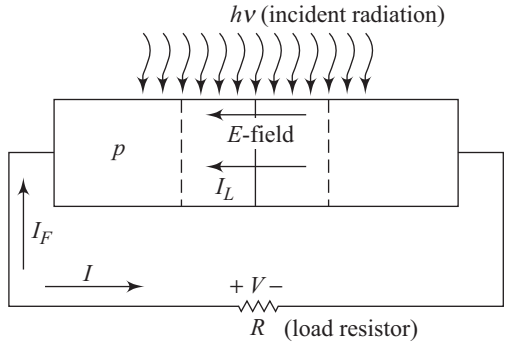


Fig. 10.5 An illuminated *p-n* junction with a load resistor

Notice that the forward bias lowers the built-in voltage and hence the built-in electric field without changing its direction. The net current, as given by Eqn (10.11a), is in the reverse-bias direction. This is because the photocurrent is itself in the reverse-bias direction and is always more than the forward-bias current due to the drop across the load resistor. Typical *I-V* characteristics of a *p-n* junction solar cell are shown in Fig. 10.6.

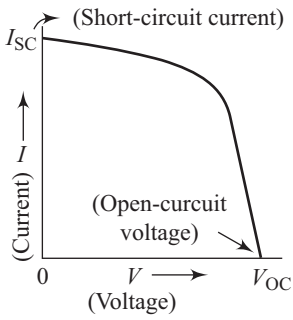


Fig. 10.6 *I-V* characteristics of a *p-n* junction solar cell

Two important points on the *I-V* characteristic curve are the short-circuit current, I_{SC}, and the open-circuit voltage, V_{OC}. Under short-circuit conditions, $R = 0$ and from Eqn (10.11a) we get

$$I = I_L = I_{SC} \tag{10.11b}$$

On the other hand, $R = \infty$ corresponds to the open-circuit condition. Putting $I = 0$ ($R = \infty$) in Eqn (10.11a) leads to

$$I = 0 = I_L - I_S\left[\exp\left(\frac{eV\text{oc}}{kT}\right) - 1\right] \tag{10.12}$$

Equation (10.12) can be rewritten in the form

$$V_{\text{oc}} = \frac{kT}{e}\ln\left(1 + \frac{I_L}{I_S}\right) \tag{10.13}$$

The complete figure generated using Eqn (10.11a) is shown in Fig. 10.6.

The power P delivered to the load R_L is given by

$$P = IV \tag{10.14}$$

which, on using Eqn (10.11a), results in

$$P = I_L V - I_S\left[\exp\left(\frac{eV}{kT}\right) - 1\right]V \tag{10.15}$$

For maximum power to be delivered to the load, we must have

$$\frac{dP}{dV} = 0 = I_L - I_S\left[\exp\left(\frac{eV_m}{kT}\right) - 1\right] - I_S V_m\left(\frac{e}{kT}\right)\exp\left(\frac{eV_m}{kT}\right) \tag{10.16}$$

where V_m gives the voltage across the solar cell at maximum power. Equation (10.16) can be rewritten in the form

$$\left(1 + \frac{eV_m}{kT}\right)\exp\left(\frac{eV_m}{kT}\right) = 1 + \frac{I_L}{I_S} \tag{10.17}$$

which leads to

$$V_m = \frac{kT}{e}\ln\left[\frac{1 + (I_L / I_S)}{1 + (eV_m / kT)}\right] \tag{10.18}$$

which yields

$$V_m \cong V_{\text{oc}} - \frac{kT}{e}\ln\left[1 + \left(\frac{eV_m}{kT}\right)\right] \tag{10.19}$$

and

$$I_m = I_S\left[\frac{eV_m}{kT}\right]e^{eV_m / kT} \simeq I_L\left[1 - \frac{1}{eV_m / kT}\right] \tag{10.20}$$

The maximum power output, P_m, is given by

$$P_m = I_m V_m \tag{10.21}$$

which on using Eqs (10.19) and (10.20) becomes

$$P_m \cong I_L\left[V_{\text{OC}} - \frac{kT}{e}\ln\left[1 + \frac{eV_m}{kT}\right] - \frac{kT}{e}\right] \tag{10.22}$$

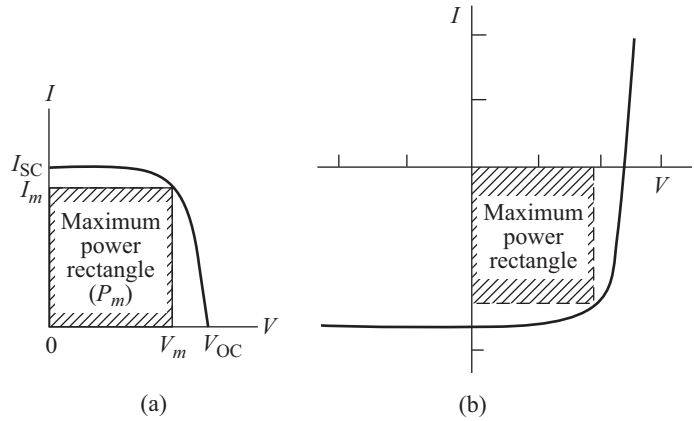


Fig. 10.7 (a) Maximum power rectangle of a photovoltaic cell (b) *I-V* characteristic in the fourth quadrant

Figure 10.7(a) shows the maximum power rectangle of a typical solar cell. Figure 10.7(b) shows the same curve with the proper choice of axes. Current I_m is the current applicable with $V = V_m$.

10.2.2 Conversion Efficiency

A solar cell converts light energy to electrical energy. The conversion efficiency of a solar cell is the ratio of the output electrical power to the incident optical or photon power. Assuming maximum power output, we can write for conversion efficiency, η,

$$\eta = \frac{P_m}{P_{\text{in}}} \times 100\% \tag{10.23}$$

which leads to

$$\eta = \frac{I_m V_m}{P_{\text{in}}} \times 100\% \tag{10.24}$$

We can now define a parameter called the fill factor (FF) as

$$\text{FF} = \frac{I_m V_m}{I_L V_{\text{OC}}} \tag{10.25}$$

Using Eqn (10.25) in Eqn (10.24) results in

$$\eta = \frac{\text{FF} \cdot I_L V_{\text{OC}}}{P_{\text{in}}} \tag{10.26}$$

FF lies in the range 0.7–0.8 for a well-designed solar cell. It is clear that all the terms in the numerator in Eqn (10.26) need to be maximized to maximize efficiency. The solar spectrum will contain photons in a wide wavelength range as shown in Fig. 10.8. The figure also gives the critical wavelength, λ_C for some common semiconductors such as Si and GaAs.

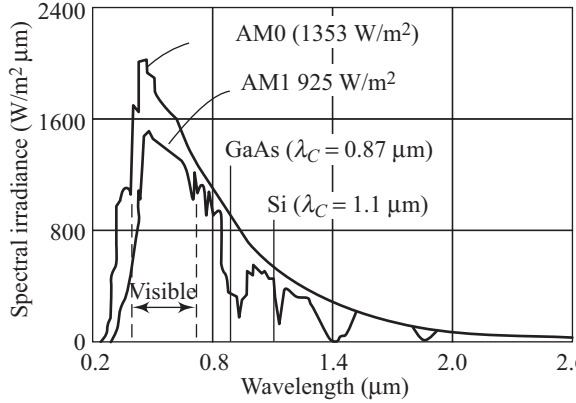


Fig. 10.8 Solar spectrum under air mass (0) and air mass (1) conditions

Solar spectrum outside the earth's atmosphere is the air mass zero condition (AM0), whereas AM1 represents the sunlight at the earth's surface with the sun at zenith. Photons with energy less than the band gap energy do not produce any photovoltaic effects whereas any energy in the excess of E_g gets wasted as heat. Considering these factors, a silicon *p-n* junction solar cell has a maximum theoretical efficiency of around 28 per cent. This ideal maximum efficiency is however not realized in practice due to dissipating factors such as series resistance and reflection losses. Various measures like anti-reflection coating and texturization are used to bring down the reflection losses. Typical efficiency values for silicon *p-n* junction solar cells fall in the range of 10–15 per cent.

One important technique of increasing the conversion efficiency is to use concentrators to increase the light intensity falling on the solar cell surface. Using such concentrators, it is possible to increase the light intensity falling on a solar cell to several hundred times. An increase in light concentration increases the short-circuit current linearly. On the other hand, the open-circuit voltage increases only slightly with increase in incident light concentration. Figure 10.9 shows the ideal solar cell efficiency at 300 K for two different values of solar concentrations, namely 1 sun and 1000 sun. The energy band gap of some common semiconductors is also shown in the same figure. Although the incident light intensity has been increased a thousand times, the corresponding increase in efficiency is not comparable. Concentration techniques are still vastly popular because the optical

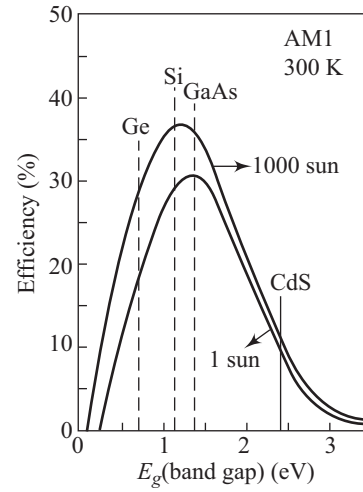


Fig. 10.9 Ideal solar cell efficiency at 300 K

lens systems used to increase the intensity of solar light are still substantially cheaper than the cost involved in increasing the surface area of solar cells to provide an equivalent increase in their efficiency.

10.2.3 Effect of Series Resistance

A finite series resistance, R_S, of a solar cell will modify the ideal *I-V* characteristics. Equation (10.10), therefore, gets replaced with

$$I = I_L - I_S\left[\exp\left\{\frac{e(V - IR_S)}{kT}\right\} - 1\right] \tag{10.27}$$

which yields

$$\ln\left[\frac{I + I_L}{I_S} + 1\right] = \frac{e}{kT}(V - IR_S) \tag{10.28}$$

Figure 10.10 shows the reduction in the maximum power achievable from a solar cell taking into account a finite series resistance value. The maximum achievable power can reduce to 25–30 per cent of its value with negligible series resistance. One can also notice that the open-circuit voltage is not affected much in comparison with the short-circuit current.

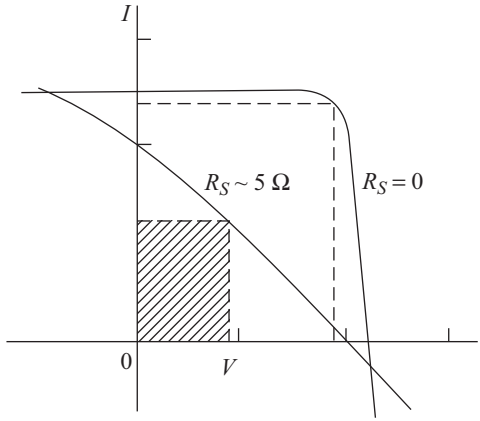


Fig. 10.10 Reduction of maximum power due to finite series resistance

10.2.4 Heterojunction Solar Cells

A heterojunction is formed between two semiconductors having different band gap energies. Figure 10.11 shows the energy band diagram of a typical *p-n* heterojunction in thermal equilibrium. The band gap of the *n*-type semiconductor is clearly more than that of the *p*-type semiconductor.

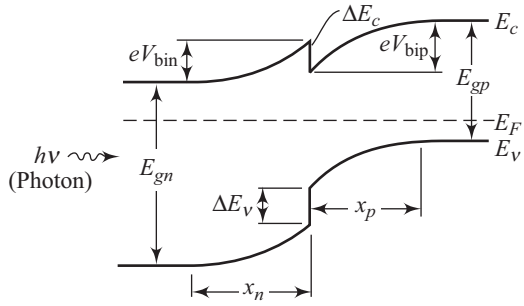


Fig. 10.11 Band diagram of *p-n* heterojunction

As can be seen from Fig. 10.11, photons are incident on the wide band gap semiconductor. Photons having energy greater than E_{gn} are absorbed in the wider band gap material and create an excess of carriers. The excess carriers generated within a diffusion length of the junction collectively contribute to the output current. Photons with energy less than E_{gn} go through the wide-band gap semiconductor till they reach the narrow band gap semiconductor. Photons with energy greater than E_{gp} are absorbed in the narrow-band gap semiconductor and generate excess charge carriers. Excess carriers created within a diffusion length from the junction once again are collected and thus contribute to the photocurrent. By selecting a large enough E_{gn}, it is possible to ensure absorption of high-energy photons in the space charge region of the narrow-band gap material. Such a heterojunction solar cell will, therefore, have a better characteristic curve at shorter wavelengths than the corresponding homojunction solar cell.

Figure 10.12 is a schematic representation of another type of heterojunction solar cell. The structure consists of a *p-n* homojunction formed in GaAs with a wide-band gap $Al_xGa_{1-x}As$ *p*-type semiconductor deposited over the *p-n* homojunction. The wide-band gap semiconductor allows photons with energies $h\nu < E_{g1}$ to pass through and reach the homojunction. In the homojunction, photons with energies in the range $E_{g2} < h\nu < E_{g1}$ create excess carriers. Photons with $h\nu > E_{g1}$ create excess carriers in the wide-band gap semiconductor. The narrow band gap semiconductor is so chosen that the absorption coefficient has a very high value. This ensures that the excess carriers created are limited to a depth within a diffusion length from the junction. This leads to a very high collection efficiency.

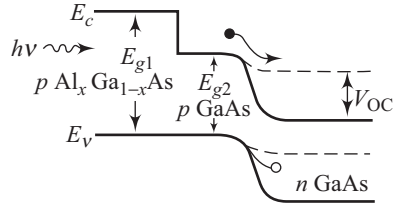


Fig. 10.12 Heterojunction solar cell with wide-band gap top layer

10.2.5 Amorphous Silicon Solar Cells

Solar cells have a variety of application areas. Some of these include space applications, consumer goods, and power generation. In some applications the cost factor plays a major role in deciding the viability. Single-crystal silicon solar cells are efficient but can prove to be expensive. Solar cells using amorphous silicon provide a viable alternative in such circumstances. Amorphous form of silicon exhibits only very short-range order with no crystalline regions. Figure 10.13 shows the density of states versus energy plot of amorphous silicon. Unlike crystalline silicon, a large number of electronic states exists within the normally empty band gap. The effective mobility is, however, small due to the short-range order. Typical values are in the range 10^{-6}–10^{-3} cm^2/V s. The mobilities in the states below E_v and above E_c are, however, between 1 and 10 cm^2/V s. Due to this widely varying mobility E_c and E_v are referred to as mobility edges instead of conduction and valence band edges and the difference between them is designated as the mobility gap. Interestingly, a modification of mobility gap is possible by incorporation of typical impurities.

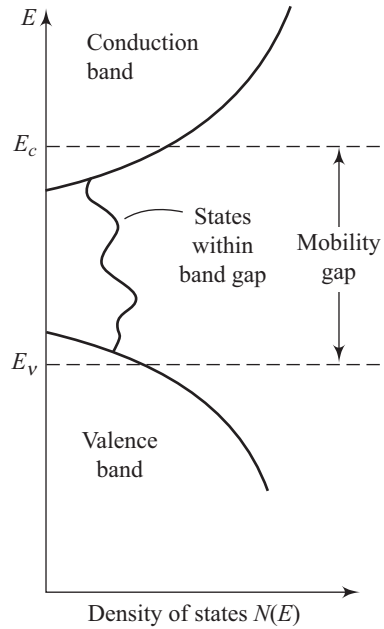


Fig. 10.13 Density of states vs energy plot for amorphous silicon

The mobility gap in amorphous silicon is around 1.7 eV. A typical PIN amorphous silicon solar cell is shown schematically in Fig. 10.14.

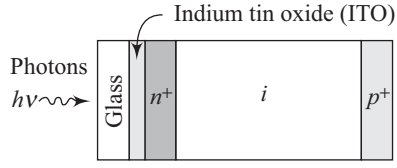


Fig. 10.14 Schematic diagram of a PIN amorphous silicon solar cell

The n^+ and p^+ regions are extremely thin whereas the intrinsic region is around 0.5 to 1.0 μm thick. Such thicknesses are usable due to the very high absorption coefficient in amorphous silicon. Figure 10.15(a) shows the energy band diagram for thermal equilibrium condition, whereas the band diagram for a device under illumination is shown in Fig. 10.15(b). As can be seen, the open-circuit voltage results due to the separation of E_{Fn} and E_{Fp}.

The EHPs generated in the intrinsic region get separated due to the built-in electric field and result in a photocurrent. Although the conversion efficiency of amorphous silicon based solar cells is low, the low cost is an extremely attractive feature.

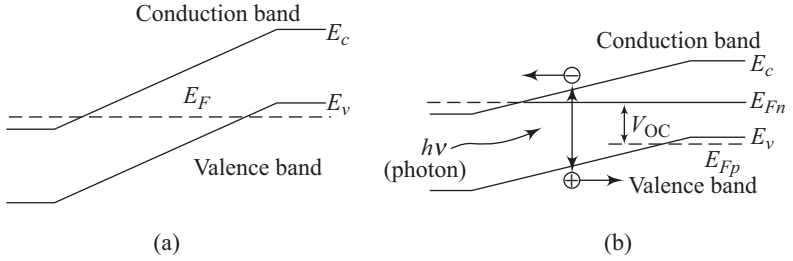


Fig. 10.15 (a) Energy band diagram of PIN solar cell (b) Energy band diagram under illumination

- Diffusion lengths in Si and GaAs increase with increasing temperature. This leads to an increase in the magnitude of I_L with temperature. The open circuit voltage, V_{OC}, rapidly decreases with increasing temperature due to the exponential dependence of reverse saturation current. This leads to an increasing *softness* in the knee of the *I-V* characteristic curve, degrading the fill factor. The overall effect is a reduction in conversion efficiency with increasing temperature.
- Thin-film solar cells use active semiconductor layers in the form of polycrystalline or disordered films on electrically active or passive substrates, such as glass, plastic, ceramic, metal, graphite, and so on. Semiconductors used are typically CdS, Si, GaAs, InP, CdTe, etc. The chief advantage of thin-film solar cells is their low cost.

10.3 Photodetectors

Photodetectors are devices that convert optical signals to electrical signals. Photons give their energies to valence band electrons and excite them to the conduction band. The resultant EHPs increase the conductivity of the material. All photodetectors use these excess carriers in some way to detect incident radiations. Photoconductors, photodiodes, and phototransistors are some examples of such devices.

10.3.1 Photoconductors

A photoconductor consists of a bar of semiconductor material with ohmic contacts at the two ends. Figure 10.16 shows a semiconductor bar of length L, cross-sectional area A, and an illumination towards the larger face area.

In the absence of illumination, the thermal equilibrium conductivity, σ_0, is given by

$$\sigma_0 = e(\mu_n n_0 + \mu_p p_0) \tag{10.29}$$

If the photogenerated excess carriers are δn and δp, then the conductivity, σ, becomes

$$\sigma = e[\mu_n (n_0 + \delta n) + \mu_p(p_0 + \delta p)] \tag{10.30}$$

Without any loss of generality, we will assume the semiconductor to be *n*-type. Using the principle of charge neutrality we can write

$$\delta n = \delta p$$

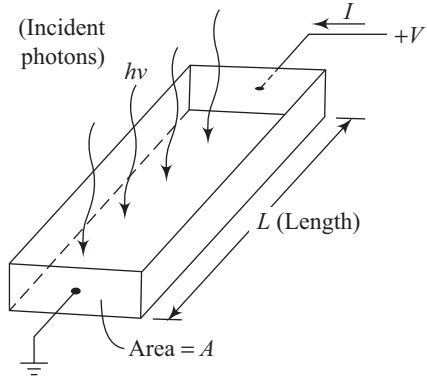


Fig. 10.16 Semiconductor bar under illumination

In the following part of the discussion we will use δp to represent excess carrier concentration. If G_L represents the photogeneration rate of excess carriers then we have.

$$G_L = \frac{\delta p}{\tau_p} \tag{10.31}$$

where τ_p is the excess minority carrier lifetime. Equation (10.30) can be rewritten in the form

$$\sigma = e(\mu_n n_0 + \mu_p p_0) + e(\delta p)\ (\mu_n + \mu_p) \tag{10.32}$$

The conductivity change due to incident light is called photoconductivity and is given by

$$\Delta\sigma = \mathrm{e}(\delta p)\ (\mu_n + \mu_p) \tag{10.33}$$

The total current density, *J*, can be written as

$$J = (J_0 + J_L) = (\sigma_0 + \Delta\ \sigma)E \tag{10.34}$$

where J_o is the initial current density and the term J_L represents the current density due to optical excitation. For uniform excess carrier generation throughout the semiconductor we have

$$I_L = J_L A = \Delta\ \sigma\, AE = eG_L \tau_p\ (\mu_n + \mu_p) AE \tag{10.35}$$

Thus the photocurrent can be seen to be directly proportional to the excess carrier generation rate, and therefore, to the incident photon flux. The electron drift velocity is expressible as

$$v_{\mathrm{de}} = \mu_n E \tag{10.36}$$

Electron transit time, t_n, is given by

$$t_n = \frac{L}{v_{\text{de}}} = \frac{L}{\mu_n E} \tag{10.37}$$

which leads to

$$E = \frac{L}{\mu_n t_n} \tag{10.38}$$

Using Eqn (10.38) in Eqn (10.35) yields

$$I_L = eG_L\left(\frac{\tau_p}{t_n}\right)\left(1 + \frac{\mu_p}{\mu_n}\right)AL \tag{10.39}$$

Photoconductor gain, G_{ph}, is defined as the ratio of the rate of collection of charge by the contacts, to the rate of generation of charge within the photoconductor. Thus

$$G_{\text{ph}} = \frac{I_L}{eG_L AL} \tag{10.40}$$

Using the expression for I_L from Eqn (10.39) in Eqn (10.40) results in

$$G_{\text{ph}} = \frac{\tau_p}{t_n}\left(1 + \frac{\mu_p}{\mu_n}\right) \tag{10.41}$$

The switching speed or frequency response is inversely proportional to minority carrier lifetime. From Eqn (10.41) one can conclude that the gain of a photoconductor is directly proportional to the minority carrier lifetime. Requirements of speed and gain are, therefore, mutually contradictory and a trade-off has to be worked out between the two.

10.3.2 Photodiodes

As outlined earlier, a *p-n* junction can be used to separate photogenerated excess carriers. A photodiode uses a reverse biased *p-n* junction diode to detect photons. Figure 10.17 shows the schematic diagram of a long reverse biased diode alongwith the associated minority carrier concentration.

Figure 10.17 uses two different origins so that the *p* and *n* regions can be treated separately. Photocurrent arises due to three different sources. There are, excess carriers in the space charge region, excess carriers in the neutral *p*-region, and excess carriers in the neutral *n*-region. We will handle each of these regions separately.

Space charge region

The excess carriers generated within the space charge region get collected due to the built-in electric field. If G_L represents the photogenerated excess carrier generation rate, the photocurrent density, J_{L1}, is given by

$$J_{L1} = e\int_W G_L dx \tag{10.42}$$

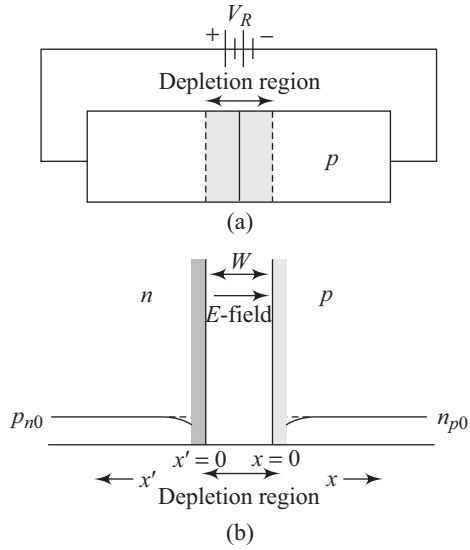


Fig. 10.17 (a) Schematic diagram of reverse biased long diode (b) Minority carrier concentration

If the generation rate is constant throughout the space charge width, we can write

$$J_{L1} = eG_L\, W \tag{10.43}$$

A comparison of Eqs (10.43) and (10.40) reveals that the *photodiode gain is unity*.

Neutral *p*-region and *n*-region

Transport equation for excess minority carrier electrons in the *p*-region is

$$D_n \frac{\partial^2(\delta n_p)}{\partial x^2} + G_L - \frac{\delta n_p}{\tau_{n0}} = \frac{\partial(\partial n_p)}{\partial t} \tag{10.44}$$

where the electric field, E, has been assumed to be zero in the neutral p-region. At steady state

$$\frac{\partial(\partial n_p)}{\partial t} = 0 \tag{10.45}$$

Thus Eqn (10.44) can be rewritten in the form

$$\frac{d^2(\delta n_p)}{dx^2} - \frac{\delta n_p}{D_n \tau_{n0}} = \frac{-G_L}{D_n} \tag{10.46}$$

Since

$$D_n\, \tau_{n0} = L_n{}^2 \tag{10.47}$$

We can write Eqn (10.46) in the form

$$\frac{d^2(\delta n_p)}{dx^2} - \frac{\delta n_p}{L_n^2} = \frac{-G_L}{D_n} \tag{10.48}$$

Equation (10.48) can be solved by adding the homogeneous and particular solutions. To find the homogeneous solution, let us drop the G_L term, giving

$$\frac{d^2(\delta n_p)}{\partial x^2} - \frac{\delta n_p}{L_n^2} = 0 \tag{10.49}$$

Equation (10.49) has solutions of the form

$$\delta n_{p1} = A_1 e^{-x/L_n} + B_1 e^{x/L_n} \tag{10.50}$$

Assuming finite values of δn_{p1}, for a long diode, we must have $B_1 = 0$. This leads to the homogeneous solution

$$\delta n_{p1} = A_1 e^{-x/L_n} \tag{10.51}$$

The particular solution for Eqn (10.48) requires that

$$\frac{\delta n_{p2}}{L_n^2} = \frac{G_L}{D_n} \tag{10.52}$$

yielding

$$\delta n_{p2} = G_L \tau_{n0} \tag{10.53}$$

Combining Eqs (10.51) and (10.53) results in

$$\delta n_p = A_1 e^{-x/L_n} + G_L \tau_{n0} \tag{10.54}$$

For a reverse biased junction, the total electron concentration, n_p, must be zero at $x = 0$. This implies

$$\delta n_p(x = 0) = -n_{p0} \tag{10.55}$$

Using Eqn (10.55) in Eqn (10.54) yields

$$\delta n_p = G_L \tau_{n0} - (G_L \tau_{n0} + n_{p0}) e^{-x/L_n} \tag{10.56}$$

Similarly, the excess hole concentration in the neutral *n*-region is given by

$$\delta p_n = G_L \tau_{p0} - (G_L \tau_{p0} + n_{n0}) e^{-x'/L_p} \tag{10.57}$$

where the notation x' has been used in keeping with Fig. 10.17. Excess carrier concentrations as given by Eqs (10.56) and (10.57) are plotted in Fig 10.18.
Minority carrier concentration gradients on either side of the junction produce diffusion currents. The diffusion current density due to minority carrier electrons at the space charge region edge is given by

$$J_{n1} = eD_n \left. \frac{d(\delta n_p)}{dx} \right|_{x=0} \tag{10.58}$$

Using Eqn (10.56) in Eqn (10.58) yields

$$J_{n1} = eD_n \frac{d}{dx} \left[G_L \tau_{n0} - (G_L \tau_{n0} + n_{p0}) e^{-x/L_n} \right] \Big|_{x=0} \tag{10.59}$$

implying

$$J_{n1} = \frac{eD_n}{L_n} (G_L \tau_{n0} + n_{p0}) \tag{10.60}$$

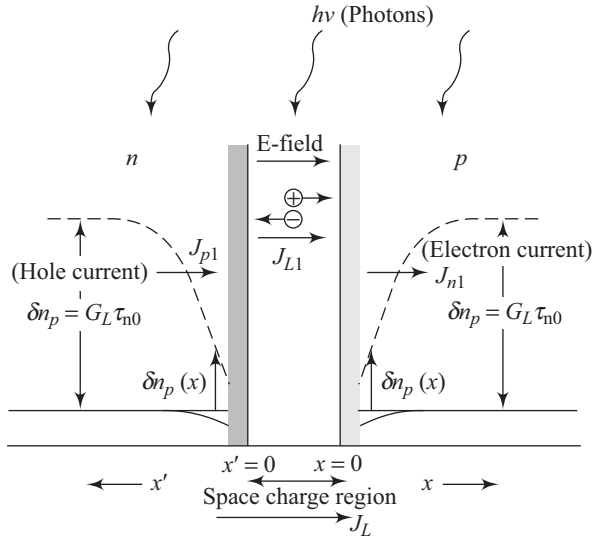


Fig. 10.18 Excess carrier concentration for a photodiode

which can be rewritten in the form

$$J_{n1} = eG_L L_n + \frac{eD_n n_{p0}}{L_n} \tag{10.61}$$

Similarly, the diffusion current density at the other edge of the space charge region, due to minority carrier holes, is given by

$$J_{p1} = eG_L L_p + \frac{eD_p p_{no}}{L_p} \tag{10.62}$$

Notice that the first terms in Eqs (10.61) and (10.62) represent the current density due to incident photons and the second terms represent the ideal reverse-saturation current density.

The total photocurrent density for a long diode is obtained from Eqs (10.43), (10.61), and (10.62). The final expression is

$$J_L = eG_L W + eG_L L_n + eG_L L_p \tag{10.63}$$

which can be simplified to

$$J_L = eG_L (W + L_n + L_p) \tag{10.64}$$

PIN photodiode

The photodiode discussed in the earlier section is inherently slow due to the dominance of diffusion based currents. In many photodetector applications such low speeds are not acceptable. Light-based communication systems are one example of such an application. The photogenerated carriers in the space charge region can however be collected very fast due to the built-in electric field. Fast photodiodes must, therefore, use large depletion widths to enhance the speed of

response. PIN photodiodes use this concept to achieve fast responses. A PIN diode consists of an intrinsic semiconductor region sandwiched between two heavily doped semiconductor regions of opposite type. Under applied reverse bias the complete intrinsic region can be swept off and constitutes the space charge region. A schematic representation of a typical PIN photodiode with a suitable biasing arrangement is shown in Fig. 10.19.

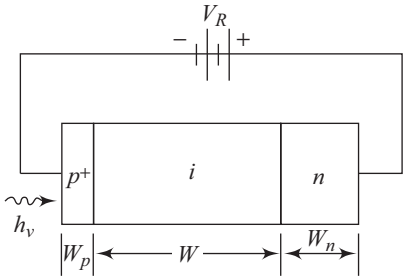


Fig. 10.19 Schematic representation of a PIN photodiode with biasing arrangement

Let us assume that photons are incident on the p^+ face of the PIN photodiode. If ϕ_i represents the incident photon flux, then, assuming a thin p^+ region width, the photon flux $\phi(x)$ at any depth x within the i-region can be written in the form

$$\phi(x) = \phi_i e^{-\alpha x} \tag{10.65}$$

where α represents the absorption coefficient. The photocurrent density due to generation of excess carriers in the intrinsic region is then given by

$$J_L = e\int_0^W G_L dx = e\int_0^W \phi_i \alpha e^{-\alpha x} dx \tag{10.66}$$

leading to

$$J_L = e\phi_i(1 - e^{-\alpha W}) \tag{10.67}$$

Equation (10.67) is valid if no electron–hole recombination takes place in the space charge region.

Avalanche photodiode

Under certain conditions, the reverse bias applied across a photodiode reaches values that are sufficient to create electron–hole pairs by impact ionization. The excess carriers initially generated due to the absorption of photons create additional electron–hole pairs through the process of impact ionization. This process leads to current gain in avalanche photodiodes.

The saturation velocity, v_s, in silicon is in the range of 10^7 cm/s. If W_d is the width of the depletion region, the transit time, τ_t, is given by

$$\tau_t = \frac{10^7}{W_d \times 10^{-4}} \tag{10.68}$$

The corresponding modulating frequency, f_m, is then expressed by

$$f_m = \frac{1}{2\tau_t} \tag{10.69}$$

Using Eqs (10.68) and (10.69) one can easily calculate that for $W_d \sim 10$ μm, $f_m \sim$ 5 GHz, and for photodiode current gain in the region 10–20, the gain–bandwidth

product will fall in the range 50–100 GHz. Thus avalanche photodiodes can be used to detect modulation of light waves at microwave frequencies.

10.3.3 Phototransistors

A schematic representation of an *n-p-n* bipolar phototransistor is shown in Fig. 10.20.

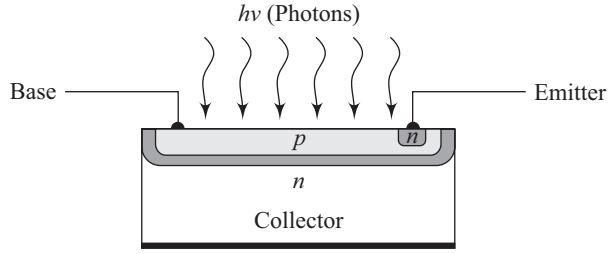


Fig. 10.20 Schematic representation of *n-p-n* phototransistor

The base terminal is generally kept open when a BJT is being operated as a phototransistor. Photons create electron–hole pairs in the large area *B-C* junction. The built-in electric field in the space charge region sweeps out these charge carriers to produce a photocurrent I_L. The holes created are swept away into the *p*-type base to make the base positive with respect to the emitter. This results in forward biassing of the *B-E* junction and electrons get emitted from the emitter region towards the base region. This leads to the usual transistor operation.

The frequency response of the phototransistor is limited by the *B-C* junction capacitance, due to its large area. Phototransistors can also be fabricated using heterostructures. Band gap differences are exploited in such devices to improve the injection efficiency.

10.4 Light-emitting Diodes

When a *p-n* junction is forward biased, injection of carriers takes place across the junction. As the injected carriers travel, they undergo recombination in the space charge region and in the neutral regions close to the junction. In the case of indirect gap semiconductors like Si and Ge this recombination releases heat to the lattice. In direct band gap materials, on the other hand, the released energy can be given off as light emitted from the junction. This process is called *injection electroluminescence* and is the fundamental principle underlying the working of light-emitting diodes (LEDs). LEDs find application in digital displays and communication systems. The wavelength of the emitted light is governed by the equation

$$\lambda = \frac{hc}{E_g} = \frac{1.24}{E_g}\,\mu\text{m} \tag{10.70}$$

where E_g is the band gap in eV.

The general process of emission of a photon due to the recombination of excess electrons and holes is referred to as *luminescence*. If the excess electrons and holes are created by photon absorption, then the resultant photon emission due to the recombination process is called *photoluminescence*.

10.4.1 LED Materials and Devices

Gallium arsenide has a band gap, $E_g = 1.42$ eV. This corresponds to $\lambda = 0.873$ μm using Eqn (10.70). Thus, an LED based on GaAs will have output outside the visible range (0.4 to 0.72 μm). The corresponding band gap energy range would be ~ 1.7 to 3.1 eV. $GaAs_{1-x}P_x$ is a direct-band gap semiconductor for $0 \leq x \leq 0.45$ as shown in Fig. 10.21.

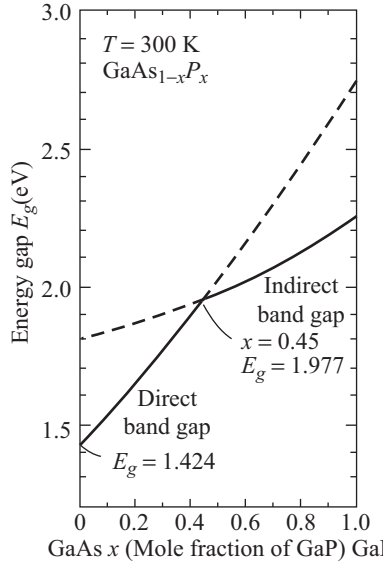


Fig. 10.21 Band gap energy versus mole fraction of GaP

As the mole fraction changes to values greater than 0.45, the material changes to an indirect-band gap semiconductor and is therefore not a very efficient choice as an LED material. For indirect-band gap semiconductors, special recombination centres need to be incorporated to enhance radiative processes. Nitrogen incorporation in the crystal lattice of $GaAs_{1-x}P_x$ can result in the formation of such a recombination centre. The nitrogen introduced replaces phosphorous atoms in the lattice sites. Difference in the electronic core structure (outer electronic structure of nitrogen and phosphorous is identical) results in the creation of an electron trap level close to the bottom of the conduction band, thereby enhancing the probability of radiative transition in an indirect-band gap semiconductor.

Quantum efficiency of an LED is defined as the number of photons generated per electron–hole pair. Figure 10.22 shows plots of quantum efficiency versus alloy composition of $GaAs_{1-x}P_x$ with and without the isoelectronic impurity, nitrogen.

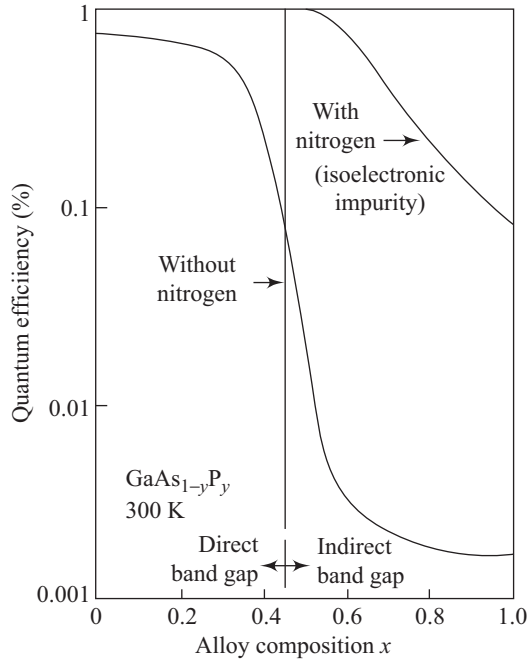


Fig. 10.22 Quantum efficiency for different compositions of Ga As_{1-x} P_x

Due to the changeover from direct band gap to indirect band gap at $x = 0.45$, the quantum efficiency of LED without nitrogen drops sharply around this alloy composition.

10.4.2 Loss Mechanisms and Structure

The quantity of photons emitted by an LED gets reduced due to three main mechanisms. These are listed below.

(i) Absorption of emitted photons within the LED semiconductor
(ii) Reflection losses as light passes from a semiconductor to air due to refractive index difference
(iii) Total internal reflection

Photons generated in the semiconductor junction are incident on the semiconductor–air interface. If the photons are incident at an angle greater than the critical angle, θ_c, for the interface, the photons go back into the semiconductor due to total internal reflection. The process is shown schematically in Fig. 10.23.

The critical angle for any pair of materials can be determined from Snell's law given by

$$\theta_c = \sin^{-1}\left(\frac{\overline{n}_1}{\overline{n}_2}\right) \tag{10.71}$$

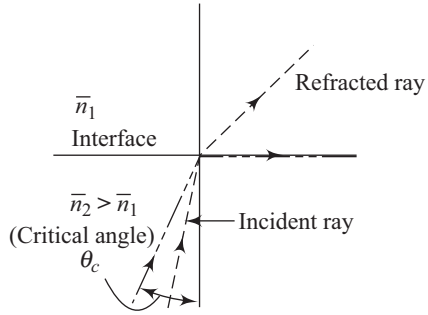


Fig. 10.23 Total internal reflection at a refracting surface

For GaAs–Air interface and for $\lambda \cong 0.8$ μm, $\overline{n}_2 = 3.66$ and $\overline{n}_2 = 1$, θ_c is about 16°. For GaP–Air interface and for $\lambda \cong 0.8$ μm, $\overline{n}_2 = 3.45$, θ_c is about 17°.

Figure 10.24 shows the schematic diagram of an LED lamp. It consists of a main LED chip and a plastic lens that is coloured to serve as an optical filter and to enhance the contrast.

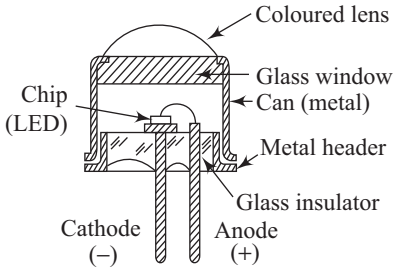


Fig. 10.24 Parts of an LED lamp

10.5 Laser Diodes

The photons produced in an LED result due to transitions of electrons from the conduction band to the valence band. The light emission is spontaneous because each band-to-band transition is in itself an independent event. The spectral output of the LED, however, has a fairly wide bandwidth. With suitable modifications in the structure and operating conditions of the LED, it is possible to operate the device in a mode that produces coherent output with wavelength bandwidths less than 0.1 nm. This special mode operated device is then called a laser diode (laser stands for light amplification by stimulated emission of radiation).

10.5.1 Materials and Structures

Semiconductor lasing can be obtained using direct band gap semiconductors. Gallium arsenide was the first material to emit laser radiations and continues to be used along with related group III–V compound alloys. Semiconductor lasers using heterostructures with negligible interface traps must use semiconductors with very low lattice mismatches. For GaAs ($a = 5.6533$ Å) substrates, the ternary compound $Al_x\ Ga_{1-x}\ As$ with lattice mismatch less than 0.1 per cent is a good choice. Similarly, for InP ($a = 5.8686$ Å) substrates, the quaternary compound ($Ga_x\ In_{1-x}\ As_y\ P_{1-y}$) with a nearly identical lattice constant, is a perfect match.

Laser structures are commonly available in three variants shown in Fig. 10.25.

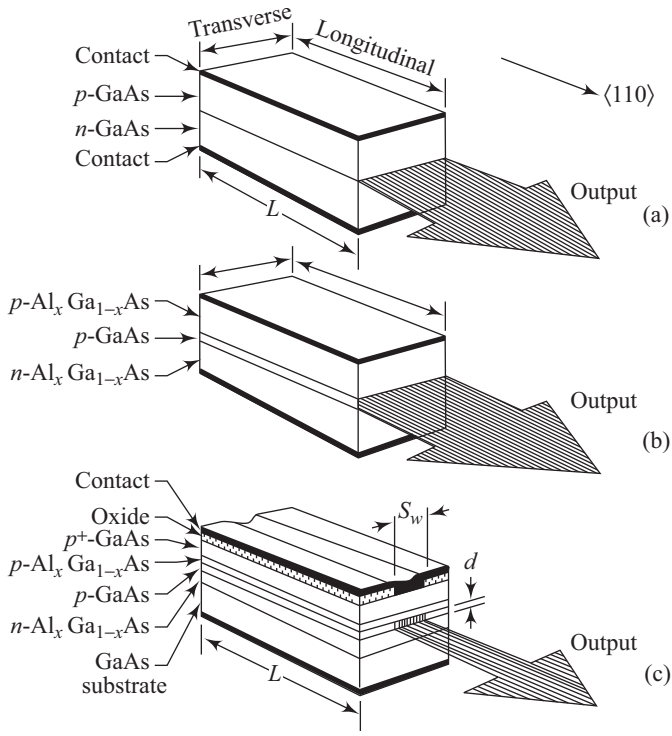


Fig.10.25 Commonly available laser structures

Figure 10.25(a) shows a *homojunction laser*. A pair of parallel facets are cleaved or polished perpendicular to the ⟨110⟩ axis. Suitable biasing conditions lead to laser lights being emitted from these planes. The two remaining facets are roughened to eliminated *lasing* action in these directions. This structure is called a Fabry–Perot cavity (typical cavity lengths are ~ 300 μm) and is extensively used. Figure 10.25(b) shows a double-heterostructure (DH) laser. The structure consists of a thin layer of a semiconductor (e.g., GaAs in Fig. 10.25(b)) sandwiched between layers of another semiconductor (e.g., Al_x Ga_{1-x} As). A DH laser requires much less current to operate than a homojunction laser with a similar geometry. The homojunction and double-heterostructure laser structures are large-area lasers since the entire area along the plane of the junction can participate in emission of radiation. Figure 10.25(c) shows a stripe geometry laser. An oxide layer is used in this structure to isolate all but the stripe contact. The lasing area gets restricted to a narrow region under the contact. The structure leads to reduced operating current, elimination of multiple-emission areas, and better reliability.

10.5.2 Population Inversion

Population inversion is an important requirement for lasing action. We will take the case of a *p-n* junction to understand the process involved in population

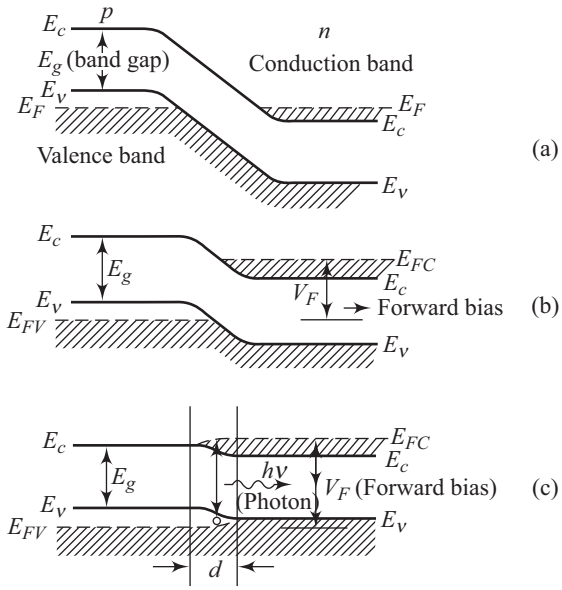


Fig. 10.26 Energy band diagram showing population inversion

inversion. Figure 10.26 shows the energy band diagram of a *p*-*n* junction formed between degenerate semiconductors.

Due to the use of degenerate semiconductors, the fermi levels are below the valence band edge and above the conduction band edge, for *p* and *n* sides respectively. As we apply a forward bias, V_F, electrons are injected from the *n*-side and holes are injected from the *p*-side. At high enough forward bias (Fig. 10.26(c)) the device reaches high injection condition with a large concentration of holes and electrons in the transition region. As can be seen, the region *d* contains a large electron concentration in the conduction band and a large hole concentration in the valence band. Thus a population inversion is achieved. From Fig. 10.26 it is clear that the essential condition for population inversion is

$$(E_{FC} - E_{FV}) > E_g \tag{10.72}$$

- LED and semiconductor laser fall in the category of luminescent devices. Luminescence refers to the emission of optical radiation due to electronic excitation in a material. Different types of luminescence include photoluminescence, cathodoluminescence, radioluminescence, and electroluminescence.
- Emission wavelength of a semiconductor laser can be changed by varying diode current, heat-sink temperature, application of pressure, application of magnetic field, etc. The heat-sink temperature changes the temperature of the semiconductor at equilibrium which, in turn, changes its band gap. Hydrostatic pressure also changes the band gap of semiconductors.

Solved Problems

10.1 Calculate the thickness of silicon required to absorb 80% of the incident photon energy at $\lambda = 1.0$ μm. Assume that the absorption coefficient of silicon for $\lambda = 1.0$ μm is 10^2 cm^{-1}.

Solution

For 80% absorption we can write

$$\frac{I_v(d)}{I_{v0}} = 0.2 = e^{-\alpha d} \tag{10.1.1}$$

where $I_v(d)$ represents the intensity of photon flux at a depth d within the semiconductor.

Putting the given value of α in Eqn (10.1.1) results in

$$d = \frac{1}{10^2}\ln\left(\frac{1}{0.2}\right)$$

$$= 0.016 \text{ cm}$$

10.2 A *p-n* junction solar cell is fabricated using silicon and has the following important parameters:

$N_a = 3 \times 10^{18}$ cm^{-3} $\quad N_d = 2 \times 10^{16}$ cm^{-3}

$D_n = 25$ cm^2/s $\quad D_p = 10$ cm^2/s

$\tau_{n0} = 4 \times 10^{-7}$ s $\quad \tau_{p0} = 10^{-7}$ s

The photocurrent density $J_L = 20$ mA/cm^2. Calculate the open-circuit voltage of the solar cell at $T = 300$ K Assume $n_i = 1.5 \times 10^{10}$ cm^{-3}.

Solution

The reverse saturation current density, J_S, is given by

$$J_S = \frac{eD_n n_{p0}}{L_n} + \frac{eD_p p_{n0}}{L_p} \tag{10.2.1}$$

which can be rewritten in the form

$$J_S = en_i^2\left(\frac{D_n}{L_n N_a} + \frac{D_p}{L_p N_d}\right) \tag{10.2.2}$$

Also

$$L_n = \sqrt{D_n \tau_{n0}} = \sqrt{25 \times 4 \times 10^{-7}} = 31.6\mu\text{m} \tag{10.2.3}$$

and

$$L_p = \sqrt{D_p \tau_{p0}} = \sqrt{10 \times 10^{-7}} = 10\ \mu\text{m} \tag{10.2.4}$$

Using Eqs (10.2.3), (10.2.4), and other given values in (10.2.2) leads to

$$J_S = (1.6 \times 10^{-19})(1.5 \times 10^{10})^2 \left[\frac{25}{(31.6 \times 10^{-4})(3 \times 10^{18})} + \frac{10}{(10 \times 10^{-4})(2 \times 10^{16})} \right] \text{A/cm}^2$$

$$= 1.81 \times 10^{-11}\ \text{A/cm}^2$$

The open-circuit voltage, V_{OC}, is given by

$$V_{OC} = \frac{kT}{e} \ln\left(1 + \frac{J_L}{J_S}\right) \tag{10.2.5}$$

Implying

$$V_{OC} = 0.026 \ln\left(1 + \frac{20 \times 10^{-3}}{1.81 \times 10^{-11}}\right)$$

$$= 0.541\ \text{V}.$$

10.3 An *n*-type silicon photoconductor has a length $L = 80$ μm, cross-sectional area $A = 2 \times 10^{-7}\text{cm}^2$ and a hole minority carrier lifetime $\tau_p = 2 \times 10^{-6}$ s. Assuming that the applied voltage $V = 12$ V, calculate the gain of the photoconductor. Take $\mu_n = 1350\ \text{cm}^2/\text{V s}$ and $\mu_p = 480\ \text{cm}^2/\text{V s}$.

Solution

The transit time, t_n, to the electron is given by

$$t_n = \frac{L}{\mu_n E} = \frac{L^2}{\mu_n V} \tag{10.3.1}$$

Putting the given values in Eqn (10.3.1) gives

$$t_n = \frac{(80 \times 10^{-4})^2}{(1350)(12)} = 3.95 \times 10^{-9}\text{s} \tag{10.3.2}$$

Photoconductor gain, G_{ph}, can be obtained using

$$G_{ph} = \frac{\tau_p}{t_n}\left(1 + \frac{\mu_p}{\mu_n}\right) \tag{10.3.3}$$

Putting the calculated value of t_n and the given values of other parameters in Eqn (10.3.3) leads

$$G_{ph} = \frac{2 \times 10^{-6}}{3.95 \times 10^{-9}}\left(1 + \frac{480}{1350}\right) \tag{10.3.4}$$

$$= 686.4$$

10.4 A silicon *p-n* junction diode has the following parameters at $T = 300$ K:

$N_a = 5 \times 10^{16}$ cm^{-3} $\quad N_d = 5 \times 10^{16}$ cm^{-3}

$D_n = 25$ cm^2/s $\quad D_p = 10$ cm^2/s

$\tau_{n0} = 6 \times 10^{-7}$ s $\quad \tau_{p0} = 2 \times 10^{-7}$ s

Calculate the steady-state photocurrent density assuming $V_R = 6$ V and $G_L = 5 \times 10^{20}$ cm^{-3} s^{-1}. Take $n_i = 1.5 \times 10^{10}$ cm^{-3}.

Solution

The steady-state photocurrent density, J_L, is given by

$$J_L = e\, G_L(W + L_n + L_p) \tag{10.4.1}$$

where

$$L_n = \sqrt{D_n \tau_{n0}} = \sqrt{25 \times (6 \times 10^{-7})} = 38.7\,\mu\text{m} \tag{10.4.2}$$

and

$$L_p = \sqrt{D_n \tau_{p0}} = \sqrt{10 \times 2 \times 10^{-7}} = 14.1\,\mu\text{m} \tag{10.4.3}$$

Also

$$V_{\text{bi}} = \frac{kT}{e} \ln\left(\frac{N_a N_d}{n_i^2}\right) \tag{10.4.4}$$

Putting the given values in Eqn (10.4.4) leads to

$$V_{\text{bi}} = 0.026 \ln\left(\frac{5 \times 10^{16} \times 5 \times 10^{16}}{(1.5 \times 10^{10})^2}\right)$$

$$= 0.78 \text{ V}$$

Finally

$$W = \left[\frac{2\varepsilon_s}{e}\left(\frac{N_a + N_d}{N_a N_d}\right)(V_{\text{bi}} + V_R)\right]^{1/2} \tag{10.4.5}$$

Putting the given values in Eqn (10.4.5) results in

$$W = \left[\frac{2 \times 11.7 \times 8.85 \times 10^{-14}}{1.6 \times 10^{-19}}\left(\frac{5 \times 10^{16} + 5 \times 10^{16}}{5 \times 10^{16} \times 5 \times 10^{16}}\right)(0.78 + 6)\right]^{1/2}$$

$$= 0.59\ \mu\text{m}$$

Using Eqn (10.4.1) we get

$$J_L = 1.6 \times 10^{-19}\,(0.59 + 38.7 + 14.1) \times 10^{-4} \times 5 \times 10^{20}$$

$$= 0.43 \text{ A/cm}^2$$

10.5 Calculate the critical angle for the GaAs–air interface. Assume $\overline{n}_2 = 3.66$ and $\overline{n}_1 = 1.0$.

Solution

Critical angle, θ_C, is given by

$$\theta_C = \sin^{-1}\left(\frac{\overline{n}_1}{\overline{n}_2}\right) \tag{10.5.1}$$

Putting the given values in Eqn (10.5.1) results in

$$\theta_C = \sin^{-1}\left(\frac{1.0}{3.66}\right) = 15.9°$$

10.6 A silicon wafer is doped with 10^{15} cm^{-3} donor atoms.

(a) Find the electron and hole concentrations and the location of the Fermi level with respect to the intrinsic Fermi level (E_i).
Assume $n_i = 1.45 \times 10^{10}$ cm^{-3} for silicon.

(b) The light illuminating the wafer leads to a steady-state photo-generated density of electrons and holes equal to 10^{12} cm^{-3}. Assume that free carriers are generated uniformly throughout its volume. Find the overall electron and hole concentrations in the wafer and calculate the positions of the quasi-fermi levels for the two carrier types.

(c) Repeat the calculations carried out in part (b) under the condition that the light intensity is increased, so that the density of photogenerated electron-hole pairs is 10^{18} cm^{-3}.

($k = 8.62 \times 10^{-5}$ eVK^{-1}; $T = 300$ K)

Solution

(a) $n = N_d = 10^{15}$ cm^{-3}

$$p = \frac{n_i^2}{N_d} = 2.1 \times 10^5 \text{cm}^{-3}$$

$$E_f - E_i = kT \ln\left(\frac{n}{n_i}\right) = 0.025 \ln\left(\frac{10^{15}}{1.45 \times 10^{10}}\right)$$

$$= 0.278 \text{ eV}$$

(b) $n = 10^{15} + 10^{12} \approx 10^{15}$ cm^{-3}

$$p = 2.1 \times 10^5 + 10^{12} \approx 10^{12} \text{ cm}^{-3}$$

$$E_{fn} - E_i = kT \ln\left(\frac{n}{n_i}\right) = 0.278 \text{ eV}$$

$$E_i - E_{fp} = kT \ln\left(\frac{p}{n_i}\right) = 0.11 \text{ eV}$$

(c) $n = 10^{15} + 10^{18} \approx 10^{18}$ cm^{-3}

$p = 2.1 \times 10^5 + 10^{18} \approx 10^{18}$ cm^{-3}

$E_{fn} - E_i = E_i - E_{fp} = 0.47$ eV

Comment: In part (b) there is a change in the minority carrier concentration, whereas the majority carrier concentration remains unchanged. Most instances of photogeneration in doped semiconductors are similar to part (b). In part (c), the concentration of both majority and minority carriers is equal and, hence, the fermi levels for majority and minority carriers are equidistant from the Fermi level.

10.7 Calculate the wavelengths of radiation needed to create electron–hole pairs in intrinsic germanium, silicon, gallium-arsenide, and SiO_2. Identify the spectrum range (e.g., infrared, visible, UV, and X-ray) for each case.

Use $h = 4.135 \times 10^{-15}$ eVs, $c = 3 \times 10^8$ m/s

Solution

$$\lambda = \frac{hc}{E_g}, E_g \text{ is in eV}$$

$$= \frac{1.24}{E_g}$$

Material	E_g(eV)	λ(μm)	Range
Ge	0.67	1.85	Infrared
Si	1.124	1.16	Infrared
GaAs	1.42	0.87	(near) Infrared
SiO_2	~9	0.14	Ultraviolet

10.8 Given an infinite number of silicon *p-n* junction solar cells, an output of 1 V has to be produced by using the solar cells in parallel or in series. The solar cells have the following parameters:

$N_a = 10^{18}$ cm^{-3}; $N_d = 10^{17}$ cm^{-3}; $\mu_p = 471$ cm^2/Vs; $\mu_n = 1417$ cm^2/Vs,

$\tau_p = 10^{-8}$ s; $\tau_n = 10^{-6}$ s; $J_L = 40$ mA/cm^2; $A = 10^{-5}$ cm^2; $R_l = 1000\ \Omega$

Assume that all the cells have the same resistance.

(a) Which configuration (parallel or series) will you use to produce the 1 V output? Justify your answer.

(b) What is the total number of solar cells required for sustaining a load of 1 kΩ at the output? Assume that each solar cell provides a drop of 0.1 V.

Solution

(a) A combination of both series and parallel configurations must be used to get an output load current of $\frac{1V}{1\,k\Omega} = 1$ mA.

Let us now calculate I_s

$$D_p = \left(\frac{kT}{e}\right)\mu_p = 12.2 \text{ cm}^2/\text{s} \quad \text{use } \frac{kT}{e} = 0.02586 \text{ V at 300 K}$$

$$D_p = \left(\frac{kT}{e}\right)\mu_n = 36.64 \text{ cm}^2/\text{s}$$

$$L_n = \sqrt{D_n \tau_n} = 6.053 \times 10^{-3} \text{cm}$$

$$L_p = \sqrt{D_p \tau_p} = 3.48 \times 10^{-4} \text{cm}$$

From Eqn 10.2.2

$$J_s = en_i^2\left(\frac{D_p}{N_d L_p} + \frac{D_n}{N_a L_n}\right)$$

$$= (1.6 \times 10^{-19})(1.45 \times 10^{10})^2\left[\frac{12.2}{10^{17} \times 3.49 \times 10^{-4}} + \frac{36.64}{10^{18} \times 6.053 \times 10^{-3}}\right]$$

$$= 1.196 \times 10^{-11} \text{A/cm}^2$$

$$I_s = 1.196 \times 10^{-16} \text{ A}$$

$$I_F = I_s\left(e^{\frac{V}{V^T}} - 1\right) = I_s\left(e^{\frac{\cdot 1}{.02586}} - 1\right) = 5.72 \times 10^{-15}$$

(b) As the voltage drop across each solar cell is 0.1 V we need 10 cells in series.

$$I_L = 40 \times 10^{-8} \text{ A}$$

$$I = I_L - I_F = (40 \times 10^{-8}) - (5.72 \times 10^{-15})$$

$$= 4 \times 10^{-7}$$

So, if every series of 10 solar cells gives a current of 4×10^{-7} A, then, to build up a current value of 1 mA we need

$$X = \frac{10^{-3}}{4 \times 10^{-7}} = 2500$$

2500 branches each having 10 solar cells.
Hence, the total number of solar cells required = $2500 \times 10 = 25000$.

10.9 Given a piece of semiconductor with an energy band gap of 1.43 eV, what is the minimum wavelength of light which must be incident on this sample so that the maximum of the light is used for transfer of electrons from the valence band to the conduction band? How will this phenomenon help if the said sample is to be used in semiconductor devices?

Solution

Before proceeding with the mathematical formulation of the solution, we must remember that the photon energy of the incident light must be greater than the band gap of the semiconductor for the photons to be able to excite the electrons from the valence band to the conduction band.

Now, the band gap of the material given = 1.43 eV.

Also

$$E = h\nu \tag{10.9.1}$$

Substituting $E = E_g$ in Eqn (10.9.1)

We get

$$E_g = h\nu$$

$$E_g = \frac{hc}{\lambda} \tag{10.9.2}$$

Substituting $E_g = 1.43$ eV

$h = 4.14 \times 10^{-15}$ eV/s, and

$c = 3 \times 10^8$ ms^{-1} in Eqn (10.9.2)

and rearranging

$$\lambda = \frac{4.14 \times 10^{-15} \times 3 \times 10^8}{1.43}$$

$$= 8.68 \times 10^{-7} \text{ m}$$

$$\simeq 8600 \text{ Å}$$

Therefore, light of wavelength ~8600 Å must be incident on the sample to ensure maximum absorption.

Application to semiconductor devices Light incident on the semiconductor is utilized by the electrons to jump from the valence band to the conduction band. When these excited electrons fall back to their equilibrium states, they emit the excess energy, which may be radiated by the material in the form of light. Many semiconductors are well suited for light emission, especially compound semiconductors with direct band gaps.

10.10 A typical LED has been so fabricated and packed that it generates 190 μW of optical power into a typical optical fibre, when a current of 25 mA is flowing through the device. Determine the overall power conversion efficiency, when the corresponding forward voltage across the diode is 1.5 V.

Solution

The power conversion efficiency, η, of an LED is defined as the ratio of the optical power coupled into the fibre, P_c, to the electrical power, P, applied at the terminals of the device.

Now, for this problem

$$P_c = 190 \text{ μW}$$

$$P = \frac{V}{I} = \frac{1.5 \text{ V}}{25 \text{ mA}}$$

$$= 37.5 \text{ mW}$$

$$\eta = \frac{P_c}{P} = \frac{190\ \mu\text{W}}{37.5\,\text{mW}}$$

$$= 5.1 \times 10^{-3}$$

$$= 0.51\%$$

Recapitulation

- Energy, E, and wavelength, λ, of a photon are related through the expression

$$\lambda = \frac{1.24}{E}\,\mu\text{m}$$

- Variation of incident photon flux with depth inside a semiconductor can be expressed as

$$I_v(x) = I_{v0}\, e^{-\alpha x}$$

- Open-circuit voltage, V_{OC}, of a solar cell can be evaluated using

$$V_{OC} = \frac{kT}{e} \ln\left(1 + \frac{I_L}{I_S}\right)$$

- Maximum power, P_m, obtainable from a solar cell is

$$P_m \simeq I_L \left[V_{OC} + \frac{kT}{e} \ln\left[1 + \frac{eV_m}{kT}\right] - \frac{kT}{e}\right]$$

- Efficiency, η, of a photovoltaic cell can be written in the form

$$\eta = \frac{\text{FF} \cdot I_L V_{OC}}{P_{in}}$$

- Photoconductor gain, G_{ph}, can be evaluated using

$$G_{ph} = \frac{\tau_p}{\tau_n}\left(1 + \frac{\mu_p}{\mu_n}\right)$$

- Total photocurrent density, J_L, for a long diode can be obtained using

$$J_L = eG_L\,(W + L_n + L_p)$$

- Critical angle for light travelling through a medium with refractive index $\bar{n}_2$ to a medium with refractive index $\bar{n}_1$ $(\bar{n}_1 < \bar{n}_2)$ is given by

$$\theta_c = \sin^{-1}\left(\frac{\bar{n}_1}{\bar{n}_2}\right)$$

- Population inversion can be realized provided

$$(E_{FC} - E_{FV}) > E_g.$$

Exercises

Review Questions

10.1 Define absorption coefficient.
10.2 Write an expression for the generation rate of electron–hole pairs due to photons.
10.3 Derive the expression for maximum power output of a photovoltaic cell.
10.4 Derive an expression for the conversion efficiency of a solar cell. Indicate the different approaches available to increase the conversion efficiency.
10.5 How does series resistance affect the conversion efficiency of a photovoltaic cell?
10.6 Why do heterojunction solar cells have a better short-wavelength response?
10.7 In what way are amorphous silicon solar cell different from those based on single crystal silicon? Explain the concept of mobility gap.
10.8 Derive an expression for photoconductor gain.
10.9 Show that the total photocurrent for a long photodiode is given by $J_L = eG_L (W + L_n + L_p)$
10.10 What are the advantages of the PIN photodiode?
10.11 How is impact ionization used in designing photodiodes?
10.12 Draw a schematic diagram of a phototransistor.
10.13 Define quantum efficiency of an LED.
10.14 Draw schematic diagrams for homojunction laser and stripe geometry laser.
10.15 Explain the concept of population inversion in a *p-n* junction.
10.16 Give the units of absorption coefficient.
10.17 Schematically represent the *I-V* characteristics of a *p-n* junction solar cell.
10.18 Derive an expression for maximum power delivered by a solar cell. What is meant by the term 'maximum power rectangle'?
10.19 Derive an expression for conversion efficiency of a solar cell. Differentiate between AM0 and AM1.
10.20 How does series resistance affect the maximum power delivered by a solar cell?
10.21 How does an amorphous silicon solar cell work?
10.22 Justify the statement that switching speed and gain of a photoconductor are mutually contradictory quantities.
10.23 Derive expressions for the different constituents of photocurrent density for a long diode.
10.24 Why are PIN photodiodes faster than *p-n* photodiodes?
10.25 Describe the working of an avalanche photodiode.
10.26 Explain the working of a phototransistor.
10.27 What are the different loss mechanisms in LEDs?
10.28 Why is the critical angle important in the design of an LED?
10.29 Draw a schematic diagram of a double-heterostructure laser.
10.30 Justify the condition for population inversion of a *p-n* junction, using a suitable band diagram.

Problems

10.1 Determine the thickness of silicon required to absorb 95% of incident photon energy. Assume $\lambda = 1.0$ μm and $\alpha = 10^2$ cm^{-1}.

$$\left[\textit{Hint: } \frac{I_v(d)}{I_{v0}} = e^{-\alpha d} \right]$$

Ans. 0.03 cm

10.2 At a particular wavelength λ, a 2 μm thick material absorbs 90% of the incident photon energy. Calculate the absorption coefficient of the material at λ.

$$\left[\textit{Hint: } \frac{I_v(d)}{I_{v0}} = 0.1 = e^{-\alpha d} \right]$$

Ans. 0.15×10^4 cm^{-1}

10.3 A *p-n* junction silicon solar cell has the following important parameters:

$N_a = 5 \times 10^{18}$ cm^{-3} $\quad N_d = 10^{16}$ cm^{-3}

$D_n = 25$ cm^2/s $\quad D_p = 10$ cm^2/s

$\tau_{n0} = 5 \times 10^{-7}$ s $\quad \tau_{p0} = 10^{-7}$ s

Calculate the open-circuit voltage at 300 K assuming $J_L = 30$ mA/cm^2 and $n_i = 1.5 \times 10$ cm^{-3}.

$$\left[\textit{Hint: } V_{OC} = \frac{kT}{e} \ln\left(1 + \frac{J_L}{J_S} \right) \right]$$

Ans. 0.534 V

10.4 An optical concentrator is used with the solar cell mentioned in Problem 10.3. The device leads to an increase of J_L to 60 mA/cm^2. Calculate the new open-circuit voltage.

Ans. 0.552 V

10.5 For the solar cell indicated in Problem 10.3, calculate the photocurrent density required to increase the open-circuit voltage to 0.6 V.

Ans. 382.7 mA

10.6 A photoconductor is fabricated using *n*-type silicon with the following important parameters:

$L = 60$ μm ; $\quad A = 10^{-7}$ cm^2

$\tau_p = 10^{-6}$ s ; $\quad V = 10$ V

Calculate the gain of the photoconductor assuming $\mu_n = 1350$ cm^2/V s and $\mu_p = 480$ cm^2/V s.

$$\left[\textit{Hint: } t_n = \frac{L^2}{\mu_n V} \text{ and } G_{ph} = \frac{\tau_p}{t_n}\left(1 + \frac{\mu_p}{\mu_n} \right) \right]$$

Ans. 507.7

10.7 Calculate the gain of the photoconductor mentioned in Problem 10.6 if the semiconductor used is *n*-type germanium. All other parameters remain same. Take $\mu_n = 3900$ cm^2/V s and $\mu_p = 1900$ cm^2/V s.

Ans. 1616.5

10.8 A PIN photodiode has an intrinsic region width $W = 15$ μm and an incident photon flux of 5×10^{16} cm^{-2} s^{-1}. Assuming an absorption coefficient of 8×10^2 cm^{-1}, calculate the photocurrent density.

[*Hint:* $J_L = q\phi_i (1 - e^{-\alpha W})$]

Ans. 5.6 mA/cm^2

10.9 Determine the critical angle for the GaP–Air interface. Take $n_2 = 3.37$ and $n_2 = 1.0$.

Ans. 17.3°.

10.10 Which semiconductor out of GaAs, CdSe, and GaP is suitable for opto-electronics applications which require maximum absorption of red colour light? Perform the necessary calculations to arrive at the conclusion.

[*Hint:* See Section 10.1.1]

Ans. CdSe

10.11 Photons of wavelength $\lambda = 0.9$ μm are incident on a semiconductor surface with absorption coefficient of 60 cm^{-1}. What percentage of incident photons is absorbed within a thickness of 0.02 cm of the semiconductor?

Ans. 70%

10.12 The open circuit voltage of a silicon *p*-*n* junction solar cell is 0.51 V. The photocurrent density is $J_L = 25$ mA/cm^2; calculate the reverse saturation current density, J_s.

Ans. 7.53×10^{-11} A/cm^2

11 Power Devices

- Bipolar power transistor
- Darlington pair configuration
- Power MOSFET
- Heat sink
- Semiconductor controlled rectifier
- Two-transistor model
- Bidirectional thyristors

Learning Objectives

After going through this chapter the student will be able to

- understand the role of current crowding in a bipolar junction transistor
- understand the concept of vertical transistor structure
- understand the characteristics of a power BJT
- understand the operating principle of a power MOSFET
- understand the importance of the Darlington pair configuration
- understand the characteristics of a power MOSFET
- solve numericals based on power BJT circuit configurations
- solve numericals based on power MOSFET circuit configurations
- understand the concept of heat sinking in power transistors
- understand the operation of a semiconductor controlled rectifier
- understand the two-transistor model of a *p-n-p-n* diode
- understand the operation of bidirectional thyristors
- understand the operation of a unijunction transistor

Introduction

We have discussed basic bipolar transistors and MOS transistors in earlier chapters. The voltages and currents in these transistors were, so far, assumed to be moderate. This is, however, not always the case. Many applications like power amplifiers and high power sources require semiconductor devices that can handle both high voltages and currents. Such conditions have ramifications in terms of geometry and special heat sinking routes to be adopted. This chapter will discuss all these topics in detail with special emphasis on power MOSFETs and bipolar transistors. One important application area of power devices is switching. Apart from power transistors, thyristors are another class of devices that finds use in power switching. On a very fundamental level, a thyristor is a four-layered *p-n-p-n* structure. The device can remain *off* even under forward bias, till an external trigger switches it *on*.

11.1 Bipolar Power Transistors

Bipolar transistors were discussed in Chapter 8. It was assumed that there are no inherent limitations on the voltage applied or the current allowed to flow through the device. This unrestricted ability to handle voltage or current, in turn, led to an unrestricted ability to handle or dissipate power. Transistors, however, have limitations placed on them on all the three counts, i.e., voltage, current, and maximum power. One thing is, however, clear that the maximum current rating and the maximum voltage rating are never applied together.

11.1.1 Current Crowding

Figure 11.1 shows the cross-section of a typical *n-p-n* bipolar transistor.
The base current is shared equally between the two base regions. The emitter

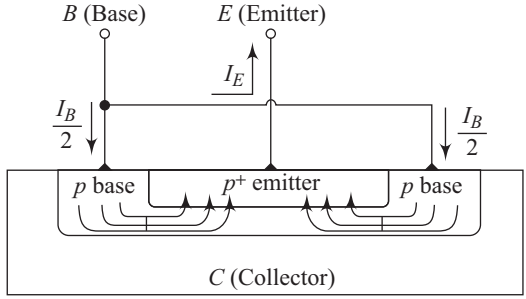


Fig. 11.1 Cross-sectional view of an *n-p-n* bipolar transistor

region is heavily doped and can, therefore, be approximated to be an equipotential surface. The base region is generally less than a micrometre in thickness and can, therefore, have a substantial base resistance. This base resistance results in a lateral potential difference under the emitter region. Thus, the potential decreases

from the edge of the emitter region towards its central region. Due to this the effective *B–E* voltage decreases as we go from the edge of the emitter, towards its centre. This results in more electrons being injected near the emitter edges than in the centre. The emitter current, therefore, gets crowded near the edges as shown in Fig. 11.2.

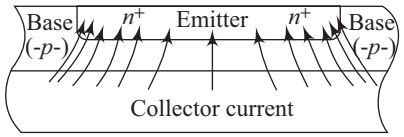


Fig. 11.2 Current crowding effects near edges

This current crowding leads to localized heating effect, non-uniform lateral base current, and localized high-injection effects.

Power transistors are required to carry large currents. This leads to the requirement of large emitter areas that can handle reasonable current densities. It is common to design power transistors with narrow emitter widths and using interdigitated design. A basic diagram of a typical interdigitated design is shown in Fig. 11.3.

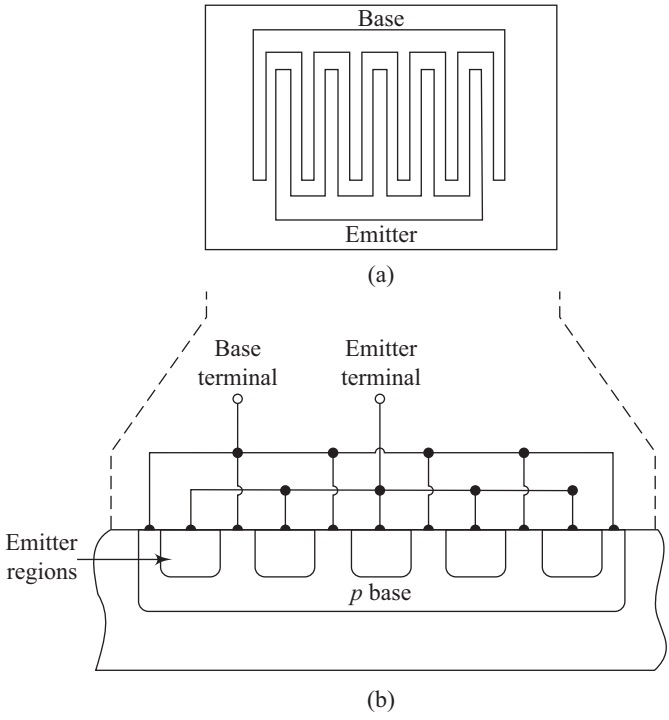


Fig. 11.3 Schematic diagram of an interdigitated design—(a) Top view (b) Device layout

The complicated geometry essentially implies many narrow emitters connected in parallel to result in an increased area.

11.1.2 Vertical Transistor Structure

Vertical transistor structures are common in switching applications. For small power handling, the collector terminal can be formed on the surface itself. For power bipolar transistors, the collector region is provided at the bottom of the device. This arrangement leads to maximization of the cross-sectional area involved in the current flow. A schematic diagram of a typical vertical *n-p-n* power BJT is shown in Fig. 11.4.

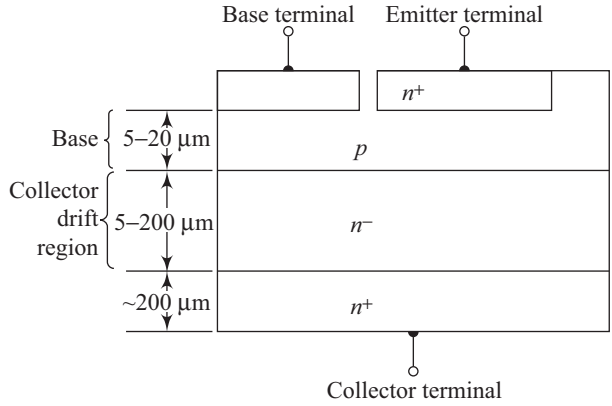


Fig. 11.4 Schematic diagram of a vertical *n-p-n* BJT

The structure has a low-doped collector drift region followed by a heavily doped collector region. The low-doped collector drift region allows the applicaton of large base-collector voltage without going into breakdown conditions. The large base-collector voltage also results in large space-charge widths on either side of the base-collector junction. The large base width helps in preventing punch-through breakdown.

11.1.3 Transistor Characteristics

General purpose small-signal BJTs differ from power BJTs in many ways. The relatively larger base width of power transistors leads to small current gain, β. The maximum current that can be handled by the lead wires connecting the semiconductor to the external terminal decides the maximum rated collector current, $I_{C(\max)}$. $I_{C(\max)}$ is also governed by the collector current at which the current gain goes below a minimum specified value or the current that results in maximum power dissipation with the transistor biased in saturation.

The *maximum rated voltage* in a BJT is decided by the onset of avalanche breakdown in the reverse-biased base-collector junction. Typical plots of I_C versus V_{CE} for different base currents are shown in Fig. 11.5.

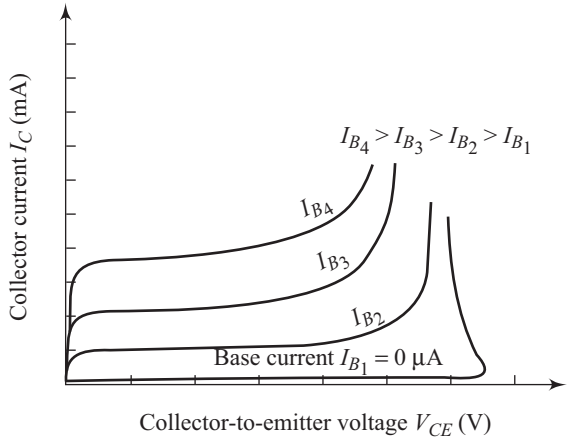


Fig. 11.5 Plots of I_C versus V_{CE} for different base currents

It can be observed that all the curves tend to merge towards the same collector–emitter voltage as the breakdown approaches. The minimum voltage necessary to sustain a transistor in breakdown is called $V_{CE(\text{sus})}$.

A power BJT is also prone to another type of breakdown process, often called the *second breakdown*. Slight non-uniformities in current densities always exist in a BJT operating at high currents and voltages. These non-uniformities lead to local regions of increased heating which are associated with increased minority carrier concentration. This increase in minority carrier concentration results in a positive feedback effect, leading to a further increase in the current in these regions. There is an accompanying increase in temperature that ultimately melts the semiconductor material and creates a short-circuit between collector and emitter. The design engineers have to ensure that the average power dissipated is kept below the specified maximum value, thereby ensuring that the temperatures remain below the maximum permitted value. For dc current and voltage values we can write for maximum rated power, P_T.

$$P_T = V_{CE} I_C + V_{BE} I_B \tag{11.1}$$

Neglecting the power dissipation component, $V_{BE} I_B$, we can rewrite Eqn. (11.1) in the form

$$P_T = V_{CE} I_C \tag{11.2}$$

Figure 11.6 shows the I_C versus V_{CE} plot for a typical power transistor. The power limitation, P_T, is a hyperbola described by Eqn (11.2). The safe operating area is bounded by $I_{C(\text{max})}$, $V_{CE(\text{sus})}$, P_T, and the second breakdown characteristic curve. Figure 11.6 is plotted on a linear scale. As already indicated, Fig. 11.6, neglects the power dissipation due to the *B-E* junction, as it is usually small.

The large area of power transistors results in a large junction capacitance. Such large junction capacitance values lower the cut-off frequency substantially in comparison to conventional switching transistors. Table 11.1 compares some important parameters of a conventional small-signal BJT with two power BJTs.

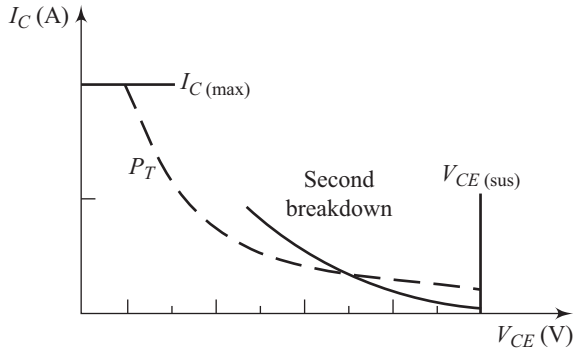


Fig. 11.6 I_C versus V_{CE} plots for a power transistor

Table 11.1 Important parameters of conventional and power transistors

Parameter	(2N2222A) Small-signal BJT	(2N3055) Power BJT	(2N6078) Power BJT
$V_{CE(max)}$ (V)	40	60	250
$I_{C\,(max)}$ (A)	0.8	15	7
$P_{D(max)}$ (W) (at $T = 25°C$)	1.2	115	45
β	35–100	5–20	12–70
f_T (MHz)	300	0.8	1

11.1.4 Darlington Pair Configuration

As discussed in the previous section, the current gain of a power BJT is relatively small due to wider base width. A Darlington pair configuration can be used to increase the effective current gain. Figure 11.7 shows the Darlington pair configuration.

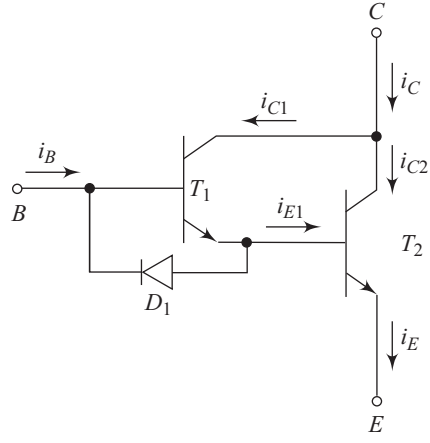


Fig. 11.7 Schematic representation of the Darlington pair configuration

The total collector current, i_C, for the configuration is given by

$$i_C = i_{C_1} + i_{C_2} \tag{11.3}$$

Assuming β_1 and β_2 to be the current gains for the individual transistors T_1 and T_2, we can rewrite Eqn (11.3) in the form

$$i_C = \beta_1 i_B + \beta_2 i_{E1} \tag{11.4}$$

which can be expressed in the form

$$\frac{i_C}{i_B} = \beta_1\beta_2 + \beta_1 + \beta_2 \tag{11.5}$$

From Eqn (11.5) it is clear that the overall current gain for the Darlington configuration is substantially higher than the individual transistors constituting the pair. The diode D_1 used in the configuration, pulls charge out of the base of the transistor T_2 in the form of its reverse current. This turns the transistor T_2 off faster than it would have switched off in the absence of the diode. Figure 11.8 shows a methodology for realizing an *n-p-n* Darlington pair in a typical integrated circuit approach. The silicon dioxide is used to isolate the base regions of the two transistors that constitute the Darlington pair.

The Darlington pair finds use in the output stage of power amplifier and also as an effective gain improvement strategy for power transistors.

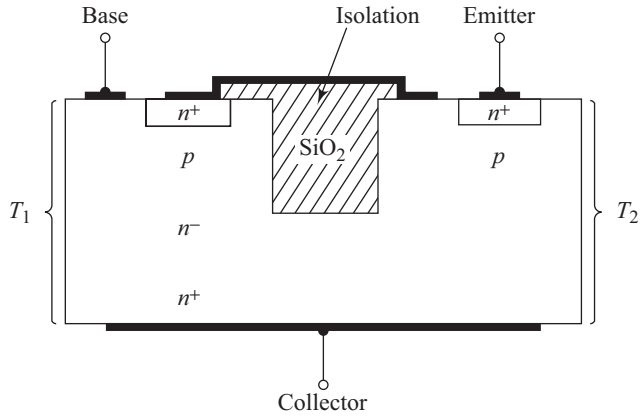


Fig. 11.8 Realization of an *n-p-n* Darlington pair configuration

11.2 Power MOSFETs

As in the case of the BJT, power MOSFETs are fundamentally identical to ordinary conventional MOSFETs. Power MOSFETs can, however, handle higher current and can accept higher impressed voltages. Drain-to-source blocking voltage of power MOSFETs is higher than conventional MOSFETs. The control signal is applied to the gate terminal of power MOSFETs. Power MOSFETs have extremely high input impedance. This allows large currents to be switched between on and off states, while maintaining low gate current values.

11.2.1 Structures

As already discussed, power MOSFETs are supposed to carry large currents and have large blocking voltages. The large current-carying capability is obtained by using large channel width, whereas large blocking voltages are achieved using vertical structures. There are basically two popular power MOSFET structures. The first of these is the DMOS device structure. This structure uses a double diffusion process and owes its name to this process (double-diffused MOS). Suitable windows are opened in the oxide mask (e.g., SiO_2) and n^+ source contact and p base or substrate regions are created by diffusion through a common window that is defined by the edge of the gate. The doping profile of the n^+-source and p-base (shown schematically in Fig. 11.9) is achieved by exploiting the higher rate of diffusion of the p-dopant (e.g., boron) in comparison to the n^+-dopant (e.g., phosphorus).

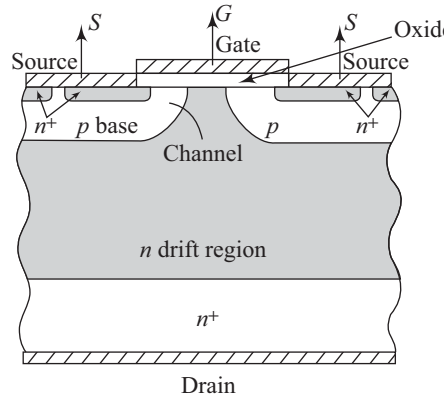


Fig. 11.9 Schematic diagram of the DMOS device

A lightly doped n-region serves as a drift region. The surface channel length is defined by the difference between the lateral diffusion distance of the p-base and the n^+-source. An inversion layer gets formed under the gate. Electrons from the source terminal flow laterally across the inversion layer and reach the n drift region. These electrons then flow vertically down through the drift region and finally reach the drain terminal. Two conditions are placed on the drift region: (i) The thickness of the drift region needs to be minimized to reduce the drain resistance. (ii) The doping in the drift region must, however, be moderate to ensure a large breakdown voltage.

The VMOS structure is also a popular choice for a power MOSFET. A schematic diagram of the VMOS structure is shown in Fig. 11.10.

The processing sequence of the VMOS structure uses the chemical etching characteristics of different crystallographic planes in a silicon single crystal. Some etchants like KOH and EDP etch silicon anisotropically. These etchants etch the ⟨111⟩ planes in silicon at a much slower rate than other planes. If a ⟨100⟩ oriented silicon is first oxidized and anisotropic

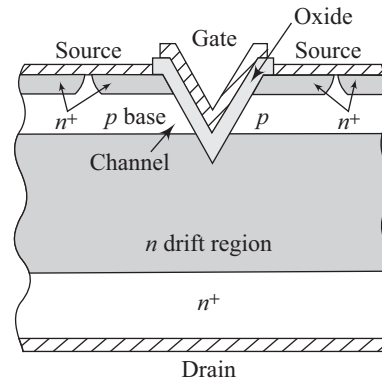


Fig. 11.10 Typical VMOS structure

etching is then carried out through windows in the surface masking oxide, then V-shaped groove can be obtained. The complete processing sequence includes *p*-base diffusion followed by n^+-source diffusion. V-grooves are then etched upto the *n* drift region. This is followed by gate oxide growth in the V-groove. Gate material is then deposited. An electron inversion region gets formed within the base region. The current from the source flows, in the vertical direction, to the drain. The drain voltage is supported by the depletion region extending into the low doped *n* drift region.

11.2.2 Power MOSFET Characteristics

On-state resistance is an important characteristic of power MOSFETs. This resistancce, R_{ON}, can be written in the form

$$R_{ON} = R_S + R_{CH} + R_D \tag{11.6}$$

where R_S represents the resistance associated with the source contact, R_{CH} represents the contribution from the channel, and R_D is the resistance due to the drain contact. Unlike conventional MOSFETs, the presence of R_S and R_D can lead to substantial power dissipation in power MOSFETs. This is due to the fact that even small values of resistances can lead to high power dissipation when large currents flow through them. For a MOSFET operating in the linear region, the channel resistance can be written in the form

$$R_{CH} = \frac{L}{W \mu_n C_{ox} (V_{GS} - V_T)} \tag{11.7}$$

where V_T is the threshold voltage and W and L represent the channel parameters.

The resistance contributions R_S and R_D are both directly proportional to the semiconductor resistivity. Since resistivity is inversely proportional to mobility, R_S and R_D are inversely proportional to mobility. Mobility decreases with increasing temperature and therefore R_S and R_D increase with increasing temperature. From Eqn (11.7), one can conclude that R_{CH} also increases with increasing temperature. Thus from Eqn (11.6) it is clear that the total on-state resistance, R_{ON}, increases with temperature. This increasing resistance has a current limiting effect in power MOSFETs. As already explained, power transistors generally consist of large number of units connected in parallel to achieve adequate power handling capability. If due to some reason, current increases assymetrically in one of the units, the increasing current leads to an increase in resistance, which in turn limits the current through the unit. Thus the total current flowing through the MOSFET gets evenly distributed amongst the constituent units. A schematic diagram of a HEXFET (Hexagonal FET) having packed units of individual DMOS devices is shown in Fig. 11.11.

In comparison to power bipolar transistor, power MOSFETs have faster switching times, no second breakdown, and relatively stable gain and response time with respect to changes in temperature. The transconductance variation with drain current for a typical high-power MOSFET (2N6757) for different temperatures is shown in Fig. 11.12.

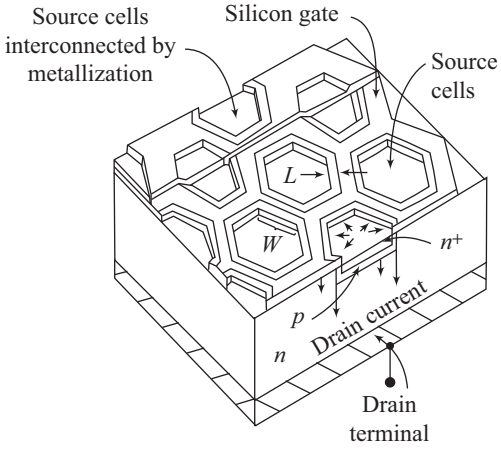


Fig. 11.11 Schematic diagram of a HEXFET structure

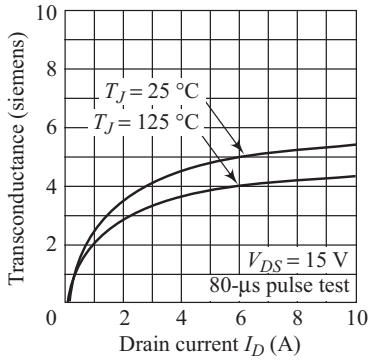


Fig. 11.12 2N6757 (MOSFET): variation in transconductance with drain current

Like power BJTs, power MOSFETs should also be operated in the safe operating area (SOA). The safe operating area is decided by three factors, namely, (i) maximum drain current, $I_{D(\max)}$, (ii) rated breakdown voltage, $BV_{DS(S)}$, and (iii) maximum power dissipation, P_T, given by $V_{DS}\, I_D$. Figure 11.13 shows the safe operating area for a typical MOSFET on a linear scale.

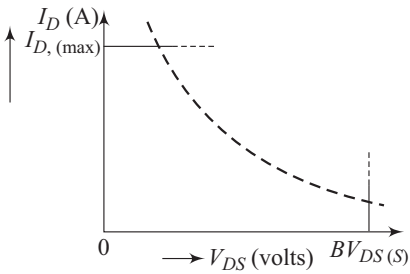


Fig. 11.13 Safe operating area of a typical MOSFET

- The transistor action ceases beyond the temperature at which the base region becomes intrinsic. When such a situation prevails, the collector effectively gets short-circuited to the emitter.
- Even distribution of current in interdigitated transistors is also achieved by adding a distributed emitter resistance. This limits the current flowing through any particular emitter. Such resistors are also referred to as stabilizing or ballistic resistors.

11.3 Heat Sink

Power transistors operate at high currents and voltages. The power dissipated within the device raises its internal temperature much beyond the ambient temperature. Excessively high device or junction temperatures can lead to permanent device failure. Suitable heat management measures are, therefore, extremely important for power transistors. Suitable packaging is used to conduct away the excess heat and maintain the operating temperature within prescribed limits.

The concept of thermal resistance, R_{th}, must be discussed for a complete understanding of this important issue. If a particular element has a temperature difference $(T_2 - T_1)$ across it, then the thermal resistance is given by

$$R_{th} = \frac{(T_2 - T_1)}{P} \tag{11.8}$$

where P represents the thermal power across the element. From Eqn (11.8) it is clear that thermal resistance, R_{th}, has the units of °C/W. In physical terms it defines the increase in temperature (or the temperature gradient) per unit increase in dissipated power. Drawing a parallel with electrical quantities, temperature difference is the driving force like voltage in electrical situation and power or heat flow is analogous to current.

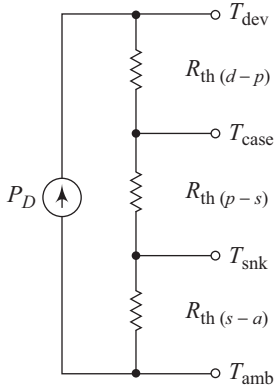


Fig. 11.14 Equivalent electrical circuit model

Figure 11.14 shows an equivalent electrical circuit model for the heat flow from the high power device to its ambient. In Fig. 11.14, $R_{th(d-p)}$ represents the thermal resistance between the device (junction) to the packaging, $R_{th(p-s)}$ represents the thermal resistance between the packaging to the heat sink, and $R_{th(s-a)}$ represents the thermal resistance between the heat sink and the ambient. Thus, the temperature difference between the device and the ambient can be written in the form

$$T_{dev} - T_{amb} = P_D[R_{th(d-p)} + R_{th(p-s)} + R_{th(s-a)}] \tag{11.9}$$

where R_D represents the dissipated power.

Power device manufacturers generally supply the maximum device or junction temperature, $T_{j(max)}$, and the thermal resistance for the device–package combination, i.e., $R_{th(d-p)}$. The maximum safe power dissipation, $P_{D(max)}$, is given by

$$P_{D(max)} = \frac{T_{j(max)} - T_{package}}{R_{th(d-p)}} \tag{11.10}$$

The plot of $P_{D(max)}$ versus $T_{package}$ is called the *power derating curve* of the transistor. A typical derating curve is shown in Fig. 11.15. $T_{j(max)}$ is the intercept on the $T_{package}$ axis. It is clear that when $T_{package}$ becomes equal to $T_{j(max)}$, the allowed power dissipation in the device becomes zero. This is because at $T_{package} = T_{j(max)}$ no further increase in temperature can be tolerated. A power device is generally operated in such a way that the device is at its maximum temperature with the package at

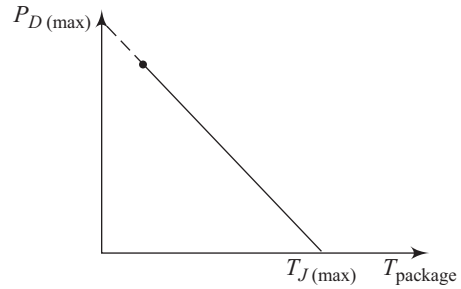


Fig. 11.15 Typical derating curve

ambient temperature. If the package can be maintained at the ambient temperature, the thermal resistance between the package and the ambient is zero and Eqn (11.9) can be rewritten in the form

$$T_{\text{dev}} - T_{\text{amb}} = P_D[R_{\text{th}(d-p)}] \tag{11.11}$$

One must however remember that generally, $R_{\text{th}(p-s)}$ and $R_{\text{th}(s-a)}$ have non-zero values and the package temperature rises above the ambient temperature. This also implies that the maximum power dissipation values as decided by Eqn (11.11) can actually never be realized in practice. In physical terms it means that an infinite heat sink can actually never be achieved. The same conclusion can also be drawn from the electrical equivalent circuit given in Fig. 11.14.

Let us suppose that the device is being operated at its maximum allowed limit, $T_{j(\text{max})}$. As the temperature T_{package} rises, the temperature difference $T_{\text{dev-package}}$ reduces, thereby lowering the power P_D through the element.

11.4 Semiconductor Controlled Rectifier

Switching is an important application of semiconductor devices. In these switching applications, it is desired that the device being used should be able to change from a blocking state (off state) to a conducting state (on state), or vice versa. As already discussed earlier, bipolar junction transistors can always be used for such applications. Suitable modifications in the base current magnitudes can drive a bipolar transistor from the cut-off to the saturation state or from the saturation state to the cut-off state. The cut-off state of a transistor corresponds to the *off* state of the device, whereas the saturation state can correspond to the *on* state. Bipolar transistors however are limited in the range of values of currents and voltages that they can handle. Thyristors or semiconductor controlled rectifiers (SCR) are semiconductor devices with operating principles closely related to the bipolar transistors. The basic operation involves both electrons and holes and there are some important differences from the switching mechanisms used in bipolar transistors. Thyristors have the ability to handle a much larger range of currents and voltages. Currents in the range of a few milliamperes to tens of thousands of amperes can be handled. Voltage ratings can exceed tens of thousands of volts. We will discuss the details of this interesting device in this section.

11.4.1 Fundamental Characteristics

A schematic cross-sectional representation of a thyristor structure is shown in Fig. 11.16(a) It is a four-layer device having a *p-n-p-n* structure that can be conceptualized as three *p-n* junctions in series. The three junctions are represented by J_1, J_2, and J_3 in the figure.

The electrical contact to the outermost *p*-layer is called the *anode* with a symbol *A* being used in Fig. 11.16(a). The outermost *n*-layer is called the cathode (*K*). The structure shown in Fig. 11.16(a) is a two terminal device but an additional

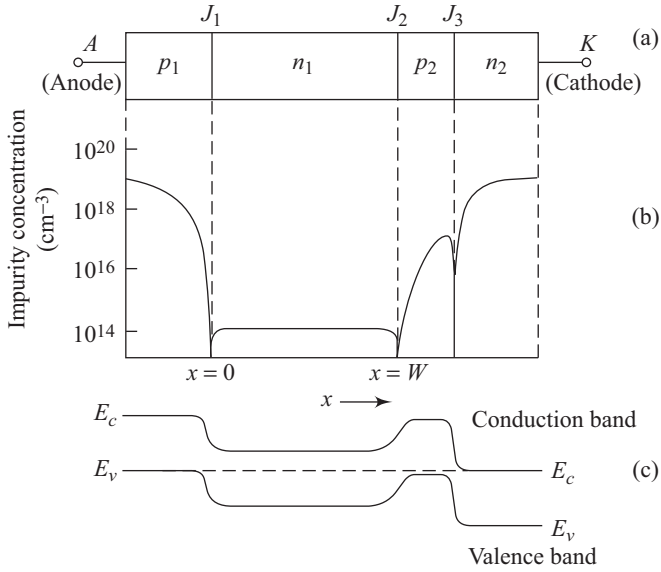


Fig. 11.16 (a) Thyristor structure (b) Doping profile of a thyristor (c) Energy band diagram

electrical contact is generally provided to the inner *p*-layer (p_2). This electrical contact is called the *gate electrode*. The four-layer diode then becomes a three-terminal device and is referred to as a *thyristor* or a semiconductor controlled rectifier (SCR). The doping profile of the device is shown in Fig. 11.16(b). The fabrication of the structure begins with the choice of the starting *n*-type high-resistivity silicon wafer. Neutron transmutation is generally used to achieve uniformity of doping in the starting substrate. This is followed by a diffusion step that results in the formation of the p_1 and p_2 layers simultaneously. As a final step the *n*-type layer (n_2) is obtained by alloying or diffusion. The energy band diagram of the structure is shown in Fig. 11.6(c). A depletion region exists at each junction. The built-in potential of each junction is dependent upon the corresponding impurity profile.

The basic current–voltage characteristics of the *p-n-p-n* diode (called so due to its two-terminal nature without the gate electrode) is shown in Fig. 11.17. The curve can be seen to exhibit five distinct regions which are briefly explained below.

Region 1 This is represented by the region between points (0) and (1). The device is said to be in the forward-blocking or off rate and offers a very high impedance.

At the $dV/dI = 0$ point, switching occurs and we can define a forward- breakover voltage V_{BF}. The corresponding switching current is I_S.

Region 2 This is characterized by the region of the current–voltage characteristics between (1) and (2). The device exhibits a negative resistance and the voltage across the device decreases very sharply.

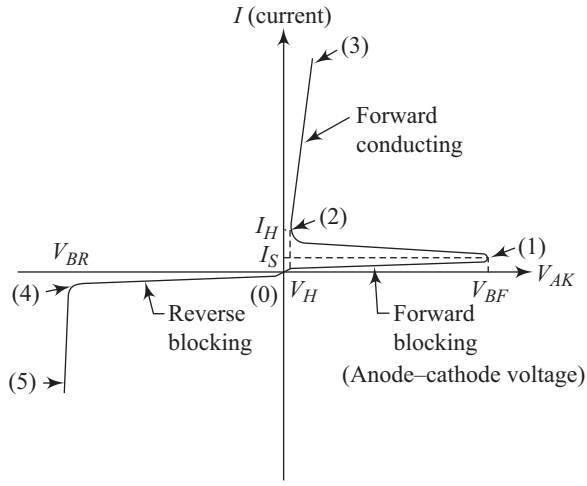


Fig. 11.17 Current–voltage characteristics of a *p-n-p-n* diode

Region 3 This region is defined by points (2) and (3). The device is now forward biased and conducting. This state of the structure is the on state and exhibits a very low impedance. $dV/dI = 0$ again, at point (2) and defines the holding current, I_H. The corresponding voltage is called the holding voltage, V_H.

Region 4 This is represented by the part of the curve between points (0) and (4). The device is said to be in the reverse-blocking state.

Region 5 This is the region between points (4) and (5). The device in this region is said to be in the reverse-breakdown state.

From the current–voltage characteristics curve shown in Fig. 11.17, it is clear that the *p-n-p-n* diode structure in the forward region is a bistable device. The device can switch from a high-impedance, low-current off state to an on state where the device offers low impedance and allows current.

11.4.2 Two-transistor Model

The forward-blocking characteristic of the *p-n-p-n* diode can be understood using a two-transistor model as shown in Fig. 11.18. In this model, the bistable device is considered to be a combination of two transistors, a *p-n-p* transistor and an *n-p-n* transistor.

The base of the *p-n-p* transistor is connected to the collector of the *n-p-n* transistor. Similarly, the base of the *n-p-n* transistor is connected

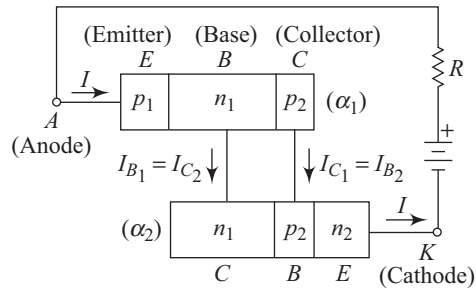


Fig. 11.18 Two-transistor model of a *p-n-p-n* diode

to the collector of the *p-n-p* transistor. The base current of the *p-n-p* transistor can be written in the form

$$I_{B_1} = I_{E_1} - I_{C_1} = (1-\alpha_1)\, I_{E_1} - I_1 \tag{11.12}$$

where α_1 is the common-base current gain of the transistor and I_1 represents the leakage current I_{CB0} for the collector–base junction with the emitter–base junction open. From Fig. 11.18, it is clear that Eqn (11.12) can be rewritten as

$$I_{B_1} = (1-\alpha_1) I - I_1 \tag{11.13}$$

This base current constitutes the collector current I_{C_2} for the *n-p-n* transistor.

Assuming the *n-p-n* transistor has a current gain α_2, the collector current for the *n-p-n* transistor can be written in the form

$$I_{C_2} = \alpha_2 I_{E_2} + I_2 \tag{11.14}$$

where I_2 represents the leakage current I_{CB0} for the *n-p-n* transistor. Equating I_{E_2} to I according to Fig. 11.18, we can rewrite Eqn (11.14) as

$$I_{C_2} = \alpha_2 I + I_2 \tag{11.15}$$

Equations (11.15) and (11.13) can be equated, yielding

$$(1 - \alpha_1)\, I - I_1 = \alpha_2 I + I_2$$

leading to

$$I = \frac{I_1 + I_2}{1 - (\alpha_1 + \alpha_2)} \tag{11.16}$$

Current gains α_1 and α_2 are functions of current I, increasing with increasing values of current. When currents are low, both α_1 and α_2 have small values. The current, I, flowing through the device is then given by the sum of the leakage currents I_1 and I_2. The current, I, increases with increase in applied bias voltage. This is accompanied by a corresponding increase in α_1 and α_2. As can be seen from Eqn (11.16) this increase in the magnitudes of α_1 and α_2 is accompanied by an increase in the magnitude of current, I. Ultimately, the term $(\alpha_1 + \alpha_2)$ approaches unity, and the current, I, increases without any limit. This is shown as the forward breakover in Fig. 11.17.

11.4.3 Depletion Layer Width and Effect of Gate Current

Figure 11.19 is a schematic diagram of the *p-n-p-n* diode in its different states. The accompanying figure shows the corresponding location of the state in the current–voltage characteristics of the diode.

(a) Figure 11.19(a) shows the depletion layer widths for the three junctions at equilibrium. With no applied bias, the depletion layer widths are decided by the impurity doping profiles of the different semiconductor regions.

(b) In Fig. 11.19(b), junctions J_1 and J_3 are forward biased, whereas the junction J_2 is reverse biased. A substantial portion of the voltage drop, therefore, occurs across the reverse biased junction J_2. The device is in a forward-blocking state.

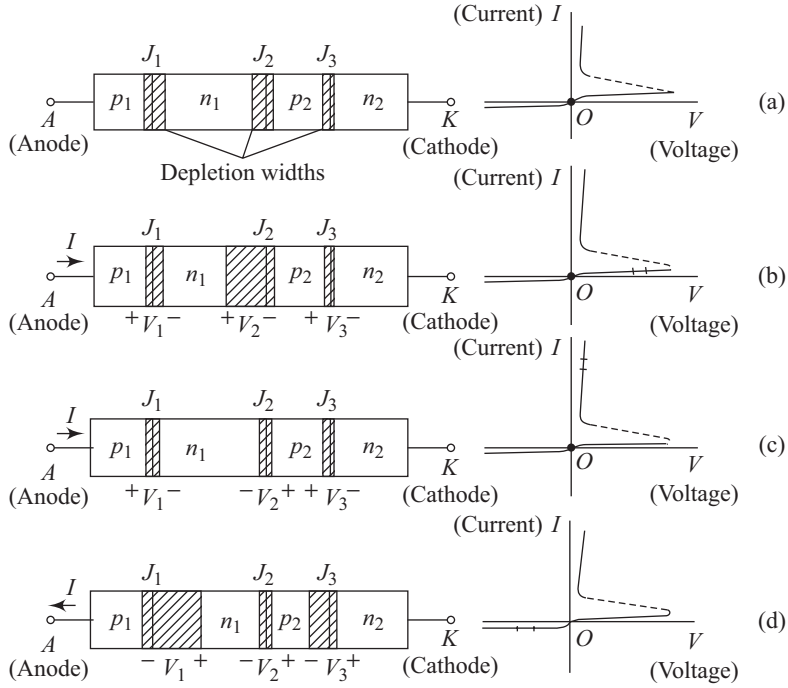


Fig. 11.19 Schematic diagram of the *p-n-p-n* diode in different states

(c) Figure 11.19(c) depicts the forward-conducting state. All the three junctions, namely J_1, J_2, and J_3 are forward biased. The two transistors formed between (p_1, n_1, p_2) and (n_1, p_2, n_2) are both in the saturation region. The total voltage drop across the device becomes ($V_1 - |V_2| + V_3$), which is approximately equal to the voltage drop across a single forward biased *p-n* junction. The voltage drop across the device is, therefore, very low.

(d) The reverse-blocking state is shown in Fig. 11.19(d). The two junctions J_1 and J_3 are reverse biased whereas the middle junction J_2 is in a forward-biased state.

A typical device configuration of a thyristor fabricated using planar processing is shown in Fig. 11.20(a). The corresponding schematic representation is shown in Fig. 11.20(b). The device includes a gate electrode provided through the region p_2. Figure 11.20(a) also identifies the three junctions, as well as the cathode (K) and the anode (A) regions. The currents I_K and I_A are not identical due to the presence of a finite gate current I_G. The masking silicon dioxide layer is also depicted.

In our discussions involving the *p-n-p-n* structure, we have so far not taken the effect of the gate terminal into account. Figure 11.21 shows the current–voltage characteristics of the *p-n-p-n* structure at different representative gate current values. To summarize the effect of gate current, it basically goes on to increase the magnitude of ($\alpha_1 + \alpha_2$). This results in a decrease in the forward-breakover voltage as the gate current increases in magnitude.

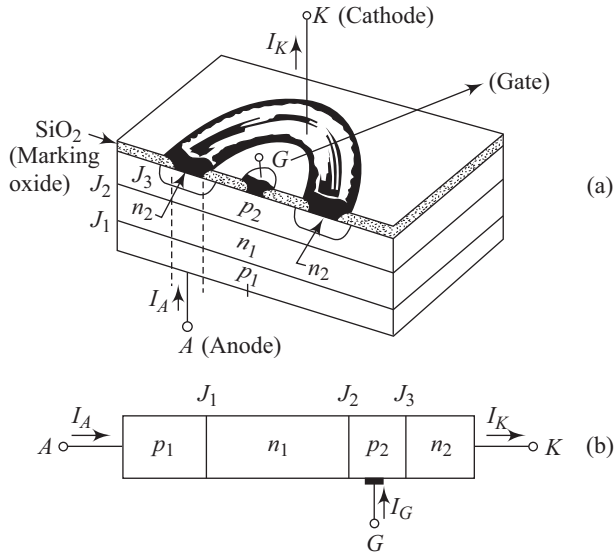


Fig. 11.20 (a) Thyristor device configuration (b) Schematic representation of the device

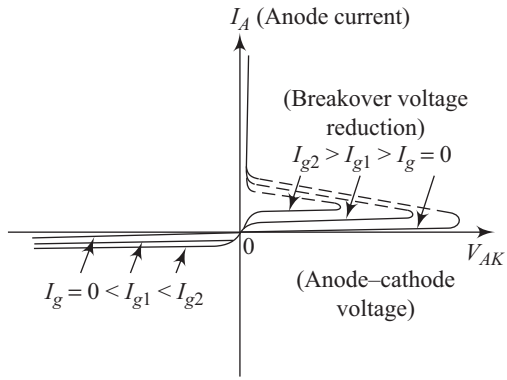


Fig. 11.21 *I-V* characteristics of the *p-n-p-n* structure for different gate currents

The timing of the gate-current pulses for a thyristor can be used to control the amount of power that is being delivered to a given load (e.g., furnace or heater). A current pulse delivered to the gate around the beginning of each cycle will ensure high power delivery to the load. Similarly, a delay of the gate current pulse will ensure a reduction in the power delivered to the load. A schematic circuit showing a typical thyristor application is shown in Fig. 11.22 along with the relevant pulse shapes of gate current, anode–cathode voltage, and load voltage.

The gate current pulse is shown to be delayed by a certain time interval, in Fig. 11.22(b). As expected, the anode–cathode voltage falls as soon as the gate current is applied and a corresponding increase in the load voltage is also observed.

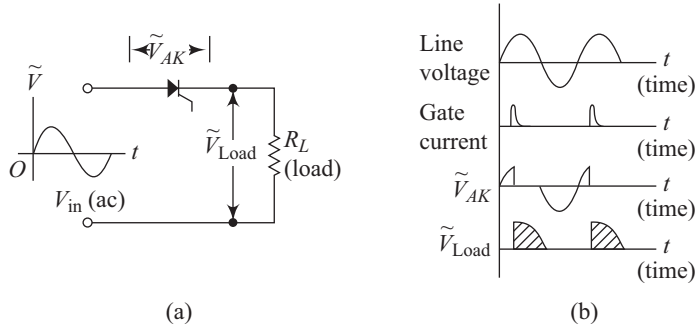


Fig. 11.22 Schematic circuit diagram showing thyristor application (a) Electrical circuit (b) Pulse shapes showing line voltage, gate current, anode–cathode voltage, and low voltage

11.4.4 Bidirectional Thyristors

The *p-n-p-n* structure discussed in the previous section has *on* and *off* states within the positive bias region. Bidirectional thyristors have *on* and *off* states in positive as well as negative bias regions. The basic bidirectional *p-n-p-n* diode switch is referred to as a *diac* (diode ac switch). Its operation can be understood as a combination of two *p-n-p-n* diodes with the anode of one of the diodes connected to the cathode of the other and vice versa. Figure 11.23 shows two possible configurations. The comprehensive device is called a *diac* and has two effective terminals, namely T_1 and T_2 as shown in the figure.

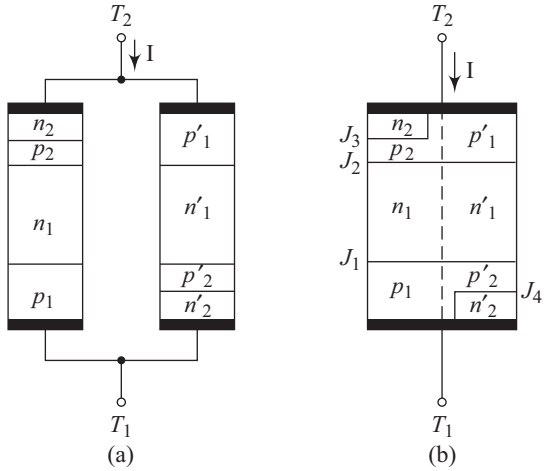


Fig. 11.23 Two configurations of bidirectional thyristors

If a positive bias is applied to the T_1 terminal with respect to T_2, junction J_4 gets reverse biased. In this cases the n_2 region does not contribute as a *p-n-p-n* diode. The *p-n-p-n* diode is then formed through the p_1, n_1, p_2, n_2 regions. On

the other hand, if a positive voltage is applied on the terminal T_2, the junction J_3 gets reverse biased. The p_1', n_1', p_2', and n_2' regions then constitute the appropriate *p-n-p-n* diode. The current–voltage characteristics of a diac are shown in Fig. 11.24.

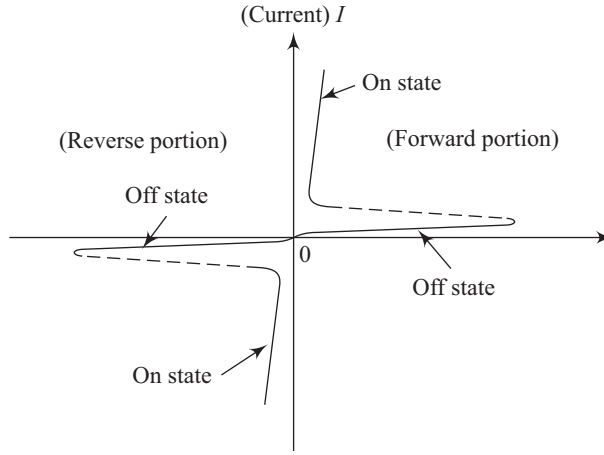


Fig. 11.24 Characteristics of a diac

The forward portion of the current–voltage characteristic curve represents the *p-n-p-n* diode and the reverse portion, the p_1' - n_1' - p_2' - n_2' diode. When suitable gate terminals are provided to a bidirectional diac, the device is called a *triac* (triode ac switch). This device can be used to switch currents in either direction by the application of low-voltage, low-current pulses of either polarity between a suitable gate electrode and one of the two terminals T_1 and T_2. A typical device is shown in Fig. 11.25.

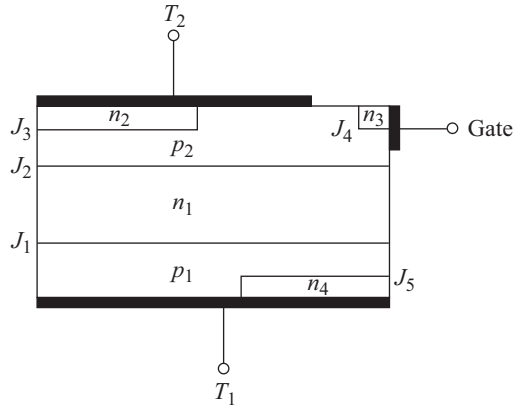


Fig. 11.25 Schematic diagram of a triac

The basic operating principle and the resultant current–voltage characteristics of a triac are identical to those of a diac. As depicted in Fig. 11.21, the breakover voltage can be varied with the adjustment of the magnitude of the gate current. Unlike the characteristic curve displayed in Fig. 11.21, this possibility exists in both polarities in the case of a triac.

11.5 Gate Turn-off Thyristor (GTO)

A gate terminal in a conventional SCR loses control over the functioning of the device once it starts to conduct. This necessitates the use of expensive commutation circuits. A gate turn-off thyristor (GTO) can be used to get over this need. A GTO is a very high power switch which is useful in industrial applications requiring high voltage-blocking and current-carrying capabilities. A GTO is basically a four-layer, three-junction. $(n^+ - p - n - p^+)$ structure. Unlike conventional thyristors, however, a GTO is designed such that it can be turned off with application of a negative voltage to the gate electrode. Thus it is not necessary to reverse the anode voltage to turn the device off. The costly commutation circuits required with conventional thyristors can thus be avoided. The turn-on features of a GTO is similar to a convential SCR. Figure 11.26 gives the cross section and symbol of a GTO.

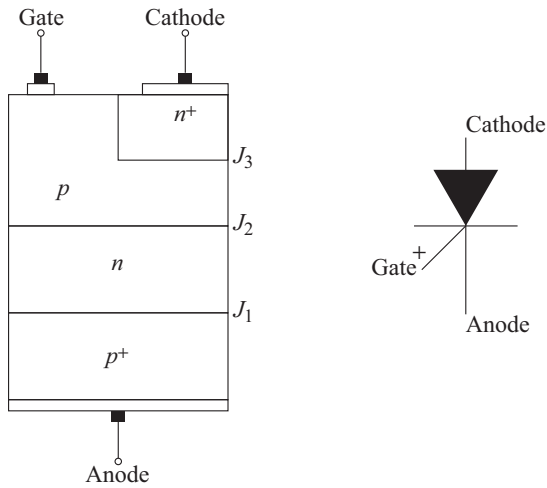


Fig. 11.26 (a) Cross section of a structure and (b) Symbol of a GTO

11.6 Insulated-gate Bipolar Transistor (IGBT)

The difficulty in switching off an SCR can be overcome using a novel device called the Insulated-gate Bipolar Transistor (IGBT). This device was invented by J. Baliga in 1979. A schematic diagram of the device is shown is Fig. 11.27.

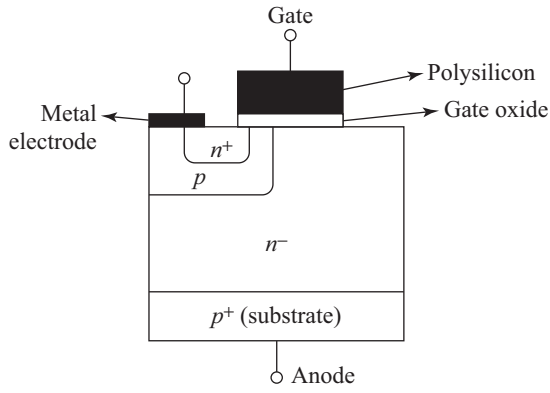


Fig. 11.27 Schematic of an IGBT

The *n*-channel device shown in Fig. 11.27 is basically a combination of an SCR and a MOSFET. The MOSFET is used to connect or disconnect the cathode (n^+) to the base (n^-) regions. When the gate voltage is either zero or below the threshold voltage, no *n*-type inversion layer is formed in the *p*-type channel region. This effectively decouples the n^+ cathode and n^- base. The structure then behaves like a conventional SCR. A low minimum current flows in either polarity and ultimately a breakdown is observed. For positive anode-to-cathode bias, this breakdown is due to the $n^- - p^+$ junction. With an applied gate bias, a significant current is observed for a positive anode-to-cathode bias. Fig. 11.28 is a schematic representation of the I-V characteristics of an IGBT.

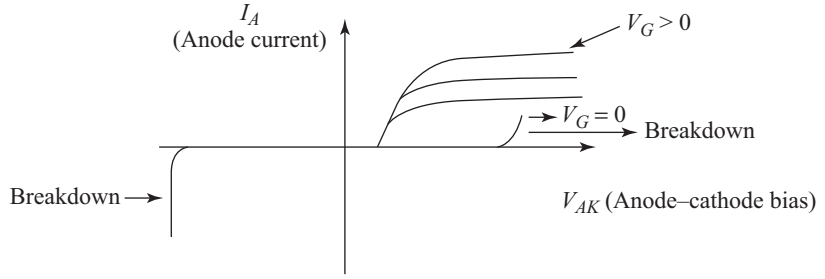


Fig 11.28 Schematic I-V characteristics of an IGBT

The *p* region is formed by diffusion of the acceptors implanted in the n^+ region. Thus unlike conventional MOSFET, the channel length is not determined using lithography but rather by diffusion of acceptor impurities. For this reason, the MOSFET structure so obtained is sometimes referred to as double-diffused MOSFET (DMOS).

The IGBT gives the best features available in MOSFETS and BJTs. Like any MOSFET, it has high input impedance and low input capacitance. In the 'on' state, the IGBT has low resistance and high current-carrying capability like any BJT. The IGBT is now the device of choice for high-power applications.

11.7 Unijunction Transistor

The *unijunction transistor* (UJT) is a solid-state device that can be used as a voltage-operated switch. A basic schematic diagram of the device is shown in Fig. 11.29(a). It consists of a bar of *n*-type silicon with two base terminals B_1 and B_2 at its two ends. A *p*-emitter region is provided on the side of this bar. The resistance between the bases B_1 and B_2 is typically in the region of 5 kΩ to 10 kΩ. Figure 11.29(b) shows the symbol for the device.

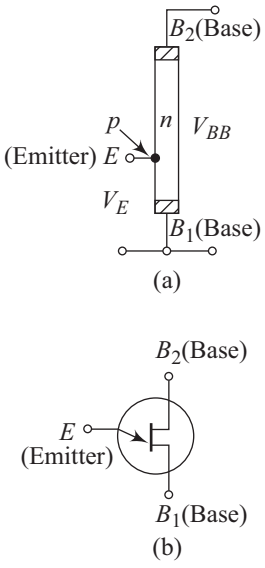


Fig. 11.29 (a) Schematic diagram of the UJT (b) Symbol for the device

During operation, the base B_2 is held positive with respect to B_1 with a total voltage drop of V_{BB}. The potential at emitter, *E*, is a fraction of V_B, called $\eta V_{BB} = V_E$. If the input voltage on the emitter terminal is less than ηV_{BB}, the emitter-to-*n*-bar silicon junction is reverse biased and only a leakage current flows through the junction. As V_E becomes more positive than ηV_{BB}, the emitter–*n*-bar junction becomes forward biased and results in a large emitter current. The current consists of injection of holes into the bar. These holes move towards B_1. The presence of these holes is accompanied by an increase in the number of electrons in the region. The process leads to an increase of the conductivity of the emitter–B_1 portion of the bar. This increase in conductivity results in a drop in the voltage across this region. Figure 11.30 shows an equivalent circuit of the unijunction transistor. From this figure it is clear that the device can be simulated as a voltage divider. Figure 11.31 shows typical UJT (unijunction transistor) emitter characteristics.

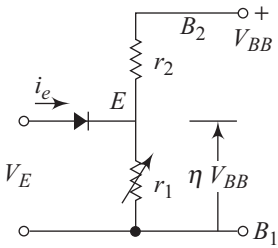


Fig. 11.30 Equivalent circuit of the UJT

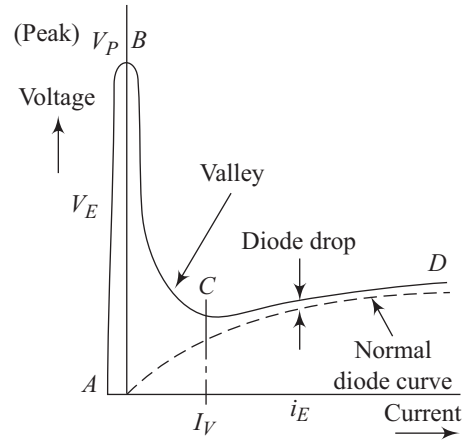


Fig. 11.31 Typical UJT emitter characteristics

The dashed curve in Fig. 11.31 is the conventional diode curve. As the emitter–bar (*n*) diode gets forward biased, the *V-I* curve follows the path *BCD* and the voltage drops in the region from *B* to *C*. Thus this region displays a negative resistance. The voltage then follows the normal diode characteristics from *C* to *D*.

Each value of V_{BB} results in its own characteristic curve. Each curve has a maximum voltage point called the peak point, V_p, and a low voltage point called the valley point V_v. As one goes to the right of the valley point, the emitter–B_1 path gets saturated with carriers and yields only a small positive resistance.

The peak voltage is given by

$$V_P = \eta V_{BB} + 0.7 \text{ V} \tag{11.17}$$

where the 0.7 V is due to the inherent emitter–base forward drop of a silicon *p-n* junction diode.

η is called the intrinsic stand-off ratio given by

$$\eta = \frac{r_1}{(r_1 + r_2)}$$

- Another member of the SCR family is a GTO or a gate turn-off switch. This device can be turned on or off by the application of a gate pulse.
- A reverse conducting thyristor (RCT) is a three-terminal, multilayer thyristor. In the forward direction, its behaviour is similar to a conventional thyristor. The device can conduct large currents in the reverse direction.
- Another interesting device of the thyristor family is a light-activated switch or LAS. It is a three-terminal four-layer reverse-blocking thyristor. The device can be switched by exceeding its light threshold. It finds widespread use in the area of fibre-optical transmission.

Solved Problems

11.1 A common-emitter circuit configuration is shown in Fig. 11.1.1.

Important parameter values are

$$R_L = 8\ \Omega;\ V_{CC} = 30 \text{ V}$$

Calculate the maximum collector current, maximum collector–emitter voltage, and the maximum power rating, P_T, for the power BJT.

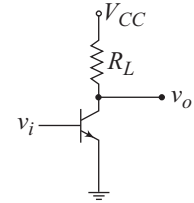


Fig. 11.1.1 Common-emitter circuit employing a BJT

Solution

Maximum collector current, $I_{C(\max)}$, is given by

$$I_{C(\max)} = \frac{V_{CC}}{R_L} \tag{11.1.1}$$

Putting the given values in Eqn (11.1.1) we get

$$I_{C(\max)} = \frac{30}{8} = 3.75 \text{ A} \tag{11.1.2}$$

The maximum collector–emitter voltage, $V_{CE(\max)}$, prevails with $I_C = 0$. We have

$$V_{CE(\max)} = V_{CC} = 30 \text{ V} \tag{11.1.3}$$

The load line can be expressed by the equation

$$V_{CE} = V_{CC} - I_C R_L \tag{11.1.4}$$

Power dissipation, P_T, is expressed as

$$P_T = V_{CE} I_C \tag{11.1.5}$$

Using Eqn (11.1.4) in Eqn (11.1.5) results in

$$P_T = (V_{CC} - I_C R_L) I_C$$

which can be simplified to

$$P_T = V_{CC} I_C - I_C^2 R_L \tag{11.1.6}$$

The maximum power dissipation condition implies

$$\frac{dP_T}{dI_C} = 0 \tag{11.1.7}$$

Using Eqs (11.1.7) and (11.1.6) yields

$$0 = V_{CC} - 2I_C R_L$$

which leads to

$$I_C = \frac{V_{CC}}{2R_L} \tag{11.1.8}$$

Putting the given values in Eqn (11.1.8) results in

$$I_C = \frac{30}{2(8)} = \frac{30}{16} = 1.88 \text{ A} \tag{11.1.9}$$

The collector–emitter voltage at maximum power dissipation point is given by

$$V_{CE} = V_{CC} - I_C R_L \tag{11.1.10}$$

Using the magnitude of I_C from Eqn (11.1.9) and the other given values in Eqn (11.1.10), yields

$$V_{CE} = 30 - (1.88 \times 8)$$

Implying

$$V_{CE} = 14.96 \text{ V} \tag{11.1.11}$$

Finally, the maximum power dissipation, P_T, can be obtained using Eqs (11.1.11) and (11.1.9), yielding

$$P_T = V_{CE} I_C = 14.96 \times 1.88 = 28.12 \text{ W}$$

One must, however, remember that the proper choice of the BJT to be used will be decided by providing adequate safety margins.

11.2 A power MOSFET inverter circuit is shown in Fig. 11.2.1. The load line for the circuit intersects the voltage axis at $V_{DD} = 25$ V. The load line is tangent to the

maximum power dissipation curve and intersects the current axis at $I_D = 4$ A.
Calculate the drain resistance, R_D, drain current at maximum power dissipation point, drain-to-source voltage at maximum power dissipation point, and the maximum power dissipation, P_T.

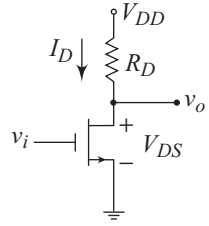


Fig. 11.2.1 Power MOSFET inverter circuit

Solution

Drain resistance, R_D, is given by

$$R_D = \frac{V_{DD}}{I_D} \tag{11.2.1}$$

Putting the given values in Eqn (11.2.1) results in

$$R_D = \frac{25}{4} = 6.25\ \Omega \tag{11.2.2}$$

Using an expression similar to Eqn (11.1.8) in Solved Problem (11.1), we can write for drain current at maximum power dissipation point

$$I_D = \frac{V_{DD}}{2R_D} \tag{11.2.3}$$

Putting the magnitude of R_D derived from Eqn (11.2.2) and the given value of V_{DD} in Eqn (11.2.3) leads to

$$I_D = \frac{25}{2 \times 6.25} = 2\ \text{A} \tag{11.2.4}$$

The corresponding drain-to-source voltage can be obtained using the expression

$$V_{DS} = V_{DD} - I_D R_D \tag{11.2.5}$$

Using Eqs (11.2.1) and (11.2.4) and the given value of V_{DD} in Eqn (11.2.5) yields

$$V_{DS} = 25 - (2 \times 6.25)$$

Resulting in

$$V_{DS} = 12.5\ \text{V} \tag{11.2.6}$$

The maximum power dissipation, P_T, can be obtained using the expression

$$P_T = V_{DS} I_D \tag{11.2.7}$$

Using Eqn (11.2.6) and (11.2.4) in Eqn (11.2.7) results in

$$P_T = 12.5 \times 2 = 25\ \text{W}$$

11.3 A Darlington pair configuration uses two identical bipolar junction transistors with gains $\beta_1 = \beta_2 = 20$. Calculate the net common-emitter current gain.

Solution

The net common-emitter current gain of a Darlington pair configuration is given by

$$\frac{i_C}{i_B} = \beta = \beta_1\beta_2 + \beta_1 + \beta_2 \tag{11.3.1}$$

where β_1 and β_2 represent the individual transistor current gains.
Putting the given values in Eqn (11.3.1) results in

$$\frac{i_C}{i_B} = 20 \times 20 + 20 + 20 = 440$$

11.4 A power BJT has the following thermal resistance parameters:

$R_{\text{th}(d-p)} = 1.7°\text{C/W}$; $R_{\text{th}(p-s)} = 1°\text{C/W}$

$R_{\text{th}(s-a)} = 4°\text{C/W}$; $R_{\text{th}(p-a)} = 40°\text{C/W}$

Assuming an ambient temperature $T_{\text{amb}} = 27°\text{C}$ and the maximum junction/device temperature to be $T_{j(\max)} = T_{\text{dev}} = 150°\text{C}$, calculate the maximum power dissipation in the transistor for cases where (a) no heat sink is used and (b) a heat sink is used.

Solution

Case (a) No heat sink used

For this case, the maximum power dissipation, $P_{D(\max)}$, is given by

$$P_{D(\max)} = \frac{T_{j(\max)} - T_{\text{amb}}}{R_{\text{th}(d-p)} + R_{\text{th}(p-a)}} \tag{11.4.1}$$

Putting the given values in Eqn (11.4.1) gives

$$P_{D(\max)} = \frac{(150 - 27)}{1.7 + 40}$$

$$= 2.95 \text{ W}$$

Case (b) Heat sink used

For this situation, the maximum power dissipation, $P_{D(\max)}$, is given by

$$P_{D(\max)} = \frac{T_{j(\max)} - T_{\text{amb}}}{R_{\text{th}(d-p)} + R_{\text{th}(p-s)} + R_{\text{th}(s-a)}} \tag{11.4.2}$$

Putting the given values in Eqn (11.4.2) yields

$$P_{D(\max)} = \frac{(150 - 27)}{1.7 + 1 + 4}$$

$$= 18.36 \text{ W}$$

11.5 A typical bipolar junction transistor has a current gain of 10 and a power limitation of 200 W for the safe operation area (SOA). Calculate the total power dissipation of the BJT for the conditions shown in Fig. 11.5.1. Also, specify the transistor's operating region and comment on whether the BJT is working in the safe operating region or not.

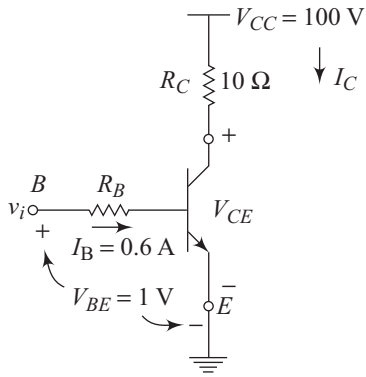


Fig. 11.5.1 Power BJT under DC bias condition

Solution

Current gain, $\beta = \dfrac{I_C}{I_B}$

Given $\beta = 10$

$\therefore$ $I_C = 10\, I_B$

It is given that $I_B = 0.6$ A and $V_{BE} = 1.0$ V.

Therefore, $I_C = 10 \times 0.6 = 6$ A

The collector–emitter voltage is given by

$$V_{CE} = V_{CC} - I_C R_C$$

$$V_{CE} = 100 - (6 \times 10) = 40 \text{ V}$$

Therefore, $V_{CE} = 40$ V

The collector–base voltage is given by

$$V_{CB} = V_{CE} - V_{BE}$$

$$V_{CB} = 40 - 1 = 39 \text{ V}$$

Since, V_{CB} is positive the CB junction is reverse biased. Hence, the BJT is operating in the active region.

Total power dissipation, $P_T = V_{CE} \times I_C + V_{BE} \times I_B$

$\Rightarrow P_T = (40 \times 6) + (1 \times 0.6) = 240 + 0.6 = 240.6$ W

Therefore, $P_T = 240.6$ W

Therefore, the BJT is working outside the SOA.

11.6 The Darlington pair configuration shown in Fig. 11.6.1 has an effective gain of 250 and the current gain of transistor T_1 is 25. Find the value of emitter current i_E of T_2, if the base of BJT T_1 is given by bias conditions of 50 mA.

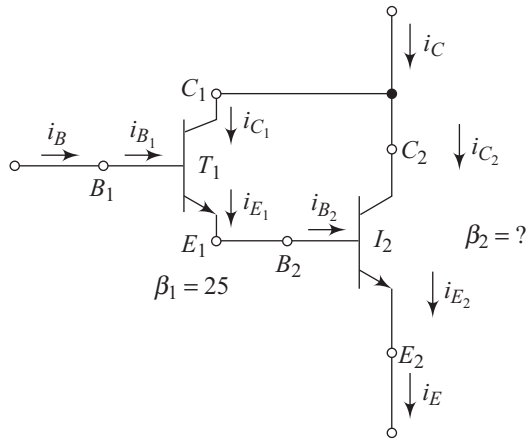


Fig. 11.6.1 Darlington pair configuration

Solution

The effective gain of a Darlington pair is given as

$$\beta_{\text{eff}} = \beta_1 \beta_2 + \beta_1 + \beta_2$$

or $250 = 25 \beta_2 + \beta_1 + \beta_2$

$= 26 \beta_2 + 25$

$\therefore \quad 225 = 26 \beta_2$

$$\therefore \quad \beta_2 = \frac{225}{26} = 8.65$$

Therefore, $\beta_2 = 8.65$

Emitter current, $i_{E_2} = i_E = i_{B_2} + i_{C_2} = i_{C_2} + \dfrac{i_{C_2}}{\beta_2} = \left(1 + \dfrac{1}{\beta_2}\right) i_{C_2}$

Collector current, $i_{C_2} = i_C - i_{C_1} = \beta_{\text{eff}} \cdot i_B - \beta_1 \cdot i_B$
$= i_B\,(250 - 25) = 225 \times 50 \times 10^{-3} = 11.25$ A

Therefore $i_{E_2} = i_E = \left(1 + \dfrac{1}{8.65}\right) i_{C_2} = 12.55$ A

11.7 A UJT operates at 24 V with a base resistance of $r_1 = 3$ kΩ and $r_2 = 5$ kΩ. Find the peak-point voltage V_P.

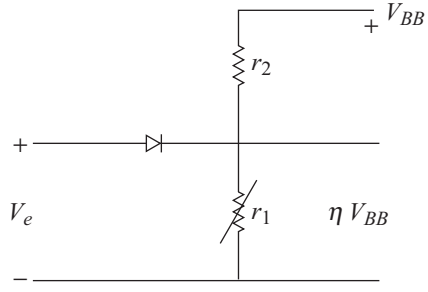


Fig. 11.7.1 UJT operation

Solution

From Eqn (11.17) in the text

$$V_P = \eta V_{BB} + 0.7$$

where $\eta = \dfrac{r_1}{(r_1 + r_2)} = \dfrac{3}{8}$

$\therefore \quad V_P = 9.7$ V

11.8 A BJT has been rated to handle a maximum of 20 W of power and 200°C of junction temperature. The ambient is at standard room temperature and the thermal resistance parameters of the device are

R_{th} (sink–ambient) = 4°C/W

R_{th} (case–sink) = 1.5°C/W

Find the actual power that will be dissipated in the transistor.

Solution

Given R_{th} (sink–ambient) = 4°C/W
and R_{th}(case–sink) = 1.5°C/W
Now, from Eqn (11.8) in the text we have

$$R_{\text{th}} = \frac{T_2 - T_1}{P} \tag{11.8.1}$$

Now, substituting $T_2 = 200$°C, $T_1 = 27$°C (room temperature), and $P_{D(\text{rated})} = 20$ W into Eqn (11.8.1) we get

$$R_{\text{th}} = \frac{200 - 27}{20}$$
$$= 8.65\text{°C/W}$$

Now from Eqn (11.9) in the text we have

$$P_D = \frac{T_{\text{dev}} - T_{\text{amb}}}{R_{\text{th}(d-p)} + R_{\text{th}(p-s)} + R_{\text{th}(s-a)}} \tag{11.8.2}$$

Substituting $T_{\text{dev}} = 200$°C, $T_{\text{amb}} = 27$°C,
$R_{\text{th}(d-p)} = 8.65$°C/W,
$R_{\text{th}(p-s)} = R_{\text{th(case-sink)}} = 1.5$°C/W, and
$R_{\text{th}(s-a)} = R_{\text{th(sink-amb)}} = 4$°C/W

into Eqn (11.8.2) we get

$$P_D = \frac{200 - 27}{8.65 + 1.5 + 4}$$

$$= \frac{173}{14.15}$$

$$= 12.22 \text{ W}$$

11.9 For a GaAs BJT, the maximum junction temperature is 400 and the supplied power is 90 W. The thermal resistance between the BJT and the heat sink is estimated to be 1.5°C/W.

(a) Determine the maximum thermal resistance of the heat sink, if the ambient operating temperature does not exceed 50°C.

(b) For a heat convection coefficient of 100 W/°Cm^2, find the required surface area.

Solution

(a) Junction temperature is calculated as

$$T_j = T_A + [(R_{\text{th}(d-p)} + R_{\text{th}(s-a)}]P$$

Here

$R_{\text{th}(d-p)}$ = 1.5°C/W is the thermal resistance from the device to the packaging. $R_{\text{th}(s-a)}$ is given as

$$R_{\text{th}(s-a)} = \frac{T_j - T_A}{P} - R_{\text{th}(d-p)}$$

$$= \left(\frac{400 - 50}{90}\right) - 1.5$$

$$= 2.38°\text{C/W}$$

(b) The surface area can be calculated as

$$A = \frac{1}{R_{\text{th}(s-a)} \times \text{convection coefficient}}$$

$$= \frac{1}{2.38 \times 100}$$

$$= 4.2 \times 10^{-3} \text{ m}^2$$

Recapitulation

- Maximum rated power, P_T, of a power BJT is given by

$$P_T = V_{CE}I_C + V_{BE}I_B$$

- For the Darlington pair configuration

$$\frac{i_C}{i_B} = \beta_1\beta_2 + \beta_1 + \beta_2$$

- On-state resistance of a power MOSFET can be expressed in the form
$$R_{ON} = R_S + R_{CH} + R_D$$
where,
$$R_{CH} = \frac{L}{W \mu_n C_{ox} (V_{GS} - V_T)}$$
- Thermal resistance, R_{th}, is given by
$$R_{th} = \frac{(T_2 - T_1)}{P}$$
where P represents the thermal power across the element.
- The maximum safe power dissipation, $P_{D(max)}$, can be expressed as
$$P_{D(max)} = \frac{T_{j(max)} - T_{Package}}{R_{th(d-p)}}$$
- The peak voltage, V_P, of a unijunction transistor is given by
$$V_P = \eta V_{BB} + 0.7 \text{ V}$$

Exercises

Review Questions

11.1 What is current crowding in a bipolar junction transistor? Show the process with the help of a schematic diagram.

11.2 Why are vertical structures preferred for a power transistor?

11.3 Explain the process of second breakdown in a BJT. What are its implications?

11.4 Why is a Darlington pair configuration useful for power transistor application?

11.5 Give a schematic diagram of a VMOS power MOSFET and explain its operation.

11.6 What is the role of crystallography in the fabrication of the VMOS structure?

11.7 How does the on state resistance of a power MOSFET help in the even distribution of current in different cells of the power transistor?

11.8 What are the three factors that decide the safe operating area of power MOSFETs? Compare these with the corresponding factors for power BJTs.

11.9 Define thermal resistance and give its units. What are the different components of thermal resistance for a power device?

11.10 Explain the operating principle of a *p-n-p-n* diode. What are the different regions in its current–voltage characteristics?

11.11 Use the two-transistor model to explain the operation of a *p-n-p-n* diode.

11.12 Draw a sketch of a typical thyristor device.

11.13 Explain the role of gate-current pulse in the functioning of a thyristor.

11.14 Explain the bidirectional property of a diac.

11.15 Draw and explain the equivalent circuit of a unijunction transistor.

11.16 Explain the process of current crowding and how it is taken care of in a power transistor.

11.17 Write an expression for gain of a Darlington pair configuration.

11.18 Explain the fabrication process of the VMOS structure in brief.

11.19 Write and explain the expression for maximum power dissipation of a power transistor.

11.20 Explain the current–voltage characteristics of an SCR.

11.21 Differentiate between a triac and a diac.

11.22 Explain the two-transistor model of a *p-n-p-n* diode.

11.23 Explain briefly the operation of the SCR in the forward-blocking region with the help of a schematic diagram depicting all the junctions and applied voltages.

Problems

11.1 Figure 11.P11.1 shows a common-emitter circuit configuration.
The values of R_L and V_{CC} are 10 Ω and 25 V respectively. Determine the maximum collector current, maximum collector–emitter voltage and the maximum power rating, P_T, for the power BJT.

$$\left[\textit{Hint: } I_{C(\max)} = \frac{V_{CC}}{R_L}, V_{CE} = V_{CC} - I_C R_L \right]$$

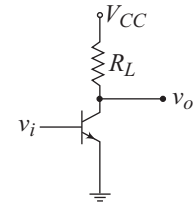


Fig 11.P11.1 Common-emitter circuit configuration

Ans. 2.5 A, 25 V, 15.63 W

11.2 In the circuit of Fig. 11.P11.1 the collector supply voltage V_{CC} is increased to twice its original value. How does this affect the magnitudes of maximum collector current and maximum collector–emitter voltage?

Ans. 5 A, 50 V

11.3 A circuit involving a power MOSFET is shown in Fig. 11.P11.3. The load line for the circuit intersects the voltage axis at V_{DD} = 30 V and is tangent to the maximum power dissipation curve, intersecting the current axis at I_D = 5 A.
Determine the drain resistance, R_D, drain current at maximum power dissipation point, drain-to-source voltage at maximum power dissipation point, and the maximum power dissipation, P_T.

$$\left[\textit{Hint : } R_D = \frac{V_{DD}}{I_D}, V_{DS} = V_{DD} - I_D R_D \right]$$

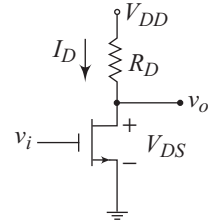


Fig 11.P11.3 Circuit using a power MOSFET

Ans. 6 Ω, 2.5 A, 15 V, 37.5 W

11.4 A MOSFET with a maximum power dissipation capability of 20 W is to be used in the circuit shown in Fig. 11.11P.3. Assume $V_{DD} = 20$ V and that the tangent to the maximum power dissipation curve intersects the current axis at $I_D = 3$ A. If the safety factor is to be 50%, can the available MOSFET be used for the application.
[*Hint:* $P_T = V_{DS} I_D$]

Ans. $P_T = 15$ W; thus the given MOSFET cannot be used.

11.5 Two bipolar junction transistors with gains $\beta_1 = \beta_2 = 15$ are being used in a Darlington configuration. Determine the net common-emitter current gain.

$$\left[\textit{Hint:} \frac{i_C}{i_B} = \beta = \beta_1\beta_2 + \beta_1 + \beta_2 \right]$$

Ans. 255

11.6 A Darlington configuration is to have a net common-emitter gain of 400. If one of the transistors in the configuration has a current gain $\beta_1 = 20$, calculate the magnitude of β_2.

Ans. 18.1

11.7 A particular power transistor has the following relevant parameters:

$R_{\text{th}(d-p)} = 2°\text{C/W}$; $R_{\text{th}(p-s)} = 1°\text{C/W}$
$R_{\text{th}(s-a)} = 3.5°\text{C/W}$; $R_{\text{th}(p-a)} = 35°\text{C/W}$

Calculate the maximum power dissipation in the transistor for (a) no heat sink (b) heat sink, cases. Assume $T_{\text{amb}} = 27°\text{C}$ and $T_{j(\max)} = T_{\text{dev}} = 150°\text{C}$.

$$\left[\textit{Hint:} P_{D(\max)} = \frac{T_{j(\max)} - T_{\text{amb}}}{R_{\text{th}}} \right]$$

Ans. (a) 3.32 W (b) 18.92 W

11.8 Calculate the magnitude of $R_{\text{th}(s-a)}$ required for the power transistor mentioned in Problem 11.7 such that the power dissipation in the transistor can reach a value of 30 W. Assume other parameters to have the same values as given in Problem 11.7.

Ans. 1.1°C/W

11.9 The peak point voltage of a unijunction transistor is 20 V. Calculate the value of coefficient η for the transistor if $V_{BB} = 30$ V.
[*Hint:* $V_P = \eta V_{BB} + 0.7$]

Ans. 0.64

11.10 The length of a typical power MOSFET is 0.4 μm and its width is 100 μm. If the gate bias voltage is 3 V, find the channel resistance of the device. Assume $V_T = 1$ V and $\mu_n C_{\text{ox}} = 240 \times 10^{-6}\ \Omega^{-1}/\text{V}$.

$$\left[\textit{Hint:} R_{\text{ch}} = \frac{L}{W \times \mu_n C_{\text{ox}} (V_{GS} - V_T)} \right]$$

Ans. 8.33 Ω

11.11 The maximum temperature of a device is 175°C and the temperature of the ambient is 27°C. The thermal resistances shown in Fig. 11.P11.1 are given as $R_{\text{th}(d-p)} = 12°\text{C/W}$, $R_{\text{th}(p-s)} = 2°\text{C/W}$, and $R_{\text{th}(s-a)} = 2.4°\text{C/W}$.

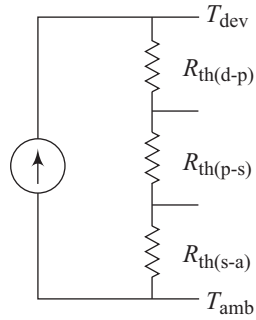


Fig. 11.P11.11 Thermal resistances

Calculate the temperatures of the package and the heat sink of the device.

$$\left[\textit{Hint}: P_{D(\max)} = \frac{T_{j(\max)} - T_{amb}}{R_{th(d-p)} + R_{th(p-s)} + R_{th(s-a)}} \right.$$

$$T_p = T_d - (R_{th(d-p)} \times P_{D(\max)})$$

$$\left. T_s = T_p - (R_{th(p-s)} \times P_{D(\max)}) \right]$$

Ans. 143.3°C; 90.44°C

11.12 A power BJT has a base doping concentration of 5×10^{15} cm^{-3} and the maximum breakdown voltage for the base–collector junction is 10^3 V. Calculate the maximum possible collector doping concentration, the minimum base width, and the minimum collector width.

Ans. 2×10^{14} cm^{-3}, 3.16 μm, 78.9 μm

11.13 A BJT is being operated in the common-emitter configuration. The maximum collector-emitter voltage is 28 V. If the maximum collector current, $I_{C(\max)}$ is 3.9 A, calculate the load resistance.

Ans. 7.2 Ω

11.14 A UJT is being operated with base resistance $r_1 = 4$ kΩ and $r_2 = 5$ kΩ. Calculate the intrinsic stand-off ratio, r.

Ans. 0.44

12 Integrated Circuits and Micro-electromechanical Systems

- Passive components
- Bipolar technology
- MOSFET technology
- Charge-coupled devices
- CMOS technology
- MESFET technology
- Micro-electromechanical systems
- Bulk micromachining
- Surface micromachining

Learning Objectives

After going through this chapter the student will be able to

- understand how passive components like resistors, capacitors, and inductors are realized in an integrated circuit (IC)
- understand junction isolation based bipolar technology
- understand oxide isolation based bipolar technology
- understand the principle of dielectric isolation
- solve numericals based on bar and meander shaped resistors
- solve numericals based on MOS structure based capacitors
- solve numericals based on *p-n* junction based capacitors
- understand details of the NMOS process
- understand the operation of charge-coupled devices
- understand CMOS technology based processing
- understand the advantages and disadvantages of MESFET technology
- understand the details of MESFET technology based processing
- understand the fundamentals of micro-electromechanical systems
- understand bulk micromachining based processing
- understand surface micromachining based processing

Introduction

The discrete devices discussed so far are important and form the basis of many electronic applications. The outstanding growth of electronics in recent years must, however, be attributed to a collection of active and passive components on a small piece of semiconductor, also known as a *chip*. This collection is also referred to as an integrated circuit or simply an IC. An integrated circuit has all the important devices discussed in previous chapters but in such a manner that the processing is totally compatible with standard technologies. This topic will be discussed in detail in this chapter. The three main technologies involved, namely bipolar technology, MOSFET technology, and MESFET technology will be presented to ensure elucidation of important ideas and concepts. Although the growth in the complexity of integrated circuits continues to be phenomenal, there are limits to its growth potential. Silicon technology has also found an excitingly novel application area in the form of micro-electromechanical systems. These systems exploit the mechanical strength of silicon and established silicon processing techniques to develop systems that can be used for a host of actuating and sensing applications. Some key aspects of the basic fabrication processes of MEMS will also be discussed in this chapter.

12.1 Photolithography

The word photolithography consists of two parts namely *photo* meaning light and *lithography*. *Lithography* is derived from two Greek words, *litho* (meaning stone) and *graphein* (meaning to write). Photolithography is a process step that uses an optical mask and a photosensitive film for producing a specific pattern on a substrate. A typical photolithography process is shown schematically in Fig. 12.1.

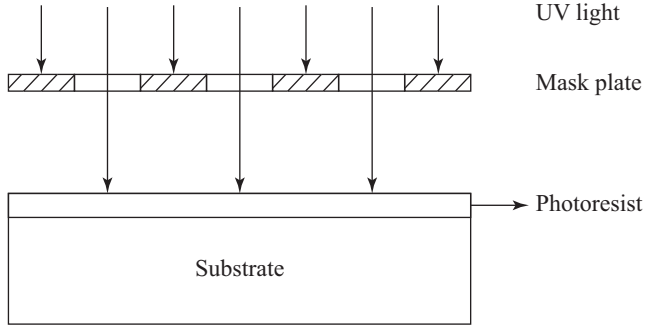
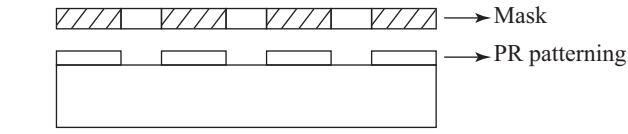
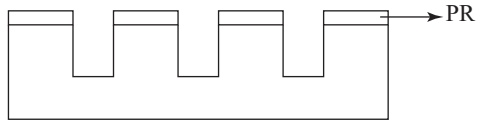


Fig. 12.1 Typical photolithography process

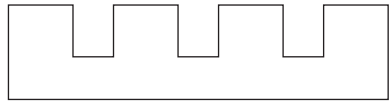
The substrate is first coated with a suitable photoresist material which serves as the photosensitive film. The photoresist material is then exposed to ultraviolet



(i) Patterning

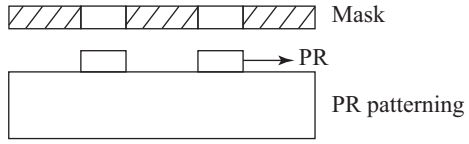


(ii) Etching

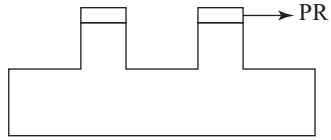


(iii) PR removal

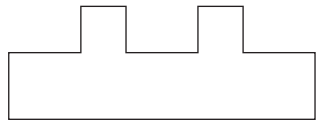
(a) Positive resist



(i) Patterning



(ii) Etching



(iii) PR removal

(b) Negative resist

Fig. 12.2 (a) Positive and (b) Negative photoresist based photolithography process

light through a suitably designed mask plate. Photoresists change their solubility in certain solvents on being exposed to ultraviolet light. There are two types of changes possible and this gives rise to two types of photoresists. Photoresists that become more soluble on being exposed to ultraviolet light are termed *positive photoresists*. *Negative photoresists* on the other hand, become more soluble under the shadow regions due to the opaque portion of the mask. The solvents used to produce a pattern in an exposed photoresist (exposed to UV light) are called developers. Positive and negative photoresists create different patterns during the photolithography process as shown schematically in Fig. 12.2.

To differentiate the patterns obtained using positive and negative photoresists, it is assumed that the mask design is similar for both the resists. The transparent portion of the mask make corresponding areas of the positive photoresist more soluble in the developer. These areas in the substrate can then be etched in suitable etchants. This step is followed by photoresist removal. In the case of negative photoresist, the area under the transparent portion of the mask becomes resistant to removal by the developer. The final etched pattern realized using the negative photoresist is shown in Fig. 12.2 (b)(iii).

The sophisticated process of photolithography requires extremely clean environment. These clean rooms are designated with class numbers, a class-100 clean room, for example, must not have more than 100 dust particles of size 0.5 μm or larger in a cubic foot of air within the room. Better than class –10 conditions are necessary to realize the microstructures required by MEMS devices and systems. Many other micro-fabrication process steps like wafer cleaning, metallization, and plating, also require clean rooms in the region of class–100. Fabrication of clean rooms is expensive and pushes up the cost factor of the devices produced.

Numerous photoresists are available for use in the photolithography process. Different photoresists have characteristics that make them suitable for a variety of applications. Photoresist thickness, viscosity, development time are some important characteristics that have to be kept in mind while making a choice for a particular application. Positive photoresists are generally of two types: (i) PMMA (polymethy methacrylate) resists and (ii) two-component DQN resists consisting of diazoquinone ester (DQ) and phenolic norolak resin (N). The PMMA resists are also used in electron beam lithography, ion beam lithography, and x-ray lithography. Positive resists can be developed using alkaline solvents like KOH (potassium hydroxide), TMAH (tetramethyl–ammonium hydroxide), ketones, or acetates.

Two common negative photoresists in use are (i) a two–component bis (aryl) azide rubber resist and (ii) azide-sensitized polyisotroprene rubber resist. In general, negative photoresists are less sensitive to exposure from optical radiation and X-rays. Negative photoresists are however more sensitive to electron beam exposure in comparison to positive photoresists. Negative resist is generally developed in xylene. Positive photoresists provide a better edge definition than negative photoresists.

The substrate to be coated with a film of photoresist is first held secure on the top of a vacuum chuck. A resist puddle is then created by dispensing some photoresist in the central portion of the substrate. The vacuum chuck forms a part of a high-speed spinner. Using this system, the substrate is allowed to spin at a rotational speed in the range from 1500 to 8000 rpm for a duration in the range from 10 to 60 seconds. The particular choice of spin speed and spin duration is dependent upon the choice of photoresist, desired thickness and uniformity of the photoresist later. Thickness in the range from 0.5 to 2μm with thickness variation in the region of ±5 nm can be obtained. Some special photoresists are available for realizing photoresist thickness in the range of 8 to 14 μm. Standardization experiments have to be carried out to obtain the optimum conditions for the requisite photoresist specifications.

A bead of resist can get formed after the process of spinning of the photoresist. These beads form at the edge of the substrate and thickness of the photoresist can be several times the thickness obtained on other locations of the substrate. An optimization of spin speed and spin time can help in the reduction of edge bead formation to a considerable extent.

Most photoresists are sensitive to light falling in the wavelength range from 300 to 500 nm. A mercury vapour lamp is the most commonly used exposure source in photolithography. The light available from a mercury vapour lamp has a wavelength spectrum from 310 to 440 nm. Normal UV radiation has wavelength range of 150 to 300 nm. For very high resolution lithography, X-rays are used as exposure radiation. X-rays have wavelengths in the range of 4 to 50Å (1 Angstrom Å = 10^{-10} m).

12.2 Etching Techniques

Selective removal of material from unprotected areas is called *etching*. Different materials can be used to protect the areas that are not to be removed away. Photoresists, metals, nitrides, and oxides are the more common choices. Etching is in general of two types, namely wet etching and dry etching. Wet etching uses chemicals to remove unwanted material. Dry etching on the other hand uses gases. We will discuss the process of etching in some detail in this chapter due to its great importance.

12.2.1 Wet Etching

In the wet etching process, chemical etchants are used to remove material from unprotected regions. A schematic representation of wet etching process is shown in Fig. 12.3.

The schematic shown in Fig. 12.3 uses photoresist as the protective mask. It must however be kept in mind that the exact contour of material after etching will depend upon the nature of etchant-material etching characteristic. Wet etching is of two types, namely isotropic etching and anisotropic etching.

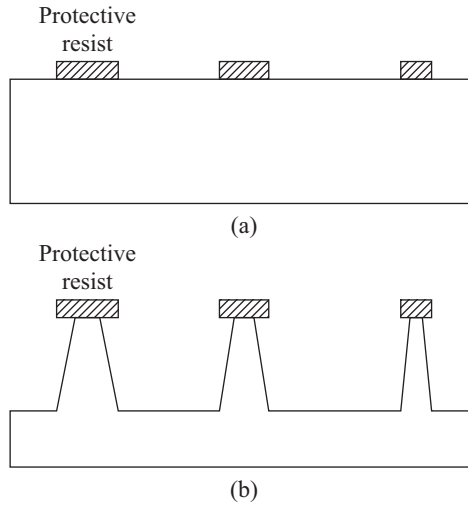


Fig. 12.3 Schematic representation of wet etching technique (a) patterned sample (b) sample after partial etching

Isotropic etching

The process of etching using etchants that attack the material uniformly in all directions is called isotropic etching. Isotropic etching is orientation-independent. Figure 12.4 shows a typical post-etch profile obtained using isotropic etching.

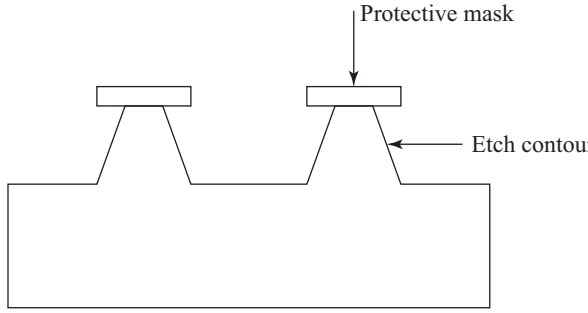


Fig. 12.4 Typical post-etch profile for istropic etching

Isotropic etching of silicon uses etchants that are mixtures of hydrofluoric acid (HF), nitric acid (HNO_3), and water or acetic acid (CH_3COOH). Due to the constituents used, they are collectively referred to as HNA etchant system. The chemical process for HNA etchant consists of hole injection and OH^- attachment to form $Si(OH)_2$. This is followed by the release of hydrogen to form SiO_2. Hydrofluoric acid is then used to dissolve away the SiO_2 by forming water soluble H_2SiF_6 according to the reaction,

$$Si + HNO_3 + 6HF \rightarrow H_2SiF_6 + H_2NO_2 + H_2O + H_2$$

Water or acetic acid is generally used as a dilutant for the isotropic etchant. Acetic acid as a dilutant helps in controlling the dissociation of the nitric acid and preserving the oxidizing power of HNO_3 for a wide usable range of dilution, thereby acting as a buffer. At high concentrations of HF and low concentrations of HNO_3, the etch rate is controlled by the concentration of HNO_3. At low HF and high HNO_3 concentrations, the etch rate is controlled by the ability of HF to efficiently remove the SiO_2 as it gets formed.

Anisotropic etching

Etchants that etch at different etch rates in different crystal orientation directions are called anisotropic etchants. Crystalline substrate materials have specific structures. The most commonly used crystalline material silicon has a diamond cubic crystal structure. There are several planes within the silicon cubic structure. Some important planes are (100), (110), and (111). It must be remembered that these planes are not uniquely defined and would depend upon the particular choice of principal axes. Figure 12.5 shows a schematic representation of the three important planes for silicon.

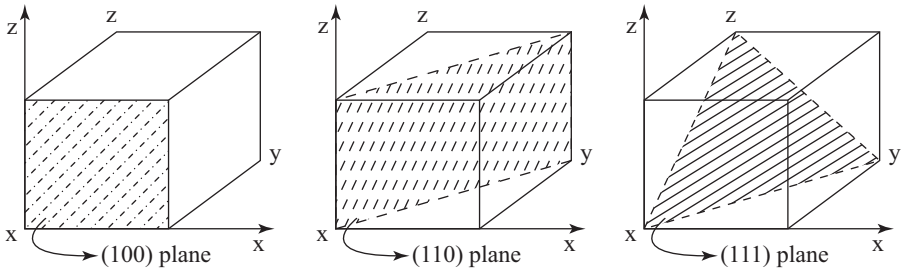


Fig. 12.5 The three important planes in silicon

The corresponding orientations <100>, <110>, and <111> are normal to the (100), (110), and (111) planes. A silicon wafer can be cleaved most cleanly along the <110> orientation. The silicon wafer can be cleaved in vertical edges only in the (110) plane. This is an important reason for preferring <110> oriented wafers for micro-machining. The IC industry, on the other hand prefers <100> and <111> orientations. The (111) plane is the most difficult plane mechanically and is therefore rarely used in micromachining. The mechanical strength of silicon along different orientation directions also gets reflected in the etching rate for certain etchants. In fact, it in possible to realize etch rate ratios in the region of 400 : 1 for <100> to <111> orientation.

The (111) plane intersects the (100) plane in silicon at an angle of 54.74°. Suppose we take a (100) surface of silicon and open windows aligned along (100) plane in a protective layer on the surface of silicon. On exposing the patterned wafer to some anisotropic etchants like potassium hydroxide (KOH), ethylene diamine pyrocatecol (EDP) or tetramethyl ammonium hydroxide (TMAH), a

pyramid shape with sidewall slope of 54.74° would be obtained. This shape would be due to the extremely slow etching rate of (111) plane with respect to the (100) plane. The process is shown schematically in Fig. 12.6.

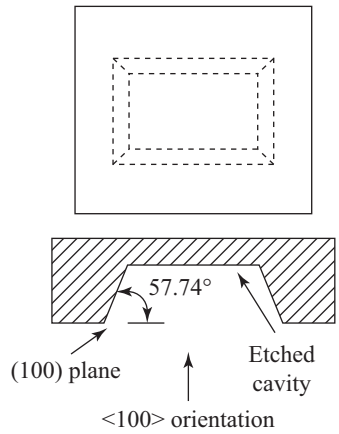


Fig. 12.6 Anisotropic etching of silicon (a) plan view (b) cross-sectional view

Potassium hydroxide (KOH) and ethylene diamine pyrocatecol based anisotropic etchants are more popular. Typically these chemicals are diluted in the ratio 1:1 with water by weight. Silicon dioxide and silicon nitride are two commonly used protective masks. The etch rate for the protective mask material should be extremely small as compared to silicon for a good choice of protective mask material. For anisotropic etching to be of practical use, the etch rate obtained for silicon should also be reasonable. Table 12.1 gives the etching rates of silicon and some of its compounds, for two popular anisotropic etchants.

Table 12.1 Etching rates of anisotropic etchants

Material	Etchant	Etch rate
Silicon <100>	EDP	0.75 μm/min
Silicon <100>	KOH	0.25 – 1.4 μm/min
Silicon dioxide	EDP	12 nm/h
Silicon dioxide	KOH	40 – 80 nm/hr
Silicon nitride	EDP	6 nm/hr
Silicon nitride	KOH	5 nm/hr

From Table 12.1 it is clear that the etching rate for silicon dioxide using KOH is 1000 times less than silicon. For silicon nitride, the etching rate is lower than silicon dioxide by one order. This property of the anisotropic etchant is important for practical applications and gets defined through **selectivity ratio. Selectivity ratio** is defined as the ratio of etching rate of silicon to the etching rate of the specific material for the same etchant. Silicon dioxide for example has a selectivity ratio of 10^3 for KOH. Table 12.2 lists the selectivity ratio of some etchants for two common silicon compounds.

Table 12.2 Selectivity ratio of common etchants

Compound	Etchant	Selectivity ratio
Silicon dioxide	KOH	10^3
	EDP	$10^3 - 10^4$
	TMAH	$10^3 - 10^4$
Silicon nitride	KOH	10^4
	EDP	$10^3 - 10^4$
	TMAH	10^4

A material with a higher selectivity ratio is a better masking choice. In practice however, other parameters have to be borne in mind in making a suitable choice of anisotropic etchant. Total etching time, etchant agitation, etchant temperature are some important parameters to be considered while making this choice.

In general, it is common to use SiO_2 masking layer and KOH as etchant for obtaining trenches of modest depth. This is because SiO_2 though inexpensive as a process step, is prone to etching if long-duration etching is involved. For higher depths, silicon nitride is the preferred choice as the protective mask.

Use of (110) oriented wafers produce a very interesting outcome. The (111) planes are perpendicular to the (110) oriented wafers. If windows in the protective mask are aligned along (110) direction, it is possible to obtain deep trenches with vertical walls. A schematic of such vertical wall trenches is presented in Fig. 12.7.

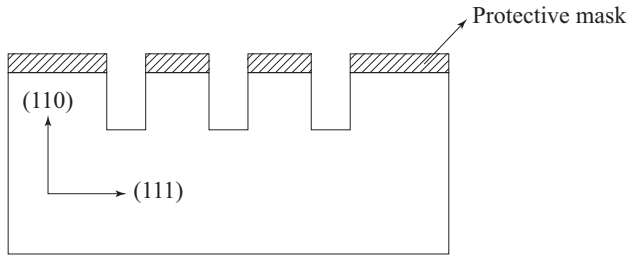


Fig. 12.7 Schematic representation of vertical wall trenches

There is however one little difficulty with the use of (110) oriented windows. The four vertical (111) planes on the (110) surface are not oriented at 90° with respect to each other. A plan view of the angle formed between the four sides is shown schematically in Fig. 12.8.

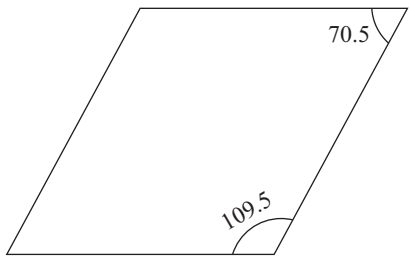


Fig. 12.8 Plan view of etched windows in (110) surface

Thus it is not possible to obtain rectangular trenches with vertical walls using (110) oriented wafers. In some applications, the need for vertical wall trenches far outstrips the non-availability of rectangular shape.

12.2.2 Dry etching

Patterns on silicon can also be obtained using gaseous etchants without exposing the material to wet chemicals. There are several dry etching techniques available. Some important techniques are plasma etching, ion milling, and reactive ion etching (RIE). A newer technique called deep reactive ion etching (DRIE) has found interesting applications.

Plasma etching

Plasma consists of an ionized gas that is neutral and carries within it a large number of free electrons along with positively charged ions. One of the common methods of producing a plasma is to use a radio-frequency (RF) source. Figure 12.9 is a schematic representation of the plasma etching process.

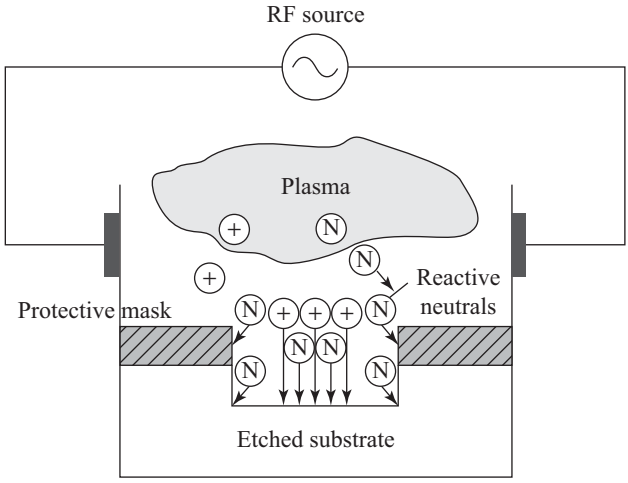


Fig. 12.9 Plasma etching process

The RF source provides the energy for generating the plasma. A chemically reactive gas like CCl_2F_2 contained in an inert carrier gas like argon is introduced into the chamber containing the plasma. The presence of the reactive gas results in the production of reactive neutrals due to the ionization within the plasma. These reactive neutrals move towards the target substrate. Reactive neutrals bombard the target substrate in sideways as well as normal directions. The charged ions however bombard the substrate only in the normal direction. The substrate is covered with a patterned protective mask and only the regions requiring etching is exposed. The high-energy ions bombard the exposed substrate surface. Simultaneously, chemical reactions are also initiated between the reactive neutrals and the substrate material. A local evaporation of the substrate material takes place leading to a selective removal of the material. Etching takes place much faster in the normal direction than the sidewalls of the substrate. This is because, only neutrals are involved in the etching process along the sidewalls whereas the normal surface is bombarded by neutrals and charged ions.

Typical plasma etching systems have etch rates in the order of 2000 Å/min. The plasma etching process is generally performed in high vacuum. Table 12.3 gives a list of etchants that are suitable for some selected materials.

Table 12.3 Etchants for selected materials

Materials	Etchants
Si, SiO_2	CCl_2F_2
	CF_4
	C_2F_6
	C_3F_8
	CHF_2/CF_4
	CHF_3/O_2
	CH_2CHF_2
Si_3N_4	Cl_2F_2
	CHF_3
	CF_4/O_2
	CF_4/H_2
	CH_3CHF_2
Polysilicon	Cl_2
	BCl_3/CF_4
	BCl_3/CCl_4
	$BCl_3/CHCl_3$
	BCl_3/CHF_3
	$SiCl_4/Cl_2$
	BCl_2/Cl_2
	$HBr/Cl_2/O_2$
	Br_2/SF_6
	SF_6
	CF_4
GaAs	CCl_2F_2
	$SiCl_4/SF_6$
	$SiCl_4/HF_3$
	$SiCl_4/CF_4$

An important parameter that decides the choice of etching technique is the aspect ratio of the structure being realized using the etching sequence. The aspect ratio (A/P) is defined as the ratio of the dimensions in the depth of the structure to that on the surface. For the schematic shown in Fig. 12.10, the aspect ratio is

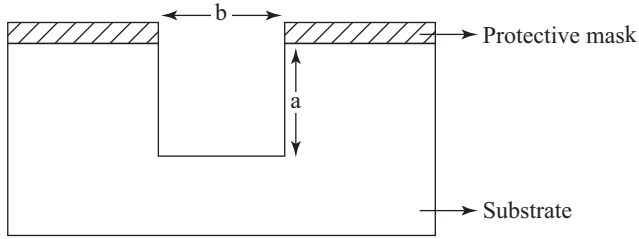


Fig. 12.10 Aspect ratio for etched structure

defined as the ratio a/b. It is possible to realize high etching rates of the order of 5 μm/min using silicon substrates and utilizing certain etchant gases in a plasma etching system. Thus a plasma etching system can in principle etch silicon at a much faster rate than wet etching systems (wet etching systems have typical etch rates of 1 – 1.5 μm/min). Both wet etching and plasma etching systems are however capable of producing only shallow trenches.

Plasma etching technique is generally used for A/P less than 15. Another disadvantage with plasma etching is the damage produced in the etched surface along with the associated contamination.

Deep reactive ion etching (DRIE)

The etch rates achievable in plasma etching is sometimes not the restrictive aspect of the plasma etching technique. Although the etch rate in the normal direction is higher than the sidewalls, a non-uniform etching of the sidewalls is unavoidable using the plasma etching route. Generally, the etched walls are at an angle with respect to the depth of the trench, as shown in Fig. 12.11. This is not always desirable for many device structures that require perfectly vertical walls.

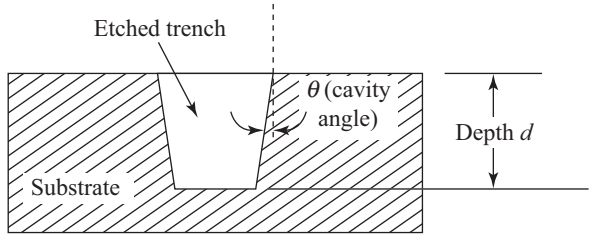


Fig. 12.11 Etch geometry using plasma etching process

Some applications require vertical wall trenches. The plasma etching process is thereby quite unsuitable for use in such applications. Systems requiring tapered walls with low aspect ratio can still use the plasma etching process.

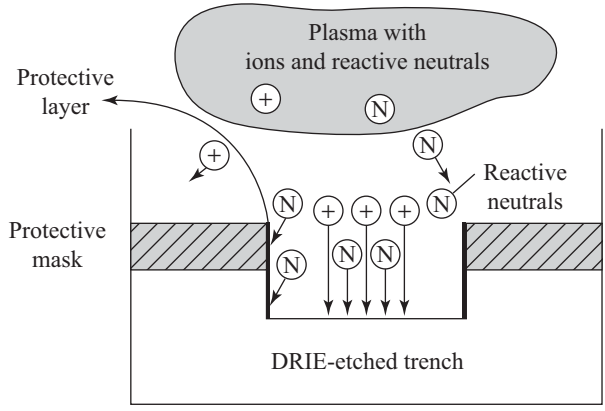


Fig. 12.12 Schematic of DRIE process

The etch angle problem can be overcome with the use of deep reactive ion etching (DRIE) technique. It is in fact possible with this technique to achieve structures with $\theta \approx$ o. The main difference between the DRIE process and the plasma etching process is the continuous formation of a protective film of a few micrometres on the sidewalls as the etching process proceeds. A high-density plasma source is used to carry out the plasma (ion) etching of the substrate material and the deposition of a protective material on the sidewalls. A schematic representation of the DRIE process is shown in Fig. 12.12.

Materials with high selectivity ratio are generally used as protective material on the sidewalls. Some commonly used protective material with their selectivity ratio and the corresponding aspect ratios are listed in Table 12.4.

Table 12.4 Characteristics of sidewall protective materials

Protective materials	Selectivity ratio	Aspect ratio A/P
Silicon dioxide	120:1	200:1
Polymer		30:1
Photoresists	50:1	100:1

Polymerization carried out during the plasma etching process is used to obtain the polymeric protective materials on the sidewalls. DRIE systems have been tried and tested using a range of reactant gases. One common reactant gas is fluoropolymer (nCF_2) added in a plasma of argon gas ions. As the etching takes place, a protective layer of a polymer gets formed on the sidewall, preventing significant etching. Etch rates in the normal direction upto 2–3 μm/min with excellent vertical trench walls are possible to be etched using this process. Etched depths in the region of 300–400 μm have been reported with modern DRIE machines. Aspect ratios over 100 with $\theta = \pm 2°$ have been incorporated in silicon structures, using DRIE technique.

12.3 Passive Components

All functional electronic circuits require passive components, such as resistors, capacitors, and inductors, for effective operation. Integrated circuit technology has methodologies to realize these components.

12.3.1 Resistors

Integrated circuit resistors are fabricated by carrying out ion-implantation or diffusion of suitable dopants through windows opened in the silicon dioxide layer, using standard photolithographic techniques. The dopant is of opposite type with respect to the conductivity type of the substrate. Two types of resistors are generally used, namely bar-shaped resistors and meander-shaped resistors.

A bar-shaped resistor is shown in Fig. 12.13. Figure 12.13(a) shows the top view and Fig. 12.13(b) is the cross-sectional view across the A_1–A_1 line. As can

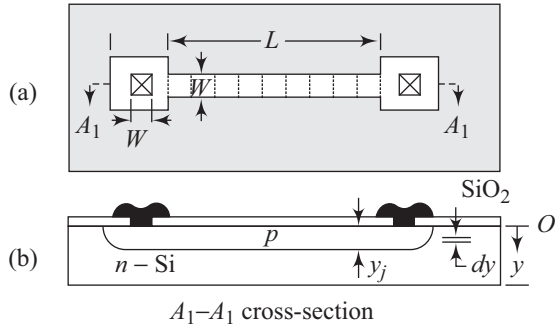


Fig. 12.13 Bar-shaped resistor (a) top view (b) cross-sectional view

be seen, the resistor has been realized by diffusing (ion-implantation can also be used) *p*-type dopant into an *n*-Si starting substrate. A junction thus gets formed at $y = y_j$. Suitable contacts are also provided to the resistor.

Let us consider a thin layer of *p*-type material of thickness *dy* at a depth *y*. The differential conductance, *dG*, of this element can be written as

$$dG = e\mu_p p(y)\frac{W}{L}dy \tag{12.1}$$

where *W* represents the width of the bar and *L*, the length of the bar. Equation (12.1) neglects the contribution due to the contact areas. The total conductance, *G*, of the implanted *p*-region can be obtained by integrating Eqn (12.1) within suitable limits

$$G = \int_0^{y_j} dG \tag{12.2}$$

Using Eqn (12.1) in Eqn (12.2) leads to

$$G = \frac{eW}{L}\int_0^{y_j} \mu_p p(y)\, dy \tag{12.3}$$

We can rewrite Eqn (12.3) in the form

$$G = G'\frac{W}{L} \tag{12.4}$$

where,

$$G' = e\int_0^{y_j} \mu_p p(y)\, dy \tag{12.5}$$

The quantity G' is the conductance of a square resistor pattern, i.e., conductance of a bar with $W = L$.

The resistance *R* of the bar is given by

$$R = \frac{1}{G} \tag{12.6}$$

which on using Eqn (12.4) yields

$$R = \frac{L}{W}\left[\frac{1}{G'}\right] \tag{12.7}$$

The quantity $\left(\frac{1}{G'}\right)$ is called the sheet resistance and a symbol $R_{\square}$ is used for it. $R_{\square}$ has units of ohms but is usually indicated in units of ohms per square (Ω/□).

Figure 12.14 shows a meander-shaped resistor. Once again, Fig. 12.14(a) shows the top view and Fig. 12.14(b) shows the cross-sectional side view along the line $A_2 - A_2$. Suitable contacts are provided to ensure electrical accessibility. The location of the contacts depends upon the effective length of the meander-shaped resistor.

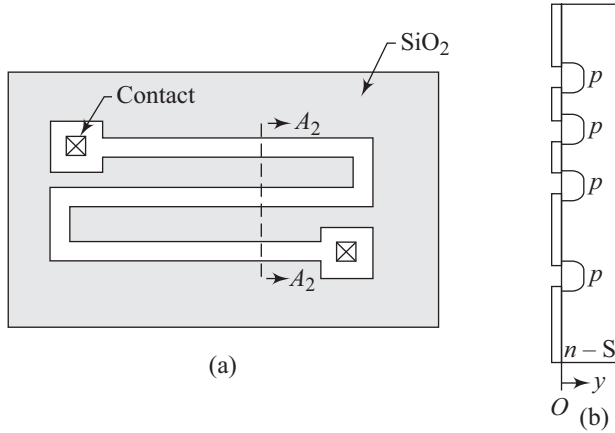


Fig. 12.14 Meander-shaped resistor (a) top view and (b) cross-sectional side

In an actual integrated circuit different resistance values can be simultaneously realized by defining windows in silicon dioxide using suitable masks. Since all the resistors use the same substrate and the same processing cycle, the sheet resistance $R_{\square}$ is expected to be the same for all the resistors in the same batch. The resistance of any particular resistor is then determined by the ratio L/W. This ratio is, in turn, decided by the mask pattern dimensions. In practice, the end contact areas introduce additional resistances. The exact magnitude of this additional resistance is dependent upon the dimensions of the contact regions. The meander-shaped resistors have another issue pertaining to the exact magnitude of resistance offered. At the bends, the electric-field lines are not uniformly spaced across the width of the resistor. More precisely, these lines are crowded towards the inside corner. This effect results in a square at the bend offering a different resistance as compared to the resistance offered by similarly sized squares in other locations along the meander-shaped resistor. Generally, the resistance offered by a square at the bend is lower than the resistance offered at other locations and an effective value of 0.65 – 0.7 times the value offered by a square at other locations, is a good approximation for a square at the bend.

12.3.2 Capacitors

Two basic types of capacitors are generally used in integrated circuits. These two types are based on MOS capacitors and *p-n* junctions. Capacitors using the MOS (metal–oxide–semiconductor) structure are fabricated by using a heavily doped semiconductor region as one plate. This heavily doped region can be realized along with the emitter region of the transistors used in the integrated circuit. A top metal electrode serves as the second electrode and the dielectric function is served by the intervening oxide layer. Figure 12.15 shows a schematic diagram of an MOS capacitor layout for a typical integrated circuit. Figure 12.15(a) shows the top layout and Fig. 12.15(b) depicts the cross-sectional side view along the line A_1-A_1. The heavily doped p^+ region serves as the bottom electrode and silicon dioxide is the dielectric material used for fabricating the capacitor.

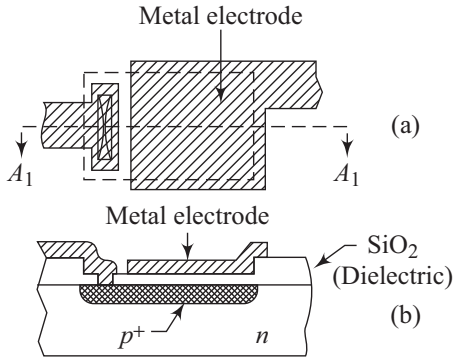


Fig. 12.15 MOS capacitor layout (a) top view (b) cross-sectional side view

A silicon dioxide layer is generally first obtained using thermal oxidation technique. The processing then involves opening of suitable windows in the oxide layer. Diffusion or ion implantation is then carried out through these windows to define the heavily doped semiconductor region. The thick thermal oxide serves as a mask during this doping stage. A thin silicon dioxide layer is then grown thermally and is the active dielectric material. The final processing involves the provision of suitable metallization contacts. Multi-layer metallization schemes are very popular at this stage of processing for a MOS-type capacitor, the capacitance per unit area, C', is given by

$$C' = \frac{\varepsilon_{ox}}{t_{ox}} \tag{12.8}$$

where t_{ox} represents the oxide thickness and ε_{ox} is the dielectric permittivity of the oxide. For silicon dioxide, $\varepsilon_{ox}/\varepsilon_0 = 3.9$, where ε_0 represents the permittivity of free space.

For certain applications, the dielectric constant of silicon dioxide may not be sufficiently high to realize the high capacitance values needed. Such applications, then, need special dielectrics like Si_3N_4 and Ta_2O_5 with dielectric constants of 8 and 22, respectively. In fact, the search for high dielectric constant materials that are compatible with silicon integrated circuit technology, is an ever-increasing

field with work being pursued using modern process tools. As we have seen in earlier chapters, the capacitance offered by the MOS structure, as given by Eqn (12.8) is independent of applied voltage since the lower plate of the capacitor is formed using a heavily doped semiconductor material. The heavily doped region also helps in reducing the series resistance associated with the capacitor.

Another form of capacitor used in the integrated circuits is the *p-n* junction capacitor. A schematic diagram of a *p-n* junction based capacitor is shown in Fig. 12.16. Figure 12.16(a) shows the top view and 12.16(b) shows the side view along the line $A_2 - A_2$.

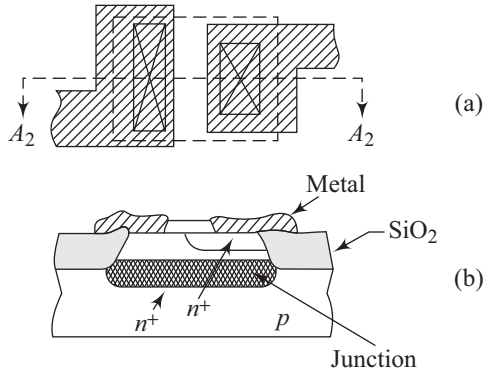


Fig. 12.16 *p-n* junction based capacitor (a) top view and (b) side view

Unlike the MOS capacitor, the *p-n* junction capacitor (n^+-*p* junction in Fig. 12.16) has a voltage dependent capacitance. This is because a reverse biased *p-n* junction has a capacitance that has a voltage dependence given by

$$C \propto (V_R + V_{bi})^{-1/2} \tag{12.9}$$

where V_R is the applied reverse bias voltage and V_{bi} is the built-in voltage. In comparison to the MOS capacitor, the *p-n* junction capacitor has a higher series resistance due to the high resistivity of the *p*-semiconductor region.

12.3.3 Inductors

Inductors are very difficult to realize in integrated circuits. Due to this, they were not a part of integrated circuits for a long time. This was coupled with the fact that there was no great need for incorporating inductors into integrated circuit based systems. The scenario has, however, changed a lot. Portable communication devices like cell phones use RF analog integrated circuits extensively. These circuits and systems use inductors. Inductors with satisfactory *Q* factors can be realized using thin metal films that are spirally wound. Modern processing techniques using photolithography and etching routes are capable of generating these spiral patterns. Furthermore the process sequence is fully compatible with IC processing. Another popular approach is to use *hybrid* IC technology. A hybrid circuit contains one or more monolithic (entire integrated circuit in a single semiconductor chip) circuits along with other circuit elements with appropriate interconnections.

12.4 Bipolar Technology

We have discussed bipolar transistors in Chapter 8. For IC applications the bipolar tansistors must be drastically reduced in size to meet the high packing density requirement of integrated circuits. The growth of integrated circuits started with small-scale integratioin (SSI), where up to 100 components were present in a chip. This was followed by medium-scale integration (MSI), where up to 1000 components were put in the same chip. Large-scale integration (LSI), which came later, had up to 100,000 components in a single chip. Even larger number of components are packed in a single chip, in very-large-scale integration (VLSI). To give an example, a typical 32-bit microprocessor chip has more than 150,000 components. If we plot the log of the number of transistors used in MOS microprocessor IC chips as a function of time, we get a straight line over three decades. This indicates that the complexity of chips has seen an exponential growth. The components have thus roughly doubled every 18 months. This fact was observed first by Gordon Moore of Intel Corporation. This periodic doubling is referred to as *Moore's law*. One of the major differences between a discreet transistor and a bipolar transistor in an IC is the location of the contact electrodes. All the contacts in an IC are located on the top surface unlike the discreet transistor, where it may not necessarily be so. Another important issue which must be ensured is that all the transistors need to be electrically isolated from each other. This prevents any interaction between adjoining devices. Various isolation methodologies are used for this purpose. Junction isolation (shown in Fig. 12.17) is a useful method.

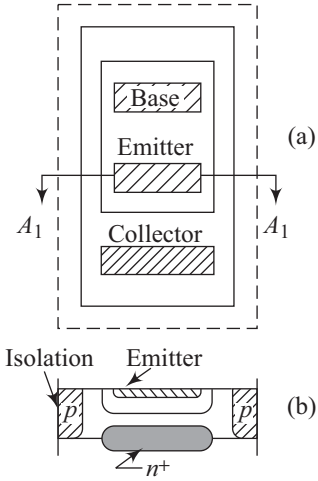


Fig. 12.17 Schematic representation of junction isolation

Figure 12.17(a) shows the top view of a bipolar transistor and 12.17(b) is a side view of the cross-section along the line $A_1 - A_1$. Lateral as well as vertical isolations can be provided using this technique. The lateral *p*-isolation region is provided with a reverse bias with respect to the *n*-type collector region. The junction isolation technique was a popular choice till around 1970. Since then, thermal oxide has been a preferred choice for lateral isolation. The basic process of oxide isolation is shown in Fig. 12.18.

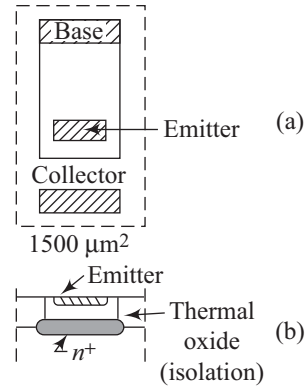


Fig. 12.18 Basic process of oxide isolation

Oxide isolation can lead to substantial reduction in device size. Further size reduction can be achieved if the emitter is extended to the walls of the oxide as shown in Fig. 12.19. The present-day technology uses vertical and lateral dimensions in the micron region.

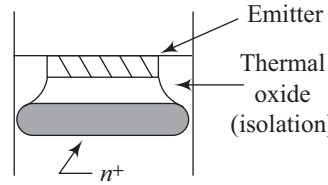


Fig. 12.19 Size reduction using extended oxide isolation

12.4.1 Basic Process

Bipolar ICs generally use *n-p-n* type transistors. This is because in *n-p-n* transistors the minority carriers in the base region are electrons, whereas for *p-n-p* devices they are holes. Since electrons have a higher mobility as compared to holes, ICs employing *n-p-n* transistors are inherently faster. Figure 12.20 shows a schematic representation of a typical *n-p-n* bipolar transistor that uses an n^+-*p* junction for vertical isolation and oxide for lateral isolation. As discussed earlier, the lateral oxide based isolation leads to substantial size reduction. Another advantage of oxide isolation is a reduction of parasitic capacitance due to a lower dielectric constant (3.9) value of the oxide as compared to the dielectric constant of silicon (11.7).

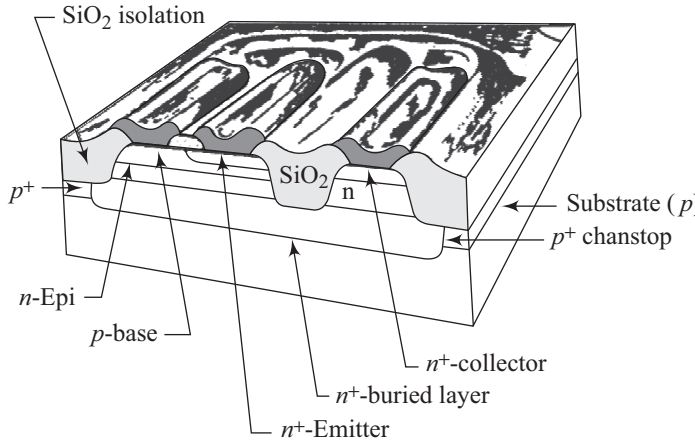


Fig. 12.20 Typical n-p-n bipolar transistor with junction and oxide isolation

The complete process sequence that can be used to realize bipolar transistors with oxide isolation is shown in Fig. 12.21. The starting material is a ⟨111⟩ or ⟨100⟩ lightly doped *p*-type polished silicon wafer. The major processing steps are described below.

Step 1 Buried layer formation

A thick oxide (0.5 μm to 1 μm) is thermally grown on the substrate. A suitable window is then opened in the oxide. Through this window a low-energy ion

implantation of arsenic ions is carried out. This ion implanted region serves as a predeposit. A high-temperature (~1100°C) drive-in step results in the n^+ buried layer. This step is shown in Fig. 12.21(a). Typical sheet resistance values in the buried layer fall in the range 10–30 Ω/□.

Step 2 Epitaxial layer deposition

The second process step is shown in Fig. 12.21(b). Here the oxide layer is first removed and then an epitaxial layer is grown in a suitable epitaxial reactor. The doping concentration and thickness of the epitaxial layer are decided by the specific application. In general, analog circuits need thicker layers with lower doping concentrations. On the other hand, digital circuits require thinner layers with higher doping concentrations. This is especially true for digital circuits using low voltages for switching. From Fig. 12.21(b) it is also clear that one problem that needs to be taken care of is the out-diffusion from the buried layer into the epitaxial layer. This out-diffusion can be minimized by using a low-temperature epitaxial process. This is also the reason for using low-diffusivity impurities like arsenic to obtain the buried layer.

Step 3 Lateral oxide isolation

The processing sequence involves first the growth of a thin thermal oxide pad over the epitaxial layer. The thickness of this pad oxide is in the region of 500 Å. The pad oxide is followed by around 1000 Å of silicon nitride deposition. The silicon nitride is not deposited directly on silicon to avoid damaging the silicon surface during further high-temperature process steps. Photolithography is then used to etch the nitride-oxide layers and around half the depth of the epitaxial layer. Boron is then implanted into the exposed areas of silicon. These process sequences are shown in Figs 12.21(c) and (d). After removing the photoresist, the wafer is placed in an oxidation furnace. The surface of the silicon having nitride layer has a very low oxidation rate, whereas much thicker oxides grow in the unprotected areas. As shown in Fig. 12.21(e), the isolation oxide thickness is such that the top surface of the oxide is coplanar with the original silicon surface. This helps in minimizing the surface topography. This stage is also sometimes referred to as planarization of the processed wafer. Figure 12.21(e) shows this stage after the removal of the nitride layer. Most of the implanted boron ions obtained in the earlier stage, are pushed below the isolation oxide, due to segregation effects. The p^+-layer thus formed is called p^+-channel stop or simply *chanstop*. The high-concentration *p*-type layer prevents the formation of surface inversion which, in turn, helps in eliminating high-conductivity paths between neighbouring buried layers.

Step 4 Base region formation

One half of the wafer is now covered with photoresist. This is followed by a boron ion implant to form the base region as shown in Fig. 12.21(f). The thin pad

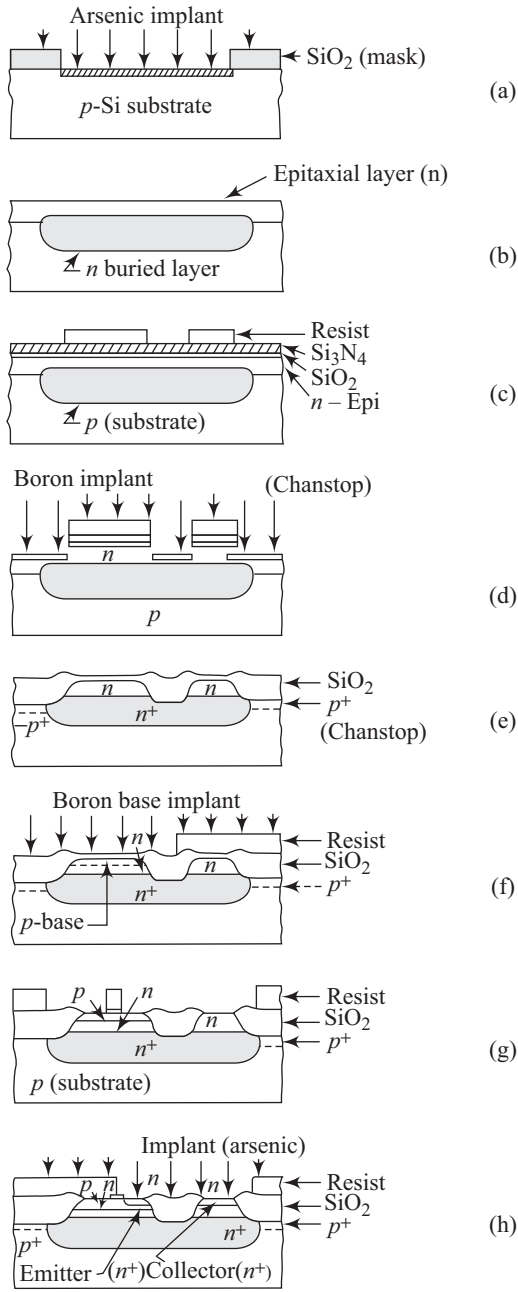


Fig. 12.21 Process sequence for realizing bipolar transistors with oxide isolation

oxide is then removed, leaving a small portion near the centre of the base region as shown in Figs 12.21(f) and (g).

Step 5 Emitter region formation

In this step, the base contact area is protected with a photoresist mask. A low-energy high-arsenic-dose implantation is then carried out to form the n^+-emitter and also the n^+-collector contact region. This stage is shown in Fig. 12.21(h). In the last processing stage, the photoresist is removed and metal contacts are provided to the base, emitter, and collector regions.

Figure 12.22 shows a typical doping profile of a completed transistor structure. From Fig. 12.22, it is clear that the emitter doping profile is quite abrupt. The collector doping is decided by the epitaxial layer doping level. At larger depths the collector region doping concentration increases as shown schematically in Fig. 12.22. This effect is due to the out-diffusion from the buried layer.

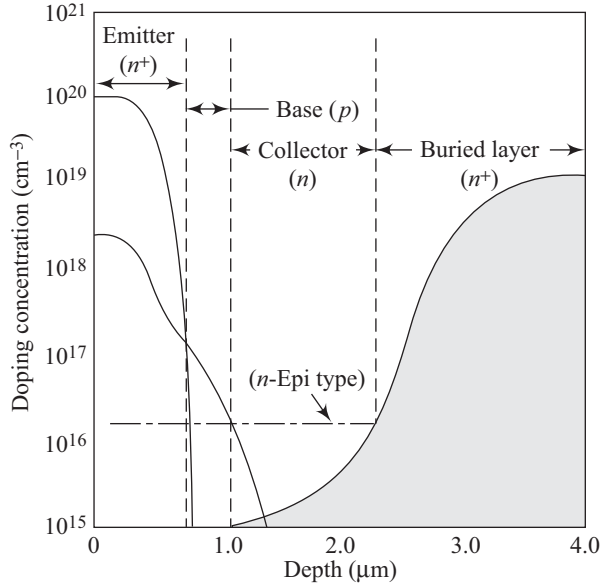


Fig. 12.22 Typical doping profile of a complete transistor

12.4.2 Dielectric Isolation

The isolation procedure described in the previous section, uses an oxide layer to isolate surrounding devices. Any single device is isolated from the common substrate using the n^+-p junction employing the buried layer. This procedure works at moderate voltages but is not effective for high-voltage applications. For high-voltage applications, a different approach is required. In this approach, insulating tubs are used to isolate a number of pockets of single-crystal semiconductor regions. This approach is called *dielectric isolation*. Any particular device is isolated from the neighbouring devices and the common substrate using a dielectric layer. A typical process sequence that can be employed for obtaining dielectric isolation is shown in Fig. 12.23. The scheme uses ⟨100⟩ oriented silicon wafer for

exploiting its crystallographic properties. A thermal oxide is first grown on the substrate. This is followed by the use of standard photolithographic techniques to open windows in the grown oxide.

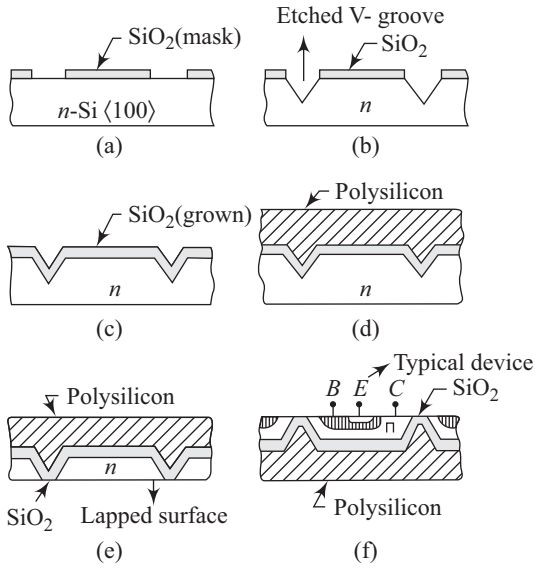


Fig. 12.23 Process sequence for realizing dielectric isolation

This stage with windows in grown oxide is shown in Fig. 12.23(a). A suitable anisotropic etchant (such as KOH and EDP) is then used to obtain V-shaped grooves as shown in Fig. 12.23(b). The masking oxide is then etched away and a thermal oxide, which is around 1 μm thick, is grown on the wafer surface. This stage is depicted in Fig. 12.23(c). Polycrystalline silicon is then grown over the oxide. The thickness of this polycrystalline layer is around 200–300 μm. This stage of the sample is shown in Fig. 12.23(d). The next stage of processing is the precision lapping of the exposed silicon surface. The lapping is carried out till the underlying oxide is revealed. The sample at this stage (shown in Fig. 12.23(e)) consists of tubs of single-crystal silicon that are separated from each other by the dielectric layer of silicon dioxide. The single crystal silicon surface can now be used to create different devices and components, with the dielectric providing the necessary isolation. The inverted sample, along with a typical bipolar transistor, is shown in Fig. 12.23(f). The precise lapping of the substrate material is one of the disadvantages of the technique and limits its use to either high-voltage devices or to develop devices that are insensitive to high-energy radiation. Any extra charge carriers created by high-energy radiations do not participate in the device operation due to the underlying dielectric oxide layer.

- A bipolar inverter circuit is one of the basic elements of a digital integrated circuit.
- A typical 1 megabit dynamic random access memory (DRAM) chip contains more than 2,000,000 components.

12.5 MOSFET Technology

MOSFET technology plays a dominant role in VLSI circuit development. This technology owes its popularity to its adaptability to size reduction which is the fundamental engine of the IC industry. There are two variants of MOSFET technology, namely NMOS (*n*-channel MOSFET) and CMOS (complementary MOSFET). In CMOS, the *n*-channel and *p*-channel MOSFETs are present on the same chip.

12.5.1 NMOS Process

The basic processing sequence for NMOS technology is shown in Fig. 12.24. The important steps of the NMOS processing technique are:

Step 1 Isolation oxide

The starting substrate is a lightly doped, *p*-type, ⟨100⟩ oriented silicon wafer. This orientation is preferred due to its lower interface trap density as compared to ⟨111⟩ oriented silicon (around one tenth). A thin oxide pad of ~ 500 Å is thermally grown on the surface of the wafer. This is followed by a silicon nitride deposition. The silicon nitride layer is typically 1000 Å. Figure 12.24(a) shows a schematic diagram of the wafer at this stage. Standard photolithography is then used to define the active area of the device. A boron chanstop is obtained by ion-implantation as shown in Fig. 12.24(b). Photoresist mask is then used to etch the silicon nitride layer, following which the masking resist layer is removed. Thermal oxidation is then carried out and an oxide layer of around 0.5 to 1 μm is grown. As shown in Fig. 12.24(c), the thermal oxide grows only in areas not covered with silicon nitride. The oxide layer is called the *field oxide*. The boron implantation carried out in the earlier stage is driven in, in the same furnace.

Step 2 Gate oxide

The silicon nitride–oxide composite layer over the active area of the device is now removed. A thin gate oxide layer is grown as shown in Fig. 12.24(c). Boron ions are implanted in the channel region, to realize the enhancement mode *n*-channel device. This increases the threshold voltage to a pre-set value. For realizing the depletion-mode *n*-channel device, arsenic ions are implanted into the channel region. This reduces the threshold voltage.

Step 3 Gate formation

A polysilicon layer suitably doped with phophorus (diffusion or implantation can be used for doping) is used as a gate material. For ICs using gate lengths greater than around 3 μm, such a gate material (sheet resistance ~ 20 Ω/□) is found to be suitable. Smaller gate lengths, however, require refractory metals (Mo, etc.) or polycides (composite layers of silicide and polysilicon). These materials have typical sheet resistances in the region of a few Ω/□. Figure 12.24(d) shows the sample with a patterned gate.

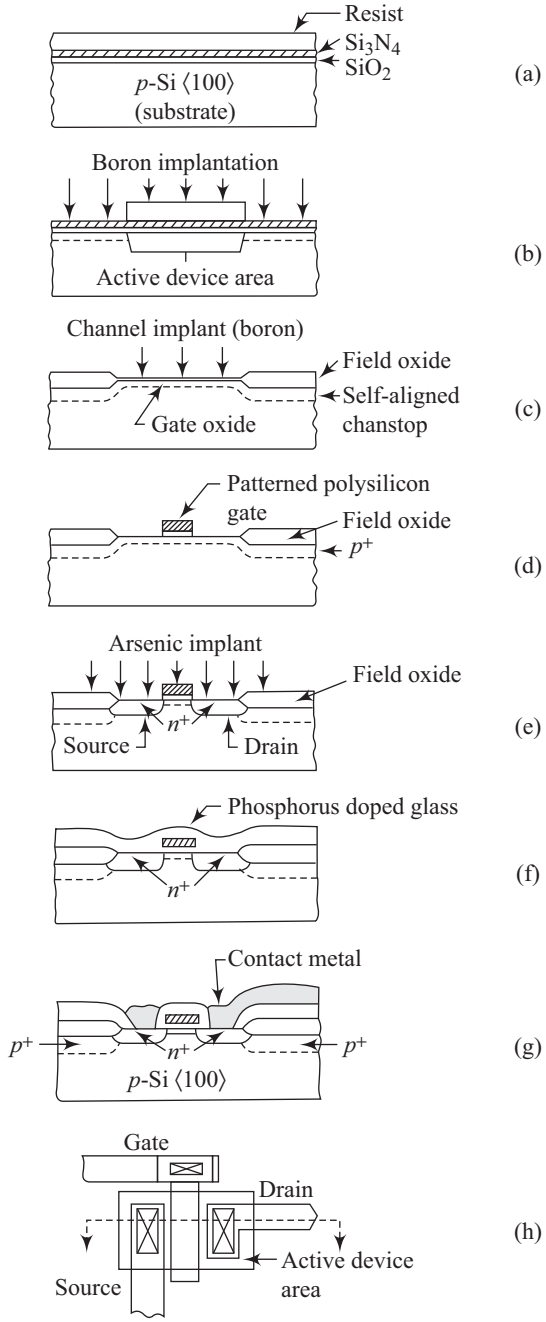


Fig. 12.24 Process sequence of NMOS technology

Step 4 Source and drain formation

Arsenic implantation is carried out to form the source and the drain regions as shown in Fig. 12.24(e). The field oxide and the patterned gate serve as mask

during this stage. The subsequent steps use low-temperature processes to minimize lateral diffusion. This helps in reducing gate–drain and gate–source coupling capacitances.

Step 5 Metallization

A phophorus-doped oxide is now deposited over the entire surface. The doped oxide is made to flow (to ensure smooth topography) by heating the wafer to adequate temperature [Fig. 12.24(f)]. Conventional photolithography is then used to define contact windows in the doped glass. A suitable metal layer (e.g. aluminium) is then deposited over the surface. The metal layer is patterned using conventional photolithography as shown in Fig. 12.24(g). A top view of the realized device is shown in Fig. 12.24(h).

12.5.2 NMOS Memory Devices

The NMOS technology described above can be used to realize devices that can store digital information in terms of *bits* (*bi*nary digi*ts*). Large memory requirements are met by random access memory (RAM) organization. A RAM consists of memory cells that are organized in a matrix structure. Any member of this matrix can be accessed in a random manner to store (write) or retrieve (read) data. Two important types of RAM are SRAM and DRAM. A static random access memory (SRAM) has the capability to retain stored data indefinitely. A basic SRAM cell consists of four enhancement-mode MOSFETs and two depletion-mode MOSFETs. In some designs, the depletion-mode MOSFETs are replaced by undoped polysilicon resistors to reduce power consumption. Dynamic random access memory (DRAM) organization is very useful in further reducing area and power consumption. A typical one-transistor DRAM cell consists of a transistor that serves as a switch and uses a storage capacitor to store one bit of information. The stored charge, however, tends to leak away due to leakage currents. This necessitates period *refreshing* of the stored charge. DRAMs have a very high component density per chip.

12.5.3 Charge-coupled Devices

A basic charge-coupled device consists of a closely spaced array of MOS diodes. The MOS diodes are fabricated using a continuous oxide layer over a semiconductor substrate. Figure 12.25(a) shows a set of three MOS structures on a *p*-type substrate. Since positive biases have been applied to all the three electrodes, the MOS devices are in the surface depletion state. As shown in Fig. 12.25(a), the central electrode has a higher positive bias. This means that the centre MOS structure is under greater depletion than the adjoining MOS structures. A potential well is formed under the central electrode as shown in Fig. 12.25(a).

Any minority carrier electrons get collected in the potential well. Suppose we use a pulsing mechanism so that the potential across the third electrode exceeds beyond the potential on the central electrode as shown in Fig. 12.25(b). A potential

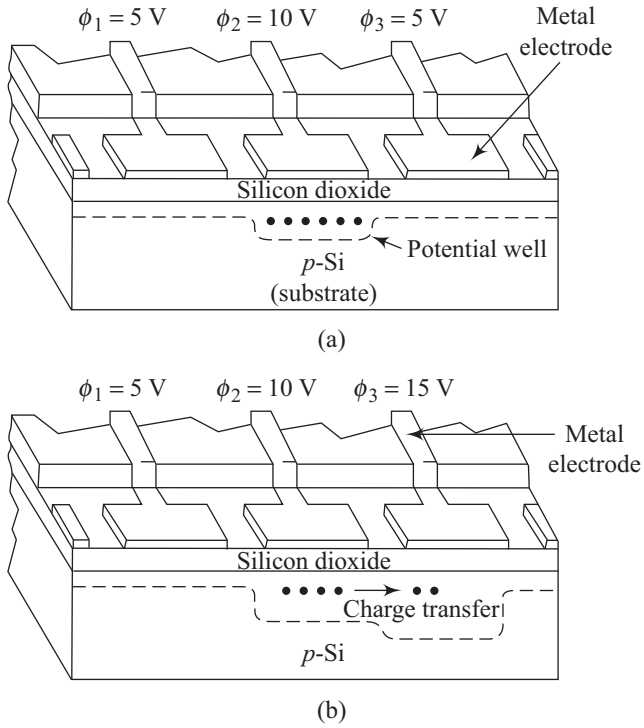


Fig. 12.25 Charge transfer in a basic charge-coupled device

lowering is created under the third electrode and the minority carrier electrons get transferred from the central electrode to the third electrode. Such a process can be continued and the minority carriers are transferred along a linear array of MOS structures. The basic CCD described above has a wide range of applications in diverse areas like image sensing and signal processing.

12.5.4 CMOS Technology

CMOS technology involves processing techniques that can realize PMOS and NMOS devices on the same chip. Figure 12.26 shows the layout of a typical CMOS inverter.

The structure consists of an upper PMOS device and a lower NMOS device. The gate of the upper device is connected to the gate of the lower device. The input voltage, V_i, is fed to this common terminal. Both the MOSFET devices are enhancement-mode devices. The PMOS device used has a threshold voltage, V_{TP}, less than zero, whereas the threshold voltage, V_{TN}, for the NMOS device is greater than zero. For input

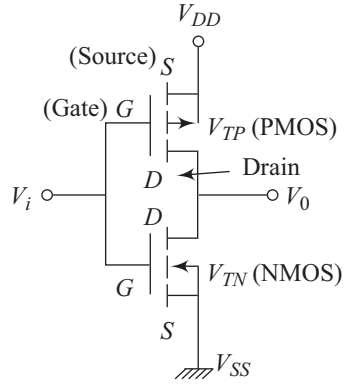


Fig. 12.26 Layout of a CMOS inverter

voltage, V_i, at ground potential or at small positive values, the PMOS device is in the *on* state and the NMOS device is *off*. The output voltage, V_o, is approximately equal to V_{DD}. Thus an input logic state 0 gets inverted to output logic state 1. Thus, in either logic state, one device is non-conducting as one goes from the supply voltage, V_{DD}, to ground in the series path. A small leakage current flows in the steady state. The average power dissipation is small when CMOS technology is used to design an inverter. The low power consumption is extremely crucial as the component density increases on a chip.

A CMOS structure can be realized in a variety of ways. In one approach, a *p*-tub or *p*-well is initially implanted and then driven into an *n*-substrate. An *n*-channel MOSFET is fabricated in this *p*-tub using the technology described in Section 12.3.1. The *p*-channel MOSFET is formed by implanting suitable ions such as B^+ into the *n*-type substrate. This implantation is used to create the source and drain regions. The structure is shown in Fig. 12.27, where the n^+-chanstop is also shown underneath the field oxide for the *p*-channel device. From the processing sequence it is clear that the fabrication of CMOS circuits requires around twice the process steps as compared to NMOS-based circuits. The relatively complex processing sequence, however, results in a substantially lower power consumption.

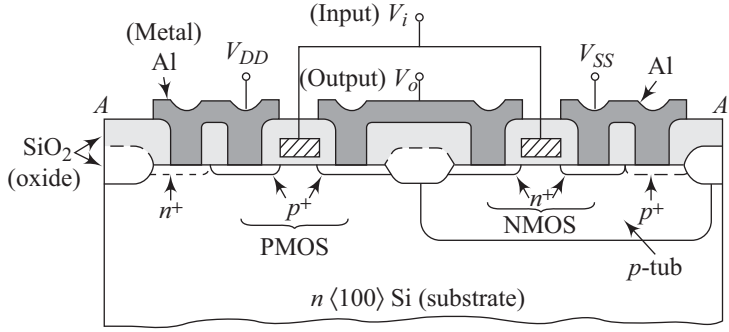


Fig. 12.27 CMOS technology based on *p*-tub

Another alternative technique is to use an *n*-type tub fabricated within a *p*-type substrate as shown in Fig. 12.28.

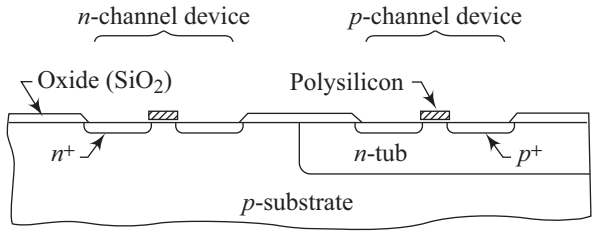


Fig. 12.28 CMOS technology based on *n*-tub

To fabricate this *n*-type tub, the dopant concentration, N_D, must be suitably higher than the acceptor concentration, N_A, in the substrate. A combination of the *p*-tub and *n*-tub approaches leads to the twin-tub approach that is shown schematically in Fig. 12.29.

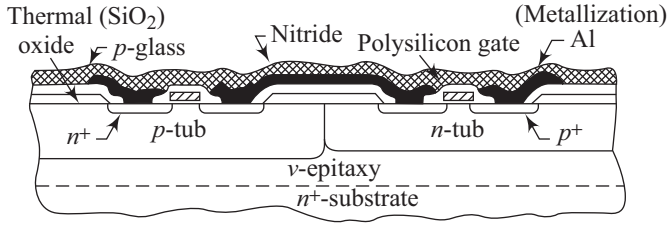


Fig. 12.29 Twin-tub approach for CMOS

In Fig. 12.29, an *n*-*p*-*n* transistor is formed with the n^+-source or drain acting as the emitter, the *p*-type tub as the base, and the *n*-type tub as the collector. Simultaneously, a *p*-*n*-*p* transistor is formed with the p^+-source or drain serving as the emitter, the *n*-tub as the base, and the *p*-tub serves as the collector. These two parasitic bipolar transistors get coupled together and behave like a *thyristor*. If the product of the current gains of the individual transistors becomes greater than unity, a large current can flow between V_{DD} and V_{SS}. This phenomenon is called *latchup* and is a cause of potential damage to CMOS circuits.

One interesting method of avoiding latchup is the use of trench isolation. A trench is formed in silicon using anisotropic reactive-sputter etching as shown in Fig. 12.30. The trench formation is followed by thermal oxidation that results in the growth of an oxide layer on the bottom and side walls of the trench.

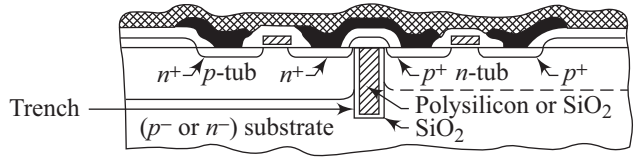


Fig. 12.30 Schematic representation of trench isolation

The trench is then refilled using deposited polysilicon or silicon dioxide. As is clear from Fig. 12.30, the refilled trench effectively isolates the *p*-channel and the *n*-channel devices.

Another approach to solve the latchup problem is the use of silicon-on-insulator or SOI structure. In this approach, a layer of silicon dioxide, around 1 μm in thickness, is first grown on a silicon substrate. This is followed by the deposition of a polycrystalline silicon layer that is recrystallized using laser annealling or similar procedures. The recrystallized silicon is then used to fabricate MOSFET devices as shown schematically in Fig. 12.31.

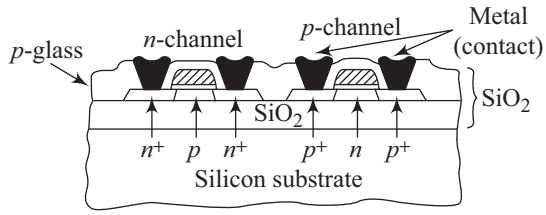


Fig. 12.31 Silicon-on-insulator technology

12.6 MESFET Technology

MESFET technology has made progress due to the advances in gallium arsenide processing techniques. Gallium arsenide offers the following major advantages in comparison with silicon:

- Electron mobility is higher in gallium arsenide. This, in turn, leads to lower series resistance.
- Gallium arsenide has higher drift velocity as compared to silicon for a given electric field. This characteristic is important in fabricating devices where speed is important.
- Gallium arsenide can be made semi-insulating. This provides a lattice-matched and dielectric-insulated substrate material.
- Monolithic integration of electronic and photonic devices on a single semiconductor substrate is possible.

These advantages however come at a price. Some of the disadvantages of gallium arsenide include the following:

- Short minority carrier lifetime
- Absence of a stable, passivating native oxide
- Crystalline defects that are many orders of magnitude greater than the vastly improved silicon material technology

The IC technology based on GaAs uses primarily the MESFET approach. The major processing steps involved are shown in Fig. 12.32.

A semi-insulating gallium arsenide substrate serves as the starting material. A suitable insulator like SiO_2 or Si_3N_4 is then deposited. Suitable photoresist pattern is then obtained and a light-dose ion implantation is carried out to form the channel and the diode regions, as shown in Fig. 12.32(a). Photolithography is then carried out to open windows to carry out n^+-implantation for forming the ohmic contact regions [Fig. 12.32(b)]. This is followed by the addition of a dielectric (SiO_2) layer before the post implantation annealling step (Fig. 12.32(c)). Figure 12.32(d) shows the step of ohmic contact metallization for the source, drain, and diode regions. The lift-off process is then used to fabricate the Schottky barrier and interconnect metallization as shown in Fig. 12.32(e). The final step consists of the deposition of another insulator and a second-level metal followed by suitable patterning to obtain the second-level interconnection [Fig. 12.32(f)].

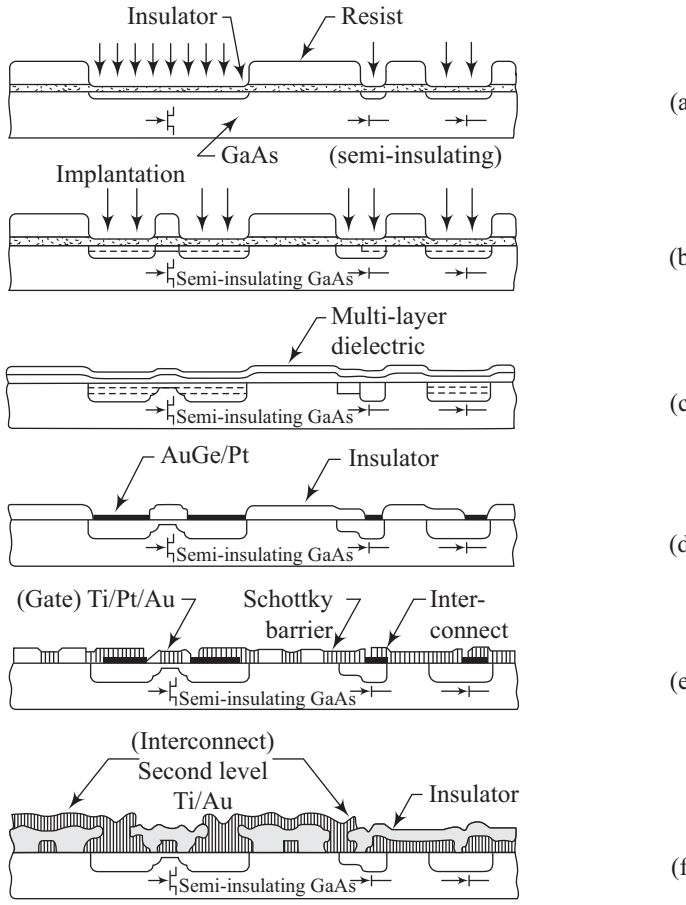


Fig. 12.32 Process steps used in the MESFET approach

12.7 Micro-electromechanical Systems

The previous section discussed the role of advances in semiconductor fabrication technology in increasing the number of electronic components on a single semiconductor substrate (also referred to as a wafer). The ICs and their increasing packing density has led to cheaper circuits due to the capability of printing a larger number of components on the same chip. A natural consequence of this integration is an effort to apply these concepts of fabrication in fields like mechanics, optics, and fluidics. This has resulted in the emergence of a new area called micro-electromechanical systems or simply MEMS. An MEMS is a batch-fabricated integrated microscale system that has the following characteristic features:

- Converts physical stimuli and parameters to electrical, mechanical, and optical signals, or vice versa
- Performs functions such as actuation, sensing, and so on.
- Comprises control, diagnostics, signal processing, and data acquisition

Any one given MEMS may not perform all the functions listed above. Though broadly speaking, MEMS utilizes the semiconductor processing technology developed for ICs, there are important differences. These differences arise due to the fact that whereas IC fabrication largely involves a two-dimensional perspective, MEMS processing must have a three-dimensional capability.

12.7.1 Basic Processes

MEMS fabrication utilizes two main micromachining processes, which are bulk and surface micromachining. Both these processes use modified CMOS technology alongwith some specific micromachining processes.

Bulk micromachining

Bulk micromachining utilizes the bulk of the semiconductor wafer to create mechanical structures. This processing of silicon was developed more than thirty years ago for the fabrication of three-dimensional structures. It uses both dry and wet etching techniques to fabricate microstructures. Suitable etch masks and etch-stop layers are used to etch predefined areas of silicon substrates. When wet etching is resorted to, both *anisotropic* and *isotropic* etching processes are used. Critical issues include crystallographic and dopant dependent etch processes. The bulk removal of silicon when combined with suitable wafer-to-wafer bonding, results in the creation of complex three-dimensional microstructures.

Anisotropic etching uses etchants, such as potassium hydroxide (KOH), sodium hydroxide (NaOH), H_2N_4, and ethylene diamine pyrocatechol (EDP), that have orientation-dependent etch rates. Some crystallographic planes like $\langle 111 \rangle$ etch very slowly and can be used to fabricate cones, pyramids, cubes, and channels. Isotropic etching is used to create hemispheres or cylindrical structures. Deep reactive ion etching (DRIE) system uses plasma to etch complicated structures in semiconductors. Figure 12.33 shows a vertical walled, combed structure that has been fabricated using the DRIE technology.

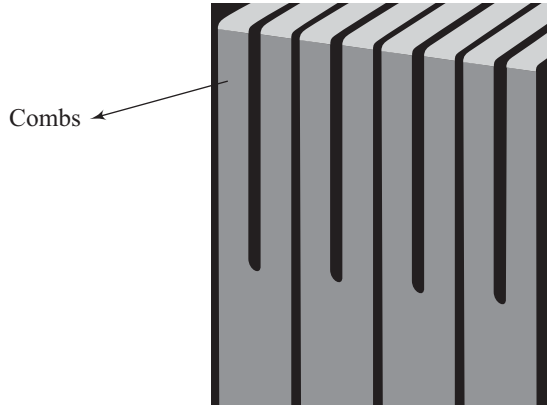


Fig. 12.33 Combed structure obtained using DRIE

The basic structure of a diaphragm-based pressure sensor using bulk micromachining process is shown in Fig. 12.34.

In recent years, a new technique called LIGA has emerged, that can be used to fabricate immensely complex three-dimensional micromechanical structures. LIGA is a German acronym for LI (Röntgen Lithographie, meaning X-ray lithography), G (Galvanik, meaning electrodeposition), and A (Abformung, meaning molding process).

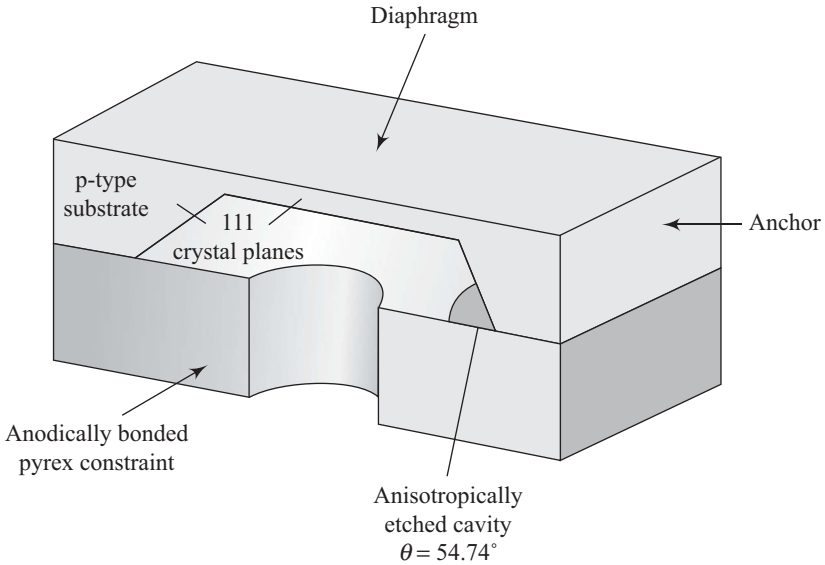


Fig. 12.34 Diaphragm based structure for pressure sensor type device

Surface micromachining

Surface micromachining utilizes a suitable thin film material in locations requiring open area or a free-standing structure. This thin film material is subsequently removed and is, therefore, called a *sacrificial layer*. Typical sacrificial layers include silicon dioxide, nitride material, photoresists, and polyimide. Surface micromachining has been used to develop microtransducers, rotational/translational microservos, accelerometers, gyroscopes, etc. Surface micromachining also offers another attractive feature. The MEMS based microstructures can be suitably integrated with microelectronics and/or optics, on the same chip, to develop a complete system. A typical process sequence based on principles of surface micromachining is shown in Fig. 12.35.

A suitable pattern of sacrificial layer is first created on the silicon substrate using standard photolithography. Two structural layers [Fig. 12.35(b)] are subsequently patterned. The microstructures are obtained using suitably designed structural layers and their numbers and choices may vary depending upon specific applications. The sacrificial layer is subsequently removed to release the microelectromechanical structure, as shown in Fig. 12.35(c). Surface machining poses three main challenges:

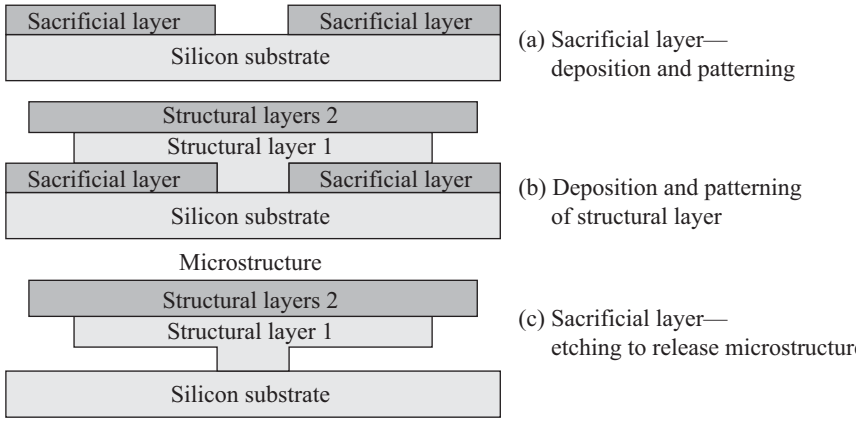


Fig. 12.35 Typical surface micromachining sequence

(i) Control and reduction of stress and stress gradient in the structural layer. This is important to avoid buckling or bending of the released microstructure.
(ii) Selectivity of sacrificial layer etchant with respect to the other process sequences and materials. For example, if silicon dioxide is used as a sacrificial layer, other suitable materials like silicon nitride need to be used as insulating materials.
(iii) Stiction of the released structure with respect to the substrate has to be avoided using suitable technology.

Silicon is the preferred material for MEMS but recently other substrates, such as GaAs, are also gaining popularity. GaAs, for example, has been used to develop MEMS devices in monolithic microwave integrated circuits (MMICs).

- Nanoelectromechanical systems are called NEMS. Molecular-scale technologies are used for manufacturing NEMS. Carbon nanotubes find a variety of applications in the area of NEMS as well as MEMS. Self- and positional-assembly concepts are the preferred routes for fabricating nanoscale structures.
- Stiction refers broadly to the process of sticking of micro-structure elements either to the substrate or to adjacent elements. Two occasions are most common where stiction can manifest. In surface micromachining, the sacrificial layer is dissolved in a solution. As the wafer with the micro-structures is pulled out of the solution, stiction can be an important issue to take care of. Sometimes the structural elements find themselves in a humid environment, making stiction sufficiently predominant. Drying the rinsed wafer with supercritical CO_2 or by freezing and then sublimating the rinsing liquid are preferred methods to circumvent stiction-related problems.

Solved Problems

12.1 A bar-shaped resistor is 100 μm long and 10 μm wide. The two end contacts contribute 0.65 square each to the effective resistance. The implanted layer has a sheet resistance of 0.9 kΩ/□. Calculate the effective net resistance of the bar resistor.

Solution

The total length can be assumed to consist of 10 squares of size 10 μm × 10 μm.
The two end contacts contribute 1.3 squares.
The total effective squares = 10 + 1.3 = 11.3 squares.
Effective resistance, R_{eff}, is then

$$R_{\text{eff}} = 11.3 \times 0.9 = 10.17 \text{ k}\Omega$$

12.2 An integrated circuit capacitor is fabricated using the MOS structure. The capacitor uses silicon dioxide of thickness 0.5 μm. Calculate capacitance per unit area of the structure. Assume dielectric constant of silicon dioxide to be 3.9 and take $\varepsilon_0 = 8.85 \times 10^{-14}$ F/cm.

Solution

Capacitance, C, per unit area is given by

$$C = \frac{\varepsilon_{\text{ox}}}{t_{\text{ox}}} \tag{12.2.1}$$

Putting the given values in Eqn (12.2.1) yields

$$C = \frac{3.9 \times 8.85 \times 10^{-14}}{0.5 \times 10^{-4}}$$
$$= 69 \times 10^{-10} \text{ F/cm}^2$$

12.3 A p-type resistor pattern is laid out for an IC with two highly conducting p-type regions contacting a 1 μm wide resistive stripe that extends 4 μm between contacts. The stripe has a junction depth, X_j, of 1 μm. The desired value of the resistance is 1 kΩ. Determine the sheet resistance and the average resistivity required to meet the given specification.

Solution

The number of squares in the resistor pattern is given by the ratio

$$\frac{\text{Length}}{\text{Width}} = \frac{4}{1} = 4$$

The sheet resistance $R_{\square}$ is given as

$$R_{\square} = \frac{R}{\text{No. of squares}} = \frac{1000}{4} = 250 \, \Omega/\square$$

Now, conductance is given by

$$G = \text{e}\mu_p N_a' (W/L) \tag{12.3.1}$$

Also $$R = \frac{1}{G} \tag{12.3.2}$$

From Eqs (12.3.1) and (12.3.2) we get

$$N_a' = \frac{G}{(e\mu_p(W/L))}$$

$$= \frac{1}{(e\mu_p R(W/L))}$$

$$= \frac{1}{(e\mu_p R_{sh})} \text{ dopant atoms/cm}^2$$

Now, the average volume density of the dopant atoms, N_a, is related to the area density, N_a', through the relation

$$\overline{N_a} = \frac{N_a'}{x_j}$$

And the average resistivity in the *p*-doped diffused resistor is given as

$$\rho = \frac{1}{(e\mu_p \overline{N_a})}$$

$$= \frac{x_j}{(q\mu_p N_a')}$$

$$= R_{\square} x_j$$

$$= 250\ \Omega/\square \times 1\ \mu\text{m}$$

$$= 0.025\ \Omega\text{cm}$$

Recapitulation

- Resistance R of a bar resistor is given by

  $$R = \frac{L}{W}\left[\frac{1}{G'}\right]$$

  where the quantity $(G')^{-1}$ is called the sheet resistance.
- Precise resistance values can be realized by defining windows in silicon dioxide using suitably designed masks. The ratio L/W then decides the resistance of any particular resistor.
- For an MOS type capacitor, the capacitance per unit area, C' is given by

  $$C' = \frac{\varepsilon_{ox}}{t_{ox}}$$

- Capacitance offered by a reverse biased *p-n* junction has a voltage dependence given by

  $$C \propto (V_R + V_{bi})^{-1/2}$$

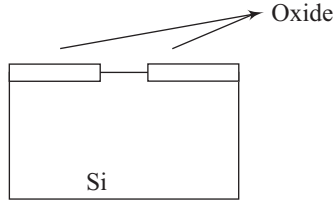
- Spiral patterns can be generated using modern photolithography techniques to realize discreet inductance values.
- SSI circuits have up to 100 components in a chip, MSI circuits up to 1000 components, and LSI circuits have up to 100,000 components in a chip.
- Junction isolation and oxide isolation are two popular techniques for the prevention of interaction between adjoining devices in bipolar IC technology.
- Dielectric isolation uses insulating tubs to isolate pockets of single-crystal semiconductor regions.
- A CMOS circuit uses *n*-channel and *p*-channel MOSFETs on the same chip.
- Small gate lengths require refractory metals or polycides as the gate material.
- A charge-coupled device consists of closely spaced array of MOS diodes. A switching of potentials leads to charge transfer between individual units.
- CMOS technology lowers the power consumption of ICs.
- Latchup takes place when the current gains of individual transistors become greater than unity in the twin-tub approach.
- Trench isolation and silicon-on-insulator approaches can prevent latchup problem.
- GaAs based ICs primarily use MESFET technology.
- MEMS processing must have a three-dimensional capability as compared to the conventional IC technology.
- DRIE technology can be used to realize complicated micro-structures.
- LIGA is a German acronym for LI (Röntgen LIthographie), G (Galvanik), and A (Abformung).
- Microstructures can be fabricated on the surface of a substrate, using the surface micromachining approach.

Exercises

Review Questions

12.1 How are resistors realized in integrated circuits?
12.2 Define sheet resistance and give its units.
12.3 Explain the use of the MOS structure in fabricating integrated circuit capacitors.
12.4 Differentiate between MSI, LSI, and VLSI.
12.5 What is the role of buried layer formation in bipolar integrated circuit technology?
12.6 Why is dielectric isolation preferred for ICs requiring insensitivity to high-energy radiation.
12.7 Explain the basic processing steps involved in typical NMOS technology.
12.8 Differentiate between isolation oxide and gate oxide.
12.9 Explain the basic principle of a charge-coupled device.
12.10 What is CMOS technology? Explain the advantages of CMOS technology using the example of a CMOS inverter.
12.11 Give a layout of the twin-tub technology.

12.12 What is the latchup problem? How can it be avoided?
12.13 Give three major advantages of MESFET technology.
12.14 Explain the salient features of a micro-electromechanical system.
12.15 What is the role of anisotropic etching in bulk micromachining?
12.16 Name some materials that can be used as sacrificial layer in surface micromachining.
12.17. Derive an expression for the resistance offered by a bar resistor. How is a meander-shaped resistor designed?
12.18. Give schematic layouts for MOS capacitor and *p-n* junction based capacitor.
12.19 Show a schematic representation of junction isolation.
12.20 What is oxide isolation?
12.21 Give a typical process sequence for realizing bipolar transistors using oxide isolation.
12.22 Explain the concept of dielectric isolation.
12.23 Indicate schematically the basic NMOS process technology.
12.24 Explain the principle of operation of a charge-coupled device.
12.25 Give a schematic diagram of a CMOS inverter.
12.26 Explain the twin-tub technology.
12.27 Mention the series of different fabrication steps that need to be implemented using a positive photoresist so as to achieve an oxide pattern on a silicon wafer similar to the one shown below.



12.28 Give the main advantages and primary disadvantages of MESFET technology.
12.29 How is anisotropic etching useful for bulk micromaching processing?
12.30 Show a typical surface micromachining in a schematic diagram.
12.31 Why is ⟨100⟩ oriented silicon preferred for fabricating BJTs and MOS circuits?
12.32 Why is oxide-based isolation preferred over junction isolation in bipolar technology?
12.33 Describe the effect of the dopant in silicon bulk micromachining.
12.34 What is meant by stiction in surface micromachining?

Problems

12.1 The length of a bar-shaped resistor is 80 μm and its width is 20 μm. The two end contacts contribute 0.65 square each to the effective resistance. The sheet resistance of the implanted layer is 0.85 kΩ/□. Calculate the effective net resistance of the bar resistor.

[*Hint:* Effective squares = 4 + 1.3 = 5.3]

Ans. 4.505 kΩ

12.2 The length of the bar-shaped resistor indicated in Problem 12.1 is increased to 160 μm. If the sheet resistance of the implanted layer is increased to 0.9 kΩ/□, calculate the change in effective resistance of the bar resistor. Assume the contribution due to the end contacts to remain the same.

Ans. $\Delta R_{\text{eff}} = 3.865\ \text{k}\Omega$

12.3 MOS structure is to be used to fabricate an integrated circuit capacitor. The insulator used is silicon dioxide of 0.4 μm thickness. Assuming dielectric constant to be 3.9 and $\varepsilon_0 = 8.85 \times 10^{-14}$ F/cm, calculate the capacitance per unit area of the structure.

$$\left[Hint: C = \frac{\varepsilon_{\text{OX}}}{t_{\text{OX}}} \right]$$

Ans. 86×10^{-10} F/cm^2.

12.4 An integrated circuit manufacturing plant uses two fablines to fabricate MOS structure based capacitors. One of the fablines uses Ta_2O_5 and the other uses Si_3N_4 as the dielectric material. Calculate the ratio of capacitance values obtained in the two fablines, assuming all other critical dimensions to remain the same. Take dielectric constant of $Ta_2O_5 = 22$ and of $Si_3N_4 = 8$.

Ans. $\dfrac{C_{\text{Ta}_2\text{O}_5}}{C_{\text{Si}_3\text{N}_4}} = 2.75$

12.5 Design a process flow for a diode given that the starting material is (100) oriented Si, *p*-type with boron doping to 1×10^{15} cm^{-3}, and the backside of the wafer is heavily doped with boron. Assume that the desired junction depth is 1 μm and the surface concentration of the dopant must be 10^{19} cm^{-3} to achieve good ohmic contact.

12.6 Consider a simple abrupt *p*-*n* junction, reverse biased from voltage $V_1 = 0$ V to $V_2 = -5$ V. The doping density of the *n*-region is $N_D = 10^{19}$ cm^{-3} and the doping density of the *p*-region is $N_A = 10^{16}$ cm^{-3}. If the junction area is 20 μm × 20 μm, find the average junction capacitance.

$$\left[Hint: C_j(V) = \frac{A \times C_{j0}}{\left(1 - \dfrac{V}{\phi_0}\right)^m} \right]$$

Ans. 69 fF

12.7 The effective resistance of a bar-shaped resistor is 5 kΩ. The total effective squares is 5.6. Calculate the sheet resistance of the implanted layer.

Ans. 0.89 kΩ/□

12.8 An integrated circuit-capacitor uses the MIS structure. Capacitance per unit area is 50×10^{-10} *F*/cm^2 and the insulator thickness is 0.4 μm. Calculate the dielectric constant of the insulator.

Ans. 2.26

13 Microwave Devices

- Gunn
- IMPATT
- TRAPATT
- BARITT

Learning Objectives

After going through this chapter the student will be able to

- understand the need for microwave devices
- understand the variety of devices available for application at microwave frequencies
- understand the basic differences between ordinary semiconductor devices and microwave devices
- understand the basic principle behind the working of microwave devices
- understand the application-specific need for devices such as Gunn, IMPATT, TRAPATT, and BARITT diodes
- solve numericals based on Gunn, IMPATT, TRAPATT, and BARITT diodes

Introduction

A quick look at the most popular electronic equipment will reveal that these devices find a large number of applications in the communications arena. A few examples of these applications are mobile phones, radio pagers, Bluetooth devices, etc. All these devices operate in a specific range of frequencies which are of the order of a few gigahertz (GHz). At such high frequencies, ordinary semiconductor devices, such as *p-n* junctions, BJTs, and FETs are not able to perform well due to relatively long lengths of devices and poor switching speeds.

To better explain the limitations of ordinary semiconductor devices at high frequencies, let us consider the Intel Pentium 4 processor which boasts a clock speed of about 3 GHz. Analytically, this means that all semiconductor devices, which constitute the core of the processor should be able to respond to the signals within a time frame of less than 3×10^{-10} s. This entails that all semiconductor devices that were triggered into processing by a clock pulse should have completed these processes by the time the next clock pulse comes in, i.e., in a time frame of 3×10^{-10} s. However, this ideal objective is not always achieved. This is mainly due to the presence of a large number of parasitic capacitances, associated with all these electronic devices, which take a finite time for charging and discharging. Similarly, there are many other factors, such as transit time, which also affect the operation of transistors at high frequencies.

For microwave-based applications, we need devices that perform well at frequencies much higher than the clock speed of the Pentium 4 processor. Therefore, the need arises for a new category of semiconductor devices which can perform electronic functions such as switching and dc-to-microwave conversion at frequencies as high as 10^{11} Hz. The devices discussed in this chapter are specifically designed to fulfil these requirements.

13.1 Types of Microwave Devices

A number of two-terminal, active microwave devices are available, which can provide good values of gain and oscillations at microwave frequencies. The most common types of active devices are Gunn, IMPATT, TRAPATT, and BARITT diodes. In this chapter, we will discuss the working and applications of these devices in detail. This section provides a cursory introduction to these devices and the following sections will describe the operation and applications of these devices in detail.

Gunn Diode

The Gunn diode was demonstrated for the first time by IBM's J.B. Gunn in 1963. It is a *transferred-electron device* (TED) and can be used for applications falling in the range of 1–100 GHz. This device uses the transferred-electron mechanism, whereby the conduction electrons of some of the semiconductors can be shifted from a state of high mobility to a state of low mobility by applying a very strong electric field. Gunn diodes are used for microwave amplification and power

generation. The Gunn diode is called a diode, even though no *p*-*n* junctions are involved, simply because it has two terminals. The basic working of the device depends upon a series of phenomena that occur in the bulk of the semiconductor and is discussed in Section 13.2.1.

IMPATT Diode

The word IMPATT is an acronym for '*imp*act-*a*valanche *t*ransit-*t*ime'. As the name indicates, the operation of this diode is related to the processes of avalanche breakdown and transit time delay. It is one of the most powerful solid state microwave generators in the 3–300 GHz range. As per the existing research in this field, IMPATT diodes have been fabricated using elemental semiconductors such as silicon and germanium and compound semiconductors such as gallium arsenide.

TRAPATT Diode

The word TRAPATT stands for '*tra*pped-*p*lasma *a*valanche-*t*riggered *t*ransit'. The maiden research paper in this area was published in 1967 in *Proceedings of the IEEE* (Prager et al. 1967). The basic operation of this device depends upon the working of a *p*-*n* junction structure, reverse biased to very high current densities. These current densities are of a range much greater than that encountered in normal avalanche operation. Due to the fact that a very complex phenomenon forms the basis of the operation of a TRAPATT diode, the properties of the device as well as those of the associated circuit need to be tightly controlled.

BARITT Diode

The BARITT (barrier-injected transit-time) diode is a relatively new addition to the family of microwave devices. One of the main advantages of BARITT diodes is that they are far less noisy than the conventional IMPATT diode structure. However, these devices are also disadvantageous in that they support relatively narrow bandwidths and their power outputs are limited to a few milliwatts.

13.2 Working Principle of Gunn and IMPATT Diodes

Section 13.1 outlined the salient features of some popular microwave devices. This section presents the detailed description of the working of Gunn and IMPATT devices.

13.2.1 Gunn Diode

Figure 13.1 shows a typical *E*-*k* diagram of the compound semiconductor GaAs, which is inherently a direct-band gap semiconductor. To understand the operation of the Gunn diode, we must first understand the implication of Fig. 13.1. In the figure, two distinct valleys have been shown for the conduction band, i.e., the *central valley* and the *satellite valley*. For normal operation of the GaAs device conduction electrons should be present in the central valley only, but, as shown in Fig. 13.1, application of very high electric fields can lead to the shifting of

electrons to the satellite valley. As discussed in Chapter 3, the curvature of energy bands has an inverse relationship with the effective mass of the particles. Therefore, once the electrons are in the satellite valley, they will have an effective mass much higher than that which they had in the central valley.

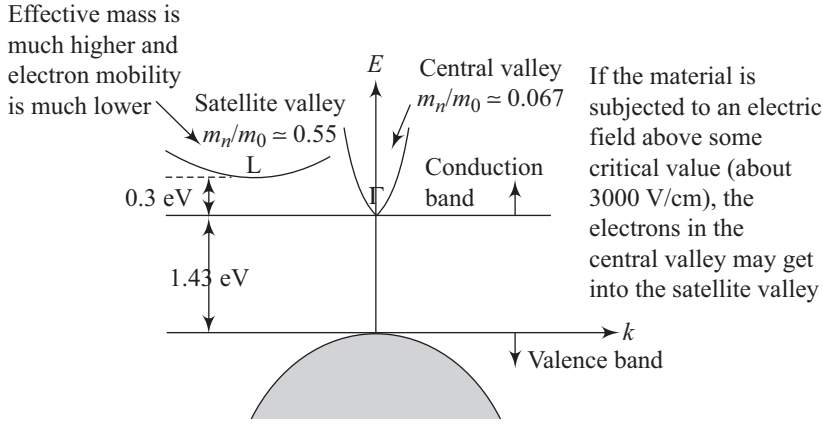


Fig. 13.1 Principle of working of a typical GaAs Gunn diode

Due to higher values of effective mass, the electrons present in the satellite valley have relatively lower values of mobility. The general equation for the current density is given as

$$J = ev_d n \tag{13.1}$$

where J is the current density, q is the charge, v_d is the drift velocity of carriers, and n is the carrier concentration. From Eqn (13.1) and the fact that electrons in the satellite valley have low mobility values, we can conclude that the current will drop at such high values of electric field. A plot showing the dependence of the drift velocity, v_d, on the electric field, E, is shown in Fig. 13.2. It must be noted that the v_d-E plot shown in Fig. 13.2 is for a material in which electron transfer is possible from the central valley to the satellite valley.

As is evident from Fig. 13.2, for low values of electric field, the electrons reside in the lower (Γ) valley of the conduction band and the associated value of mobility is high. At high electric field values, the electrons are transferred to the satellite valley, where their drift velocity, and, hence, their mobility is reduced. Marking the transition between the lower valley and the satellite valley, is a region of negative slope, which indicates a negative differential mobility.

$$\frac{dv_d}{dE} = -\mu *$$

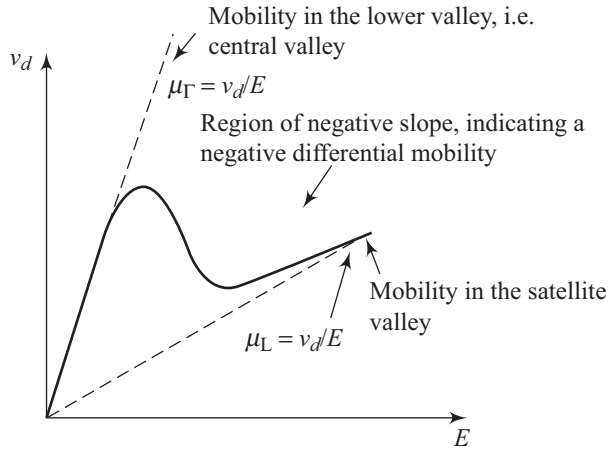


Fig. 13.2 v_d versus E plot for electrons in the satellite and central valleys

The effective masses and mobilities of electrons in the satellite valley and the central valley have been tabulated in Table 13.1.

Table 13.1 Effective mass and mobility in central and satellite valleys

Valley	Effective mass in terms of electronic mass ($m_e = 9.1 \times 10^{-31}$ kg)	Mobility (in cm^2/Vs)
Central	0.068 m_e	8000
Satellite	1.2 m_e	180

Electron densities in the satellite and the central valleys remain the same under equilibrium conditions. Now, biasing a sample of GaAs while keeping the value of electric field in the negative conductance region, leads to space charge instabilities which may enable the use of the structure for microwave amplification or generation.

13.2.2 IMPATT Diode

Impact-avalanche transit-time (IMPATT) diodes are based on a combination of carrier injection and transit time effects. Carriers generated by the injection process are swept through a drift region to the terminals of the device. Under proper biasing conditions and device configuration, the ac component of the final resulting current will be out of phase with the applied voltage by almost 180°. This gives rise to negative conductance and oscillations in a resonant circuit.

The principle behind the operation of an IMPATT diode can be explained with the help of Fig. 13.3. As shown in Fig. 13.3(a), a high electric field is generated at the interface between the n^+ and p layers resulting in an avalanche of carriers through impact ionization as shown in Fig. 13.3(b).

The IMPATT diode essentially consists of two regions—(a) the n^+-p region in which avalanche multiplications occurs and (b) the wider intrinsic (I) region

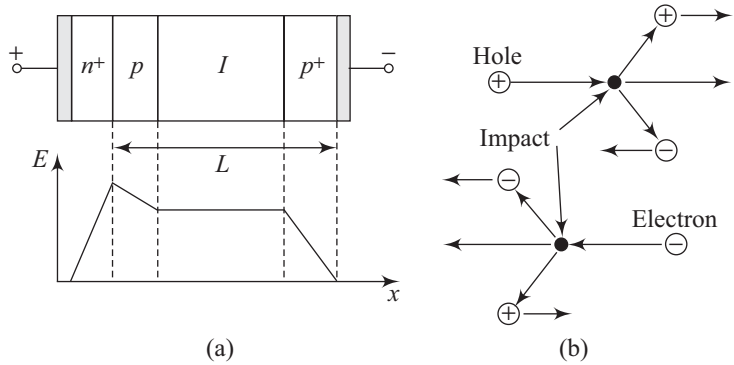


Fig. 13.3 (a) Layer structure and electric field profile of an IMPATT diode, (b) the phenomenon of impact ionization

through which the generated holes must drift on their way to the p^+ contact. Similar devices can be built based on the p^+-n-I-n^+ configuration, in which electrons generated through avalanche multiplication drift through the I-region. These devices are advantageous because electrons have higher mobility as compared to holes.

Like the Gunn diode, the IMPATT diode also operates in the negative conductance mode, i.e., the ac component of current is negative over a portion of the cycle during which the ac voltage is positive and vice versa. Negative conductance is a result of two factors which cause the current to lag behind the voltage. These factors are the delay due to the avalanche process and the delay due to the transit time of carriers across the drift region. If the sum of these two delays is approximately equal to one half cycle of the operating frequency, then negative conductance occurs enabling the device to be used for oscillation and amplification.

The situation discussed above results in a negative ac conductance, if the ac component of the carrier flow drifts opposite to the influence of the ac electric field. Therefore, with a dc reverse bias on the device (Fig. 13.4), holes drift from left to right (in the direction of the field) as expected. Now, if an ac voltage is superimposed such that E decreases during the negative half cycle, then it is expected that the drift of holes will also decrease. However, in IMPATT devices, the drift of holes through the I-region *increases* while the ac field is decreasing.

Figure 13.4 shows the time dependence of drift of carriers during a complete cycle of applied voltage for the phase angle going from 0 to $3\pi/2$. Figure 13.4(a) shows the n^+-p space charge region where all avalanche multiplication takes place provided that the dc bias is conducively set. Electrons generated due to avalanche multiplication move towards the n^+-region and the holes enter the I drift region. The hole pulse reaches its peak value not when the voltage is maximum (at $\omega t = \pi/2$) but at $\omega t = \pi$ [Fig. 13.4(c)]. Therefore, a phase delay of $\pi/2$ is inherent in the avalanche process itself. A further delay is provided by the drift region in the form of carrier transit time. Once avalanche multiplication ceases ($\omega t > \pi$), the pulse of holes simply drifts towards the p^+ contact. But during this period

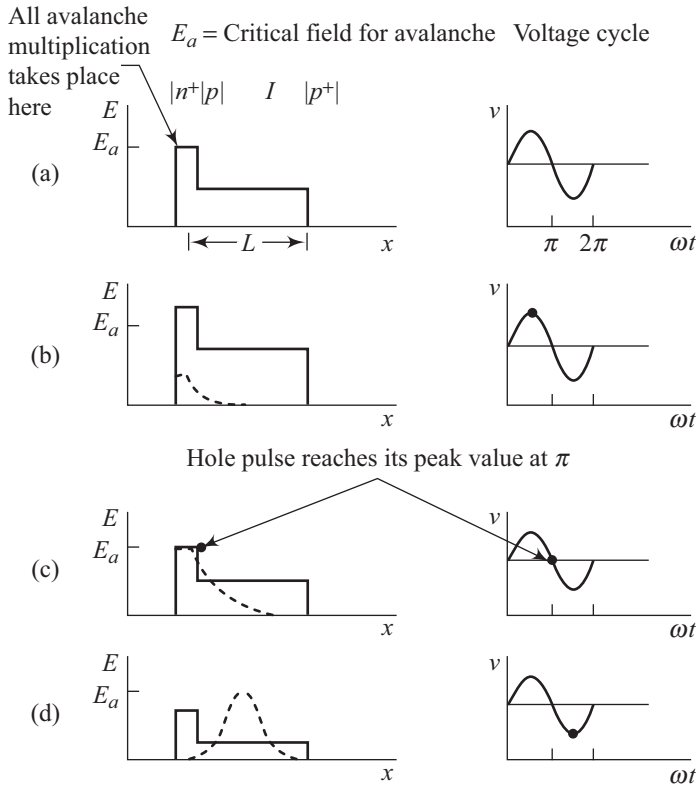


Fig. 13.4 Time dependence of drift of carriers during a cycle of applied voltage for (a) $\omega t = 0$, (b) $\omega t = \pi/2$, (c) $\omega t = \pi$, and (d) $\omega t = 3\pi/2$.

[Fig. 13.4(d)] the ac terminal voltage is negative. Therefore, the dynamic conductance is negative and energy is supplied to the ac field.

If the length of the drift region is chosen properly, the pulse of holes is collected at the p^+ contact just as the voltage cycle is completed and the cycle then repeats itself. The pulse will drift through the length, L, of the I-region during the negative half cycle, if we choose the transit time to be one half the oscillation period

$$\frac{L}{v_d} = \frac{1}{2}\frac{1}{f} \tag{13.2}$$

or

$$f = \frac{v_d}{2L} \tag{13.3}$$

where f is the operating frequency and v_d is the drift velocity of holes. Carriers generally traverse the drift region at saturation velocity, v_s. Equation (13.3) can then be written in the form

$$f = \frac{v_s}{2L} \tag{13.4}$$

IMPATT diodes are largely manufactured from Si or GaAs, although the structure is relatively more difficult to fabricate as compared to any other diode structure. The main application of IMPATT diodes falls in the range of frequencies above 3 GHz. The disadvantage of IMPATT diodes lies in the high levels of noise associated with them.

- The original design of the IMPATT diode was proposed by Read. Due to advances in technology, exotic structures, such as modified Read diode and structures incorporating highly doped regions called bumps, are now feasible. Fabricated IMPATT diodes are packaged in suitable microwave packages.
- InP-based transferred-electrons devices (TED) are being extensively studied for suitable application along with the more popular GaAs-based devices.

13.3 Operation of TRAPATT and BARITT Diodes

The detailed working principles of TRAPATT and BARITT diodes will be discussed in this section.

13.3.1 TRAPATT Diode

As explained in Section 13.1, 'TRAPATT' is an acronym for trapped plasma avalanche triggered transit mode. It is a high-efficiency microwave generator capable of operating in a wide range of frequencies varying from several hundred megahertz to several gigahertz.

The basic device is a semiconductor *p*-*n* junction diode reverse biased to current densities much greater than those encountered in normal avalanche operation. High-peak-power diodes are typically silicon n^+-p-p^+ or p^+-n-n^+ structures with the intermediate *p*-type or *n*-type depletion region width within a range of about 2.5 to 12.5 μm. The doping levels of the depletion region are so chosen that diodes are well punched through at breakdown. This ensures that the value of the dc electric field in the depletion region just prior to breakdown is well above that required to reach the saturated drift velocity level. The device's n^+- or p^+-region is kept as thin as possible at 2.5 to 7.5 μm. The TRAPATT diode's diameter ranges from as small as 50 μm for CW operation to 750 μm for operation at low frequencies for high-peak-power devices. Analysing the operation of a TRAPATT diode has been a challenging exercise for researchers, but has been successfully accomplished by some scientists such as Clorfeine and DeLoach during 1969–70.

These analyses have shown that a high-field avalanche-zone propagates through the diode and fills the depletion layer with a dense plasma of electrons and holes that become trapped in the low-field region behind the zone. A typical voltage waveform for the TRAPATT mode of an avalanche p^+-n-n^+ diode assumed to be operating with a square-wave current drive is shown in Fig. 13.5.

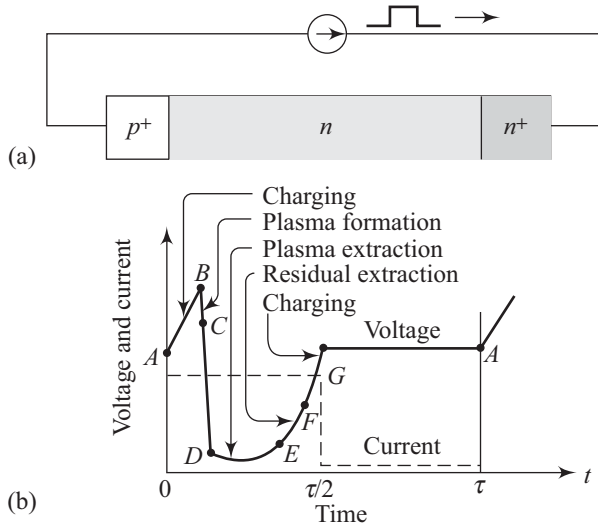


Fig. 13.5 (a) Structure of a p^+-n-n^+ TRAPATT diode (b) Voltage and current waveform for the TRAPATT diode

The overall operation of the TRAPATT diode has been explained in the following steps with a discussion of the series of events occurring in the device as depicted in Fig. 13.5.

Step 1 At point *A* the electric field is uniform throughout the sample and its magnitude is large but less than the value required for avalanche breakdown. The current density is expressed by

$$J = \varepsilon_s \frac{dE}{dt} \tag{13.5}$$

where ε_s is the dielectric permittivity of the semiconductor.

At the instant of time corresponding to point *A*, the diode current is turned on. Since the only charge carriers present are those caused by thermal generation, the diode initially charges up like a linear capacitor, driving the magnitude of the electric field above the breakdown voltage.

Step 2 When a sufficient number of carriers have been generated, the particle current exceeds the external current and the electric field is depressed throughout the depletion region, causing the voltage to decrease. The portion of the cycle is shown by the curve from point *B* to point *C*. During this time interval the electric field is sufficiently large for the avalanche to continue and a dense plasma of electrons and holes is created.

Step 3 As some of the electrons and holes drift out of the ends of the depletion layer, the field is further depressed and *traps* the remaining plasma. The voltage decreases to point *D*. A long time is required to remove the plasma because the total plasma charge is larger compared to the charge per unit time in the external current.

Step 4 At point *E*, the plasma is removed but a residual charge of electrons remains at one end of the depletion layer and a residual charge of holes at the other end. As the residual charge is removed, the voltage increases from point *E* to point *F*. At point *F*, all the charge that was generated internally is removed. This charge must be greater than or equal to that supplied by the external current, otherwise the voltage at point *F* will exceed that at point *A*.

Step 5 From point *F* to point *G*, the diode charges up again like a fixed capacitor. At point *G* the diode current goes to zero for half a period and the voltage remains constant at V_A until the current comes back on and the cycle repeats. The electric field can be expressed as

$$E(x, t) = E_m - \frac{eN_d}{\varepsilon_s} x + \frac{Jt}{\varepsilon_s} \tag{13.6}$$

where N_D is the doping concentration of the *n*-region and *x* is the distance. Thus the value of *t* at which the electric field reaches E_m at a given distance *x* into the depletion region is obtained by setting $E(x, t) = E_m$, yielding

$$t = \frac{eN_d}{J} x$$

$$\Rightarrow x = \frac{J}{eN_d} t \tag{13.7}$$

Differentiating Eqn (13.7) with respect to *t* results in

$$v_z \equiv \frac{dx}{dt} = \frac{J}{eN_d} \tag{13.8}$$

where v_z is the avalanche-zone velocity.

The avalanche-zone velocity is much larger than the scattering-limited velocity. Thus the avalanche zone (or the avalanche shock front) will quickly sweep across most of the diode, leaving the diode filled by a highly conducting plasma of holes and electrons, the space charge of which depresses the voltage to low values. Because of the dependence of the drift velocity on the field, the electrons and holes will drift at velocities determined by the low-field mobilities and the transit time of the carriers can become much longer than

$$\tau_s = \frac{L}{v_s} \tag{13.9}$$

where v_s is the saturated carrier drift velocity.

Thus the TRAPATT mode can operate at comparatively low frequencies, since the discharge time of the plasma (given by the ratio (Q/I) of its charge to its current) can be considerably greater than the nominal transit time, τ_s, of the diode at high field. The TRAPATT mode is still a transit-time mode in the sense that the time delay of carriers in transit (that is, the time between injection and collection) is utilized to obtain a current phase shift favourable for oscillation.

When the diode is placed in a circuit with a load, RF power is delivered by the diode to the external load. The main function of this circuit is to match the

diode's effective negative resistance to the load at the output frequency while reactively terminating (trapping) frequencies above the oscillation frequency in order to ensure TRAPATT operation. To date, the highest pulse power of 1.2 kW has been obtained at 1.1 GHz (five diodes in series) and the highest efficiency of 75 per cent has been achieved at 0.6 GHz. Table 13.2 lists the TRAPATT oscillator parameters at different operating frequencies.

Table 13.2 TRAPATT oscillator parameters

Frequency GHz	Peak power (W)	Average power (W)	Operating voltage (V)	Efficiency (%)
0.5	600	3	150	40
1.0	200	1	110	30
1.0	400	2	110	35
2.0	100	1	80	25
2.0	200	2	80	30
4.0	100	1	80	20
8.0	50	1	60	15

TRAPATT operation is a rather complicated means of achieving microwave oscillation and requires a good control on both device and circuit properties. Moreover, the TRAPATT mode generally exhibits a considerably higher noise figure than the IMPATT mode and the upper operating frequency appears to be practically limited to below the millimetre wave region.

13.3.2 BARITT Diode

Barrier-injected transit-time (BARITT) diodes are the latest addition to the family of active microwave diodes. They have long drift regions similar to those of IMPATT diodes. The carriers traversing the drift regions of BARITT diodes, however, are generated by minority carrier injection from forward-biased junctions instead of being extracted from the plasma of an avalanche region. Several different structures have been operated as BARITT diodes, including *p*-*n*-*p*, *p*-*n*-*v*-*p*, *p*-*n*–metal, and metal–*n*–metal. For a *p*-*n*-*v*-*p* BARITT diode, the forward-biased *p*-*n* junction emits holes into the *v*-region. These holes drift with saturation velocity through the *v*-region and are collected at the *p* contact. The diode exhibits a negative resistance for transit angles between π and 2π. The optimum transit angle is approximately 1.6π.

BARITT diodes are much less noisy than IMPATT diodes. Noise figures are as low as 15 dB at *C*-band (3.9 to 6.2 GHz) frequencies with silicon BARITT amplifiers. The major disadvantages of BARITT diodes are relatively narrow bandwidths and power outputs limited to a few milliwatts.

One typical method of fabricating a BARITT diode is to start with selection of an *n*-type silicon crystal wafer with a resistivity of about 11 Ω cm and a doping of 4×10^{14} cm^{-3}. Then this *n*-type silicon wafer is sandwiched between two PtSi Schottky barrier contacts of about 0.1 μm thickness. A schematic diagram of a

metal–*n*–metal structure is shown in Fig. 13.6(a). The energy band diagram at thermal equilibrium is shown in Fig. 13.6(b), where ϕ_{n1} and ϕ_{n2} are the barrier heights for the metal–semiconductor contacts. For the PtSi–Si–PtSi structure mentioned previously, $\phi_{n1} = \phi_{n2} = 0.85$ eV. The hole barrier height ϕ_{p2} for the forward-biased contact is about 0.15 eV. Figure 13.6(c) shows the energy band diagram when a voltage is applied. The mechanisms responsible for microwave oscillation are derived from the following factors:

- The rapid increase of the carrier injection process caused by the decreasing potential barrier of the forward-biased metal–semiconductor contact.
- An apparent $3\pi/2$ transit angle of the injected carrier that traverses the semiconductor depletion region.

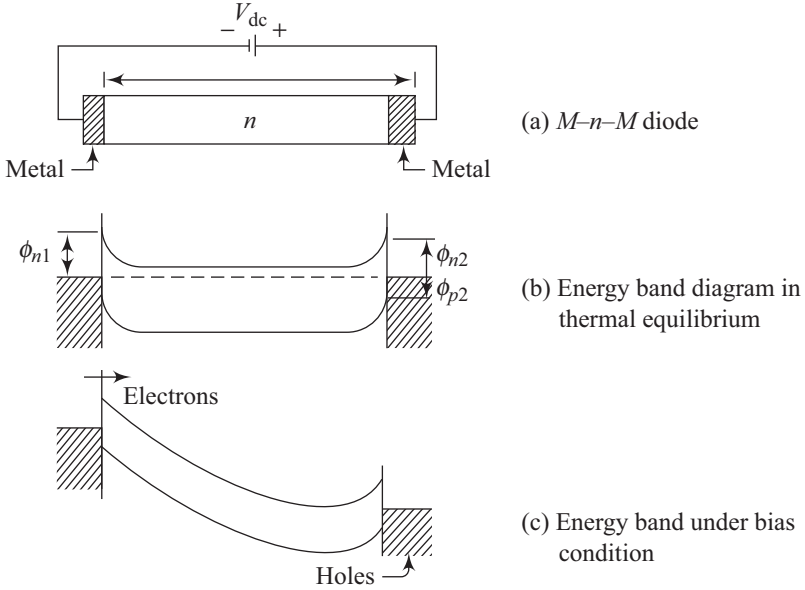


Fig. 13.6 Structure of a typical *M-n-M* diode along with the energy band diagrams

The rapid increase in terminal current with applied voltage (above 30 V) as shown in Fig. 13.7 is caused by thermionic hole injection into the semiconductor as the depletion layer of the reverse-biased contact spreads through the entire device thickness. The critical voltage is approximately given by

$$V_c = \frac{eN_d L^2}{2\varepsilon_s} \tag{13.10}$$

where N_d is the doping concentration, L is the semiconductor thickness, and ε_s is the semiconductor dielectric permittivity (= 11.7 in case of silicon).

The current–voltage characteristic of the silicon MSM structure (PtSi–Si–PtSi) measured at 77 K and 300 K are shown in Fig. 13.7. The increase in current is not due to avalanche multiplication, as is apparent from the magnitude of

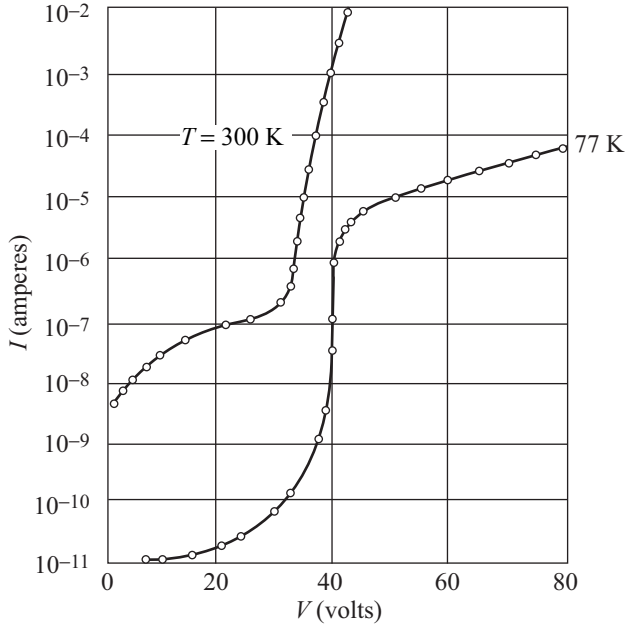


Fig. 13.7 Current–voltage characteristics of a typical BARITT diode (PtSi–Si–PtSi) structure

the critical voltage and its negative temperature coefficient. At 77 K the rapid increase is stopped at a current of about 10^{-5} A. This saturated current is expected in accordance with the thermionic emission theory of hole injection from the forward-biased contact with a hole barrier height (ϕ_{n2}) of about 0.15 eV.

The noise measure of a BARITT diode is substantially lower than that of a silicon IMPATT diode and is comparable to that of a GaAs transferred-electron oscillator.

- The low noise levels and the stability of the BARITT diode make it a suitable device for low-power applications such as local oscillators and doppler detectors.

Solved Problems

13.1 A typical *n*-type GaAs diode with an *n*-type doping concentration of 2×10^{14} cm^{-3} is subjected to an electric field of 3200 V/cm. If the threshold electric field value of the device is 2800 V/cm, its overall length is 10 μm, and the device is operated at a frequency of 10 GHz, calculate the electron drift velocity and the current density. Also estimate the value of the negative electron mobility.

Solution

The drift velocity of the device can be calculated as the product of the operating frequency and the length of the device.

$$\therefore \quad v_d = 10 \times 10^9 \times 10 \times 10^{-6}$$
$$= 10^5 \text{ ms}^{-1}$$
$$= 10^7 \text{ cms}^{-1}$$

The current density is given as

$$J = env_d$$
$$= 1.6 \times 10^{-19} \times 2 \times 10^{14} \times 10^7$$
$$= 320 \text{ A/cm}^2$$

The negative electron mobility is given by

$$\mu = -\frac{v_d}{E} = -\frac{10^7}{3200}$$
$$= -3100 \text{ cm}^2/\text{Vs}$$

13.2 A GaAs IMPATT diode has a drift length of 2 μm and a carrier drift velocity of 2×10^7 cm/s. Determine the drift time of the carriers and the operating frequency for this IMPATT structure.

Solution

$$\text{Drift time} = \frac{\text{drift length}}{\text{drift velocity}}$$
$$= \frac{2 \times 10^{-4}}{2 \times 10^7}$$
$$= 10^{-11} \text{ s}$$

$$\text{Operating frequency} = \frac{\text{drift velocity}}{2 \times \text{drift length}}$$
$$= \frac{2 \times 10^5}{2 \times 2 \times 10^{-6}}$$
$$= 50 \text{ GHz}$$

13.3 A p^+-n-n^+ TRAPATT diode has an n-region doping concentration of 2×10^{15} cm^{-3}. Calculate the avalanche-zone velocity, if the device comes across a current density (J) of 20 kA/cm^2.

Solution

From Eqn (13.8) we know that the avalanche-zone velocity for a TRAPATT diode is given as

$$v_z = \frac{J}{eN_d} \tag{13.3.1}$$

Substituting $J = 20$ kA/cm^2, $N_d = 2 \times 10^{15}$ cm^{-3}, and $e = 1.6 \times 10^{-19}$ C in Eqn (13.3.1) we get

$$v_z = \frac{20 \times 10^3}{1.6 \times 10^{-19} \times 2 \times 10^{15}}$$
$$= 6.25 \times 10^7 \text{ cm/s}$$

13.4 An *M*–Si–*M* BARITT diode has a silicon donor doping concentration of 2.8×10^{21} m^{-3}. If the length of the diode is 6 μm, determine the breakdown voltage and the breakdown electric field. Take the value of the relative dielectric constant (ε_r) of silicon to be 11.8.

Solution

The breakdown voltage V_{bd} is double the critical voltage given by Eqn (13.10). Therefore

$$V_{bd} = \frac{eN_d L^2}{\varepsilon_s}$$
$$= \frac{1.6 \times 10^{-19} \times 2.8 \times 10^{21} \times (6 \times 10^{-6})^2}{8.854 \times 10^{-12} \times 11.8}$$
$$= 154.36 \text{ V}$$

The breakdown electric field is given as

$$E_{bd} = \frac{V_{bd}}{L}$$
$$= \frac{154.36}{6 \times 10^{-6}}$$
$$= 2.573 \times 10^7 \text{ V/m}$$
$$= 2.57 \times 10^5 \text{ V/cm}$$

Recapitulation

- Gunn, IMPATT, TRAPATT, and BARITT diodes are important microwave devices.
- Gunn diode is based on the principle of electron transfer from the lower (central) valley of the conduction band to the satellite valley.
- In IMPATT diode, the total delay time is approximately one half cycle of the operating frequency.
- Operating frequency, *f*, of an IMPATT diode is given by

  $$f = \frac{v_d}{2L}$$

  where *L* is the drift region length.
- For a TRAPATT diode, the avalanche-zone velocity, v_z, is given by

  $$v_z = \frac{J}{eN_d}$$

- Injected carriers have an apparent transit angle of $3\pi/2$ in BARITT diodes.

Exercises

Review Questions

13.1 Why do we need microwave devices?

13.2 Give the full forms for IMPATT, TRAPATT, and BARITT.

13.3 Compare the advantages and disadvantages of some common microwave devices.

13.4 Explain the existence of multiple valleys in the conduction bands of some semiconductors.

13.5 Why is there a difference in the mobility of electrons in the satellite valley and the central valley?

13.6 How does the difference in the mobility of electrons in the central and satellite valleys give rise to the phenomenon of negative conductance?

13.7 What kind of semiconductors exhibit the transferred-electron mechanism?

13.8 Justify the name given to IMPATT diodes by explaining the series of phenomena that take place in the device.

13.9 Why is it important to select a correct length for the *I*-region of an IMPATT diode structure?

13.10 What are the factors which may limit the maximum operating frequency of an IMPATT diode?

13.11 Why is there an inherent noise associated with the IMPATT structures?

13.12 Justify the name given to TRAPATT devices by explaining the series of phenomena that occurs within them.

13.13 What is the significance of selecting the correct level of doping in the depletion region of a typical TRAPATT diode?

13.14 What are the factors that may limit the maximum operating frequency of a TRAPATT diode?

13.15 How will you compare the noise levels of an IMPATT and a TRAPATT diode?

13.16 Why are BARITT diodes so called? Explain by detailing the phenomena that occur inside the device.

13.17 On what factors does the critical voltage of BARITT diode structure depend?

13.18 Give a method for comparing the noise levels in IMPATT, TRAPATT, and BARITT diodes.

13.19 How does the size of microwave devices depend upon the operating frequency?

13.20 Indicate the role of impact ionization in the operation of IMPATT diodes.

13.21 What is a satellite valley in the conduction band?

13.22 How does the mobility of a charge carrier depend upon its effective mass?

13.23 Explain the phenomenon of Gunn effect.

13.24 Why is GaAs an important semiconductor for the fabrication of the Gunn diode?

13.25 Explain the physical process involved in the operation of IMPATT diodes.

13.26 What is the role of saturation velocity in the operation of IMPATT diodes?

13.27 Explain the operating principle of TRAPATT devices with a suitable diagram.

13.28 What is the source of noise in TRAPATT diodes?

13.29 Indicate some typical applications of BARITT diodes.

Problems

13.1 A GaAs Gunn diode is fabricated using *n*-type material with a doping concentration of 10^{14} cm^{-3}. An electric field of 3000 V/cm is applied across the diode of length 8 μm. The threshold field is 2800 V/cm. The device is being operated at 12 GHz. Calculate the electron drift velocity, current density, and the negative electron mobility.

$$\left[Hint: v_d = \frac{L}{1/f} \text{ and } J = e\, n\, v_d \right]$$

Ans. 9.6×10^4 ms^{-1}, 1.536×10^6 A/m^2, -3200 cm^2/Vs.

13.2 A Gunn diode is being operated at 8 GHz and has a length of 10 μm. Assuming all other parameters to be the same as in Problem 13.1, evaluate the drift velocity and the negative electron mobility.

Ans. 8×10^4 m/s, -2666.7 cm^2/Vs

13.3 A GaAs IMPATT diode has a drift length of 1.8 μm and a saturation drift velocity of 2×10^7 cm/s. Calculate the drift time of the carriers and the approximate operating frequency of the diode.

$$\left[Hint: \text{ drift time } = \frac{\text{drift length}}{\text{drift velocity}} \text{ and } f = \frac{v_d}{2L} \right]$$

Ans. 0.9×10^{-11} s, 55.6 GHz.

13.4 The material being used to fabricate a batch of IMPATT diodes has a saturation drift velocity of 4×10^6 cm/s. Calculate the approximate optimum operating frequency.

Ans. ~11.1 GHz.

13.5 A p^+-n-n$^+$ TRAPATT structure is being operated at a current density of 18 kA/cm^2. The n-region doping concentration is $N_d = 4 \times 10^{15}$ cm^{-3}. Determine the avalanche-zone velocity.

$$\left[Hint: v_z = \frac{J}{eN_d} \right]$$

Ans. 2.81×10^7 cm/s.

13.6 The operating current density for the TRAPATT mentioned in Problem 13.5 is changed to 10 kA/cm^2. Calculate the percentage change in the avalanche-zone velocity.

Ans. – 44.5%

13.7 Silicon with a doping of 3×10^{21} m^{-3} is used to fabricate a *M*–Si–*M* BARITT diode. The length of the device is 5 μm. Calculate the breakdown voltage and the breakdown electric field for the device. Assume $\varepsilon_s = 11.8$.

$$\left[Hint: V_{\text{bd}} = \frac{eN_d L^2}{\varepsilon_s} \right]$$

Ans. 114.86 V, 2.30×10^7 V/m.

13.8 A *n*-type GaAs diode is being operated at a frequency of 12 GHz. The electron drift velocity under the operating condition is 10^7 cm sec^{-1}. Calculate the overall length of the device needed.

Ans. 8.33 μm

13.9 A GaAs IMPATT is to be operated at 40 GHz. If the carrier drift velocity is 2×10^7 cm/s, calculate the drift length.

Ans. 2.5 μm

14 Rectifiers and Power Supplies

- Single-phase rectifiers
- Filter circuits
- Voltage regulators
- Switched-mode power supply

Learning Objectives

After going through this chapter the student will be able to

- understand the operation of half-wave and full-wave rectifiers
- derive expressions for I_m, I_{dc}, I_{rms} for rectifiers
- understand the advantages of a bridge rectifier
- understand the concept of ripple factor
- derive expressions for the ripple factor for various filter circuits
- understand the operation of shunt-capacitor filter, π filter, and RC filter circuits
- solve numericals based on half-wave and full-wave rectifier circuits
- solve numericals based on filter circuits
- understand the operation of zener diode regulator and series regulator
- understand the salient features of a switched-mode power supply
- solve numericals based on regulator circuits

Introduction

All electronic equipment and instruments require power supplies of various types and specifications. Rectifiers are systems that convert the alternating input voltages into direct voltages. Since most electronic devices require dc voltages for their operation, rectifiers are very useful circuits to study. Rectifiers can be either half-wave rectifiers, where only half the cycle of the alternating input voltage is usefully utilized, or full-wave rectifiers, where the fully cycle is used. The output of a rectifier is generally not smooth and uniform enough to be put to direct use. The circuits that are used to smoothen the outputs obtained from rectifiers are called filters. This chapter will discuss some commonly employed filter circuits like capacitor filter, π filter, and RC filter. Regulation of the output voltage is crucial for optimum operation of all electronic circuits. This important topic will also be covered in this chapter. Switched mode power supplies are finding numerous applications in electronic circuits due to their wide range of operation and flexibility. This chapter will also present some important aspects of the switched mode power supply (SMPS).

14.1 Single-phase Rectifiers

Rectifiers are circuits that convert alternating current to direct current. Nearly all types of electronic equipment require dc input to power them. The most common form of rectifiers operate with single phase inputs. We will discuss some single-phase rectifier circuits in this section.

14.1.1 Half-wave Rectifier

Figure 14.1 shows a half-wave rectifier circuit employing an ideal *p-n* junction diode.

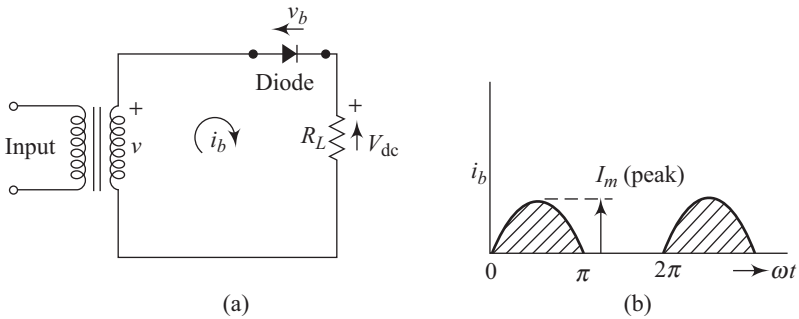


Fig. 14.1 (a) Half-wave rectifier circuit (b) Current waveform

An ideal *p-n* junction diode, it may be recollected, has a negligible forward voltage drop and zero reverse leakage current. The diode in the circuit shown in

Fig. 14.1 allows current to pass through it when it is forward biased and does not allow any current under reverse bias conditions. Thus the junction diode connects the ac source to the load when its *p*-region is positive with respect to the *n*-region and disconnects the source and the load under opposite polarity conditions.

The current pulses during the two half-cycles of the voltage are given by

$$i_b = \frac{V_m \sin \omega t}{R_L} \text{ for } 0 \le \sin \omega t \le 1 \tag{14.1}$$

and

$$i_b = 0 \text{ for } -1 \le \sin \omega t \le 0 \tag{14.2}$$

The current waveform is shown in Fig 14.1(b). From Fig. 14.1(a) it is clear that the voltage across the load, v_L is given by

$$v_L = i_b R_L \tag{14.3}$$

Thus, the load voltage has the same pulse shape as the current i_b sketched in Fig. 14.1(b). If V_m represents the peak voltage across the load, then a Fourier analysis of the half-sinusoid voltage pulses across the load R_L leads to

$$v_L = \frac{V_m}{\pi} + \frac{V_m}{2} \sin \omega t - \frac{2V_m}{3\pi} \cos 2\omega t - \frac{2V_m}{15\pi} \cos 4\omega t - \cdots \tag{14.4}$$

The dc voltage across the load resistor, R_L, is the average of the above series. This results in

$$V_{\text{dc}} = \frac{V_m}{\pi} \tag{14.5}$$

The corresponding current I_{dc} is given by

$$I_{\text{dc}} = \frac{V_m}{\pi R} = \frac{I_m}{\pi} \tag{14.6}$$

where

$$I_m = \frac{V_m}{R} \tag{14.7}$$

represents the peak current magnitude.

Equation (14.4) also contains the harmonic frequencies that are not a part of the dc output. The harmonic terms, however, result in power dissipation in resistive loads, reduced power efficiency, and lead to *ripple* in the rectifier output waveform. Circuits like the half-wave rectifier have too high a ripple to be of any direct use in electronic equipment. More efficient rectifier circuits are available for such applications.

14.1.2 Full-wave Rectifier

A full-wave rectifier circuit is shown in Fig. 14.2. The corresponding current waveform at the resistive load is also shown in the figure. It can be seen that the load gets current during the entire cycle of the input waveform. This is a commonly used circuit for supplying dc input to electronic circuits.

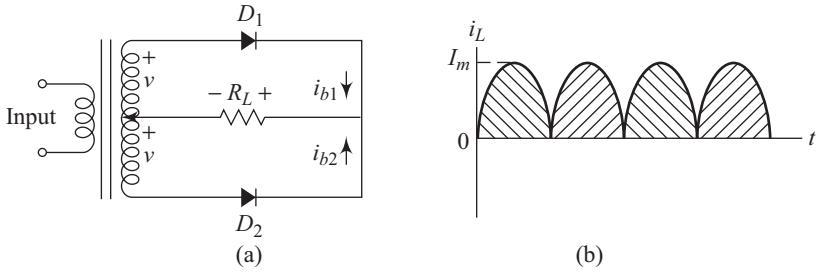


Fig. 14.2 (a) Full-wave rectifier circuit (b) Current waveform

The diodes conduct when forward biased and, therefore, one of the two diodes D_1 or D_2 is conducting during either of the two half-cycles during any particular cycle of the input. For ideal diodes, we can write for the currents in the circuit

$$\left.\begin{aligned} i_{b1} &= \frac{V_m \sin \omega t}{R_L} \\ i_{b2} &= 0 \end{aligned}\right\} \text{ for } 0 \le \sin \omega t \le 1 \tag{14.8}$$

and

$$\left.\begin{aligned} i_{b1} &= 0 \\ i_{b2} &= -\frac{V_m \sin \omega t}{R_L} \end{aligned}\right\} \text{ for } -1 \le \sin \omega t \le 0 \tag{14.9}$$

Due to the configuration of the circuit shown in Fig. 14.2, the current flowing through the load, R_L, is in the same direction in either of the two half-cycles.

Using Fourier analysis, the voltage waveform across the load, v_L, can be written in the form

$$v_L = \frac{2V_m}{\pi} - \frac{4V_m}{\pi}\left(\frac{1}{3}\cos 2\omega t + \frac{1}{15}\cos 4\omega t + \frac{1}{35}\cos 6\omega t + \cdots\right) \tag{14.10}$$

A comparison of Eqs (14.9) and (14.10) will reveal that the lowest frequency component for a full-wave rectifier output load current is twice that for a half-wave rectifier circuit.

The desired rectifier output is dc or the zero frequency component given by

$$V_{\text{dc}} = \frac{2V_m}{\pi} \tag{14.11}$$

Thus, the load current, I_{dc}, can be written in the form

$$I_{\text{dc}} = \frac{2V_m}{\pi R_L} = \frac{2I_m}{\pi} \tag{14.12}$$

where, I_m represents the peak current. A comparison of Eqs (14.6) and (14.12) will help us conclude that the dc current in a full-wave rectifier is twice that in a half-wave rectifier.

The effective current, I_{rms}, is given by the equation

$$I_{\text{rms}} = \sqrt{\frac{1}{2\pi}\int_0^{2\pi} i_b^2 \, d(\omega t)} \tag{14.13}$$

Using Eqn (14.8) in Eqn (14.13) and utilizing the inherent symmetry of the load current, we can rewrite Eqn (14.13) as

$$I_{\text{rms}} = \frac{2}{2\pi}\left[\int_0^{\pi} I_m^2 \sin^2 \omega t \, d(\omega t)\right]^{\frac{1}{2}} \tag{14.14}$$

yielding

$$I_{\text{rms}} = \frac{I_m}{\sqrt{2}} \tag{14.15}$$

The ac power input from the transformer can be expressed as

$$P_{\text{ac}} = \left(\frac{I_m}{\sqrt{2}}\right)^2 R_L \tag{14.16}$$

leading to

$$P_{\text{ac}} = \frac{I_m^2 R_L}{2} = \frac{V_m^2}{2R_L} \tag{14.17}$$

The dc power output is given by

$$P_{\text{dc}} = I_{\text{dc}}^2 R_L \tag{14.18}$$

The conversion efficiency, η, of the full-wave rectifier for a resistive load is expressible in the form

$$\eta = \frac{P_{\text{dc}}}{P_{\text{ac}}} \tag{14.19}$$

Using Eqs (14.11), (14.17), and (14.18) we get

$$\eta = \left(\frac{2V_m}{\pi R_L}\right)^2 R_L \times \frac{2R_L}{V_m^2}$$

leading to

$$\eta = \frac{8}{\pi^2} = 0.810 \tag{14.20}$$

Thus the conversion efficiency of a full-wave rectifier is 81 per cent. The reduction in the efficiency value from its ideal value of 100 per cent is due to the harmonic power losses in the resistive load.

From Fig. 14.2, we can see that during the interval in which diode D_1 is conducting, the voltage in the outer loop of the full-wave rectifier circuit is given by

$$v_{b2} = 2V_m \sin \omega t \tag{14.21}$$

The maximum value of v_{b2} is called the *peak reverse voltage* or simply PRV From Eqn (14.21), one can see that the maximum value of v_{b2} occurs for $\omega t = \pi/2$ for which sin $\omega t = 1$ and we get

$$\text{PRV} = 2V_m \tag{14.22}$$

Thus the PRV for a full-wave rectifier is the peak of the total transformer secondary voltage. For comparison, one can recollect that the PRV for a half-wave rectifier circuit is V_m, i.e., half of the PRV value for a full-wave rectifier circuit.

14.1.3 Bridge Rectifier

The full-wave rectifier circuit shown in Fig. 14.2 uses a centre tapping on the supply transformer. Full-wave rectification can also be achieved without such a tapping. One such circuit is the bridge rectifier circuit shown in Fig. 14.3.

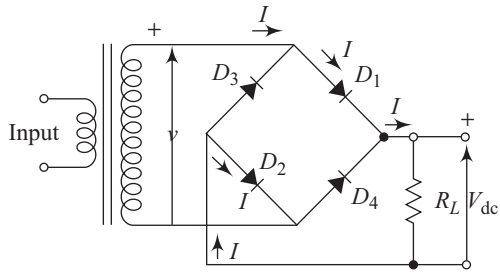


Fig. 14.3 Bridge rectifier

If the upper terminal of the secondary of the transformer is positive, diodes D_1 and D_2 are in the forward-biased state. As already discussed, in this state the diodes are conducting. The direction of current for this situation is shown with arrows in Fig. 14.3. Diodes D_3 and D_4 are in the reverse-biased state and, therefore, remain in the non-conducting mode. In the next half-cycle, the lower terminal of the secondary of the transformer turns positive, leading to the forward biasing of diodes D_3 and D_4. Interestingly, the current through the load resistor, R_L, continues to flow in the same direction in both the half-cycles. A full-wave current waveform is thus generated across the load.

From Fig. 14.3 one can conclude that for the same dc voltage, the transformer incorporated in the bridge rectifier circuit, needs only half the secondary turns as compared to the centre-tapped full-wave rectifier circuit transformer. Also, current flows through the transformer in the bridge rectifier circuit in both the half-cycles, whereas the centre-tapped transformer circuit allows current to flow through a particular portion of the transformer during one of the half-cycles. Due to this, the current rating of the secondary turns of the transformer is around 40 per cent larger in the bridge rectifier circuit in comparison to the circuit shown in Fig. 14.2. Thus with the requirement of half the turns and only a 40 per cent escalation in the current rating, the transformer is more efficiently utilized in the bridge rectifier circuit. The peak reverse voltage, PRV, across a reverse-biased diode in Fig. 14.3 equals the peak of the transformer voltage. Thus, for a specific

dc voltage, the bridge rectifier circuit needs diodes with reverse rating values that are one-half of the PRV rating required by the centre-tapped circuit. One must, however, remember that the PRV is finally restricted to values substantially lower than the avalanching voltage of the diodes and a lowering by a factor of one-half is very significant. Thus, the bridge rectifier circuit is preferred by circuit designers.

14.1.4 Ripple Factor

The output from a rectifier is a direct current with some residual pulsation that is superimposed on it. This residual pulsation is referred to as ripple. A quantitative measure of ripple is provided by a parameter called ripple factor, designated by the symbol, γ. The ripple factor, γ, is defined as the ratio of the effective value of the ac components to the dc voltage (or current). Thus,

$$\text{Ripple factor} = \gamma = \frac{\text{Effective value of ac components}}{\text{Average or dc components}} \tag{14.23}$$

As we have already seen, the ac components at the load of any rectifier circuit, include the time-varying terms in the Fourier series expansion of the relevant wave form. These are terms indicated in expressions such as Eqn (14.4) and Eqn (14.10). If the effective value of these ac components is designated as I_{ac}, the total load current, I_{rms}, is given by

$$I_{rms} = \sqrt{I_{dc}^2 + I_{ac}^2} \tag{14.24}$$

where I_{dc} is the corresponding dc component. From Eqn (14.24), we can write

$$I_{ac} = \sqrt{I_{rms}^2 - I_{dc}^2} \tag{14.25}$$

Using the definition of ripple factor, γ, from Eqn (14.23), we can write

$$\gamma = \frac{I_{ac}}{I_{dc}} \tag{14.26}$$

Using Eqn (14.25) in Eqn (14.26) yields

$$\gamma = \frac{\sqrt{I_{rms}^2 - I_{dc}^2}}{I_{dc}}$$

amounting to

$$\gamma = \sqrt{\left(\frac{I_{rms}}{I_{dc}}\right)^2 - 1} \tag{14.27}$$

Using Eqs (14.12) and (14.15) in Eqn (14.27), we get for a full-wave rectifier

$$\gamma = \sqrt{\left(\frac{I_m}{\sqrt{2}} \times \frac{\pi}{2I_m}\right)^2 - 1}$$

leading to

$$\gamma = \sqrt{\frac{\pi^2}{8} - 1} \tag{14.28}$$

yielding

$$\gamma = 0.48 \text{ or } 48\ \%.$$

For a half-wave rectifier we get

$$\frac{I_{\text{rms}}}{I_{\text{dc}}} = \frac{I_m/2}{I_m/\pi}$$

resulting in

$$\frac{I_{\text{rms}}}{I_{\text{dc}}} = \frac{\pi}{2} \tag{14.29}$$

Putting Eqn (14.29) in Eqn (14.27) results in

$$\gamma = \sqrt{\frac{\pi^2}{4} - 1}$$

gives us

$$\gamma = 1.21 \text{ or } 121\ \%$$

The ripple factors for both half-wave and full-wave rectifiers are too high for supply in most electronic equipment, which require ripple factors lower than 0.001. Such low ripple factors can be achieved with the help of filter circuits. These are discussed in Section 14.2.

14.2 Filter Circuits

As discussed in the previous section, the output waveforms obtained from half-wave and full-wave rectifiers are not smooth dc. Smoothness of the output voltage of rectifiers is obtained using different types of filter circuits. We will discuss some popular filter circuits in this section.

14.2.1 Shunt-capacitor Filter

A shunt-capacitor filter circuit is shown schematically in Fig. 14.4. The capacitor C is chosen to be of a value such that $X_C \ll R_L$. With this condition placed on the reactance of the capacitor, the alternating components get bypassed around the load resistor.

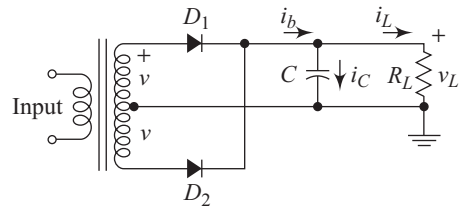


Fig. 14.4 Shunt-capacitor filter circuit

During the half-cycle when the upper end of the secondary of the transformer is positive, D_1 is in the forward-biased conducting mode. As the output of the diode D_1 increases after $\omega t = \theta_1$ as shown in Fig. 14.5(a), the capacitor C receives energy from the transformer and gets charged to the peak of the input voltage waveform, V_m.

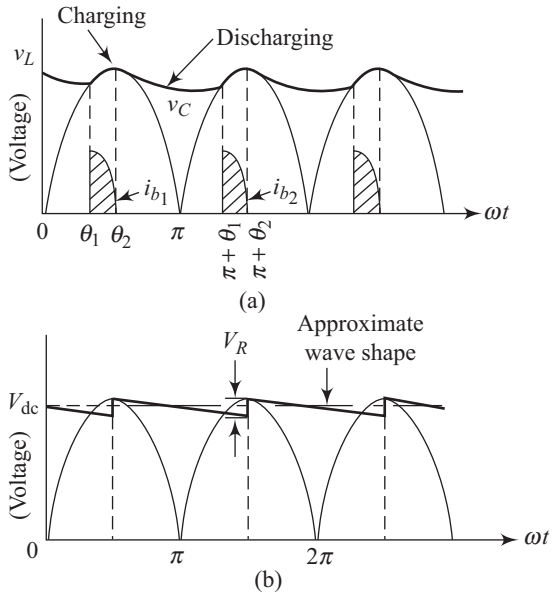


Fig. 14.5 (a) Exact voltage waveform (b) Approximated waveform

The ac waveform peaks and then the voltage from the transformer starts falling. Due to the condition placed on the capacitor–resistance discharge circuit, the fall of the transformer voltage is much faster than the rate at which the capacitor, C, can get discharged. The diode D_1 gets reverse-biased at the point shown by θ_2 in Fig. 14.5(a). During this phase, the capacitor keeps supplying the load current till the transformer voltage starts rising during the next half-cycle. This phase is shown at $\pi + \theta_1$ in Fig. 14.5(a).

During operation, as each diode gets forward biased, a large-amplitude current pulse is supplied to the capacitor and then the diodes get reverse biased, thereby effectively disconnecting the transformer and the load. The exact shape of the current pulse is decided by the magnitude of ωCR. A basic current pulse along with the transformer voltage waveform is shown in Fig. 14.6

If the load is an open circuit, the output voltage reaches the peak voltage V_m and then remains constant. For finite load values, the capacitor supplies the load current and, during this interval, also gets discharged through the finite resistance. The output voltage thus falls below V_m and a ripple appears in the voltage across the load. If the diode is assumed to open at $\omega t = \theta_2$ (disconnecting the

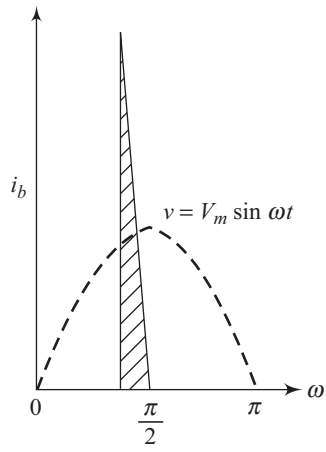


Fig. 14.6 Current pulse and transformer voltage

load with the source), then from Fig. 14.4, we can see that

$$i_R = -i_C \tag{14.30}$$

which can be written in the form

$$\frac{v_C}{R_L} = -C\frac{dv_C}{dt}$$

leading to

$$\frac{dv_C}{dt} + \frac{v_C}{R_L C} = 0 \tag{14.31}$$

Equation (14.31) has solutions of the form

$$v_C = Ae^{-t/RC}$$

At $\omega t = \theta_2$, we have

$$v_0 = V_m \sin \theta_2$$

and, therefore, the exact form of solution is

$$v_C = V_m \sin \theta_2 \, e^{-(\omega t - \theta_2)/\omega RC} \quad \text{for} \quad \theta_2 \le \omega t \le (\pi + \theta_1) \tag{14.32}$$

Equation (14.32) gives the load voltage for the period when the capacitor supplies the load current. From Fig. 14.5(a) it is clear that the load voltage falls exponentially with time in accordance with Eqn (14.32).

The parameter $\omega CR = 2\pi CR/T$, thus represents the ratio of the filter-load time constant to the period of the wave, T. ωCR should have large values by design to ensure low ripple values.

The dc voltage across the load is obtained by summing of the average of the sine voltage between θ_1 and θ_2 and Eqn (14.32) between θ_2 and $\pi + \theta_1$. The final result is

$$V_{\text{dc}} = \frac{V_m}{\pi}\sqrt{1 + \omega^2 C^2 R^2}\ \left[1 - \cos(\theta_2 - \theta_1)\right] \tag{14.33}$$

For large values of ωCR and small values of ripple, the conduction angle is very short and the discharge period can be approximated to $1/2f$. The output waveform can be approximately represented as shown in Fig. 14.5(b). In this figure, the output voltage is approximated to be made of straight lines with a peak value of V_m and a peak-to-peak ripple magnitude of V_R. From Fig. 14.5(b) it is clear that the dc voltage V_{dc} can be written in the form

$$V_{\text{dc}} = V_m - \frac{V_R}{2} \tag{14.34}$$

During the discharge interval, the capacitance voltage falls by V_R as the capacitor loses charge at an average rate given by

$$\frac{dq}{dt} \cong I_{\text{dc}} \tag{14.35}$$

The average rate of fall of voltage can be written as

$$\frac{V_R}{1/2f} = \frac{d}{dt}v_C \tag{14.36}$$

If the capacitor has a capacitance, C, Eqn (14.36) can be rewritten in the form

$$\frac{V_R}{1/2f} = \frac{1}{C}\frac{dq}{dt} \tag{14.37}$$

Using Eqn (14.35) in Eqn (14.37) results in

$$V_R = \frac{1}{2fC} I_{\text{dc}}$$

implying

$$V_R = \frac{\pi I_{\text{dc}}}{\omega C} \tag{14.38}$$

The effective value of the triangular ripple wave shape shown in Fig. 14.5(b) is given by

$$V_{R(\text{rms})} = \frac{V_R}{2\sqrt{3}} \tag{14.39}$$

Using Eqn (14.38) in Eqn (14.39) leads to

$$V_{R(\text{rms})} = \frac{\pi\, I_{\text{dc}}}{2\sqrt{3}\,\omega C} \tag{14.40}$$

Also

$$I_{\text{dc}} = \frac{V_{\text{dc}}}{R_L} \tag{14.41}$$

Employing Eqn (14.41) in Eqn (14.40) results in

$$V_{R(\text{rms})} = \frac{\pi\, V_{\text{dc}}}{2\sqrt{3}\,\omega C R_L} \tag{14.42}$$

The ripple factor, γ, is then expressible in the form

$$\gamma = \frac{V_{R(\text{rms})}}{V_{\text{dc}}} = \frac{\pi}{2\sqrt{3}\,\omega C R_L} \tag{14.43}$$

Since the ripple factor is inversely proportional to the load resistance, R_L, it decreases with the load resistance. Figure 14.7 shows the variation of ripple factor with load resistance. The dashed curve represents a more accurate variation that takes into account the conduction angle of the diode.

Using Eqs (14.34) and (14.38) we can write

$$V_{\text{dc}} = V_m - \frac{\pi I_{\text{dc}}}{2\omega C} \tag{14.44}$$

which can be rewritten as

$$V_{\text{dc}} = V_m - \frac{\pi I_{\text{dc}} R_L}{2\omega C R_L} \tag{14.45}$$

which, on using Eqn (14.41), leads to

$$V_{\text{dc}} = V_m - \frac{\pi\, V_{\text{dc}}}{2\omega C R_L} \tag{14.46}$$

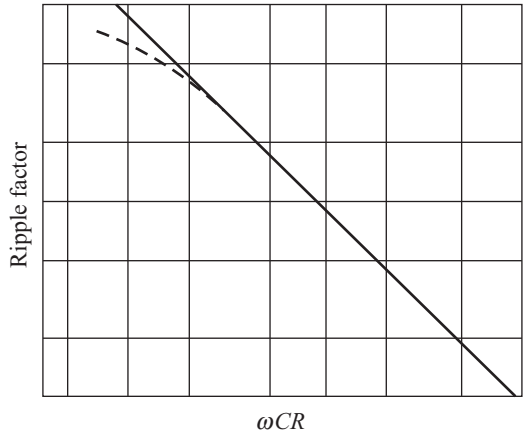


Fig. 14.7 Variation of ripple factor with load

Eqn (14.46) can be rewritten in the form

$$\frac{V_{\text{dc}}}{V_m} = \frac{1}{1 + (\pi/2\,\omega CR_L)} \tag{14.47}$$

A typical variation of the ratio $\frac{V_{\text{dc}}}{V_m}$ as a function of the filter parameter ωCR_L is shown in Fig. 14.8.

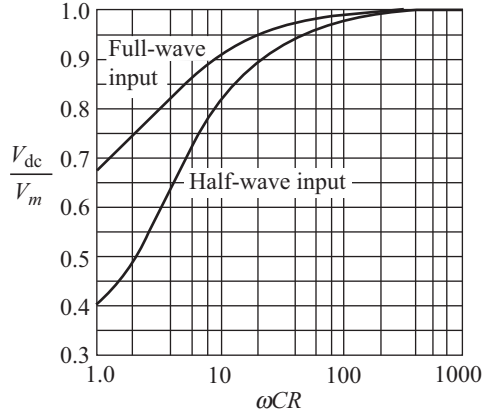


Fig. 14.8 Variation of $\left(\frac{V_{\text{dc}}}{V_m}\right)$ with the filter parameter

An important issue for the design of power supply for amplifiers is the constancy of the output dc voltage as the load undergoes variations. From Fig. 14.8 one can notice that ωCR greater than around 20 would limit the operation to the upper plateau region for both half-wave and full-wave input waveforms. A nearly constant dc voltage with variations limited to a few per cent is possible using such methodologies in design.

The rms voltage of the transformer in a bridge rectifier circuit or the rms voltage from the centre tap to the diode for a full-wave rectifier circuit, can be written in the form

$$V = \frac{V_m}{\sqrt{2}} \tag{14.48}$$

Using Eqn (14.47) in Eqn (14.48) leads to

$$V = \frac{V_{\text{dc}}}{\sqrt{2}} \left(1 + \frac{\pi}{2\omega C R_L}\right) \tag{14.49}$$

Equation (14.49) can be used to obtain the transformer voltage for a specified dc load voltage. If the ripple factor is γ, the capacitor to be used must fulfil the condition

$$C \geq \frac{1}{2\sqrt{3}\,\omega R_L \gamma} \tag{14.50}$$

14.2.2 π Filter

Addition of an LC voltage divider to the shunt-capacitor circuit, results in a circuit that can be used to realize ripple factors much lower than 1 per cent. A circuit incorporating this feature is shown schematically in Fig. 14.9.

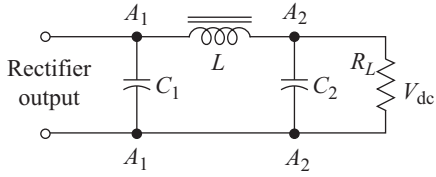


Fig. 14.9 π filter circuit

Using Eqn (14.40), the ripple voltage $V_{R(\text{rms})}$ at A_1, A_1 can be written as

$$V_{A_1 A_1} = V_{R(\text{rms})} = \frac{\pi I_{\text{dc}}}{2\sqrt{3}\,\omega C_1} \tag{14.51}$$

The circuit shown in Fig. 14.9 uses the reactance of L to suppress the effective ripple current passing through the load R_L. The reactance offered by C_2 is kept small with respect to the load R_L by design. The ripple frequency is 2ω and, using voltage division rules, the ripple voltage remaining across the load at A_2, A_2 can be written in the form

$$V_{A_2 A_2} = -\frac{j/2\omega C_2}{j2\omega L - j/2\omega C_2} V_{R(\text{rms})} \tag{14.52}$$

implying

$$V_{A_2 A_2} = \frac{\pi I_{\text{dc}}}{2\sqrt{3}\,\omega C_1 (4\omega^2 L C_2 - 1)} \tag{14.53}$$

If L and C_2 are large numbers then we can assume $4\omega^2 LC_2 \gg 1$ and we can write Eqn (14.53) in the form

$$V_{A_2A_2} = \frac{\pi\, I_{\text{dc}}}{2\sqrt{3}\,\omega C_1 (4\omega^2 LC_2)} \tag{14.54}$$

The ripple factor, γ, across the load is then given by

$$\gamma = \frac{V_{A_2A_2}}{V_{\text{dc}}} = \frac{\pi\, I_{\text{dc}}}{I_{\text{dc}}\, R_L} \frac{1}{8\sqrt{3}\omega^3 C_1 C_2 L}$$

leading to

$$\gamma \cong \frac{0.216}{R_L} \frac{1}{\omega^3\, C_1 C_2\, L} \tag{14.55}$$

Equation (14.55) can be written in the form

$$\gamma \cong \frac{0.216}{R_L} \frac{\left(\frac{1}{\omega C_1}\right)\left(\frac{1}{\omega C_2}\right)}{(\omega L)} \tag{14.56}$$

resulting in

$$\gamma \cong \frac{0.216}{R_L} \frac{X_{C_1}\, X_{C_2}}{X_L} \tag{14.57}$$

where X_{C_1}, X_{C_2}, and X_L represent the reactances offered by C_1, C_2, and L at the supply frequency, respectively. From Eqn (14.57) one can conclude that the ripple factor is inversely proportional to load resistance and, therefore, increases with load current.

Using Eqn (14.34), we can write for the voltage at A_1, A_1

$$V_{\text{dc}} = V_m - \frac{V_R}{2} \tag{14.58}$$

Suppose the inductor offers a resistance R_l, in that case, Eqn (14.58) gets further modified to

$$V_{\text{dc}} = V_m - \frac{V_R}{2} - I_{\text{dc}} R_l \tag{14.59}$$

Using Eqn (14.38) in Eqn (14.59) leads to

$$V_{\text{dc}} = V_{\text{m}} - I_{\text{dc}} \left(\frac{\pi}{2\omega C_1} + R_l \right) \tag{14.60}$$

Since, $V_{\text{dc}} = I_{\text{dc}} R_L$, we can rewrite Eqn (14.60) in the form

$$V_{\text{dc}} = \frac{V_m}{1 + (\pi/2\omega C_1 R_L) + R_l/R_L} \tag{14.61}$$

From Eqs (14.55) and (14.61) we can see that the ripple factor is a function of the product C_1C_2 whereas the dc voltage is a function of C_1 only. Thus the individual value of C_1 is not important for deciding the ripple factor of the π filter, as long as the product C_1C_2 is maintained.

14.2.3 RC Filter

The iron-cored inductor that is to be used in the π-filter circuit, pushes its cost and space requirement. It is always desirable to reduce both these constraining factors. The RC filter circuit shown in Fig. 14.10 is often a good alternative.

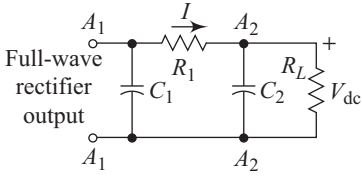


Fig. 14.10 RC filter circuit

A comparison of Figs 14.9 and 14.10 will reveal that the resistance R_1 replaces the inductance L in the RC filter circuit. This results in an increase in the internal dc voltage drop.

The ripple factor for the RC filter can be obtained by replacing X_L by R_1 in Eqn (14.55). Thus, we get

$$\gamma = \frac{0.216}{R_L} \frac{1}{\omega^2 C_1 C_2 R_1} \tag{14.62}$$

which can be rewritten in the form

$$\gamma = \frac{0.216}{R_L} \frac{X_{C_1} X_{C_2}}{R_1} \tag{14.63}$$

where the symbols have their usual meanings. On similar lines, the load dc voltage can be obtained from Eqn (14.61) by replacing R_l by R_1, resulting in

$$V_{dc} = \frac{V_m}{1 + (\pi/2\omega C_1 R_L) + R_1/R_L} \tag{14.64}$$

Thus large values of C_1 and C_2 result in low ripple factors.

- The simplest polyphase rectifier circuit is the three-phase half-wave form with star-connected secondary connection in the transformer. The ripple factor, γ, for such a circuit is 17 per cent.
- More than one filter sections can be incorporated in a power supply to reduce the ripple factor to very low values.

14.3 Voltage Regulators

Electronic circuits very often require a constant voltage for efficient operation. The power supply output voltage can vary due to changes in load resistance or supply-line voltage. A regulator is a circuit that controls the output voltage variations caused due to changes in the load resistance or the supply-line voltage.

14.3.1 Zener Diode Regulator

Figure 14.11 is a schematic diagram of a zener diode voltage regulator circuit. From the circuit diagram it is clear that the voltage across the input of the regulator needs to be greater than the output voltage.

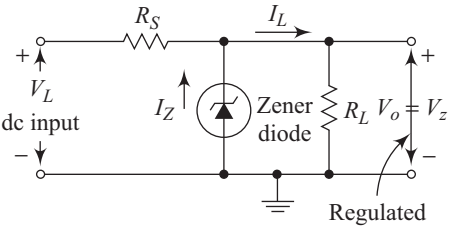


Fig. 14.11 Zener diode voltage regulation

A current equal to the sum of the load current and the zener diode current flows through the series resistor R_S. The currents are chosen so that the voltage across the zener diode and the load combination equals the zener breakdown voltage. Any change in the input voltage or the output current, produces a corresponding change in the zener diode current such that the output voltage is maintained at the desired value $V_o = V_z$. This is ensured by producing the necessary change in the voltage drop across the series resistor R_S. Thus, the input voltage, V_i, must satisfy the condition

$$V_i = (I_Z + I_L)\, R_S + V_o \tag{14.65}$$

The maximum power, $P_{\max}$, is decided by the equation

$$P_{\max} = V_Z I_{Z(\text{rated})} \tag{14.66}$$

where V_Z is the nominal zener voltage. In actual operation, a safety factor is included, so that the maximum zener diode current is taken to be

$$I_{Z(\max)} = 0.8\, I_{Z(\text{rated})} \tag{14.67}$$

A second lower extreme point is chosen to avoid the rounded knee of the current–voltage characteristic curve at low zener currents. A minimum zener current is chosen to be

$$I_{Z(\min)} = 0.2\, I_{Z(\text{rated})} \tag{14.68}$$

The safe operating region is enclosed between the two extremes decided by Eqs (14.67) and (14.68). Figure 14.12 is a schematic representation of the safe operating region.

Suppose that the dc input voltage varies within the extreme cases $V_{i(\max)}$ and $V_{i(\min)}$. We will now discuss two worst case design considerations.

Case I

One worst case would be maximum input voltage and no load. This situation is shown schematically in Fig. 14.13.

For this case we have

$$\frac{V_{i(\max)} - V_Z}{R_s} \le 0.8\, I_{Z(\text{rated})}$$

giving

$$R_S \ge \frac{V_{i(\max)} - V_Z}{0.8\, I_{Z(\text{rated})}} \tag{14.69}$$

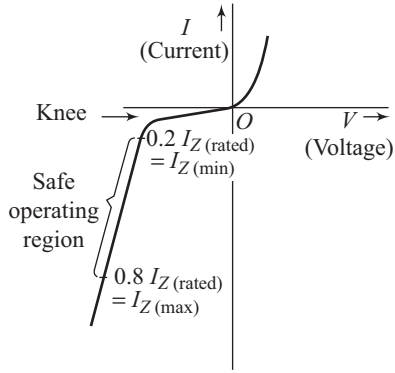


Fig. 14.12 *I-V* characteristics of zener diode showing safe operating region

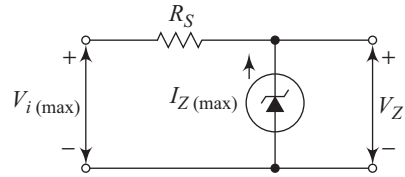


Fig. 14.13 Maximum input voltage with no load for zener diode regulator

Thus the minimum series resistance $R_{S(\min)}$ is expressed through the equation

$$R_{S(\min)} = \frac{V_{i(\max)} - V_Z}{0.8\, I_{Z\,(\text{rated})}} \tag{14.70}$$

Case II

The second worst case corresponds to rated load resistance and minimum input voltage. This situation is shown schematically in Fig. 14.14.

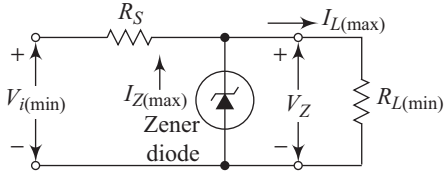


Fig. 14.14 Minimum input voltage and maximum load current for zener regulator

The governing equation is

$$[I_{Z(\min)} + I_{L(\max)}]\, R_S = V_{i(\min)} - V_Z \tag{14.71}$$

resulting in

$$I_{Z(\min)}\left[\frac{V_{i(\min)} - V_Z}{R_s}\right] - I_{L\ (\max)} \tag{14.72}$$

Using Eqn (14.68) in Eqn (14.72) leads to

$$0.2\, I_{Z\,(\text{rat})} = \frac{[V_{i(\min)} - V_Z]}{R_S} - I_{L(\max)} \tag{14.73}$$

Designating the corresponding series resistance as $R_{S(\max)}$, we can rewrite Eqn (14.73) in the form

$$R_{S(\max)} = \frac{V_{i(\min)} - V_Z}{0.2\, I_{Z(\text{rat})} + I_{L(\max)}} \tag{14.74}$$

We must have

$$R_{S(\max)} > R_{S(\min)} \tag{14.75}$$

Using Eqn (14.70) and Eqn (14.74) in Eqn (14.75) leads to

$$\frac{V_{i(\min)} - V_Z}{0.2\, I_{Z(\text{rated})} + I_{L(\max)}} > \frac{V_{i(\max)} - V_Z}{0.8\, I_{Z(\text{rated})}}$$

resulting in

$$0.8\left[\frac{V_{i(\min)} - V_Z}{V_{i(\max)} - V_Z}\right] > \frac{0.2\, I_{Z(\text{rated})} + I_{L(\max)}}{I_{Z(\text{rated})}}$$

giving us

$$0.8\left[\frac{V_{i(\min)} - V_Z}{V_{i(\max)} - V_Z}\right] - 0.2 > \frac{I_{L(\max)}}{I_{Z(\text{rated})}} \tag{14.76}$$

The current handling capability of the zener diode must satisfy the relation

$$I_{Z(\max)} > I_{L(\max)} + I_{Z(\min)} \tag{14.77}$$

We also have

$$I_{Z(\max)} = 0.8\, I_{Z(\text{rated})}$$

and

$$I_{Z(\min)} = 0.2\, I_{Z(\text{rated})}$$

Equation (14.77), therefore, results in

$$0.6\, I_{Z(\text{rated})} > I_{L(\max)} \tag{14.78}$$

14.3.2 Series Voltage Regulator

Any general series voltage regulator consists of five basic elements. These basic elements are shown in a block diagram in Fig. 14.15.

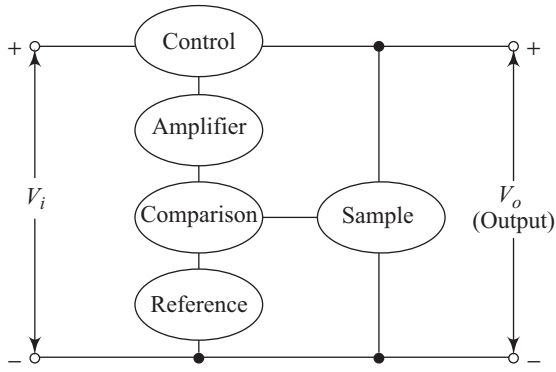


Fig. 14.15 Basic elements of a series voltage regulator

We will briefly discuss the role of each element to understand its role in functioning of the series regulator.

Sampling

A sample of the output voltage is compared with a reference and the generated error is amplified and subsequently used to control a series element. The simplest form of sampling element is a voltage divider across the regulated output as shown in Fig. 14.16.

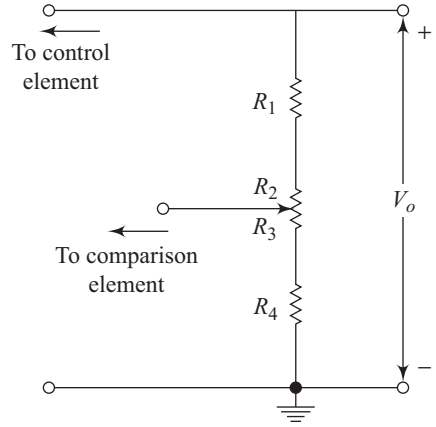


Fig. 14.16 Basic sampling element

In Fig. 14.14, the voltage being fed to the comparison element is given as

$$V = \frac{R_3 + R_4}{R_1 + R_2 + R_3 + R_4} V_o \tag{14.79}$$

where V_o represents the regulated output voltage.

Reference element

Breakdown diodes are generally used as voltage references in transistor-based regulators. This choice is dictated by the fact that the breakdown voltage is relatively constant over a wide range of reverse current. Figure 14.17 shows a typical reference element along with a series resistance R_5.

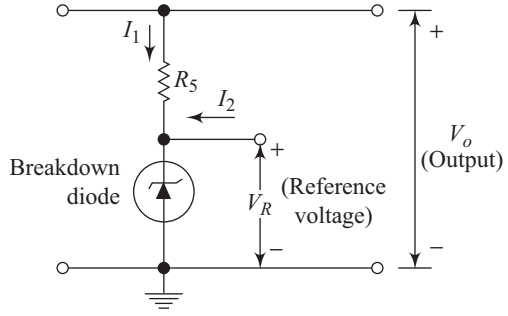


Fig. 14.17 Typical reference element

Assuming $I_1 \gg I_2$ and ΔI_1 to be a small quantity, we have

$$\frac{\Delta V_R}{\Delta V_o} \cong \frac{R_Z}{R_5 + R_Z} \tag{14.80}$$

where R_Z is the resistance of the breakdown diode.
Equation (14.80) shows that the change in reference voltage, ΔV_R, for a change in output voltage can be made very small by selecting a breakdown diode with a low resistance and by maintaining a nearly constant reverse current.

Comparison element

A basic comparison element that can be used for moderate output voltage is shown schematically in Fig. 14.18. The comparison element takes a sample of the generated output voltage and compares it with the reference element discussed in Section 14.3.1. A signal is then produced that is proportional to the difference of the two.

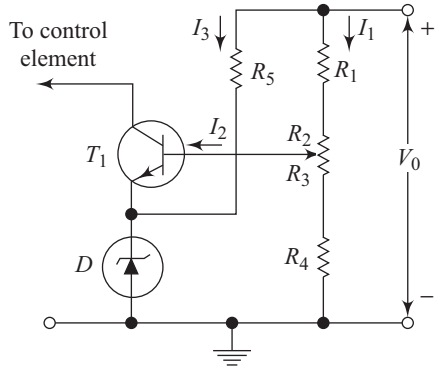


Fig. 14.18 Basic comparison element

DC amplifier

The dc amplifier is used to raise the magnitude of the difference signal from the comparison element to a sufficiently high level to drive the control element. At the same time, the gain of the amplifier must not be so high so as to affect the circuit stability.

Control element

The purpose of the control element is to interpret the signal from the dc amplifier and carry out the necessary adjustment to maintain a constant output voltage. Figure 14.19 shows a basic control element for a series regulator.

The base drive for the series regulator comes from the dc amplifier as shown in Fig. 14.17. Variations in the base drive produce corresponding changes in the output current I_o, thereby maintaining a nearly constant output voltage.

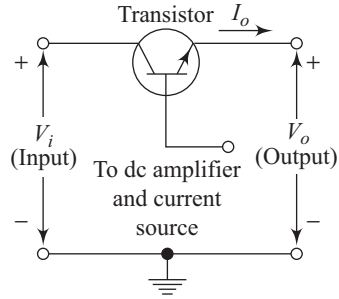


Fig. 14.19 Basic control element of a series regulator

14.4 Switched-mode Power Supply

A basic linear power supply is shown schematically in Fig. 14.20. The isolation transformer pushes up the frequency being fed to the rectifier. The rectifier usually consists of high surge-rated silicon diodes. The series-pass element indicated in Fig. 14.20 helps in maintaining the output dc voltage constant by changing its effective resistance as a response to the base signal as discussed in Section 14.3. The load current flows through the series-pass element and, for large differences between the input dc voltage and the output dc voltage, the power dissipated by the series-pass transistor can become appreciably large. This results in a low overall efficiency of the power supply.

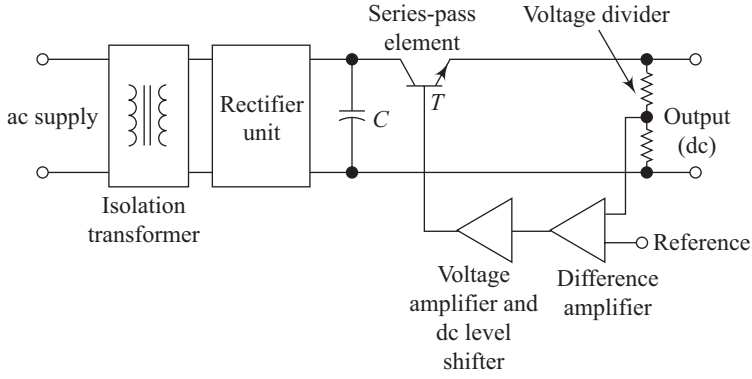


Fig. 14.20 Basic linear regulated power supply

The switched-mode power supply (SMPS) replaces the series-pass transistor with a low-loss switching element. Common options are bipolar power transistors and MOSFETs. With these devices operated in the switching mode, energy dissipation takes place only during switching on and off. Figure 14.21 shows the schematic diagram of a typical step-down SMPS. A signal is derived from the voltage-divider network and compared to a reference voltage. The voltage difference between the two is amplified and fed to a voltage controlled pulse-width modulator.

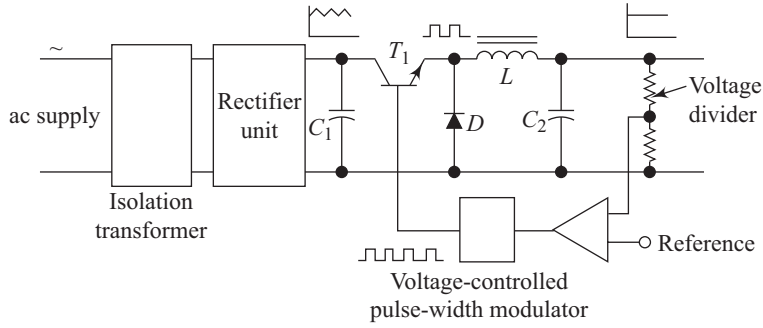


Fig. 14.21 Basic SMPS circuit

Any change in the voltage being fed to the pulse-width modulator leads to a corresponding change in the ratio of t_{ON} to the periodic time T. This produces corresponding switching of the switching transistor T_1, keeping the output voltage constant. Switching frequencies in SMPS are generally kept in the range of 3 to 50 kHz. The diode D in Fig. 14.21 maintains current through the inductor L and limits the induced voltage in the inductor due to high values of di/dt as the switching transistor turns off. For efficient operation, diode D should be capable of switching off in less than one microsecond, thereby necessitating the use of a fast recovery diode. Input ripple is less of a problem for SMPS as compared to a linear power supply. Therefore, a smaller input filter capacitor can be used in the circuit. Efficiency of a well-designed SMPS can reach 90 per cent and, therefore, heat dissipation is considerably less. This results in smaller heat sinks and cooling fans. The absence of resistors and use of smaller inductors and capacitors, helps in reducing the size of the overall packaging arrangement. SMPS circuits do suffer from self-induced noise and radio frequency interference (RFI). These can be maintained at acceptable levels by using suitable designs and input filters.

- The output voltage of a zener diode regulator cannot be chosen at random and is decided by the available zener diode breakdown voltages.
- Emitter–follower regulator and shunt regulator are two other frequently used regulator circuits.

Solved Problems

14.1 A half-wave rectifier circuit has input voltage of the form 100 cos (100t). The load resistance value is 4 kΩ. The diode used in the circuit can be represented by an ideal diode with a resistance of 1 kΩ in series. Calculate (a) maximum current, (b) dc component of current, and (c) rms value of current.

Solution

(a) Maximum value, I_m, of the current is given by

$$I_m = \frac{V_m}{(R_f + R_L)} \tag{14.1.1}$$

where, R_f is the resistance offered by the diode. Putting given values in Eqn (14.1.1) yields

$$I_m = \frac{100}{(1+4)\times 10^3}$$

$$= 20 \text{ mA}$$

(b) The dc value of current, I_{dc}, can be obtained using the equation

$$I_{dc} = \frac{I_m}{\pi} \tag{14.1.2}$$

Putting the evaluated value of I_m in Eqn (14.1.2) results in

$$I_{\text{dc}} = \frac{20 \times 10^{-3}}{\pi}$$

$$= 6.37 \times 10^{-3} \text{ A}$$

(c) The rms value of current, I_{rms}, is evaluated using

$$I_{\text{rms}} = \frac{I_m}{2} \tag{14.1.3}$$

which, on using the calculated value of I_m, gives us

$$I_{\text{rms}} = \frac{20 \times 10^{-3}}{2} = 10 \text{ mA.}$$

14.2 The load resistance of a full-wave rectifier is R_L = 1 kΩ. Diodes used in the circuit have idealized characteristics with a slope that corresponds to a resistance of 500 Ω. The input voltage applied to each diode is 200 sin (50*t*). Calculate (a) peak value of current, (b) dc value of current, (c) rms value of current, and (d) ripple factor.

Solution

(a) Peak value of current, I_m, is given by

$$I_m = \frac{V_m}{(R_f + R_L)} \tag{14.2.1}$$

Putting the given values in Eqn (14.2.1), results in

$$I_m = \frac{200}{(500 + 1000)}$$

$$= 0.133 \text{ A}$$

(b) The dc value of current, I_{dc}, can be evaluated using the equation

$$I_{\text{dc}} = \frac{2I_m}{\pi} \tag{14.2.2}$$

Putting the evaluated value of I_m into Eqn (14.2.2), results in

$$I_{\text{dc}} = \frac{2 \times 200}{\pi \times 1500}$$

$$= 0.0849 \text{ A}$$

(c) The rms value of current I_{rms} can be obtained using the equation

$$I_{\text{rms}} = \frac{I_m}{\sqrt{2}} \tag{14.2.3}$$

Substituting the evaluated value of I_m in Eqn (14.2.3) leads to

$$I_{\text{rms}} = \frac{200}{1500 \times \sqrt{2}}$$

$$= 0.094 \text{ A}$$

(d) Ripple factor, γ, is given by

$$\gamma = \sqrt{\left(\frac{I_{\text{rms}}}{I_{\text{dc}}}\right)^2 - 1} \tag{14.2.4}$$

Putting the values of I_{rms} and I_{dc} in Eqn (14.2.4), we get

$$\gamma = \sqrt{\left(\frac{0.094}{0.0849}\right)^2 - 1}$$

$$= 0.475$$

14.3 A π filter circuit has components with values $C_1 = 100$ μF, $C_2 = 50$ μF, $L = 5$ H, and $R_L = 500$ Ω. Evaluate the ripple factor of the filter if the operating frequency is 50 Hz.

Solution

The ripple factor γ is given by

$$\gamma \cong \frac{0.216}{R_L} \frac{1}{\omega^3 C_1 C_2 L} \tag{14.3.1}$$

Putting the given values in Eqn (14.3.1) leads to

$$\gamma = \frac{0.216}{500} \times \frac{1}{(2\pi \times 50)^3 \times 100 \times 10^{-6} \times 50 \times 10^{-6} \times 5}$$

$$= 0.00055$$

14.4 A voltage regulator is to be designed to provide an output voltage, V_o, of 25 V. The maximum current drain to the load is likely to be $I_{L(\max)} = 100$ mA. The input voltage being fed to the regulator is to vary between $V_{i(\min)} = 40$ V to $V_{i(\max)} = 70$ V. Determine the minimum power rating required for the zener diode to be used.

Solution

From Eqn (14.76), we know

$$0.8\left[\frac{V_{i(\min)} - V_Z}{V_{i(\max)} - V_Z}\right] - 0.2 > \frac{I_{L(\max)}}{I_{Z(\text{rated})}} \tag{14.4.1}$$

Putting the given values in Eqn (14.4.1) results in

$$0.8\left[\frac{40 - 25}{70 - 25}\right] - 0.2 > \frac{100 \times 10^{-3}}{I_{Z(\text{rated})}}$$

yielding

$$I_{Z(\text{rated})} > \frac{100 \times 10^{-3}}{0.067}$$

giving us

$$I_{Z(\text{rated})} > 1492.5 \text{ mA} \tag{14.4.2}$$

From Eqn (14.78), we have

$$0.6\, I_{Z(\text{rated})} > I_{L(\max)}$$

resulting in

$$I_{Z\,(\text{rated})} > \frac{100 \times 10^{-3}}{0.6}$$

or

$$I_{Z(\text{rated})} > 166.7 \text{ mA} \qquad (14.4.3)$$

From Eqs (14.4.2) and (14.4.3) it is clear that the minimum value of $I_{Z(\text{rated})}$ is

$$\text{Min. } I_{Z(\text{rated})} = 1492.5 \text{ mA}$$

Minimum power rating of the zener diode, $P_{\min}$ is thus

$$P_{\min} = V_Z I_{Z(\text{rated})}$$
$$= 25 \times 1492.5 \times 10^{-3} \cong 37.3 \text{ W}$$

Recapitulation

- For a half-wave rectifier

$$V_{\text{dc}} = \frac{V_m}{\pi} \text{ and } i_b = \frac{V_m \sin \omega t}{R_L} \text{ for } 0 \le \sin \omega t \le 1.$$

- For a full-wave rectifier

$$V_{\text{dc}} = \frac{2V_m}{\pi} \text{ and } I_{\text{rms}} = \frac{I_m}{\sqrt{2}}$$

- Conversion efficiency, η, of a full-wave rectifier is $\frac{8}{\pi^2} = 0.810$.
- Ripple factor, γ, is given by

$$\gamma = \sqrt{\left(\frac{I_{\text{rms}}}{I_{\text{dc}}}\right)^2 - 1}$$

- For a capacitance filter

$$V_{\text{dc}} = \frac{V_m}{\pi}\sqrt{1 + \omega^2 C^2 R^2}\,[1 - \cos(\theta_2 - \theta_1)]$$

and the ripple factor γ is

$$\gamma = \frac{\pi}{2\sqrt{3}\omega C R_L}$$

- For a π filter, ripple factor, γ, can be expressed in the form

$$\gamma \cong \frac{0.216}{R_L}\frac{X_{C_1} X_{C_2}}{X_L}$$

- For an RC filter the ripple factor, γ, is expressed as

$$\gamma = \frac{0.216}{R_L} \frac{1}{\omega^2 C_1 C_2 R_1}$$

- For a zener diode voltage regulator

$$0.8\left[\frac{V_{i(\min)} - V_Z}{V_{i(\max)} - V_Z}\right] - 0.2 > \frac{I_{L(\max)}}{I_{Z(\text{rated})}}$$

Exercises

Review Questions

14.1 Sketch the input and output waveforms for a half-wave rectifier.
14.2 Derive an expression for conversion efficiency of a full-wave rectifier.
14.3 Why is a bridge rectifier circuit sometimes preferred over a centre-tapped full-wave rectifier circuit?
14.4 What is ripple factor? Explain the concept using a suitable diagram.
14.5 Describe the operation of a shunt-capacitor filter circuit.
14.6 Sketch a π-filter circuit and explain the role of the major components.
14.7 Explain the operation of a zener diode regulator using a suitable circuit diagram.
14.8 Derive an expression for the minimum series resistance to be used in a zener diode regulator.
14.9 Why is a series voltage regulator called so?
14.10 Explain the term *sampling* in the context of series voltage regulator.
14.11 List some advantages of switched-mode power supplies.
14.12 Sketch the input and output waveforms for a full-wave rectifier.
14.13 Derive an expression for I_{dc} for a half-wave rectifier.
14.14 Compare PRV (peak reverse voltage) across a reverse-biased diode for a bridge rectifier and a centre-tapped full-wave rectifier circuit.
14.15 Derive an expression for the ripple factor of a full-wave rectifier.
14.16 Derive a mathematical expression for V_{dc} for a shunt-capacitor filter.
14.17 Explain the functioning of an RC filter circuit.
14.18 Draw a typical *I*-*V* curve for a zener diode showing the safe operating region of a zener diode regulator.
14.19 Derive the relation

$$0.8\left[\frac{V_{i(\min)} - V_Z}{V_{i(\max)} - V_Z}\right] - 0.2 > \frac{I_{L(\max)}}{I_{Z(\text{rated})}}$$

14.20 Explain the role of reference element and comparison element in a series regulator circuit.
14.21 State some disadvantages of an SMPS.
14.22 Draw a typical SMPS circuit and explain its operation.

Problems

14.1 The input voltage to a half-wave rectifier circuit is of the form, 120 cos (100 t). The load resistance value is 5 kΩ. The diode used in the rectifier circuit can be approximated by an ideal diode with a resistance of 500 Ω in series. Calculate (a) maximum current, (b) dc component of current, and (c) rms value of circuit.

$$\left[Hint: I_{\text{m}} = \frac{V_m}{(R_f + R_L)}, I_{\text{dc}} = \frac{I_m}{\pi}, I_{\text{rms}} = \frac{I_m}{2} \right]$$

Ans. 21.82 mA, 6.95 mA, 10.91 mA.

14.2 A half-wave rectifier is to generate a dc current, I_{dc}, of magnitude 5 mA. The input voltage waveform is 120 cos (100t) and the load resistance is 6 kΩ. Determine the resistance offered by the diode to achieve this result.

Ans. 2.64 kΩ.

14.3 A full-wave rectifier has an input waveform of the shape 100 sin (40t). The load resistance value is 1 kΩ and the effective diode resistance is 400 Ω. Determine (a) peak value of current, (b) dc value of current, (c) rms value of current, and (d) ripple factor.

$$\left[Hint: I_m = \frac{V_m}{(R_f + R_L)}, I_{\text{dc}} = \frac{2I_m}{\pi}, I_{\text{rms}} = \frac{I_m}{\sqrt{2}}, \text{and } \gamma = \sqrt{\left(\frac{I_{\text{rms}}}{I_{\text{dc}}} \right)^2 - 1} \right]$$

Ans. 0.071 A, 0.045 A, 0.05 A, 0.48.

14.4 The diode used in Problem 14.3 is replaced by an ideal diode with $R_f = 0$. Which of the four calculated quantities will remain unchanged? Show actual calculations to demonstrate the result.

Ans. Ripple factor remains unchanged; $I_m = 0.1$ A, $I_{\text{dc}} = 0.064$ A, $I_{\text{rms}} = 0.071$ A giving $\gamma = 0.48$.

14.5 A π filter is fabricated using components with the following values:
$C_1 = C_2 = 100$ μF, $L = 4$ H, and $R_L = 400$ Ω
Calculate the ripple factor of the filter at 50 Hz.

$$\left[Hint: \gamma \cong \frac{0.216}{R_L} \frac{1}{\omega^3 C_1 C_2 L} \right]$$

Ans. 0.00044

14.6 Calculate the ratio of ripple factor for a π filter when used in two countries having operating frequencies $f_1 = 50$ Hz and $f_2 = 60$ Hz.

Ans. $\dfrac{\gamma_{(f_1)}}{\gamma_{(f_2)}} = 1.728$

14.7 The inductance of the π filter of Problem 14.5 is replaced with a resistance $R_1 = 5$ Ω. Calculate the ripple factor of the corresponding RC filter.

$$\left[Hint: \gamma = \frac{0.216}{R_L} \frac{1}{\omega^2 C_1 C_2 R_1} \right]$$

Ans. 0.11

14.8 The output voltage of a regulator is to be 30 V. The maximum current to be passed through the load is $I_{L(\max)} = 100$ mA. The input voltage swing is likely to be between 50 V and 70 V. Calculate the minimum power rating required for the zener diode needed in the circuit.

$$\left[\textit{Hint:}\ 0.8\left[\frac{V_{i(\min)} - V_Z}{V_{i(\max)} - V_Z}\right] - 0.2 > \frac{I_{L(\max)}}{I_{Z(\text{rated})}} \right]$$

Ans. 15 W

14.9 Calculate the safe operating region of the zener diode used in the circuit indicated in Problem 14.8.
[*Hint:* Safe operating region is between 0.8 $I_{Z(\text{rated})}$ and 0.2 $I_{Z(\text{rated})}$]

Ans. 100 mA to 400 mA.

14.10 A half-wave rectifier circuit has input voltage of the form 50 cos(100t). The diode used in the circuit can be represented of 0.8 kΩ in series. If the maximum current required is 30 mA, calculate the load resistance.

Ans. 0.87 **kΩ**

14.11 The peak current, I_m in a full-wave rectifier circuit is 0.14 A. The load resistance $R_L = 1$ kΩ and the diodes used can be represented with an idealized resistance of 400 Ω. Calculate the amplitude of the input voltage.

Ans. 196 V

Important Formulae and Expressions

1. $E = h\nu$ (10.1)
2. $\nu = \frac{c}{\lambda}$ (10.2)
3. $I_\nu(x) = I_{\nu 0} e^{-\alpha x}$ (10.8)
4. $I = I_L - I_S\left[\exp\left(\frac{eV}{kT}\right) - 1\right]$ (10.11a)
5. $V_{\text{OC}} = \frac{kT}{e}\ln\left(1 + \frac{I_L}{I_S}\right)$ (10.13)
6. $FF = \frac{I_m V_m}{I_L V_{\text{OC}}}$ (10.25)
7. $J_S = \frac{eD_n n_{p0}}{L_n} + \frac{eD_p p_{n0}}{L_p}$ (10.2.1)
8. $L_n = \sqrt{D_n \tau_{n0}}$ (10.2.3)
9. $L_p = \sqrt{D_p \tau_{p0}}$ (10.2.4)
10. $\sigma = e[\mu_n(n_0 + \delta_n) + \mu_p(p_0 + \delta_p)]$ (10.30)
11. $v_d = \mu_n E$ (10.36)
12. $\theta_c = \sin^{-1}\left(\frac{\overline{n}_1}{\overline{n}_2}\right)$ (10.71)
13. $R_{\text{th}} = \frac{(T_2 - T_1)}{p}$ (11.8)
14. $C' = \frac{\varepsilon_{\text{ox}}}{t_{\text{ox}}}$ (12.8)
15. $I_{\text{rms}} = \left[\frac{1}{2\pi}\int_0^{2\pi} i^2 d(wt)\right]^{\frac{1}{2}}$ (14.13)
16. $\eta = \frac{P_{\text{dc}}}{P_{\text{ac}}}$ (14.19)
17. Ripple factor $= \frac{\text{effective value of ac components}}{\text{average or dc components}}$ (14.23)
18. $u = \sqrt{\frac{2Ve}{m}}$ (1.38)
19. $\left(p + \frac{a}{V^2}\right)(V - b) = RT$ (2.2)

20. $E = \dfrac{p^2}{2m}$ (3.2)

21. $E = \dfrac{\hbar^2}{2m} k^2$ (3.4)

22. $m_{\text{eff}} = \dfrac{\hbar^2}{d^2E/dk^2}$ (314)

23. $f(E) = \dfrac{1}{1 + e^{(E - E_F)/kT}}$ (3.24)

24. $N_c = 2\left[\dfrac{2\pi m_n^* kT}{h^2}\right]^{3/2}$ (3.31)

25. $N_v = 2\left[\dfrac{2\pi m_p^* kT}{h^2}\right]^{3/2}$ (3.36)

26. $n_0 p_0 = n_i^2$ (3.40)

27. $J = (nq\mu_{n+} pq\mu_p)E$ (3.68)

28. $\dfrac{D_n}{\mu_n} = \dfrac{D_p}{\mu_p} = \dfrac{kT}{e}$ (3.103)

29. $\dfrac{\partial p}{\partial t} = D_p \dfrac{\partial^2 p}{\partial x^2} - \mu_p\left(E\dfrac{\partial p}{\partial x} + p\dfrac{\partial E}{\partial x}\right) + G_p - \dfrac{p}{\tau_p}$ (4.66)

$\dfrac{\partial n}{\partial t} = D_n \dfrac{\partial^2 n}{\partial x^2} + \mu_n\left(E\dfrac{\partial n}{\partial x} + n\dfrac{\partial E}{\partial x}\right) + G_n - \dfrac{n}{\tau_n}$ (4.67)

30. $n_0 = n_i \exp\left[\dfrac{E_F - E_{Fi}}{kT}\right]$ (5.77)

31. $p_0 = n_i \exp\left[\dfrac{E_{Fi} - E_F}{kT}\right]$ (4.78)

32. $C(x, t) = C_S \operatorname{erfc} \dfrac{x}{2\sqrt{Dt}}$ (5.9)

33. $C(x, t) = \dfrac{Q}{\sqrt{\pi Dt}} e^{-x^2/4Dt}$ (5.13)

34. $V_{bi} = \dfrac{kT}{e} \ln\left[\dfrac{N_a N_d}{n_i^2}\right]$ (5.25)

35. $W = \left\{\dfrac{2\varepsilon_s V_{\text{bi}}}{e}\left[\dfrac{N_a + N_d}{N_a N_d}\right]\right\}^{1/2}$ (5.50)

36. $J = J_S\left[\exp\left(\dfrac{eV}{kT}\right) - 1\right]$ (6.34)

References

Anderson, R.L. 1962, *Solid-state Electronics*, vol. 5, No. 5, pp. 341–351.

Baliga, B.J. 1987, *Modern Power Devices*, Wiley, New York.

Baliga, B.J. 1996, *Power Semiconductor Devices*, PWS Publishing Co., Boston, Massachusetts.

Bhattacharya, Pallab 1994, *Semiconductor Optoelectric Devices*, Prentice Hall, Englewood Cliffs, New Jersey.

Chatterjee, R. 1988, *Microwave Engineering*, East–West Press, Delhi.

Cooper, William David 1982, *Electronic Instrumentation and Measurement Techniques*, Prentice Hall of India, Delhi.

Epifanov, G.I. 1979, *Solid State Physics*, Mir Publications, Moscow.

Esaki, L. 1958, *Physical Review*, vol. 109, p. 603.

Esaki, L. 1974, *Proceedings of the IEEE*, vol. 62, p. 825.

Fonash, S.J. 1981, *Solar Cell Device Physics*, Academic Press, New York.

Gandhi, S.K. 1977, *Semiconductor Power Devices: Physics of Operation and Fabrication Technology*, Wiley, New York.

Gandhi, S.K. 1983, *VLSI Fabrication Principles: Silicon and Gallium Arsenide*, Wiley, New York.

Grove, Andrew S. 1967, *Physics and Technology of Semiconductor Devices*, John Wiley and Sons, New York.

Hall, R.H. 1952, *Physical Review*, vol. 87, no. 2, p. 387.

Haynes, J.R. and W. Shockley 1951, *Physical Review*, vol. 81, no. 5, pp. 835–843.

Kingston, R.H. 1957, *Semiconductor Surface Physics*, University of Pennsylvania Press, Philadelphia.

Kittel, C. 1993, *Introduction to Solid State Physics*, Springer–Verlag, Berlin.

Lyshevski, Sergey Edward 2002, *MEMS and NEMS: Systems, Devices, and Structures*, CRC Press, Boca Raton, Florida.

Mathur, S.P., D.C. Kulshreshtha, and P.R. Chaddha 1980, *Electronic Devices Applications and Integrated Circuits*, Umesh Publications, New Delhi.

Mckelvey, J.P. 1993, *Solid State Physics for Engineering and Material Science*, Krieger, Malabar, Florida.

Millman and Halkias 1972, *Integrated Electronics: Analog and Digital Circuits and Systems*, McGraw–Hill Kogakusha, Tokyo.

Milnes, A.G. 1980, *Semiconductor Devices and Integrated Electronics*, Van Nostrand Reinhold, New York.

Milnes, A.G. and D.L. Feucht 1972, *Heterojunctions and Metal–Semiconduictor Junctions*, Academic Press, New York.

Moll, J.L. 1964, *Physics of Semiconductors*, McGraw–Hill, New York.

Moorthi, V.R. 2005, *Power Electronics: Devices, Circuits, and Industrial Applications*, Oxford University Press, New Delhi.

Neaman, Donald A. 2005, *Semiconductor Physics and Devices: Basic Principles*, Tata McGraw–Hill, Delhi.

Nicollian, E.H. and A. Goetzberger 1965, *Applied Physics Letters*, vol. 7, p.216.
Nicollian, E.H. and A. Goetzberger 1967, *Bell Systems Technology Journal*, vol. 46, p. 1055.
Nicollian, E.H. and J.R. Brees 1982, *MDS Physics and Technology*, Wiley, New York.
Pearman, Richard A. 1984, *Solid State Industrial Electronics*, Reston Publishing Company, Reston, Virginia.
Prager, H.J., K.K.N. Chang, and S. Weisbrod 1967, *Proceedings of the IEEE*, vol. 55, pp. 586–587.
Read, W.T. 1958, *Bell Systems Technology Journal*, vol. 37, p. 401.
Rideonbt, V.L. 1978, *Thin Solid Films*, vol. 48, no. 3, pp. 261–291.
Runyan, W.R. and K.E. Bean 1990, *Semiconductor Integrated Circuits Processing and Technology*, Addison Wesley, Boston, Massachusetts.
Ryder, John D. 1976, *Electronic Fundamentals and Applications: Integrated and Discreet Systems*, Prentice–Hall of India, Delhi.
Schroder, D.K. 1987, *Advanced MOS Devices*, Addison Wesley, Reading, Massachusetts.
Shockley, W. 1950, *Electrons and Holes in Semiconductors*, D. Van Nostrand, New York.
Shockley, W. and W.T. Read, Jr 1952, *Physical Review*, vol. 87, no. 5, pp. 835–842.
Shur, M. 1990, *Physics of Semiconductor Devices*, Prentice Hall, Englewood Cliffs, New Jersey.
Shur, M. 1996, *Introduction to Electronic Devices*, John Wiley and Sons, New York.
Singh, J. 1994, *Semiconductor Devices: An Introduction*, McGraw–Hill, New York.
Sonde, B.S. 1980 *Introduction to System Design Using Integrated Circuits*, Wiley Eastern, Delhi.
Streetman, Ben G. and Sanjay Banerjee 2002, *Solid State Electronic Devices*, Pearson Education Asia, Hong Kong.
Sze, S.M. 1988, *VLSI Technology*, McGraw–Hill, New York.
Sze, S.M. 2002, *Physics of Semiconductor Devices*, John Wiley and Sons, New York.
Tay, Francis E.H. (Ed.) 2002, *Materials and Process Integration for MEMS*, Kluwer Academic Publishers, Norwell, Massachusetts.
Van der Ziel, A. 1968 *Solid State Physical Electronics*, Prentice Hall, New York.

Appendix A

IMPORTANT PHYSICAL CONSTANTS

Quantity	Symbol	Value
Angstrom unit	Å	$1\ Å = 10^{-4}\ \mu m = 10^{-8}$ cm
Avogadro constant	N_{AVO}	6.02204×10^{23} mol^{-1}
Bohr radius	a_B	0.52917 Å
Boltzmann constant	k	1.38066×10^{-23} J/K (R/N_{AVO})
Elementary charge	q	1.60218×10^{-19} C
Electron rest mass	m_0	0.91095×10^{-30} kg
Electron volt	eV	$1\ eV = 1.60218 \times 10^{-19}$ J $= 25.053$ kcal/mol
Gas constant	R	1.98719 cal mol^{-1} K^{-1}
Permeability in vacuum	μ_0	1.25663×10^{-8} H/cm ($4\pi \times 10^{-9}$)
Permittivity in vacuum	ε_0	8.85418×10^{-14} F/cm ($1/\mu_0 c^2$)
Planck's constant	h	6.62617×10^{-34} J s
Reduced Planck's constant	$\hbar$	1.05458×10^{-34} J s ($h/2\pi$)
Proton rest mass	M_p	1.67264×10^{-27} kg
Speed of light in vacuum	c	2.99792×10^{10} cm/s
Standard atmosphere		1.01325×10^5 N/m^2
Thermal voltage at 300 K	kT/e	0.0259 V
Wavelength of 1 eV quantum	λ	1.23977 μm

Appendix B

IMPORTANT LATTICE CONSTANTS

Compound type	Element or compound	Name	Crystal structure	Lattice constant at 300 K (Å)
Element	C	Carbon (diamond)	Diamond	3.56683
	Ge	Germanium	Diamond	5.64613
	Si	Silicon	Diamond	5.43095
	Sn	Grey Tin	Diamond	6.48920
IV–IV	SiC	Silicon carbide	Wurtzite	$a = 3.086$, $c = 15.117$
III–V	AlAs	Aluminium arsenide	Zincblende	5.6605
	AlP	Aluminum phosphide	Zincblende	5.4510
	AlSb	Aluminum antimonide	Zincblende	6.1355
	BN	Boron nitride	Zincblende	3.6150
	BP	Boron phosphide	Zincblende	4.5380
	GaAs	Gallium arsenide	Zincblende	5.6533
	GaN	Gallium nitride	Wurtzite	$a = 3.189$, $c = 5.185$
	GaP	Gallium phosphide	Zincblende	5.4512
	GaSb	Gallium antimonide	Zincblende	6.0959
	InAs	Indium arsenide	Zincblende	6.0584
	InP	Indium phosphide	Zincblende	5.8686
	InSb	Indium antimonide	Zincblende	6.4794
II-VI	CdS	Cadmium sulphide	Zincblende	5.8320
	CdS	Cadmium sulphide	Wurtzite	$a = 4.16$, $c = 6.756$
	CdSe	Cadmium selenide	Zincblende	6.050
	CdTe	Cadmium telluride	Zincblende	6.482
	ZnO	Zine oxide	Rock salt	4.580
	ZnS	Zinc sulphide	Zincblende	5.420
	ZnS	Zinc sulphide	Wurtzite	$a = 3.82$, $c = 6.26$
IV–VI	PbS	Lead sulphide	Rock salt	5.9362
	PbTe	Lead telluride	Rock salt	6.4620

Appendix C

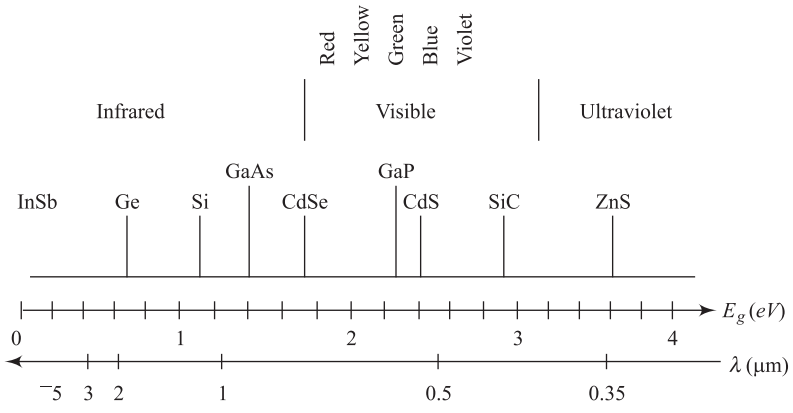
PROPERTIES OF SOME COMMON SEMICONDUCTORS

Semiconductor		Band gap (eV)		Mobility at 300 K ($cm^2/V\ s$)[1]			Effective Mass m^*/m_0		
		300 K	0 K	Electrons	Holes	Band[2]	Elec.	Holes	$\varepsilon_s/\varepsilon_0$
Element	C	5.47	5.48	1800	1200	i	0.2	0.25	5.7
	Ge	0.66	0.74	3900	1900	i	1.64[3]	0.64[5]	16.0
							0.082[4]	0.28[6]	
	Si	1.12	1.17	1500	450	i	0.98[3]	0.16[5]	11.9
							0.19[4]	0.49[6]	
	Sn		0.082	1400	1200	d			
IV–V	α-SiC	2.996	3.03	400	50	i	0.60	1.00	10.0
III–V	AlSb	1.58	1.68	200	420	i	0.12	0.98	14.4
	BN	~7.5							
	BP	2.0							
	GaN	3.36	3.50	380			0.19	0.60	12.2
	GaSb	0.72	0.81	5000	850	d	0.042	0.40	15.7
	GaAs	1.42	1.52	8500	400	d	0.067	0.082	13.1
	GaP	2.26	2.34	110	75	i	0.82	0.60	11.1
	InSb	0.17	0.23	80000	1250	d	0.0145	0.40	17.7
	InAs	0.36	0.42	33000	460	d	0.023	0.40	14.6
	InP	1.35	1.42	4600	150	d	0.077	0.64	12.4
II–VI	CdS	2.42	2.56	340	50	d	0.21	0.80	5.4
	CdSe	1.70	1.85	800		d	0.13	0.45	10.0
	CdTe	1.56		1050	100	d			10.2
	ZnO	3.35	3.42	200	180	d	0.27		9.0
	ZnS	3.68	3.84	165	5	d	0.40		5.2
IV–VI	PbS	0.41	0.286	600	700	i	0.25	0.25	17.0
	PbTe	0.31	0.19	6000	4000	i	0.17	0.20	30.0

[1]The values are for drift mobilities obtained in the purest and most perfect materials available to date.
[2]i = indirect, d = direct.
[3]Longitudinal effective mass.
[4]Transverse effective mass.
[5]Light-hole effective mass.
[6]Heavy-hole effective mass.

Appendix D

BAND GAPS OF SOME SEMICONDUCTORS RELATIVE TO THE OPTICAL SPECTRUM



Appendix E

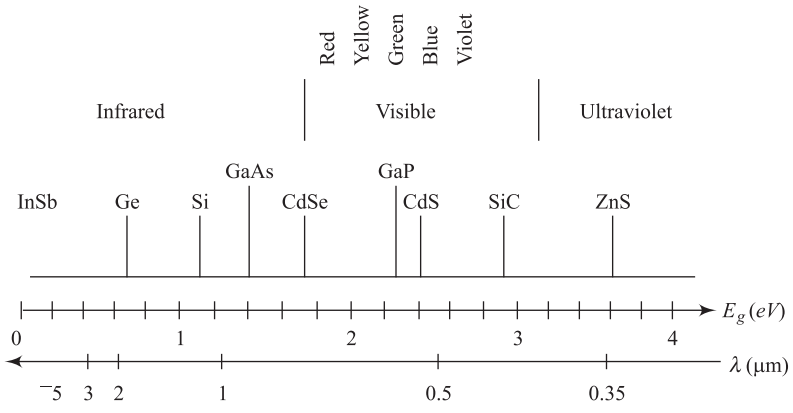
PROPERTIES OF SILICON, GERMANIUM, AND GALLIUM ARSENIDE AT 300 K

Properties	Si	Ge	GaAs
Atoms/cm^3	5.0×10^{22}	4.42×10^{22}	4.42×10^{22}
Atomic weight	28.09	72.60	144.63
Breakdown field (V/cm)	$\sim 3 \times 10^5$	$\sim 10^5$	$\sim 4 \times 10^5$
Crystal structure	Diamond	Diamond	Zincblende
Density (g/cm^3)	2.328	5.3267	5.32
Dielectric constant	11.9	16.0	13.1
Effective density of states in conduction band, N_C (cm^{-3})	2.8×10^{19}	1.04×10^{19}	4.7×10^{17}
Effective density of states in valence band, N_V (cm^{-3})	1.04×10^{19}	6.0×10^{18}	7.0×10^{18}

(*Contd*)

Appendix D

BAND GAPS OF SOME SEMICONDUCTORS RELATIVE TO THE OPTICAL SPECTRUM



Appendix E

PROPERTIES OF SILICON, GERMANIUM, AND GALLIUM ARSENIDE AT 300 K

Properties	Si	Ge	GaAs
Atoms/cm^3	5.0×10^{22}	4.42×10^{22}	4.42×10^{22}
Atomic weight	28.09	72.60	144.63
Breakdown field (V/cm)	$\sim 3 \times 10^5$	$\sim 10^5$	$\sim 4 \times 10^5$
Crystal structure	Diamond	Diamond	Zincblende
Density (g/cm^3)	2.328	5.3267	5.32
Dielectric constant	11.9	16.0	13.1
Effective density of states in conduction band, N_C (cm^{-3})	2.8×10^{19}	1.04×10^{19}	4.7×10^{17}
Effective density of states in valence band, N_V (cm^{-3})	1.04×10^{19}	6.0×10^{18}	7.0×10^{18}

(*Contd*)

(*Contd*)

Properties	Si	Ge	GaAs
Effective Mass, m^*/m_0			
Electrons	$m_l^* = 0.98$	$m_l^* = 1.64$	0.067
	$m_l^* = 0.19$	$m_l^* = 0.082$	
Holes	$m_{lh}^* = 0.16$	$m_{lh}^* = 0.044$	$m_{lh}^* = 0.082$
	$m_{hh}^* = 0.49$	$m_{hh}^* = 0.28$	$m_{lh}^* = 0.45$
Electron affinity, χ(V)	4.05	4.0	4.07
Energy gap (eV) at 300 K	1.12	0.66	1.424
Intrinsic carrier concentration (cm^{-3})	1.45×10^{10}	2.4×10^{13}	1.79×10^6
Intrinsic Debye length (μm)	24	0.68	2250
Intrinsic resistivity (Ω cm)	2.3×10^5	47	10^8
Lattice constant (Å)	5.43095	5.64613	5.6533
Linear coefficient of thermal expansion, $\Delta L/L\Delta T$ ($°C^{-1}$)	2.6×10^{-6}	5.8×10^{-6}	6.86×10^{-6}
Melting point (°C)	1415	937	1238
Minority carrier lifetime (s)	2.5×10^{-3}	10^{-3}	~10^{-8}
Mobility (drift) (cm^2/V s)	1500	3900	8500
	450	1900	400
Optical-phonon energy (eV)	0.063	0.037	0.035
Phonon mean free path λ_0 (Å)	76 (electron)	105	58
	55 (hole)		
Specific heat (J/g °C)	0.7	0.31	0.35
Thermal conductivity at 300 K (W/cm°C)	1.5	0.6	0.46
Thermal diffusivity (cm^2/s)	0.9	0.36	0.44
Vapour pressure (Pa)	1 at 1650°C	1 at 1330°C	100 at 1050°C
	10^{-6} at 900°C	10^{-6} at 760°C	1 at 900°C

Appendix F

IMPORTANT PROPERTIES OF SI3N4 AND SIO2 AT 300K

Insulator	Si_3N_4	SiO_2
Structure	Amorphous	Amorphous
Melting point (°C)	—	~1600
Density (g/cm^3)	3.1	2.2
Refractive index	2.05	1.46
Dielectric constant	7.5	3.9
Dielectric strength (V/cm)	10^7	10^7
Infrared absorption band (μm)	11.5–12.0	9.3
Energy gap (eV)	~5.0	9
Thermal-expansion coefficient ($°C^{-1}$)	—	5×10^{-7}
Thermal conductivity (W/cm) K	—	0.014
dc resistivity (Ω cm)		
at 25°C	$\sim 10^{14}$	$10^{14} - 10^{16}$
at 500°C	$\sim 2 \times 10^{13}$	—

Appendix G

TABLE OF THE ERROR FUNCTION

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, dt$$

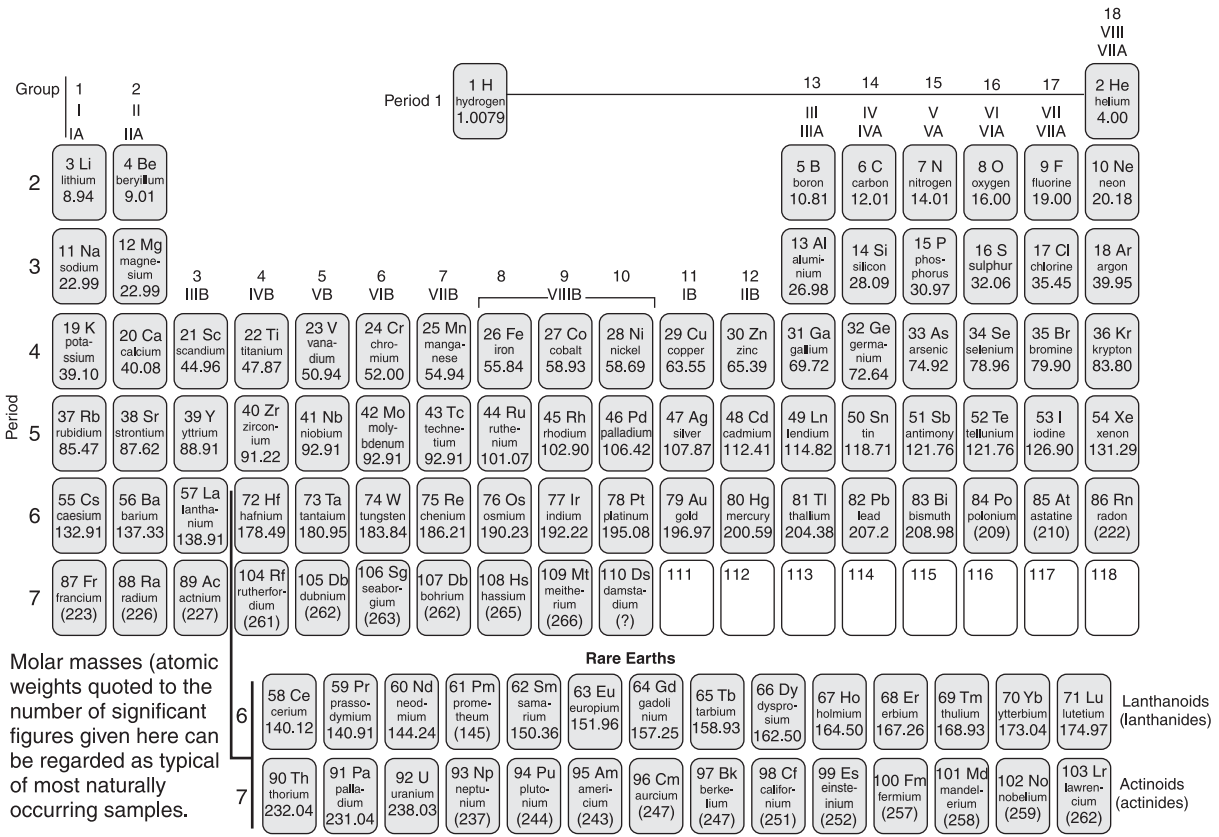
$$\text{erf}(0) = 0 \quad \text{erf}(\infty) = 1$$

$$\text{erfc}(z) = 1 - \text{erf}(z)$$

z	**erf(z)**	z	**erf(z)**
0.00	0.00000	1.00	0.84270
0.05	0.05637	1.05	0.86244
0.10	0.11246	1.10	0.88021
0.15	0.16800	1.15	0.89612
0.20	0.22270	1.20	0.91031
0.25	0.27633	1.25	0.92290
0.30	0.32863	1.30	0.93401
0.35	0.37938	1.35	0.94376
0.40	0.42839	1.40	0.95229
0.45	0.47548	1.45	0.95970
0.50	0.52050	1.50	0.96611
0.55	0.56332	1.55	0.97162
0.60	0.60386	1.60	0.97635
0.65	0.64203	1.65	0.98038
0.70	0.67780	1.70	0.98379
0.75	0.71116	1.75	0.98667
0.80	0.74210	1.80	0.98909
0.85	0.77067	1.85	0.99111
0.90	0.79691	1.90	0.99279
0.95	0.82089	1.95	0.99418
1.00	0.84270	2.00	0.99532

Appendix H

THE PERIODIC TABLE OF ELEMENTS



Molar masses (atomic weights quoted to the number of significant figures given here can be regarded as typical of most naturally occurring samples.

Appendix I

INTERNATIONAL SYSTEM OF UNITS

Quantity	Unit	Symbol	Dimension
Length	metre	m	
Mass	kilogram	kg	
Time	second	s	
Temperature	kelvin	K	
Current	ampere	A	
Frequency	hertz	Hz	1/s
Force	newton	N	kg m/s^2
Pressure	pascal	Pa	N/m^2
Energy	joule	J	N m
Power	watt	W	J/s
Electric charge	coulomb	C	A s
Potential	volt	V	J/C
Conductance	siemens	S	A/V
Resistance	ohm	Ω	V/A
Capacitance	farad	F	C/V
Magnetic flux	weber	Wb	V s
Magnetic flux density	tesla	T	Wb/m^2
Inductance	henry	H	Wb/A

INDEX

C

D

E

F

G

M

Q