UNIT 1: SEMICONDUCTOR BASICS

UNIT STRUCTURE

1.1 Learning Objectives
1.2 Introduction
1.3 Bonding forces and Energy bands
    1.3.1 Conductor
    1.3.2 Insulator
    1.3.3 Semiconductors
1.4 Semiconductor Types
    1.4.1 Direct and Indirect Semiconductors
    1.4.2 Electrons and Holes
    1.4.3 Intrinsic and Extrinsic Semiconductors
1.5 Charge Transport and Mobility in Semiconductors
    1.5.1 Conductivity and Mobility
    1.5.2 Drift and Resistance
    1.5.3 Effects of Temperature and Doping on Mobility
    1.5.4 Hall Effect
1.6 Let Us Sum Up
1.7 Answers to Check Your Progress
1.8 Further Reading
1.9 Model Questions

1.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- explain the concept of bonding forces and energy bands in conductor, insulator and semiconductor.
- discuss about the types of semiconductors
- familiarize with the carrier transport of semiconductors.

1.2 INTRODUCTION

This is the first unit of the course on Introduction to Semiconductor Devices. This unit will provide you the fundamental concepts of semiconductor materials and behavior. It covers bonding forces and energy band theory and explains the classification of materials into conductors, insulators, and semiconductors. The unit also provides different types of semiconductors, including direct and indirect semiconductors, electrons and holes, and intrinsic and extrinsic materials. Key aspects of charge transport and mobility are discussed, including conductivity, drift, resistance, and the effects of temperature and doping on carrier mobility. The unit concludes with the Hall effect, an important technique for analyzing semiconductor properties.

1.3 BONDING FORCES AND ENERGY BANDS

An atom is the basic unit of matter. It has a central nucleus made up of protons and neutrons, and electrons revolve around it in different orbits. Protons carry a positive charge, electrons carry a negative charge, and neutrons have no charge. The number of protons in an atom determines its atomic number and the type of element. Atoms are electrically neutral because they have equal numbers of protons and electrons. Electrons occupy different energy levels or shells, with each shell having a maximum capacity.

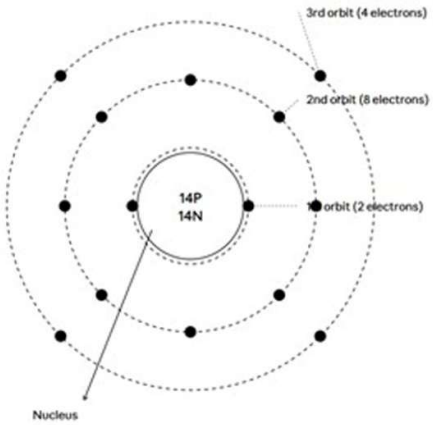


Figure 1.1 (a) Atomic structure of Si (14)

The figure 1.1 (a) depicts the atomic structure and energy levels of silicon (Si) atoms. In the atomic structure of silicon, electrons revolve around the nucleus in discrete orbits (shells), each having a fixed number of electrons. These orbits represent specific energy levels, and electrons can only exist in these permitted levels, not in between. The energy of electrons increases as they move to outer orbits. The maximum number of electrons that the orbits can have follows the following mathematical formula-

$$e = 2n^2$$

where, n = count of the orbit (orbits are counted from the inner side)

e = maximum number of electrons that the orbit can have

It means that for the first orbit, n=1 so e=2, so it cannot have more than 2 electrons. Similarly, the second orbit can have a maximum number of 8 electrons; the third one can have 18 and so on. When the number of electrons and protons are same then they neutralise each other's charge so as a whole the atom is neutral. If the case is such that the outermost orbit of the atom contains more electrons than the fixed value, then the atom is called negatively charged and if the number of electron is less in the outer most orbits than the maximum value then the atom is called positively charged. The idea of energy levels explains that electrons closer to the nucleus possess lower energy, whereas electrons farther away in outer shells have higher energy.

Bonding forces are the forces of attraction that hold atoms together to form molecules and solids. These forces determine the physical and electrical properties of materials.

There are two main types of bonding forces: primary (chemical) bonds and secondary (molecular) bonds. Primary bonds are strong and include ionic bonds which is formed by transfer of electrons, covalent bonds formed by sharing of electrons, and metallic bonds formed by a sea of free electrons around positive ions. Secondary bonds are weaker and include van der Waals forces, which arise due to temporary dipole interactions between atoms or molecules. These bonding forces play an important role in determining whether a material behaves as a conductor, semiconductor, or insulator.

1.3.1 Conductor

Conductors are materials that allow electric current to flow very easily due to the presence of a large number of free electrons. They have very low resistivity**,** typically in the range of 10^{-8} to 10^{-6} Ω·m and hence very high conductivity**,** around 10^{6} to 10^{8} S/m. This means electrons can move freely with minimal opposition when an electric field is applied. Common examples include metals like copper and aluminum, which are widely used in electrical wiring and circuits.

The energy band diagram explains how electrical conduction varies in different materials based on the arrangement of energy bands. In solids, individual atomic energy levels combine to form two main bands: the valence band and the conduction band. The valence band is the highest energy band occupied by electrons under normal conditions and contains bound electrons, while the conduction band lies above it and contains free electrons that can move and responsible for electrical conduction.
The separation between these bands is called the band gap or the forbidden energy gap (E_g)**,** where no electron states exist. Its size determines whether a material behaves as a conductor, semiconductor, or insulator.

In a conductor, the valence band and conduction band overlap, meaning there is no forbidden energy gap between them. As a result, electrons can move freely even at very low temperatures, leading to high electrical conductivity. The conduction band contains a large number of free electrons, and even a small applied electric field can cause these electrons to drift easily. Since the valence band is either partially filled or overlaps with the conduction band, electrons do not require additional energy to participate in conduction. This overlapping of bands is the key reason why metals are good conductors of electricity.

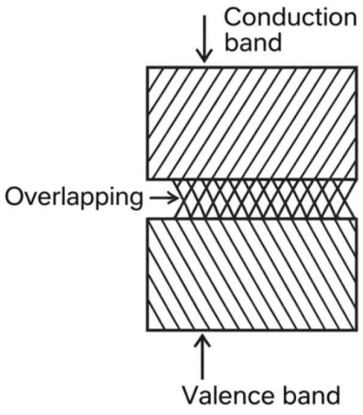


Figure 1.2 (a) Energy band diagram of Conductor

1.3.2 Insulator

Insulators are materials that strongly resist the flow of electric current because they have very few or no free charge carriers. Their resistivity is extremely high**,** ranging from 10^7 to 10^{16} Ω·m, and their conductivity is very low, in the range between 10^{-16} and 10^{-7} S/m. As a result, electrons cannot move easily through them. Materials such as rubber, glass, and plastic are common insulators and are used to protect and isolate conductive components in electrical systems.

In insulators, the band gap is very large (about 5 eV or more), so electrons cannot easily move to the conduction band, resulting in no conduction. In metals, the valence and conduction bands overlap, allowing electrons to move freely and giving high conductivity.

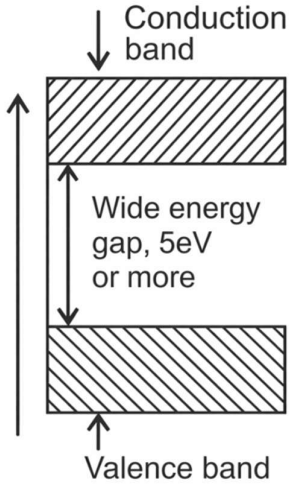


Figure 1.2 (b) Energy band diagram of Insulator

1.3.3 Semiconductor

Semiconductors have electrical properties that lie between conductors and insulators. Their resistivity ranges from about 10^{-5} to 10^3 Ω·m**,** while their conductivity varies from 10^{-3} to 10^5 S/m. Unlike conductors, their conductivity can be controlled by factors such as temperature and doping. Materials like silicon and germanium are example semiconductors, and they form the foundation of electronic devices such as diodes, transistors, and integrated circuits.

Semiconductor is a material whose electrical properties lie in between the conductor and insulator. Two usually used semiconductors are Silicon (Si) and Germanium (Ge). They require very small energy to release an electron from their valance band. For Ge, it is 0.7 ev and for Si 1.1 ev. The resistivity of semiconductor is in the range of 0.4 to 2500 Ωm. In terms of energy band, they have partially filled valence band and nearly empty conduction band. The for bidden gap is very small (1ev) that is why smaller electric field is required to push an electron from

the valence band to the conduction band. In semiconductors, the band gap is small (around 1 eV), so at room temperature some electrons gain enough energy to jump from the valence band to the conduction band, enabling moderate conductivity. As temperature increases, more electrons cross the band gap, increasing conductivity in semiconductors.

At 0 K, the valence band is completely filled and the conduction band is totally empty. So at absolute zero temperature, a semiconductor behaves like an insulator. When the temperature increases, the forbidden gap is decreased so that few electrons can be injected into the conduction band. This means, the conductivity of semiconductors increases with temperature. That is why they have negative temperature coefficient of resistance. At 300 K the semiconductor acts like a conductor.

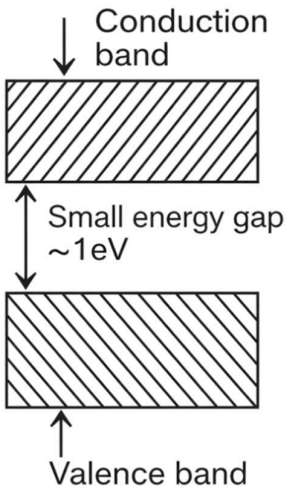


Figure 1.2 (c) Energy band diagram of Semiconductor

1.4 SEMICONDUCTOR TYPES

Semiconductors are materials whose electrical conductivity lies between that of conductors and insulators, and their behaviour can be controlled by factors such as temperature, light, and impurities. Semiconductors are classified based on band structure and impurity doping. Based on their energy band structure, semiconductors are classified as direct and indirect semiconductors. Again, based on their purity and composition, semiconductors are mainly classified into intrinsic and extrinsic types. This classification helps in understanding how charge carriers i.e. electrons and holes are generated and how conductivity can be modified for different applications. Now, let us discuss the different types of semiconductors in detail below.

1.4.1 Direct and Indirect Semiconductor

A direct semiconductor is a material in which the lowest point of the conduction band and the highest point of the valence band lie at the same momentum (k-value) in the energy–momentum diagram.

Because the two energy levels occur at the same k-value, the electron can transition between them without any change in momentum. When an electron in the conduction band loses energy, it can directly drop into the valence band and recombine with a hole.

During this recombination process, the energy difference between the two bands is released in the form of a photon (light). Since no change in momentum is required, the transition occurs without the involvement of phonons or any additional particles. As a result, the process is highly efficient for light emission. Example of direct semiconductor is Gallium Arsenide (GaAs). It is used in LEDs and laser diodes because they efficiently convert electrical energy into light. In basic terms, direct semiconductors are excellent light emitters because electrons can transition straight across the band gap without any extra assistance.

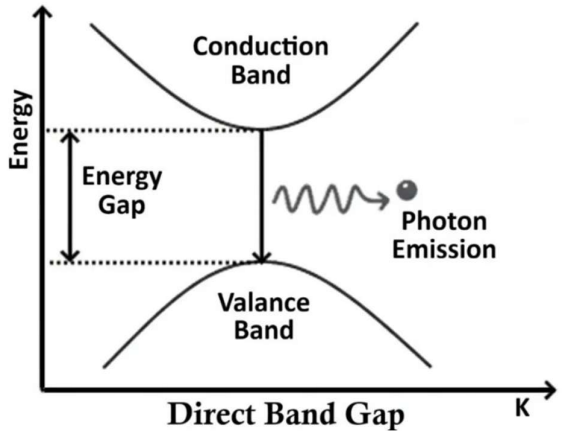


Figure 1.3 (a) Direct band gap semiconductor

An indirect semiconductor is a material in which the minimum of the conduction band and the maximum of the valence band occur at different momentum (k-values) in the energy–momentum diagram. Because these points are not aligned, an electron cannot directly transition from the conduction band to the valence band; instead, it must undergo a change in momentum, which requires the assistance of a phonon (a lattice vibration). During this process, the electron interacts with a phonon to conserve momentum, and the energy released is predominantly in the form of heat rather than light, with only a very small fraction emitted as photons. Typical examples of such materials include silicon (Si) and germanium (Ge), which are widely used in transistors and integrated circuits (ICs). In simple terms, indirect semiconductors are inefficient light emitters because electron recombination requires phonon assistance which make the process less favorable for light production.

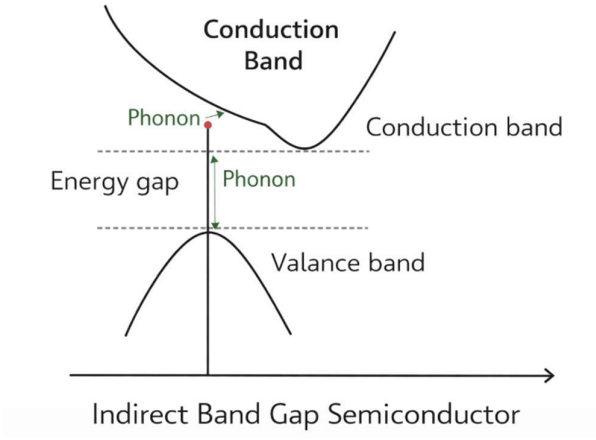


Figure 1.3 (b) Indirect band gap semiconductor

1.4.2 Electrons and Holes

In semiconductors, an electron is a negatively charged particle that occupies energy levels within the material. When sufficient energy (thermal or electrical) is supplied, some electrons gain enough energy to move from the valence band to the conduction band. These free electrons can move easily under an applied electric field and contribute to electrical conduction. In n-type semiconductors, electrons are the majority charge carriers and dominate the current flow.

A hole is created when an electron leaves its position in the valence band, resulting in a vacancy that behaves like a positive charge. Although it is not a physical particle, a hole effectively moves as nearby electrons shift to fill the vacancy. This movement appears as the flow of positive charge. In p-type semiconductors, holes are the majority carriers and play a key role in conduction. Both electrons and holes together determine the electrical behaviour of semiconductors.

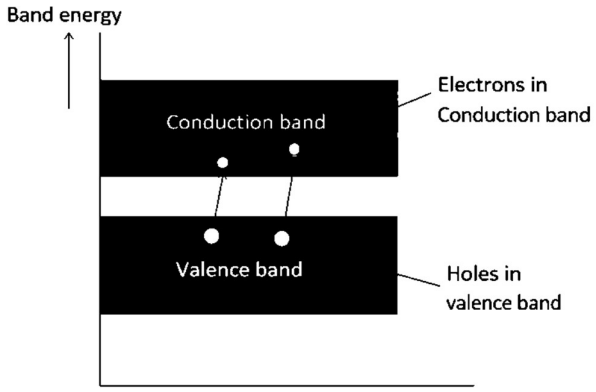


Figure 1.4: Energy Band Diagram of Electrons and Holes

1.4.3 Intrinsic and Extrinsic Semiconductor

Semiconductors are classified into two types. They are–

- Intrinsic or Pure semiconductors.
- Extrinsic or Doped Semiconductor.

The extrinsic semiconductors are also two types.

- N-type and
- P-type.

Intrinsic Semiconductors: The semiconductor material in its pure form is called intrinsic semiconductor. Pure Silicon and Germanium are the intrinsic semiconductors. Each atom of the semiconductor has only four electrons in its outermost orbit. They require four more electrons in the valence shell to become a stable structure. Hence each of the four valence electrons forms a covalent bond with another four valance electrons, one from each of the four neighbouring atoms. The two electrons of the covalent bond are shared by the two atoms.

At absolute zero temperature, all valence electrons are strongly bounded to the parent atoms; hence no free electrons are available for conduction. The intrinsic semiconductor acts like a perfect insulator at absolute zero temperature.

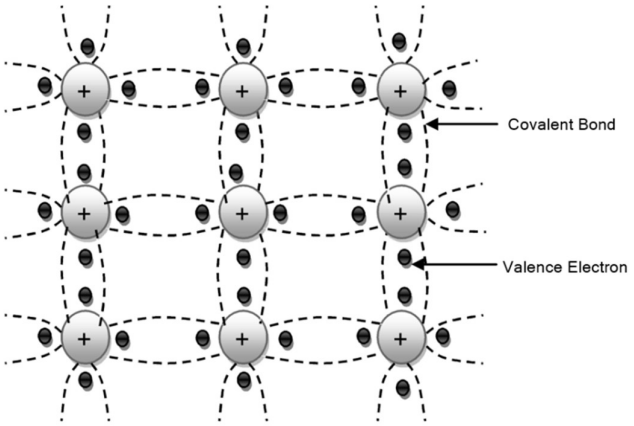


Figure1.5 (a): Structure of intrinsic semiconductor at absolute zero temperature

At room temperature (300K) some of the covalent bonds of the semi- conductor breaks due to thermal energy and valence electrons of the atom moves away from the nucleus. They become free to move in the crystal. When the electron moves away by breaking a covalent bond, a vacancy is created in that covalent bond which is called a hole. For every free electron, a hole is generated. Hence in intrinsic semiconductors the number of free electrons is equal to the number of holes. A hole is a positively charged particle.

Since, an electron is a negatively charged and a hole is positively charged, when a potential difference is applied in an intrinsic semiconductor, electrons will move towards the positive terminal and the holes will flow towards the negative terminal of the battery. Therefore, the total current in the intrinsic semiconductor is the sum of currents due to free electrons and holes.

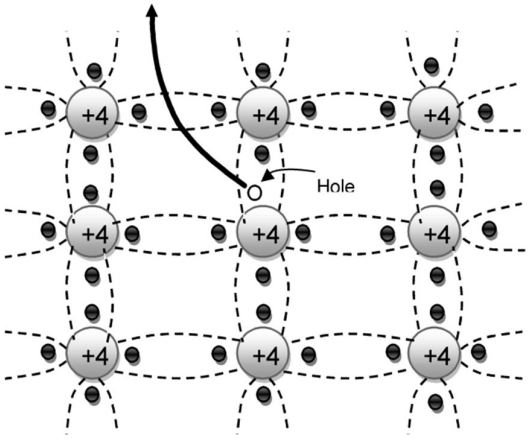
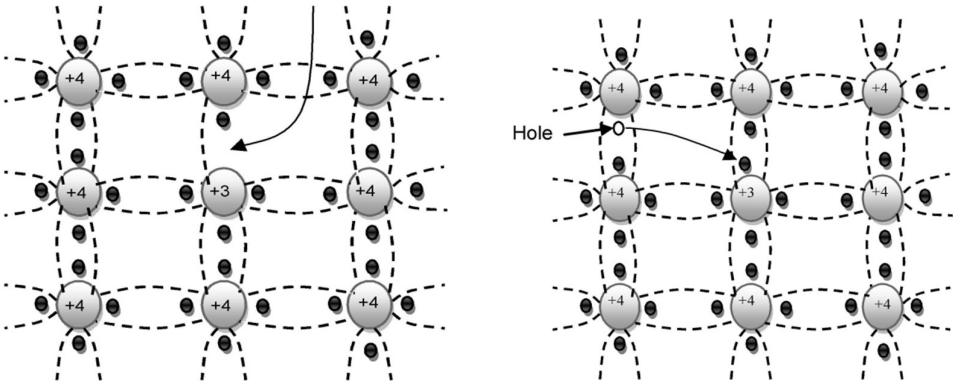


Figure 1.5 (b): Generation of electron-hole pair in an intrinsic semiconductor

Extrinsic Semiconductors: The conductivity of intrinsic semi- conductor can be increased by adding some suitable impurity atoms to the semiconductor. When a small amount of impurity atom is added to the intrinsic semiconductor, it is called extrinsic or impurity semi-conductor and the process of adding impurities to the semiconductor is known as doping. The conductivity of semiconductor increases means either the number of free electrons or holes increases in the semiconductor crystal. When a trivalent impurity atom is added to the semiconductor, a large number of holes are produced in the semiconductor. On the other hand, addition of pentavalent impurity atoms creates a large number of electrons in the semiconductor crystal.
Depending upon the type of impurity added, extrinsic semiconductors are classified into two types. They are P–type and N-type.

P-type Semiconductor: When a trivalent impurity atom is added to an intrinsic semiconductor, it is called P-type semiconductor. Some typical examples of trivalent impurities are Boron, Aluminum, Gallium and Indium. They are known as acceptor impurities because the holes created can accept the electrons.

Let us consider that a Boron impurity is added to pure Silicon atom. Since the boron atom has three valence electrons, three covalent bonds will be formed with the three neighbouring Silicon atoms. The fourth covalent bond is incomplete as Boron has no valence electron to share with Silicon atom to form the covalent bond. This missing electron is called a hole. Therefore, for each Boron atom added, one hole is created. To fill this vacancy, electrons from adjacent covalent bond can jump due to thermal energy by creating another hole in the adjacent bond. This hole is positively charged and the Boron atom has a negative ionas it accepts the electron. Thus, a P-type semiconductor has holes in majority and free electrons in minority in addition, negative ions.



a) Boron added to Silicon b) Creation of a hole

Figure 1.6(a): P-type Semiconductor

N-type Semiconductor: When a pentavalent impurity atom is added to an intrinsic semiconductor, it is called N-type semiconductor. Examples of pentavalent atoms are phosphorus, arsenic and antimony. They are called donor impurities because they donate one valence electron to the semiconductor crystal.

We consider a phosphorous impurity atom is added to a silicon atom. The phosphorous atom has five valence electrons, four of these form covalent bonds with four neighbouring silicon atoms. The fifth electron is excessive and loosely bound to the nucleus. Hence, it is easily energized from the valence band to the conduction band by increase in the thermal energy. Thus, at room temperature each impurity atom donates one electron to the conduction band and achieved a positive ion. Due to the thermal energy, there will also some electron-hole pair generated because of the breaking of the covalent bonds. The number of free electrons generated by the pentavalent impurity is very large than the number of holes. Hence, an N-type semiconductor has electrons as majority carriers and holes as minority carriers, in addition positive ions.

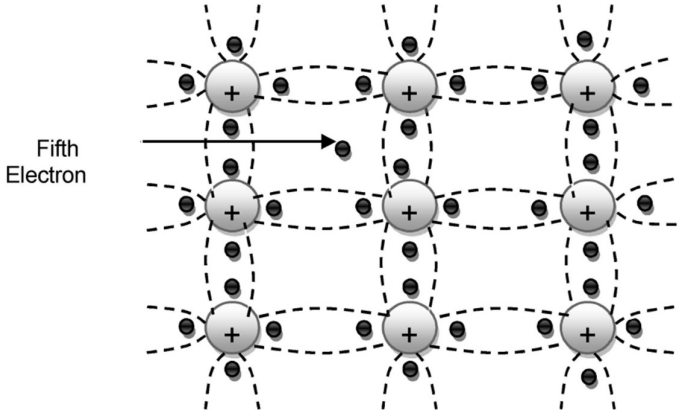


Figure1.6(b): N-type semiconductor

1.5 CHARGE TRANSPORT AND MOBILITY IN SEMICONDUCTORS

Charge transport in semiconductors occurs through the movement of electrons and holes under an applied electric field. These charge carriers acquire a drift velocity due to the field, resulting in an electric current. Mobility describes how easily the carriers move through the material and determines the conductivity of the semiconductor.

1.5.1 Conductivity and Mobility

The conductivity of a material is related to the concentration of free electrons. When a constant electric field E is applied to a metal, the free electrons are accelerated and acquire velocity. However, due to frequent collisions with atoms, the electrons do not accelerate indefinitely; instead, they attain a steady average velocity called the drift velocity. The drift velocity is proportional to the applied electric field and acts in a direction opposite to it. It is given by:

$$v_d = \mu E \qquad (1)$$

where μ is the mobility of electrons in m^2/volt.sec.

Thus, the steady drift motion of electrons is superimposed on their random thermal motion. This directed motion results in an electric current. If the number of free electrons per unit volume is n and q is the charge of each electron, then the current density J is given by:

$$J = nqv_d \qquad (2)$$

Substituting the value of v_d in Equation (2),

$$J = nq\mu E \qquad (3)$$

Since current density is also expressed as,

$$J = \sigma E \qquad (4)$$

Where σ is the electrical conductivity of the material.

$$\sigma = nq\mu \qquad (5)$$

From equation (4), we can write,

$$\sigma = \frac{J}{E} \qquad (6)$$

Thus, Conductivity is the property of a material that measures its ability to conduct electric current. It is defined as the ratio of current density to the applied electric field. It can also be expressed as given in equation (6). A material has high conductivity if it has many charge carriers and/or high mobility.

From Equation (1), we have, $\mu = \frac{v_d}{E}$ (7)

Thus, mobility is the ability of charge carriers (electrons or holes) to move through a material when an electric field is applied. It tells us how quickly the carriers drift in response to the field. Higher mobility means charge carriers move more easily through the material.

The relation between conductivity and mobility is that conductivity is directly proportional to mobility. Even with a small number of charge carriers, high mobility can lead to significant conductivity.

1.5.2 Drift and Resistance

Drift is the directional motion of charge carriers (electrons or holes) in a material when an electric field is applied. Under this field, carriers gain an average velocity called the drift velocity, which causes an electric current to flow. Despite their random thermal motion, this steady movement in response to the field is what generates usable current.

The drift velocity is given by:

$$v_d = \mu E$$

where μ is the mobility of electrons/holes in m^2/volt.sec and E is the electric field in volt per metre (V/m).

Resistance is the property of a material that opposes the flow of electric current. It quantifies how much a material resists the movement of charge carriers. The higher the resistance, the lower the current for a given voltage. Resistance depends on the material's nature, length, cross-sectional area, and temperature.

Resistance R is defined as the ratio of applied voltage V to the resulting current I.

$$R = \frac{V}{I}$$

For a uniform conductor of length L and cross-sectional area A, resistance can be expressed as:

$$R = \rho \frac{L}{A}$$

where ρ is the resistivity of the material. Since conductivity σ is the reciprocal of resistivity ($\sigma = \frac{1}{\rho}$)

$$R = \frac{L}{\sigma A}$$

The Figure 1.7 illustrates the relationship between drift velocity, electric current, and resistance in a conductor. When a voltage is applied across the conductor, an electric field (E) is established from the positive to the negative terminal, causing free electrons to move in the opposite direction. This motion of electrons is called drift, and although individual electrons move slowly, their collective motion produces an electric current (I) flowing through the conductor. As electrons move, they collide with atoms and impurities within the material, which opposes their motion; this opposition is known as resistance (R). The resistor symbol in the diagram represents this obstruction to current flow. Thus, the figure shows that current arises due to the drift of charge carriers under an electric field, while resistance limits the magnitude of this current.

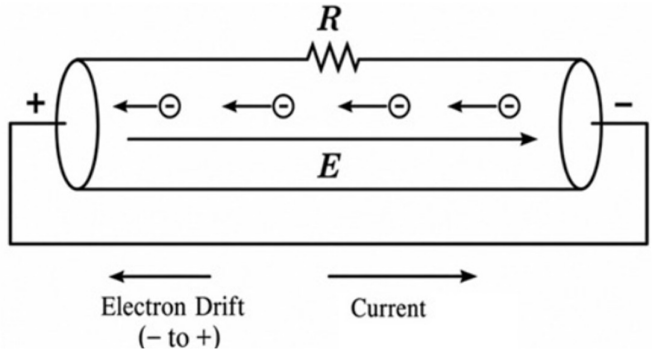


Figure 1.7: Representation of Drift Velocity, Electric Current, and Resistance in a Conductor

1.5.3 Effects of temperature and doping on mobility

The mobility of charge carriers in a semiconductor is strongly affected by both temperature and doping concentration. As temperature increases, the lattice atoms vibrate more intensely which causes more frequent collisions between charge carriers (electrons and holes) and the lattice. This lattice (phonon) scattering reduces the mobility, so carrier mobility generally decreases with increasing temperature. However, at very low temperatures, impurity scattering dominates, and mobility may initially increase with temperature before decreasing at higher temperatures.

Doping also influences mobility by introducing impurity atoms into the semiconductor. As the doping concentration increases, the number of ionized impurity centers rises, leading to more impurity scattering of charge carriers. This reduces their ability to move freely, thereby decreasing mobility. Thus, lightly doped semiconductors have higher mobility, while heavily doped semiconductors exhibit lower mobility due to increased collisions with impurity ions.

1.5.4 Hall Effect

When a current-carrying material (a metal or a semiconductor) is placed in a magnetic field that is perpendicular (transverse) to the direction of the current, an electric field is generated inside the material. This induced electric field is perpendicular to both the current and the magnetic field. This effect is called hall effect where an electric field appears perpendicular to both the current and the applied magnetic field.

The figure 1.8 illustrates the Hall effect in a semiconductor bar, where the current (I) flows along the X-axis (left to right) and the magnetic field (B) is applied along the Z-axis, perpendicular to the current. Due to the presence of the magnetic field, the moving charge carriers experience a Lorentz force ($F = q(v \times B)$), which deflects them toward one side of the bar depending on their type. This deflection leads to the accumulation of charges on opposite faces by making one side positively charged and the other negatively charged. As a result, a transverse electric field is established along the Y-axis, generating a measurable potential difference known as the Hall voltage.

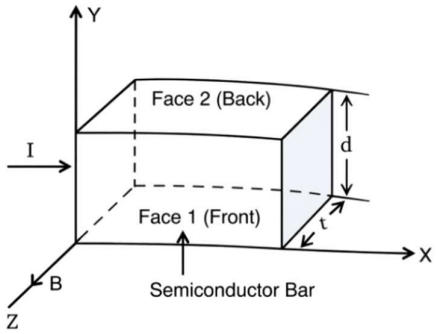


Figure 1.8: Hall Effect in a Semiconductor Bar

The Hall Effect is used for the following purposes:

- to identify the type of semiconductor (N-type or P-type) based on the dominant charge carriers.
- to determine the carrier concentration, since the Hall voltage depends on the number of charge carriers.
- to measure conductivity and calculate carrier mobility using Hall Effect and conductivity measurements.

WCHECK YOUR PROGRESS

Q1. Why do semiconductors have moderate conductivity compared to conductors and insulators?
Q2. Why are direct band gap semiconductors used in LEDs but indirect ones are not?
Q3. What happens to conductivity of a semiconductor when temperature increases? Why?
Q4. Why does mobility decrease at high temperatures?
Q5. How does doping improve conductivity in semiconductors?
Q6. Why are holes considered positive charge carriers?
Q7. What will happen if a semiconductor is heavily doped?
Q8. How does the Hall effect help identify carrier type?
Q9. What is drift and how is it related to resistance?
Q10. Why is silicon preferred over germanium in electronics?

21.6 LET US SUM UP

- Atomic bonding in solids leads to the formation of energy bands that determine electrical properties.
- Conductors have overlapping bands allowing free electron movement and high conductivity.
- Insulators have a large band gap preventing electron flow, resulting in very low conductivity.
- Semiconductors have a small band gap with controllable conductivity via temperature and doping.

- Semiconductors are classified based on band structure and impurity doping.
- Direct and indirect semiconductors are defined based on their band structure, where a direct semiconductor has the conduction band minimum and valence band maximum at the same momentum allowing direct electron transition.
- An indirect semiconductor has the conduction band minimum and valence band maximum at different momentum, requiring a phonon for transition.
- Electrons and holes are charge carriers responsible for current flow in semiconductors.
- Intrinsic are pure, while extrinsic are doped to enhance conductivity.
- Charge transport explains how carriers move and contribute to current.
- Conductivity depends on carrier concentration and their mobility.
- Drift is carrier motion under an electric field that causes electrical resistance.
- Mobility decreases with increasing temperature and doping due to scattering.
- The Hall effect measures carrier type and concentration using a magnetic field.

K 1.7 ANSWERS TO CHECK YOIUR PROGRESS

Answer to Q1. Because they have a small band gap, allowing limited charge carrier generation that can be controlled.

Answer to Q2. Direct semiconductors emit light efficiently, while indirect ones lose energy as heat.

Answer to Q3. Conductivity increases because more electrons gain energy to jump to the conduction band.

Answer to Q4. Due to increased lattice vibrations causing more scattering of charge carriers.

Answer to Q5. It increases the number of charge carriers (electrons or holes).

Answer to Q6. Because the absence of an electron behaves like a positive charge in current flow.

Answer to Q7. Its conductivity increases but mobility decreases due to impurity scattering.

Answer to Q8. The polarity of Hall voltage indicates whether carriers are electrons or holes.

Answer to Q9. Drift is the motion of charge carriers under an applied electric field, and resistance is the opposition offered by a material to this drift motion, limiting the flow of current.

Answer to Q10. Because silicon has better thermal stability and lower leakage current.

O 1.8 FURTHER READING

- Neamen, D. A. (2021). *Semiconductor physics and devices* (4th ed.). McGraw Hill Education (India).
- Streetman, B. G., & Banerjee, S. K. (2015). *Solid state electronic devices* (7th ed.). Pearson India.
- Sze, S. M., & Lee, M. K. (2021). *Semiconductor physics and devices* (3rd ed., Indian adaptation). Wiley India.

- Boylestad, R. L., & Nashelsky, L. (2012). *Electronic devices and circuit theory* (11th ed.). Pearson

M 1.9 MODEL QUESTIONS

Q1. Explain the classification of semiconductors and discuss intrinsic and extrinsic semiconductors in detail.
Q2. Discuss the P type and N type semiconductor.
Q3. Explain direct and indirect band gap semiconductors with the help of energy band diagrams and compare their properties.
Q4. Discuss the generation and recombination of electron-hole pairs in semiconductors.
Q5. Derive the relation between current, drift velocity, and charge carrier density.

*** ***** ***

UNIT 2: P-N JUNCTION DIODE

UNIT STRUCTURE

2.1 Learning Objectives
2.2 Introduction
2.3 Overview of the PN Junction
2.4 Forward and Reverse Biased Junction
2.5 I-V characteristics of the Diode
2.6 Reverse Bias Breakdown
2.6.1 Zener Breakdown
2.6.2 Avalanche Breakdown
2.7 Let Us Sum Up
2.8 Answers to Check Your Progress
2.9 Further Reading
2.10 Model Questions

2.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- explain the formation and basic structure of a PN junction
- distinguish between forward bias and reverse bias in a PN junction
- describe the current–voltage relationship of a diode
- differentiate between different types of breakdown mechanisms in a diode.

2.2 INTRODUCTION

In the previous unit, we learned the basics of semiconductors. This unit introduces the basic concepts of a PN junction and its role in semiconductor devices. It explains how a PN junction is formed, the behavior of charge carriers, and the operation of a diode under forward and reverse bias conditions. The unit also discusses the I–V characteristics of a diode and important parameters such as cut-in voltage, forward current, and reverse saturation current. Finally, it focuses on the concept of reverse bias breakdown, including the mechanisms of Zener breakdown and Avalanche breakdown, and explains how these phenomena occur in semiconductor devices.

2.3 OVERVIEW OF THE PN JUNCTION

A PN junction is a fundamental structure in semiconductor physics and forms the basis of many electronic devices such as diodes, transistors, and integrated circuits. It is created by developing p-type and n-type semiconductor regions within a single crystalline semiconductor material, such as silicon or germanium. The electrical properties of the PN junction depend on the behavior of charge carriers and the equilibrium established at the interface between the p-type and n-type regions. The electrical behavior of a diode is determined by the properties of this PN junction.

When a PN junction is formed, three key phenomena occur simultaneously and in a mutually consistent way. They are diffusion, formation of depletion region and built in electric field.

In a p-type semiconductor, the majority charge carriers are holes, produced by doping the semiconductor with acceptor impurities such as boron. On the other hand, an n-type semiconductor contains electrons as the majority carriers, introduced by donor impurities such as phosphorus or arsenic. When the p-type and n-type regions are brought into contact, electrons from the n-side move to the p-side, and holes from the p-side move to the n-side, just because of the difference in concentration. Thus a concentration gradient exists across the junction. As a result, electrons from the n-region diffuse into the p-region, while holes from the p-region diffuse into the n-region. Therefore, diffusion is the process by which charge carriers, such as electrons and holes, move from a region of higher concentration to a region of lower concentration. This is similar to how the smell of perfume spreads across a room or how ink spreads evenly in water.

When a PN junction is formed, carriers diffuse and recombine near the junction. Due to the recombination process, it leaves behind immobile ions (positively charged donor ions on the n-side and negatively charged acceptor ions on the p-side). That means, this process leaves behind positively charged donor ions on the n-side and negatively charged acceptor ions on the p-side near the junction. This region near the junction where lacking of free charge carriers, is called the depletion region (DR) or the Space Charge Region (SCR). It acts as a barrier that prevents further diffusion of electrons and holes and creates an internal electric field across the junction. So, the depletion region is the area around the PN junction where free charge carriers (electrons and holes) are absent**.** The width of the depletion region depends on the doping level of impurities in the p-type and n-type semiconductors. As the doping level increases, the depletion region becomes narrower. This occurs because a highly doped PN junction contains a larger number of electrons and holes, which leads to faster recombination near the junction.

Due to the presence of immobile positive and negative ions on the either sides of the junction, an internal electric field is created across the depletion region. This electric field opposes the further movement of charge carriers and establishes equilibrium in the junction. It is called the built-in electric field. Due to the built-in electric field potential difference (voltage) developed across the depletion region which is known as barrier potential or junction potential. This is called as barrier potential because it acts as a barrier to oppose the flow of majority charge carriers across the junction. The barrier potential for Silicon is about 0.7 volt and for Germanium its value is 0.3 volt at room temperature of 300 K. Figure 2.1 shows the formation of depletion region in PN junction.

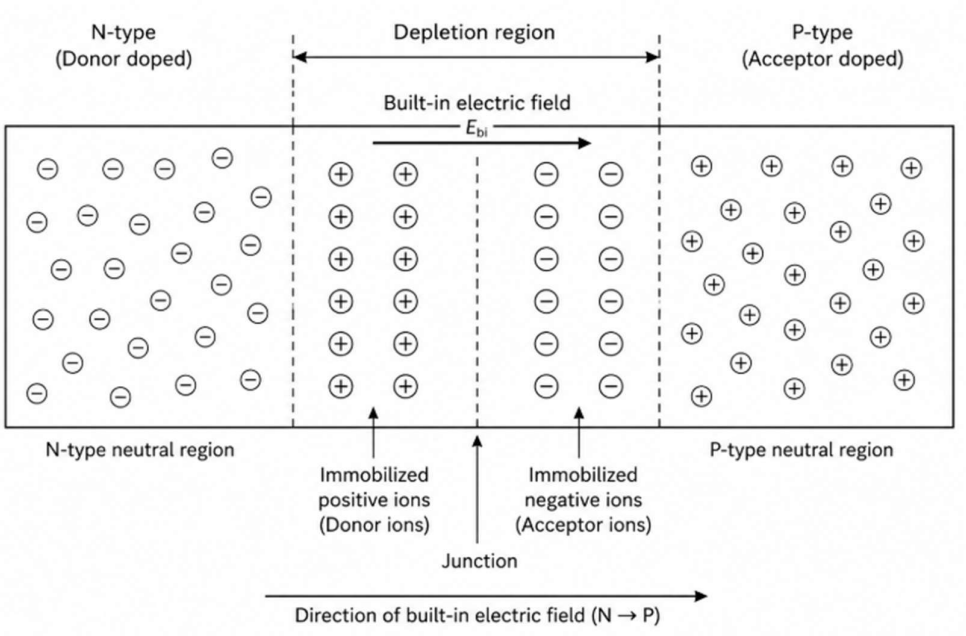


Figure 2.1: Depletion region formation in PN Junction

The built-in electric field present in the depletion region causes minority carriers to move across the junction, resulting in drift current. Thus, the drift current is the current produced due to the motion of charge carriers under the influence of an electric field. Whereas, diffusion current is the current produced due to the movement of charge carriers from a region of higher concentration to a region of lower concentration.

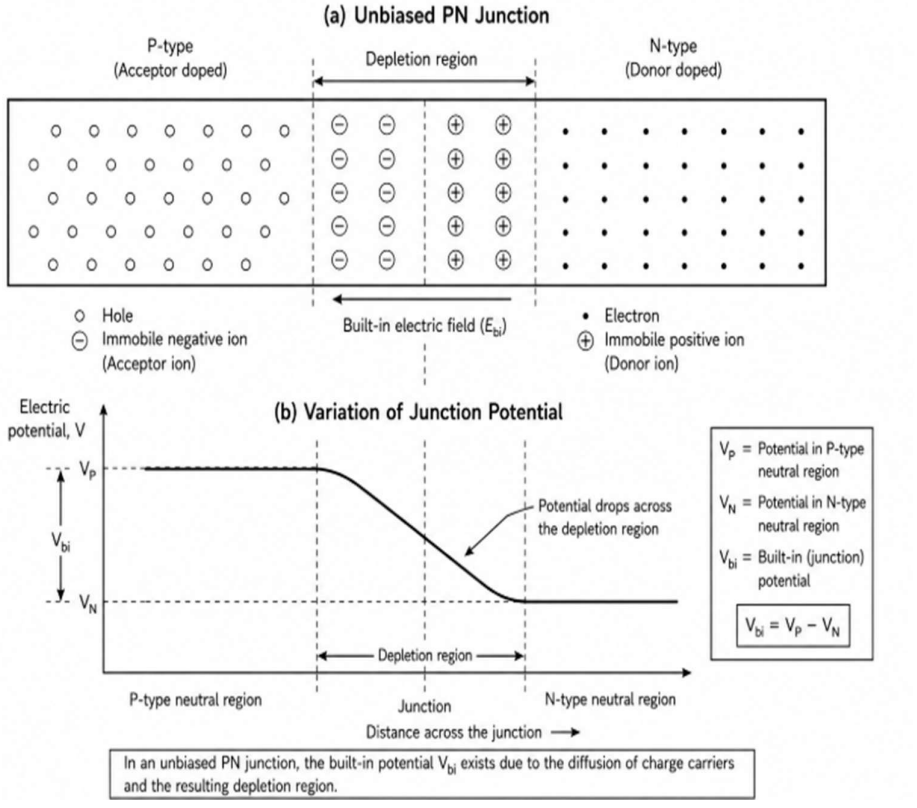


Figure 2.2: Unbiased PN Junction and variation of junction potential

2.3.1 Energy Diagrams of PN Junction

In a PN junction, the valence band and conduction band energy levels differ slightly between the P-type and N-type materials due to their distinct atomic structures. The valence and conduction bands in the N-type region lie at lower energy levels than those in the P-type region at thermal equilibrium. This occurs due to the built-in potential formed by the diffusion of electrons and holes across the PN junction. The resulting depletion region creates an electric field from the N-side to the P-side, causing the energy bands to bend. At equilibrium, the Fermi level remains constant throughout the junction, establishing the equilibrium band alignment. The energy band diagram of the PN junction at the moment of its formation illustrates this difference in Figure 2.3.

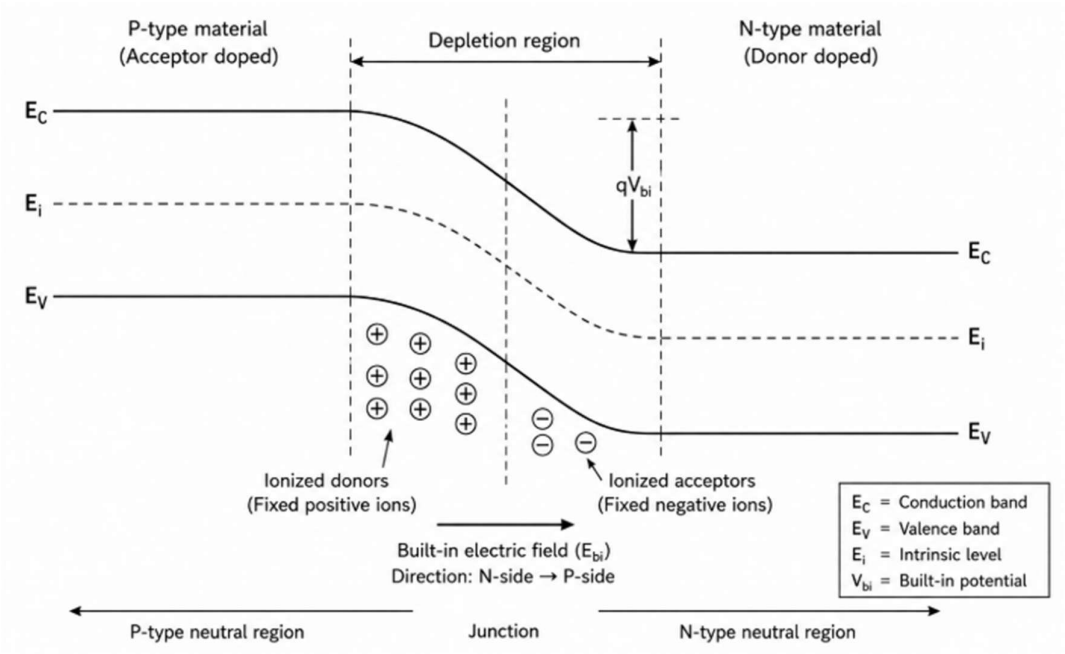


Figure 2.3: Energy diagram of PN Junction at the moment of its formation

In the P-region, the majority carriers, i.e., holes, occupy the valence band, while the thermally generated minority carriers, i.e., electrons, occupy the conduction band. In the N-region, the majority carriers, i.e., electrons, are present in the conduction band, whereas the minority carriers, i.e., holes, occupy the valence band.

2.4 FORWARD AND REVERSE BIAS PN JUNCTION

When a dc voltage source (battery) is connected across a PN junction, the process is called biasing of a PN junction. By applying an external voltage source, the width of the depletion region can be controlled and hence the current conduction through the PN junction.

A PN junction can be biased in two ways:

- Forward Bias
- Reverse Bias

Forward Biasing of PN Junction

In forward biasing of a PN junction, the positive terminal of the battery is connected to the P-side and the negative terminal is connected to the N-side. In this condition, holes in the P-region are repelled by the positive terminal of the battery and move towards the junction, while electrons in the N-region are repelled by the negative terminal and also move towards the junction. As a result, some holes and electrons enter the depletion region, which leads to a decrease in the potential barrier. Consequently, the width of the depletion region is reduced. Because the depletion layer becomes thinner, a large number of majority charge carriers are able to cross the junction, resulting in the flow of a large current through the PN junction. Thus, in forward bias, the depletion layer decreases, the potential barrier reduces, and the current through the junction increases.

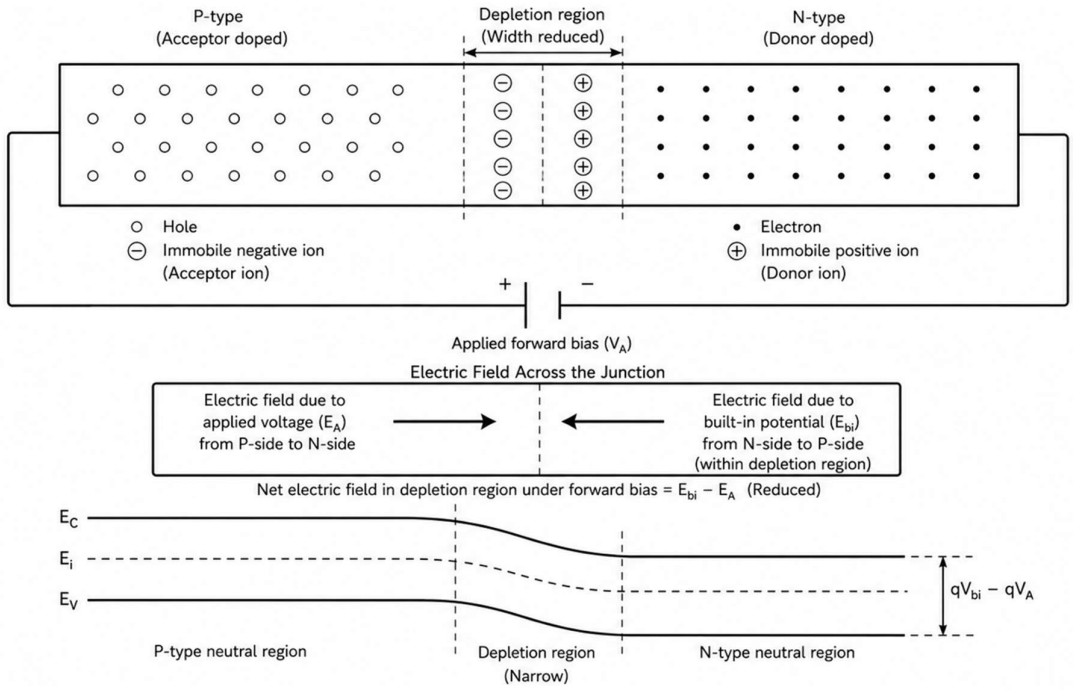


Figure 2.4: Forward Biased PN Junction

Reverse Biasing of PN Junction

In reverse biasing of a PN junction, the positive terminal of the battery is connected to the N-side and the negative terminal is connected to the P-side. In this condition, holes in the P-region are attracted toward the negative terminal of the battery and move away from the junction, while electrons in the N-region are attracted toward the positive terminal and also move away from the junction. As a result, the majority charge carriers move away from the junction region, causing the depletion region to widen. Consequently, the potential barrier across the junction increases and the width of the depletion layer becomes larger. Due to the widening of the depletion region, the flow of majority carriers across the junction is prevented and only a very small current, known as the reverse saturation current due to minority carriers, flows through the junction. Thus, in reverse bias, the depletion layer increases, the potential barrier increases, and only a very small current flows through the PN junction.

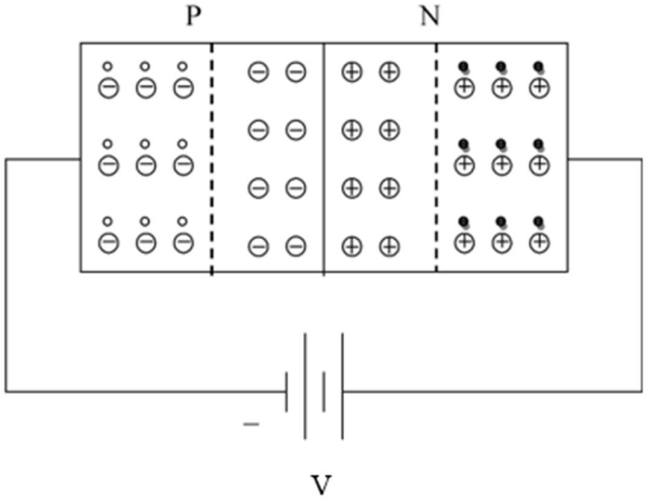


Figure 2.5: Reverse Biased PN Junction

2.5 I-V CHARACTERISTICS OF THE DIODE

A PN junction diode is a two-terminal semiconductor device formed by joining P-
type and N-type semiconductor materials together that allows current to flow in one direction. It has two terminals: Anode (P-side) and Cathode (N-side). Figure 2.6 represents the symbol of PN junction diode.

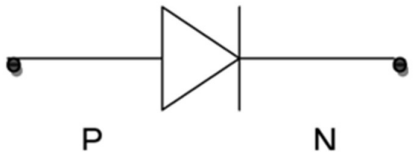


Figure 2.6 Symbol of PN Junction Diode

The behavior of a P–N junction diode can be studied by observing the relationship between the voltage applied across the diode and the current flowing through it. This relationship is represented graphically by the voltage–current (V–I) characteristics of the P–N junction diode. The V–I characteristics provide important information about how the diode responds under different biasing conditions, such as forward bias and reverse bias. By analyzing this graph, we can understand the conduction properties of the diode and the conditions under which it allows or restricts the flow of electric current.

The graph between the voltage applied across the two terminals of the diode, namely the anode and cathode, and the current that flows through the P–N junction is known as the V–I characteristics of a P–N junction diode. In this graph, the applied voltage is represented along the X-axis, while the corresponding current is represented along the Y-axis.

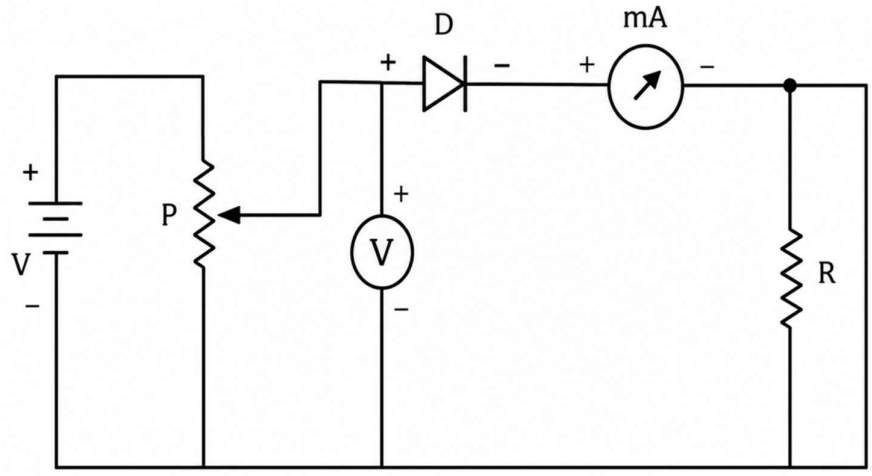


Figure 2.7 (a) The circuit to determine V-I characteristic curve of P-N junction diode

To obtain these characteristics experimentally, a DC battery (V) is connected to the diode through a potentiometer (P). The potentiometer is used to vary the voltage applied across the diode in a controlled manner. A milliammeter (mA) is connected in series with the circuit to measure the current flowing through the diode, while a voltmeter is connected across the diode to measure the potential difference between its terminals. By gradually changing the voltage using the potentiometer and recording the corresponding current values, the V–I characteristic curve of the P–N junction diode can be plotted. The V–I characteristics of a P–N junction diode describe how the current flowing through the diode changes with the applied voltage across it. This curve helps in understanding the electrical behavior of the diodeunder different operating conditions.

Generally, the characteristics of a P–N junction diode can be studied under three conditions: zero external voltage, forward bias, and reverse bias.

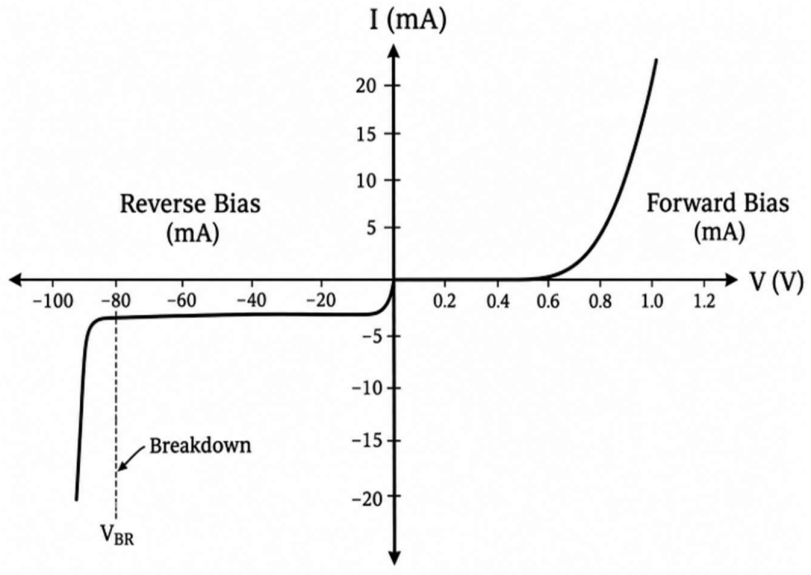


Figure 2.7 (b) The V-I characteristic curve of P-N junction

Zero External Voltage:
When no external voltage is applied across the P–N junction, the diode is said to be in an unbiased condition. At this stage, a potential barrier exists at the junction due to the formation of the depletion region. This barrier prevents the free movement of majority charge carriers (electrons and holes) across the junction. Although a small amount of carrier movement occurs due to thermal energy, the drift current and diffusion current balance each other. As a result, there is no net current flow in the external circuit, and the circuit current remains zero. This condition is represented by the origin point (O) on the V–I characteristic graph.

Forward Bias:
When the diode is forward biased, the positive terminal of the battery is connected to the P-side and the negative terminal to the N-side. This connection reduces the potential barrier and narrows the depletion region at the junction. Initially, when the applied voltage is smaller than the barrier potential, only a small current flows through the circuit because the barrier still restricts the movement of charge carriers. However, as the forward voltage increases and reaches a certain value known as the cut-in voltage (or threshold voltage), the potential barrier is effectively overcome. After this point, a large number of majority carriers cross the junction and the current increases rapidly with a small increase in voltage. In this region, the V–I curve rises steeply and becomes almost linear, indicating that the diode offers very low resistance in forward bias.

Reverse Bias:
When the diode is reverse biased, the positive terminal of the battery is connected to the N-side and the negative terminal to the P-side. Under this condition, the potential barrier at the junction increases and the depletion region becomes wider. This makes the junction resistance very high and prevents the flow of majority carriers across the junction. As a result, practically no current flows through the circuit. However, a very small current known as the reverse saturation current flows due to the movement of minority carriers present in the semiconductor material. This current is extremely small and remains almost constant with increase in reverse voltage. If the reverse voltage is increased beyond a certain critical value called the breakdown voltage, the junction undergoes breakdown, resulting in a sudden increase in reverse current and a sharp decrease in the resistance of the junction.

2.6 REVERSE BIAS BREAKDOWN

In a P–N junction diode operating under reverse bias, the applied voltage increases the potential barrier and widens the depletion region. Under normal reverse bias conditions, only a very small current called the reverse saturation current flows through the junction due to the movement of minority charge carriers. However, when the reverse voltage is increased beyond a certain critical value known as the breakdown voltage, the junction suddenly begins to conduct a large current. This phenomenon is known as breakdown of a P–N junction. During breakdown, the resistance of the junction drops sharply and the reverse current increases rapidly. If the current is not limited by an external resistance, the diode may be damaged. The breakdown of a P–N junction occurs mainly due to two mechanisms: Zener breakdown and Avalanche breakdown.

2.6.1 Zener Breakdown

Zener breakdown occurs in a heavily doped P–N junction, where the depletion region is very thin. When a reverse voltage is applied across the junction, a strong electric field is set up in the narrow depletion region. As the reverse voltage increases to a certain value, the electric field becomes extremely strong and is capable of breaking the covalent bonds of the semiconductor atoms. As a result, a large number of electron–hole pairs are generated. These charge carriers contribute to a large reverse current known as the Zener current (I_Z). This current is almost independent of the applied voltage and is mainly limited by the external resistance of the circuit. The zener breakdown process is reversible, meaning that the diode can return to its normal condition when the voltage is reduced. This effect is commonly used in zener diodes for voltage regulation in electronic circuits.

2.6.2 Avalanche Breakdown

Avalanche breakdown occurs in a lightly doped P–N junction, where the depletion region is comparatively wide. When a high reverse voltage is applied, the electrons and minority carriers in the depletion region gain a large amount of kinetic energy due to the strong electric field. These high-energy electrons collide with the atoms of the semiconductor lattice. During these collisions, the covalent bonds are broken, producing additional electron–hole pairs. The newly generated carriers also gain energy and collide with other atoms, producing more carriers. This chain reaction multiplies the number of charge carriers rapidly, resulting in a large reverse current. Because the process resembles a chain reaction where one collision leads to many others, it is called avalanche breakdown. This process is generally irreversible and may damage the diode permanently if the current is not controlled.

In general, the type of breakdown that occurs depends on the breakdown voltage of the diode. If the breakdown occurs at a voltage below about 5 V, it is mainly due to the Zener effect, whereas if the breakdown voltage is above about 5 V, it is mainly due to the avalanche effect. Thus, both mechanisms lead to a sudden increase in reverse current, but they differ in their doping level, depletion layer width, and physical process involved.

Table 2.1 presents the basic comparison between Zener breakdown and Avalanche breakdown.

Zener Breakdown	Avalanche Breakdown
Occurs in heavily doped P–N junctions.	Occurs in lightly doped P–N junctions.
The depletion region is very thin.	The depletion region is wide.
Caused by strong electric field breaking the covalent bonds in the depletion region.	Caused by collision of high-energy electrons with atoms, producing more electron–hole pairs.
Occurs at low reverse voltage (generally below 5 V).	Occurs at higher reverse voltage (generally above 5 V).
The reverse current produced is called Zener current.	The large reverse current is called avalanche current.
The process is reversible if current is properly limited.	The process may damage the diode if current is not controlled.

W CHECK YOUR PROGRESS

Q1. What is barrier potential?
Q2. What happens in forward bias PN junction?
Q3. Why is reverse current very small?
Q4. What is threshold voltage?
Q5. Why does current increase rapidly after threshold voltage?
What happens to the depletion region width when temperature increases? Why?
Q6. Why does Zener breakdown occur at low voltage?
Q7. What is avalanche breakdown?

2 2.7 LET US SUM UP

- A P–N junction is formed by joining p-type and n-type semiconductor materials which leads to the diffusion of charge carriers across the junction.
- Due to the recombination of electrons and holes near the junction a depletion region is formed which contains immobile ions and acts as a barrier to the movement of charge carriers.
- The potential barrier developed across the depletion region prevents the free movement of majority charge carriers across the P–N junction.
- In forward bias, the positive terminal of the battery is connected to the P-side and the negative terminal to the N-side which reduces the barrier potential and decreases the width of the depletion region**.**
- In reverse bias, the positive terminal of the battery is connected to the N-side and the negative terminal to the P-side which increases the barrier potential and widens the depletion region.
- Under reverse bias condition only a very small current called the reverse saturation current, due to minority carriers, flows through the diode.
- The V–I characteristics of a diode represent the relationship between the applied voltage across the diode and the resulting current through it.
- Breakdown in reverse bias is the condition in which a P–N junction diode suddenly conducts a large current when the reverse voltage exceeds the breakdown voltage.
- Zener breakdown occurs in heavily doped P–N junctions at low reverse voltage due to the strong electric field breaking the covalent bonds and generating electron–hole pairs.
- Avalanche breakdown occurs in lightly doped P–N junctions at higher reverse voltage when high-energy carriers collide with atoms and create additional electron–hole pairs, resulting in a sudden increase in current.

K 2.7 ANSWERS TO CHECK YOIUR PROGRESS

Answer to Q No. 1: Barrier potential in PN junction is the voltage that prevents further diffusion of charge carriers. It is ≈0.7V for silicon and 0.3V for germanium.

Answer to Q No. 2: In forward bias PN junction, depletion region narrows and current flows easily.

Answer to Q No. 3: Because it is mainly due to minority carriers, which are very few in number.

Answer to Q No. 4: It is the minimum voltage required for conduction in diode. (≈0.7V for silicon).

Answer to Q No. 5: Because the barrier is overcome and a large number of charge carriers can move across the junction.

It slightly decreases because increased thermal energy generates more charge carriers.

Answer to Q No. 6: Due to heavy doping, which creates a strong electric field even at low voltages.

Answer to Q No. 7: Avalanche breakdown Occurs at high reverse voltage due to collision of carriers.

O 2.8 FURTHER READING

- Boylestad, R. L., & Nashelsky, L. (2013). *Electronic devices and circuit theory* (11th ed.). Pearson Education.
- Malvino, A. P., & Bates, D. J. (2015). *Electronic principles* (8th ed.). McGraw-Hill Education.
- Sedra, A. S., & Smith, K. C. (2015). *Microelectronic circuits* (7th ed.). Oxford University Press.
- Floyd, T. L. (2012). *Electronic devices* (9th ed.). Pearson Education.
- Sarma, S. (2012). *Electronic principles*. S. K. Kataria & Sons.
- Streetman, B. G., & Banerjee, S. K. (2015). *Solid state electronic devices* (7th ed.). Pearson Education.

M 2.9 MODEL QUESTION

Q1. Explain the formation of a P–N junction and depletion region.
Q2. Describe the forward and reverse biasing of a P–N junction diode.
Q3. Why does a P–N junction conduct easily in forward bias but not in reverse bias?
Q4. Define the V–I characteristics of a diode and explain how they are obtained experimentally.
Q5. Why is the depletion region wider in reverse bias and narrower in forward bias??
Q6. What is reverse saturation current? Explain its origin.
Q7. Why does current increase rapidly after the cut-in voltage in forward bias?
Q8. Describe the zener breakdown and avalanche breakdownmechanisms in a P–N junction diode.
Q9 Explain the behavior of a P–N junction under zero external voltage.
Q10. A diode is reverse biased. If the temperature increases, what happens to the reverse saturation current, and why?

*** ***** ***

UNIT 3: DIODE APPLICATIONS

UNIT STRUCTURE

3.1 Learning Objectives
3.2 Introduction
3.3 Power Supply
3.4 Filter Circuits and Voltage Regulation
3.5 Zener Diode
3.6 Switching Applications
    3.6.1 Clipping and Clamping Circuits
3.7 Let Us Sum Up
3.8 Answers to Check Your Progress
3.9 Further Reading
3.10 Model Questions

3.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- describe the concept and components of a power supply system
- explain the working of filter circuits and methods of voltage regulation
- discuss the characteristics and operation of a zener diode
- apply the zener diode for voltage regulation in practical circuits
- illustrate the use of diodes in switching applications.

3.2 INTRODUCTION

In the previous unit, we studied the pn junction diode. Building on this foundation, this unit introduces the concept and components of a power supply system. It explains the operation of filter circuits and various methods of voltage regulation, along with the characteristics and working of the zener diode. The unit also discusses the application of zener diodes in voltage regulation circuits. In addition, learners will explore the role of diodes in switching applications.

3.3 POWER SUPPLY

A power supply is an electrical device that converts electrical energy from a source into a suitable form of a regulated voltage and current that required to operate a load. In most electronic systems, the primary function of a power supply is to transform the alternating current (AC) into a stable direct current (DC) with the desired voltage level. In general, power supplies are classified into two main categories:

- Unregulated Power Supply
- Regulated Power Supply

Unregulated Power Supply: An unregulated power supply provides a DC output, but the voltage is not constant. It varies with changes in input supply, load conditions, and environmental factors such as temperature. Fluctuations in the AC mains or variations in operating conditions can lead to instability in the output voltage. Due to the

limitations of unregulated power supplies, it becomes necessary to incorporate a mechanism to maintain a constant output voltage. This is achieved by introducing a feedback-controlled regulating stage, which converts an unregulated supply into a stable and reliable one.

Regulated Power Supply: A regulated power supply is a device designed to deliver a constant DC voltage of predetermined value across the load, irrespective of variations in input voltage, load current, or environmental conditions such as temperature. The four basic components of a regulated power supply are transformer, rectifier, filter and regulator. The block diagram of a regulated power is shown in figure 3.1. The functions of each block are discussed and below.

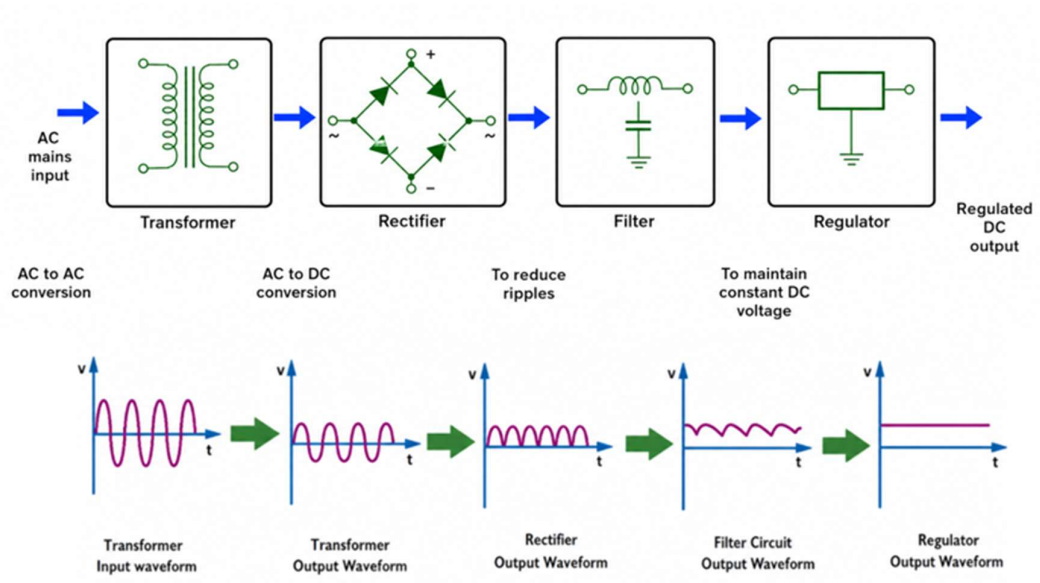


Figure 3.1: Block diagram representation of regulated power supply

Transformer:

A transformer is a device that transfers electrical energy from one alternating current circuit to another by increasing or decreasing the voltage level. When it increases the voltage, it is called a step-up transformer, and when it decreases the voltage, it is known as a step-down transformer.

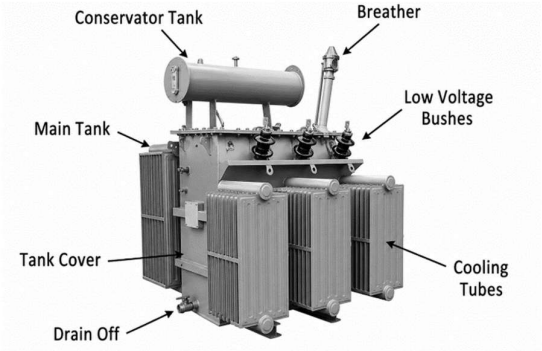


Figure 3.2: A commercial transformer

The working of a transformer is based on the principle of electromagnetic induction. It consists of two coils, known as the primary coil and the secondary coil. The number of turns or windings in these coils determines the voltage transformation. In a step-down transformer, the number of turns in the secondary coil is less than that in the primary coil, so the ratio of secondary to primary turns is less than one. This type of transformer reduces high voltage to a lower voltage and is commonly used in power distribution to minimize energy loss over long distances.

In contrast, a step-up transformer has more turns in the secondary coil than in the primary coil, enabling it to increase the voltage. Such transformers are typically used in applications like inverters and voltage stabilizers where a higher output voltage is required.

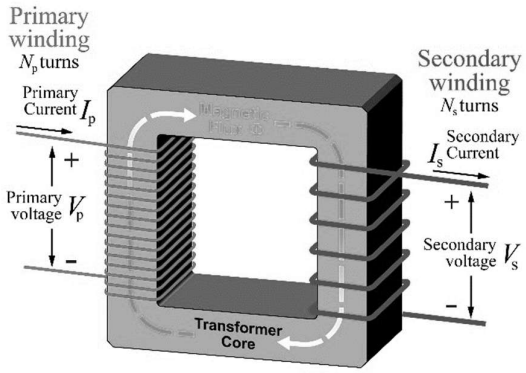


Figure 3.3: Transformer with Primary and Secondary Coils

When an alternating voltage is applied to the primary winding**,** it produces an alternating current, which in turn generates a time-varying magnetic flux ϕ(t) in the transformer core. This flux links both the primary and secondary windings, inducing EMFs in them. Thus, the electromotive force (EMF) is the voltage induced in the primary or secondary winding due to the time-varying magnetic flux linking the coils.
Mathematically, the induced EMF is given by,

$$e = -N\frac{d\phi}{dt}$$

Where e is the induced EMF, N is the number of turns in the winding, ϕ is the magnetic flux.
The negative sign indicates the direction of the induced EMF, meaning that it opposes the change in magnetic flux that produces it.

Rectifier:
Rectifier is also a device that helps to convert AC to DC. It uses PN junction diode. About the PN junction we have already learnt in previous unit. A PN junction diode is a device that allows electric current to flow when it is forward biased and blocks the current when it is reverse biased. The figure 3.5 shows a PN junction diode in forward bias. In this condition, the positive terminal of the battery is connected to the p-type region and the negative terminal to the n-type region. Without an applied voltage, no current flows through the diode. When forward bias is applied, the depletion layer narrows which allow electrons from the n-side and holes from the p-side to move

across the junction. The battery supplies electrons to the n-region, which then combine with holes from the p-region and results the flow of current.

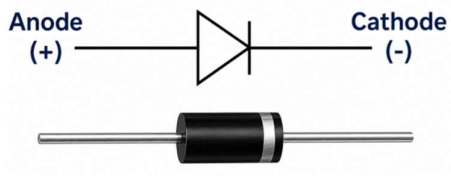


Figure 3.4: PN Junction diode

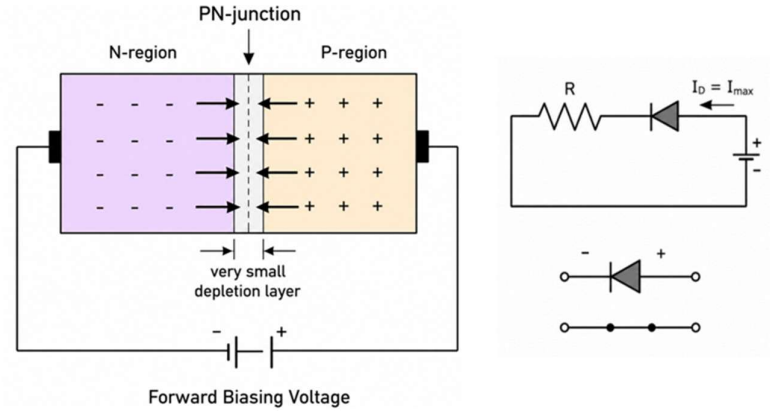


Figure 3.5: Forward-biased PN junction diode with circuit representation

When a PN junction is reverse biased, the p-type semiconductor is connected to the negative terminal of the battery and the n-type semiconductor to the positive terminal. Due to this polarity, holes in the p-region and electrons in the n-region move away from the junction. This causes the depletion region to widen. As a result, current flow is effectively blocked, with only a very small current due to minority carriers.

Thus, a PN junction allows current to flow only in the forward-biased condition. This principle is utilized in rectifier circuits, which convert alternating current (AC) into direct current (DC), a process known as rectification. As AC is represented by a sinusoidal waveform, it varies periodically in both magnitude and direction with time.

Rectifiers are broadly classified into two types. They are half-wave and full-wave rectifiers. A half-wave rectifier converts only one half of the AC cycle and operates during the positive half cycle whereas, a full-wave rectifier converts both halves of the AC cycle which produces a more continuous and efficient DC output. For better understanding, consider the following diagrams of half wave and full wave rectifier circuits.

Half wave rectifier:

In the figure 3.6 a half-wave rectifier circuit shows that consists an AC input supply, a step-down transformer, a PN junction diode, and a load resistance. The alternating voltage from the source is first applied to the primary winding of the transformer, and a reduced AC voltage is obtained across the secondary winding. This secondary voltage is then fed to the diode, while the other terminal of the diode is connected to the load resistor. The remaining end of the load is connected back to the secondary winding to complete the circuit.

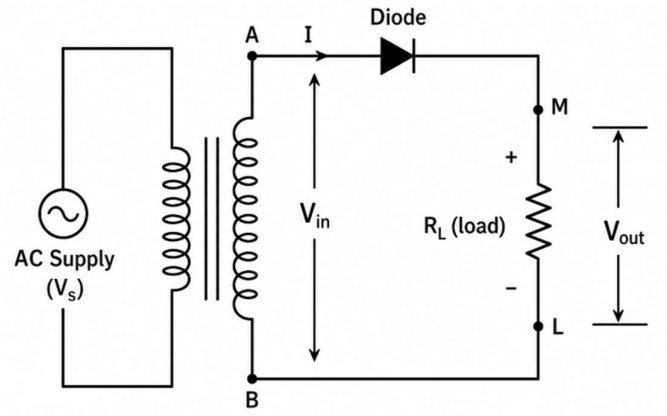


Figure 3.6: Half wave rectifier

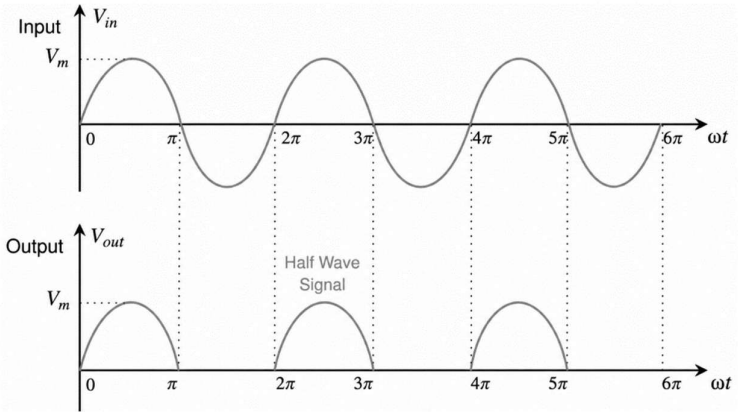


Figure 3.7: Input and output waveforms of a half-wave rectifier

During the positive half-cycle of the AC input i.e., from 0 to π, the upper end of the transformer secondary becomes positive. This forward-biases the PN junction diode and current flows through the circuit. As a result, a output voltage V_{out} develops across the load resistor R_{L}..

During the negative half-cycle of the AC input i.e., from π to 2π, the polarity of the transformer secondary reverses and becomes negative. As a result, the PN junction diode is reverse-biased and blocks current. Consequently, no output voltage across the load is obtained. Hence, the circuit allows only the positive half-cycle of the input waveform to pass which produce a pulsating DC output while the negative half-cycle is blocked by the diode. This process is known as half-wave rectification.

Disadvantage of Half wave rectifier:

Full wave rectifier:

In a full-wave rectifier, a centre-tapped transformer is used where the secondary winding of the transformer has a central tap that divides the winding into two halves. Here two PN junction diodes are used instead of one. The centre tap of the transformer's secondary winding is connected to ground, which helps in obtaining a rectified output during both half cycles of the AC input. During the positive half cycle, one end of the secondary (say point M) becomes positive with respect to the centre

tap, while the other end (point N) becomes negative. As a result, diode D1 is forward biased and conducts, while diode D2 is reverse biased and does not conduct.

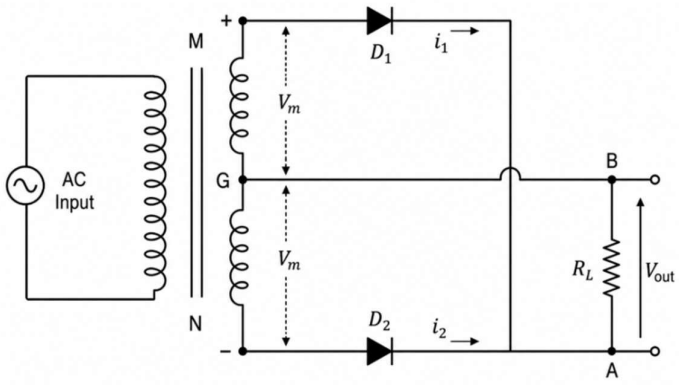


Figure 3.8: Full wave rectifier circuit

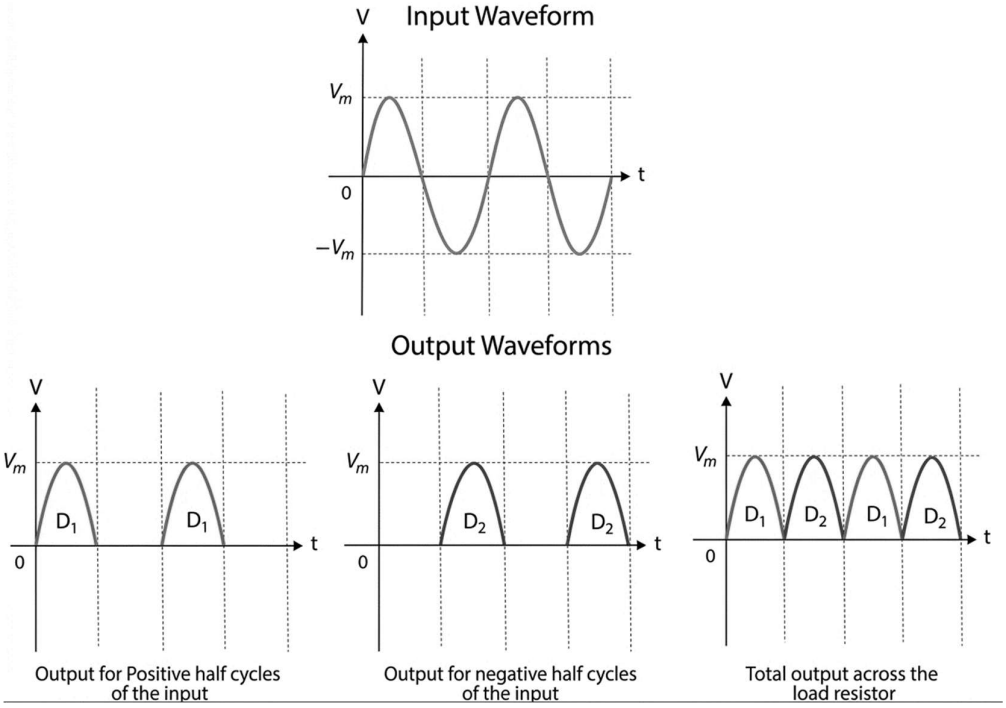


Figure 3.9 (a): Output waveform of full wave rectifier

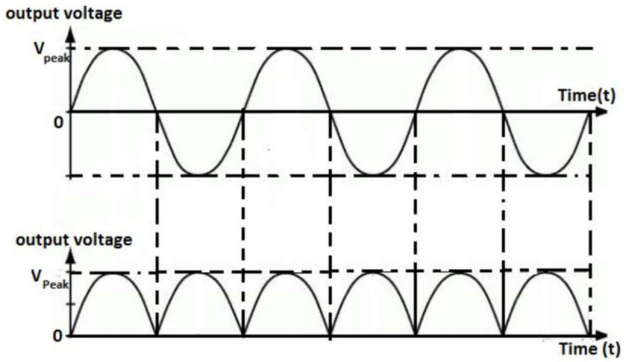


Figure 3.9 (b): Output waveform of full wave rectifier

In figure 3.9 (a), the outputs of diodes D1 and D2 during the positive and negative half cycles of the AC signal are shown, and the final rectified output is obtained by combining these two results as shown in figure 3.9 (b).

3.4 FILTER CIRCUITS AND VOLTAGE REGULATION

After rectification, the output is passed through a filter circuit. A filter circuit in a power supply is an electronic circuit used to reduce or remove the AC fluctuations called ripple from the rectified output and provide a smooth and steady DC voltage. The output from the rectifier circuit contains both the desired DC component and some AC components. But we need pure DC output at the load. Hence we have to filter out the AC components. The filter circuit uses components such as capacitors, inductors, or a combination of both to minimize variations in the output signal.
Here we will discuss about three important filter circuits.

1. Shunt capacitor based filter
2. LC filter and
3. CLC filter or π (pi) filter

First, let us discuss about a shunt capacitor based filter circuit.

Shunt Capacitor Based Filter Circuit:
A capacitor filter circuit contain a capacitor and the load connected in parallel. The output of the rectifier circuit becomes the input for the capacitor filter circuit. A capacitor blocks DC and allows AC components to pass.
The rectified output is applied across the capacitor which has a low reactance. If $X_c$is the capacitive reactance in Ω then:

$$X_c = \frac{1}{\omega c} = \frac{1}{2\pi f c} \quad (1)$$

Where, ω is the angular frequency of the AC signal in radians per second and is defined as ω=2πf.

f is the frequency of the AC signal in hertz (Hz) and C is the capacitance of the capacitor in farads (F).

From equation (1), we can write,

$$X_c \propto \frac{1}{f} \quad (2)$$

Thus, the frequency of the current f is inversely proportional to the reactance for a given capacitance C.

The rectified output has both AC and DC components and the AC component has its own frequency f while the DC components has zero frequency. Hence from the equation (2), we can see that for DC, the reactance is infinite. That is why the capacitor offers infinite resistance to DC and hence DC will not pass through the Capacitor. So the capacitor will only allow the AC to pass through it. Hence the remaining DC signal will pass through the Load resistance as shown here. But still we can have some part of the AC component going along with the DC to the load, which has to be removed by deploying another circuit element.

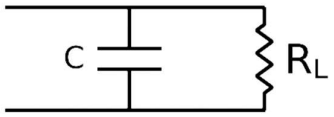


Figure 3.10: Shunt Capacitor Filter

LC filter or Choke filter:

In LC filter circuit, we apply the output form the rectifier across the inductor. The reactance of a inductor is given by:

$$X_l = \omega L = 2\pi f L$$

We can see that for a DC signal which have f=0, X_L is also zero. Hence it will smoothly pass the low frequency or DC signals while blocks high frequency signals. So the inductor has to be connected in series so that it allows DC and blocks AC. As a result, the inductor filters out the majority of the alternating current component, and the remaining alternating current component is now bypassed by the capacitor (C) due to its low reactance for the alternating current component.

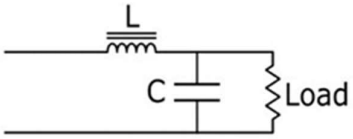


Figure 3.11: LC filter circuit

CLC Filter

The CLC filter circuit uses two capacitors and one inductor. Here is the circuit:

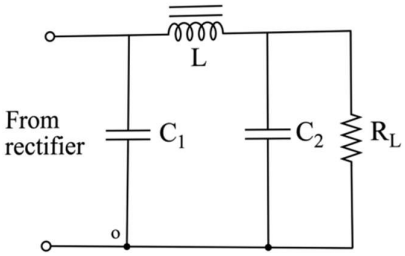


Figure 3.12: CLC filter circuit

This filter is also known as Pi section filter or Capacitor input filter. In this circuit, the major filter is done by the C_1 capacitor as it offers low reactance to the ripple. The remaining filtering is done by the combination of L and C_2. This offers the best filter circuit as the output produced by this circuit is almost perfect dc.

A comparison of the output produced by the above circuits is presented below:

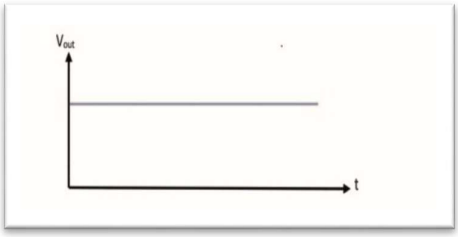


Figure 3.13(a): Expected output of filter circuit

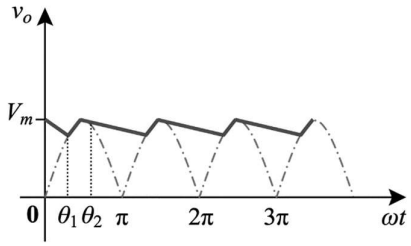


Figure 3.13(b): Output of Shunt capacitor filter

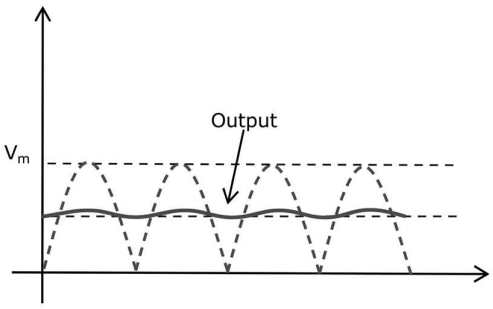


Figure 3.13(c): Output of LC filter

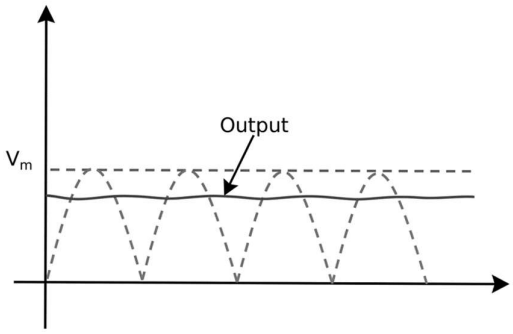


Figure 3.13(d): Output of CLC filter

The LC filter provides better ripple reduction by smoothing the current. Its performance is largely independent of load current variations, which is a significant advantage. Despite this, it has several limitations: it is bulky, heavier, and produces a relatively lower output voltage. Additionally, it is not suitable for half-wave rectifiers, restricting its application.

The CLC filter (π-filter) offers superior ripple reduction and higher output voltage compared to the other filters. It can be used with both half-wave and full-wave rectifier circuits, making it more versatile. However, these advantages come at the cost of being expensive, large in size, and heavy in weight, which limits its practical use in compact systems.

The last Block is the regulator block. It detects and corrects variations in the power supply's output voltage. The output voltage may alter or fluctuate as a result of changes in the input AC voltage, the load, or any physical factors of the circuit, such as temperature. The unregulated DC output from filter circuit is fed into this regulating stage. The voltage regulator maintains a steady DC voltage at the output terminals.

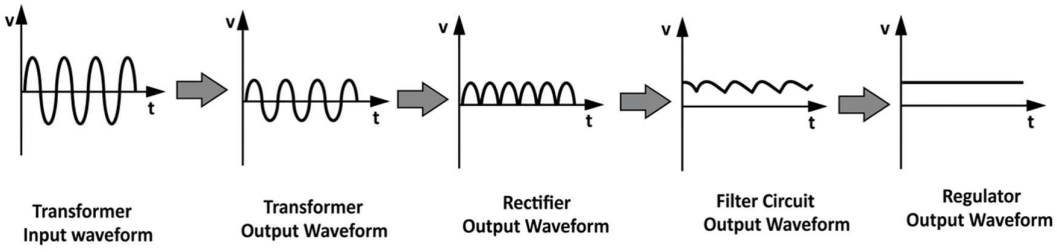


Figure 3.14: Output of each block of power supply

Figure 3.14 shows the change of the AC input at each block of the power supply. Starting at the AC input, we finally get the constant DC output.

Applications of regulated power Supply:

A regulated power supply is a reliable embedded circuit that provides a stable DC output with low noise, reduced circuit complexity, and fast response to changes in input or load conditions. Many electronic circuits require regulated power because unregulated supply can cause damage or improper operation. It is widely used in mobile charging circuits, computers, oscillators, and amplifier systems. In addition, regulated power supplies are commonly used as bench power supplies in laboratories,

where they provide adjustable output voltages for testing and development of electronic circuits.

3.5 ZENER DIODE

A Zener diode is a special type of semiconductor diode designed to operate in the reverse breakdown region. It allows current to flow not only in the forward direction like a normal diode, but also in the reverse direction when the applied voltage reaches a specific value called the Zener breakdown voltage. A Zener diode is formed from a heavily doped p-n junction, unlike a conventional diode. Due to this heavy doping, it has a very thin depletion layer.



Figure 3.15: Circuit symbol of Zener diode

This unique structure allows the diode to undergo breakdown at a precisely defined and controlled voltage. In forward bias, a Zener diode behaves like a normal diode and begins to conduct current after approximately 0.7 V in the case of silicon. In reverse bias, only a small leakage current flows initially. However, when the reverse voltage reaches the Zener voltage (VZ), breakdown occurs, causing a sharp increase in current while the voltage across the diode remains nearly constant. This unique property makes the Zener diode highly useful for voltage regulation applications.

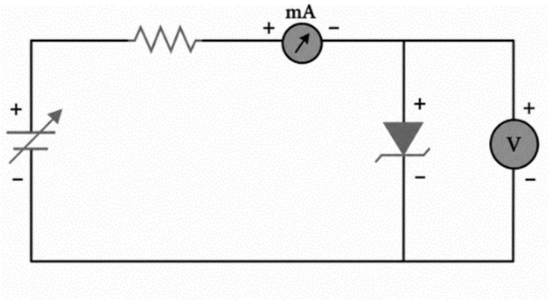


Figure 3.16 (a) Zener diode in forward bias

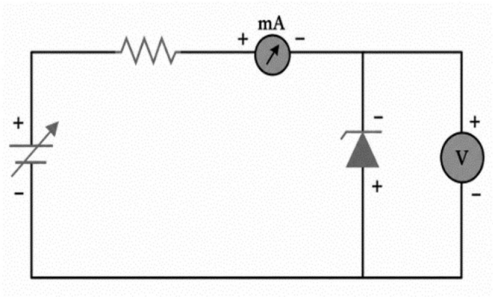


Figure 3.16 (b) Zener diode in reverse diode

V-I Characteristics of Zener diode:

The V–I characteristics of a diode describe how the current varies with applied voltage. In the forward region, the diode behaves like a normal diode, where current increases significantly after the threshold voltage (about 0.7 V for silicon). In the reverse region, only a very small leakage current flows initially due to minority charge carriers. When the reverse voltage reaches the breakdown value known as the Zener voltage (V_Z), the diode enters breakdown mode. At this point, the current increases sharply, but the voltage across the diode remains nearly constant. This stable voltage regulation beyond breakdown is a key feature of Zener diodes.

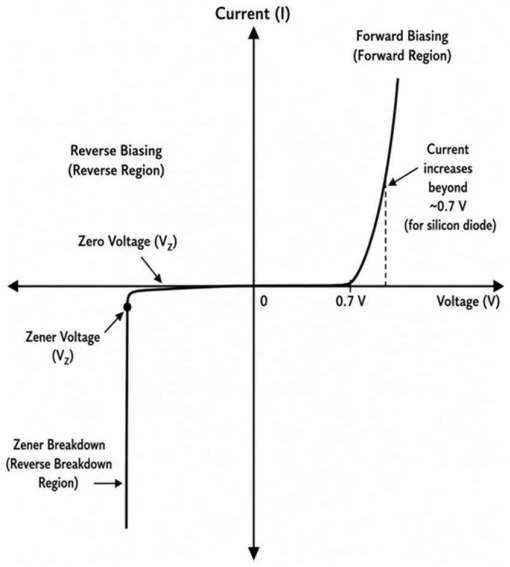


Figure 3.16 (c): V-I Characteristics of Zener diode

Diode as Voltage Regulator

A Zener diode acts as a voltage regulator when it is connected in reverse bias across the load. In this configuration, it maintains a nearly constant output voltage even when there are variations in the input voltage or changes in the load current. When the input voltage increases beyond the Zener breakdown voltage, the diode conducts in the reverse direction and stabilizes the voltage across the load. As a result, the output voltage remains approximately equal to the Zener voltage (V_Z) and provide a stable and regulated supply.

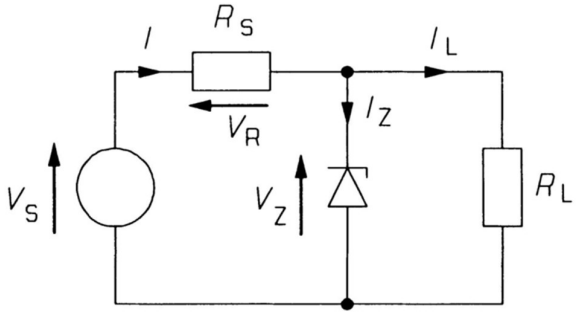


Figure 3.17: Zener diode shunt regulator

Advantage and Disadvantage of Zener Diode:

A Zener diode voltage regulator offers several advantages. It is simple in design and inexpensive, making it easy to implement in basic electronic circuits. It provides reliable voltage regulation and responds quickly to changes in load or input conditions, ensuring stable output. However, it also has certain limitations. It is suitable only for low-power applications, as it cannot handle large power levels efficiently. In addition, power dissipation can be high, especially when regulating larger currents, which results in lower efficiency in high-current applications.

3.6 SWITCHING APPLICATION

A PN junction diode is widely used in switching applications due to its ability to rapidly transition between ON and OFF states based on the applied voltage. Under forward bias, the diode conducts current and behaves like a closed switch, allowing current to flow through the circuit. Under reverse bias, it blocks current flow except for a small leakage current and acts like an open switch. This quick switching between conducting and non-conducting states makes the PN junction diode highly suitable for applications such as rectifiers, digital logic circuits, clippers, and clamping circuits, where controlled signal switching is essential.

3.6.1 Clipping and clamping circuits

A clipping circuit is an electronic circuit that removes (clips) a portion of an input waveform without distorting the remaining part. It is mainly used to limit voltage levels in a signal. There are three basic types of clippers. They are positive clipper, negative clipper and biased clipper.

In a positive clipper removes the positive half cycles of the input voltage waveform. There are two types of positive clipper. Series and shunt clipper. When the diode is forward biased, it conducts and allows the signal to pass. When the diode is reverse biased, it blocks the signal, thereby clipping it. Figure 3.18 shows the positive clipper circuit.

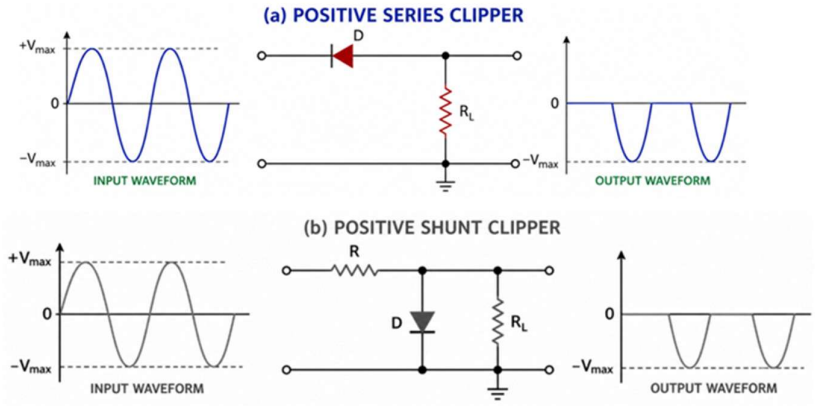


Figure 3.18 (a): Positive clipper circuit

A negative clipper removes the negative half cycles of the input voltage waveform. Like positive clipper, negative clipper is also two types. Series and shunt clipper. In negative clipping, the diode is connected in a direction opposite to that of a positive clipper. In the series negative clipper, during the positive half cycle of the input signal, the diode conduct and work as a short circuit. Again during the negative half cycle of the input signal, the diode does not conduct and acts as an open circuit. The negative half cycle will not appear at the output and it is clipped off.

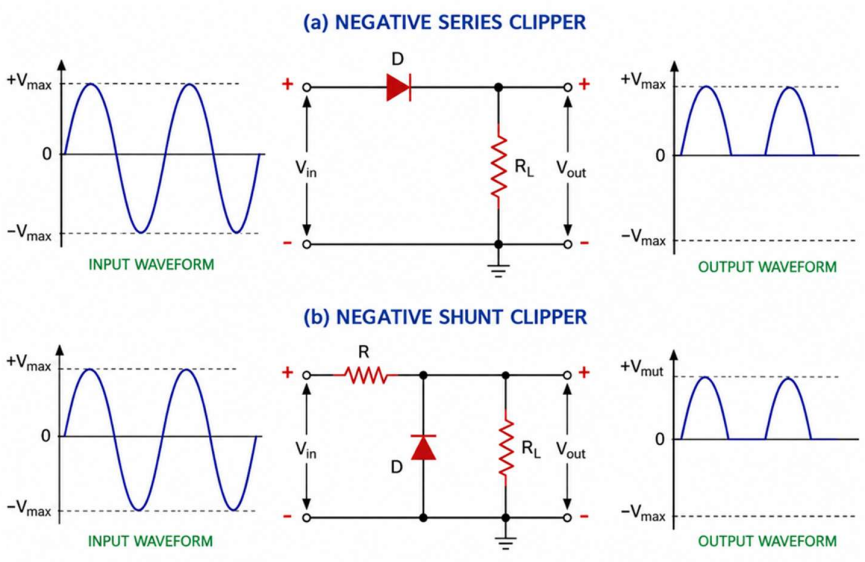


Figure 3.18 (b): Negative clipper circuit

A biased clipper includes an additional DC source along with the diode, which allows clipping to occur at a specified voltage level rather than at zero. For example, the circuit can be designed to clip signals above +5 V or below −3 V, providing more precise control over the output waveform.

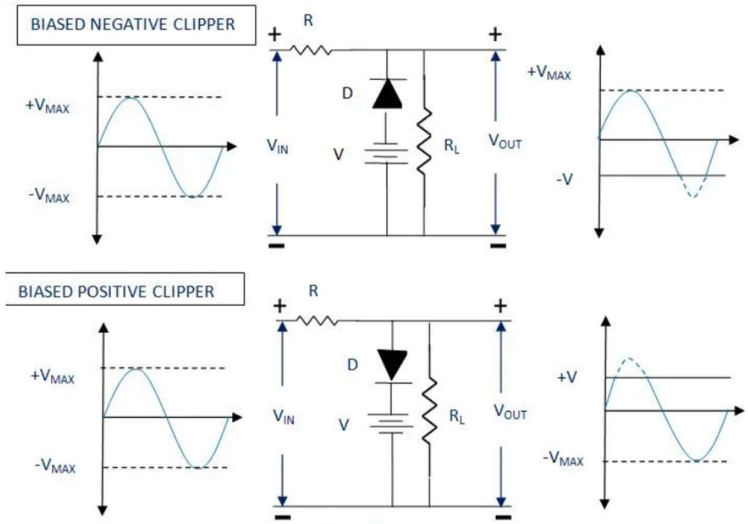


Figure 3.18 (c): Biased clipper circuit

Clamper circuit:

A clamper circuit is an electronic circuit that adds a DC level to an AC signal, shifting the entire waveform upward or downward without changing its shape.

Clamper circuit is classified as:

Positive Clamper, Negative Clamper and Biased Clamper

A positive clamper is a circuit that shifts the entire input waveform upward so that all parts of the signal lie above the zero reference level. It adds a positive DC level to the signal without changing its shape, typically using a diode, capacitor, and resistor.

A negative clamper is a circuit that shifts the entire input waveform downward so that all parts of the signal lie below the zero reference level. It adds a negative DC level to the signal without altering its shape, using components such as a diode, capacitor, and resistor.

A biased clamper is a circuit that shifts the input waveform to a desired DC level other than zero by using an additional DC voltage source along with a diode, capacitor, and resistor. This allows the signal to be clamped at a specific positive or negative voltage level without changing its shape.

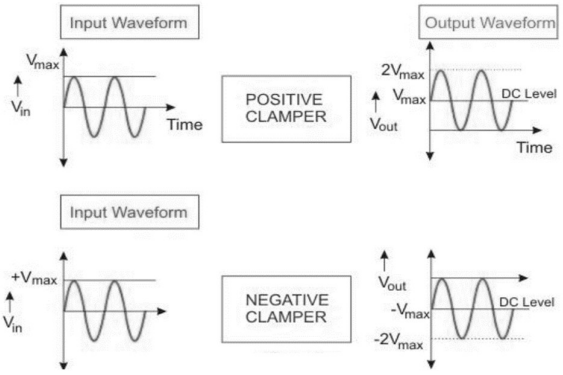


Figure 13.19 (a) Positive and negative clamper

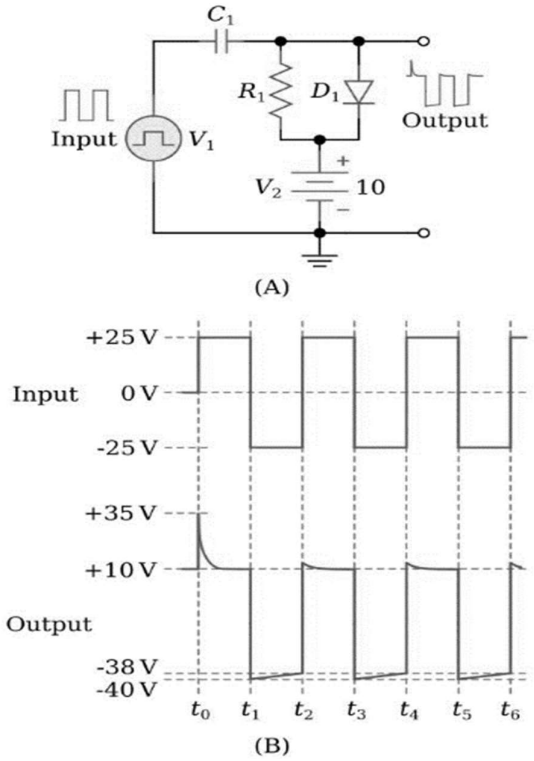


Figure 13.20: Biased clamper

The given circuit is a biased clamper that uses capacitor C_1, diode D_1, resistor R_1, and a DC source of 10 V to shift the input waveform. The input is a square wave varying between +25 V and −25 V. During the positive half cycle (+25 V), the diode D_1 becomes forward biased and conducts, allowing the capacitor C_1 to charge rapidly to approximately the sum of the input peak and the bias voltage. As a result, the output is clamped near the +10 V level. During the negative half cycle (−25 V), the diode becomes reverse biased and turns OFF. The capacitor retains its charge and effectively adds its voltage to the input signal, causing the output to shift downward to around −38 V to −40 V. Thus, the overall effect is that the waveform is shifted in the negative direction, with the positive peak clamped at +10 V, while the shape of the waveform remains unchanged.

W CHECK YOUR PROGRESS

Q1. What is a regulated power supply?
Q2. What is the function of a Zener diode in a power supply?
Q3. Why is a Zener diode operated in reverse bias in power supplies?
Q4. What is a clipper circuit?
Q5. What is a clamper circuit?
Q6. What is the main difference between a clipper and a clamper?

2 3.7 LET US SUM UP

- A power supply converts AC input into a stable DC output required for electronic circuits.
- Filter circuits are used to reduce ripples in the rectified output and produce a smoother DC signal.
- Voltage regulation maintains a constant output voltage despite variations in input supply or load conditions.
- A Zener diode is used in reverse bias to provide stable voltage regulation in power supply circuits.
- Diodes are widely used in switching applications, acting as ON/OFF devices based on biasing conditions.
- Clipping circuits are used to remove unwanted portions of a waveform without affecting the remaining signal.
- Clamping circuits shift the entire waveform to a desired DC level without altering its shape.

K 3.8 ANSWERS TO CHECK YOIUR PROGRESS

Answer to Q1. A regulated power supply provides a constant DC output voltage despite variations in input voltage or load conditions.

Answer to Q2. A Zener diode is used for voltage regulation by maintaining a constant output voltage in the reverse breakdown region.

Answer to Q3. Because it maintains a constant voltage across it after breakdown, which helps in stabilizing the output voltage.

Answer to Q4. A clipper circuit is used to remove or limit a portion of an input waveform without affecting the remaining part.

Answer to Q5. A clamper circuit shifts the entire waveform to a different DC level without changing its shape.

Answer to Q6. A clipper removes part of the signal, while a clamper shifts the whole signal up or down.

O 3.9 FURTHER READING

- Boylestad, R. L., & Nashelsky, L. (2013). *Electronic devices and circuit theory* (11th ed.). Pearson Education.
- Malvino, A. P., & Bates, D. J. (2015). *Electronic principles* (8th ed.). McGraw-Hill Education.
- Sedra, A. S., & Smith, K. C. (2015). *Microelectronic circuits* (7th ed.). Oxford University Press.
- Floyd, T. L. (2012). *Electronic devices* (9th ed.). Pearson Education.
- Sarma, S. (2012). *Electronic principles*. S. K. Kataria & Sons.
- Streetman, B. G., & Banerjee, S. K. (2015). *Solid state electronic devices* (7th ed.). Pearson Education.

M 3.10 MODEL QUESTIONS

Q1. Explain the working of a diode-based rectifier and its role in a DC power supply.

Q2. Describe the types of filter circuits used in power supplies and explain their operation.

Q3. Explain the working of a Zener diode as a voltage regulator with a neat circuit diagram.

Q4. Discuss the characteristics and applications of a Zener diode in electronic circuits.

Q5. Explain the switching action of a PN junction diode and its applications in electronic circuits.

Q6. Describe the role of diodes in regulated power supplies, including rectification, filtering, and voltage regulation.

*** ***** ***

UNIT 4: THERMISTORS

UNIT STRUCTURE

4.1 Learning Objectives
4.2 Introduction
4.3 Introduction to Thermistor
4.4 Principle of Operation of Thermistors
4.5 Types of Thermistors
4.6 Characteristics and Application of Thermistors
4.7 Let Us Sum Up
4.8 Answers to Check Your Progress
4.9 Further Reading
4.10 Model Questions

4.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- explain the physical principles governing thermistor operation
- derive and interpret the resistance–temperature relation of thermistors
- differentiate between NTC and PTC thermistors based on material physics
- analyse the important characteristics of thermistors
- evaluate thermistor performance in sensing and protection circuits
- analyse the practical applications of thermistors.

4.2 INTRODUCTION

In the previous unit, we studied the diode and its characteristics, applications, and temperature-dependent behaviour, which forms the basis for understanding temperature-sensitive semiconductor components. In this unit, we shall study thermistors, which are temperature-dependent resistors made from semiconductor materials. Unlike ordinary resistors, thermistors show a large change in resistance with temperature for which they are highly useful in sensing and protection circuits. We will discuss the principle of operation of thermistors, their resistance-temperature relationship, and the role of semiconductor conduction mechanisms. Further, we will study the two main types of thermistors namely NTC and PTC thermistors, their characteristics, and their practical applications in temperature measurement, compensation, and circuit protection.

4.3 INTRODUCTION TO THERMISTOR

A thermistor is a temperature-sensitive resistor whose resistance changes significantly with temperature. The term thermistor is derived from the words “thermal” and “resistor.” Thermistors are generally made from semiconductor metal oxides such as manganese oxide, nickel oxide, cobalt oxide, copper oxide, and iron oxide. These materials are mixed in suitable proportions and heated at high temperatures to form ceramic semiconductor elements. Their electrical conductivity strongly depends on temperature due to thermally activated charge carriers.

Based on the variation of resistance with temperature, thermistors are classified into two main types:

1. Negative Temperature Coefficient (NTC) thermistors
2. Positive Temperature Coefficient (PTC) thermistors

In NTC thermistors, resistance decreases with increase in temperature, whereas in PTC thermistors, resistance increases with temperature. Owing to their high sensitivity and compact size, thermistors are widely used in temperature sensing, temperature compensation, current limiting, and protection circuits. In this unit, we shall study the operating principles, types, characteristics, and important applications of thermistors.

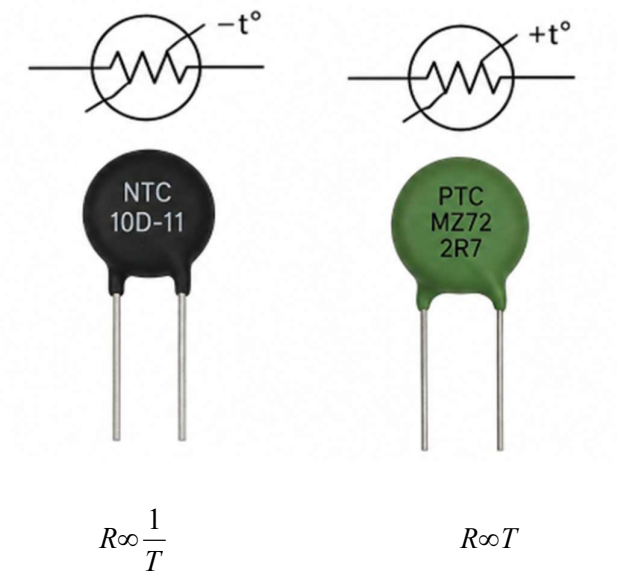


Figure 4.1: Types of Thermistors

The main materials used in NTC thermistors include manganese (Mn), nickel (Ni), and copper (Cu). They are widely used for inrush current limiting, temperature sensing and measurement, temperature compensation, and temperature control. NTC thermistors are manufactured by heating semiconductor ceramic materials made from mixtures of metallic oxides such as cobalt, nickel, and manganese. These materials exhibit small-polaron conduction, which causes the electrical resistance to decrease as the temperature increases. At normal temperatures, electrons require sufficient energy to move from one site to another because of an energy barrier. As the temperature increases, the available thermal energy increases, allowing more electrons to overcome this barrier. As a result, the resistance of the material decreases, producing the negative temperature coefficient of resistance (NTCR) behaviour.

PTC thermistors are made mainly from doped barium titanate ($BaTiO_3$), a semiconducting ceramic material whose resistance increases significantly with a small rise in temperature. Their high sensitivity and ease of use make them suitable for a wide range of applications, such as resistance compensation and temperature measurement. PTC thermistors are also widely used for overcurrent, overload, and

short-circuit protection; telecommunication protection; soft switching and time-delay circuits; motor starting; temperature sensing and protection; and self-regulating heating applications.

4.4 PRINCIPLE OF OPERATION OF THERMISTORS

Thermal and electrical effects play an important role in the practical operation of thermistors. Since thermistors are temperature-sensitive devices, both heat transfer and electrical power dissipation influence their performance. Important practical factors include self-heating, thermal equilibrium, thermal time constant, and stability under environmental conditions. These factors determine measurement accuracy and response behaviour.

Semiconductor Basis of Thermistor Action

The operation of a thermistor is based on the temperature-dependent electrical conduction of semiconductor materials. In a semiconductor, the valence band and conduction band are separated by an energy gap called the band gap. At low temperatures, the valence band is nearly filled, and the conduction band is almost empty. As the temperature increases, some electrons gain sufficient thermal energy to cross the band gap and move into the conduction band, leaving behind holes in the valence band. Both electrons and holes contribute to electrical conduction.

The number of charge carriers generated depends strongly on temperature.

The intrinsic carrier concentration is given by:

$$n_i = AT^{\frac{3}{2}} \exp(-\frac{E_g}{2kT})$$

where

A is a material-dependent constant,

T is the absolute temperature

E_g is the band gap energy of the semiconductor,

k is Boltzmann's constant, and

This equation shows that the carrier concentration increases exponentially with temperature.

The electrical conductivity of a semiconductor is given by:

$$\sigma = nq\mu$$

where

n is the carrier concentration,

q is the electronic charge, and

μ is the carrier mobility.

Although carrier mobility decreases slightly with increasing temperature due to enhanced lattice vibrations, the exponential increase in carrier concentration dominates, which results in an overall increase in conductivity.

Since electrical resistance R is inversely proportional to conductivity,

$$R \propto \frac{1}{\sigma}$$

the resistance decreases with increase in temperature. This negative temperature dependence forms the basic operating principle of Negative Temperature Coefficient (NTC) thermistors. In practical thermistor materials made of polycrystalline metal oxides, conduction involve impurity levels, grain boundary effects, and hopping conduction between localized states. However, the strong temperature dependence of carrier transport remains the fundamental basis of thermistor action.

Resistance–Temperature Relationship

The temperature dependence of resistance is one of the most important characteristics of a thermistor. Since electrical resistance is inversely related to conductivity, it can be written as:

$$R \propto \frac{1}{\sigma}$$

In NTC thermistors, the increase in charge carrier concentration with temperature causes the conductivity to increase, and hence the resistance decreases. For [practical thermistor materials, the resistance–temperature relationship is commonly represented by the empirical relation:

$$R(T) = R_0 \exp\left[\beta\left(\frac{1}{T} - \frac{1}{T_0}\right)\right]$$

where

R(T) is the resistance at temperature T,
R_0 is the resistance at reference temperature T_0, and
β is the material constant known as the beta constant.

The beta constant is related to the effective activation energy E_a by:

$$\beta = \frac{E_a}{k}$$

where $\boldsymbol{k}$ is Boltzmann's constant. For analysis and calibration, the resistance equation can be written in logarithmic form as:

$$\ln R = \ln R_0 + \beta\left(\frac{1}{T} - \frac{1}{T_0}\right)$$

This shows that a plot of $ln\ R$ versus $1/T$ gives a straight line with slope β. This method is commonly used for experimental determination of the beta constant.

The temperature coefficient of resistance (TCR) is defined as:

$$\alpha = \left(\frac{1}{R}\right)\left(\frac{dR}{dT}\right)$$

For an NTC thermistor, substituting the resistance relation gives:

$$\alpha = -\frac{\beta}{T^2}$$

This shows that the temperature coefficient is negative, its magnitude decreases with increasing temperature, and the thermistor exhibits higher sensitivity at lower temperatures. Thus, the resistance–temperature relationship of thermistors is strongly nonlinear, which makes them highly sensitive to small temperature variations and suitable for temperature sensing applications.

Grain Boundary and Conduction Mechanisms

In practical thermistor materials, especially ceramic thermistors, electrical conduction is influenced not only by intrinsic band-to-band excitation but also by microstructural effects such as grain boundaries and localized impurity states. Most thermistors are polycrystalline and consist of many semiconductor grains separated by grain boundaries. These grain boundaries introduce localized energy states due to impurities, lattice defects, and structural discontinuities, which can trap charge carriers and form potential barriers.

At low temperatures, these grain boundary barriers are relatively high, restricting the movement of charge carriers from one grain to another and resulting in high resistance. As the temperature increases, charge carriers gain sufficient thermal energy to overcome these barriers, enhancing intergranular conduction and reducing the effective resistance.

The conduction process in such materials is often described as thermally activated transport, and the resistivity is expressed as:

$$\rho = \rho_0 \exp\left(\frac{E_a}{kT}\right)$$

where

ρ is the resistivity,
ρ_0 is the pre-exponential factor,
E_a is the activation energy associated with carrier transport,
k is Boltzmann's constant, and
T is the absolute temperature.

Since resistance is proportional to resistivity, thermistors exhibit an exponential temperature dependence of resistance.

In many oxide-based NTC thermistors, conduction occurs through small polaron hopping, where charge carriers hop between neighbouring ions of different valence states. For example, in manganese-based thermistors, electron hopping may occur between Mn^{3+} and Mn^{4+} ions. The hopping probability increases with temperature, further strengthening the negative temperature coefficient behaviour. Thus, electrical conduction in ceramic thermistors arises from a combination of intrinsic carrier excitation, impurity ionization, thermally activated hopping, and grain boundary barrier modulation. The relative contribution of each mechanism depends on the material composition, doping level, and microstructure.

Thermal and Electrical Considerations

Heat moves in and out of these sensors during use, shaping how they respond. Not just material science matters - what happens around them plays a role too. Electricity flows through while temperature shifts nearby. Their behaviour ties back to tiny internal structures, yet outside conditions pull influence. Accurate readings come only when both power delivery and warmth transfer are weighed. Even solid theory needs grounding in physical context for reliability.

One of the most important practical phenomena in thermistors is self-heating.

When an electrical current flows through a thermistor, electrical power is dissipated as heat according to Joule's law. The power dissipated is given by:

$$P = I^2R$$

where

P is the power dissipated,
I is the current through the thermistor,
V is the applied voltage, and
R is the resistance.

This power causes the thermistor temperature to rise above the ambient temperature. If the current is sufficiently large, the temperature increase may be significant, leading to measurement error. In temperature sensing applications, the current must therefore be kept small to minimize self-heating.

The thermal equilibrium condition for a thermistor can be expressed as:

$$P = K(T - T_a)$$

where

K is the dissipation constant of the thermistor,
T is the thermistor temperature, and
T_a is the ambient temperature.

The dissipation constant K represents the power required to raise the thermistor temperature by one degree Celsius above the ambient temperature. It depends on:

- Thermistor size and shape
- Surface area
- Mounting configuration
- Surrounding medium (air, liquid, vacuum)

A higher dissipation constant implies better heat transfer to the environment and reduced self-heating effects.

Thermal Time Constant

Another important parameter is the thermal time constant, which determines how quickly the thermistor responds to a change in temperature.

The thermal time constant is defined as the time required for the thermistor to reach approximately 63% of the total temperature change when subjected to a step change in ambient temperature.

It can be expressed as:

$$\tau = C/K$$

where

τ (tau) is the thermal time constant,
C is the thermal capacitance (heat capacity) of the thermistor, and
K is the dissipation constant.

Small bead thermistors, having low thermal mass, exhibit very small time constants and therefore respond rapidly to temperature changes. Larger disc-type thermistors respond more slowly due to higher thermal mass.

4.5 TYPES OF THERMISTORS

Thermistors are classified according to the nature of their temperature coefficient of resistance. Depending on whether the resistance decreases or increases with

temperature, thermistors are categorized into Negative Temperature Coefficient (NTC) and Positive Temperature Coefficient (PTC) thermistors. Both types differ in their material properties, operating principles, and applications.

Negative Temperature Coefficient (NTC) Thermistors

NTC thermistors are the most widely used type of thermistors. In these devices, resistance decreases as temperature increases. This behaviour is characteristic of semiconductors in which conductivity rises due to increased thermal excitation of charge carriers.

Physical Basis

As discussed earlier, the intrinsic carrier concentration in a semiconductor increases exponentially with temperature. The resistance of an NTC thermistor can be expressed as:

$$R(T) = R_0 \exp[\beta(\frac{1}{T} - \frac{1}{T_0})]$$

Here,

R(T) is the resistance at temperature T,

R_0 is the resistance at reference temperature T_0,

β is the material constant known as the beta constant.

Since temperature appears in the denominator inside the exponential term, an increase in temperature reduces the exponential value, thereby reducing resistance.

The temperature coefficient of resistance for an NTC thermistor is negative and is defined as:

$$\alpha = (\frac{1}{R})(\frac{dR}{dT})$$

For NTC thermistors, substituting the exponential form yields:

$$\alpha = -\beta/T^2$$

This clearly shows that α is negative and its magnitude decreases with increasing temperature. Therefore, NTC thermistors are more sensitive at lower temperatures.

Materials and Fabrication

NTC thermistors are fabricated by mixing semiconductor metal oxides such as manganese oxide, nickel oxide, cobalt oxide, copper oxide, and iron oxide in suitable proportions. The mixture is then shaped into beads, discs, or chips and sintered at high temperatures to form ceramic semiconductor structures. Electrical conduction in these materials occurs mainly through hopping of electrons between ions of different valence states.

Electrical Characteristics

The resistance–temperature curve of an NTC thermistor is highly nonlinear. At low temperatures, resistance is high due to limited carrier availability and grain boundary barriers. As temperature increases, resistance falls rapidly.

If we take the natural logarithm of the resistance equation:

$$ln\, R = ln\, R_0 + \beta(\frac{1}{T} - \frac{1}{T_0})$$

A plot of ln R versus 1/T yields a straight line with slope β. This graphical method is often used to determine the beta constant experimentally.

Applications of NTC Thermistors

Because of their high sensitivity, NTC thermistors are extensively used in:

- precision temperature measurement
- digital thermometers
- temperature compensation circuits
- automotive coolant temperature sensing
- battery temperature monitoring
- inrush current limiting in power supplies

In inrush current limiting, the high initial resistance limits current when the device is first powered. As current flows, the thermistor heats up, its resistance decreases, and normal current operation is established.

Positive Temperature Coefficient (PTC) Thermistors

PTC thermistors exhibit an increase in resistance with increasing temperature. However, unlike metallic conductors, the resistance increase in PTC thermistors is often highly nonlinear and may occur abruptly near a specific transition temperature.

PTC thermistors can be divided into two main categories:

A. Silistor-type PTC thermistors (based on doped silicon)
B. Switching-type PTC thermistors (based on ferroelectric ceramics)

Silistor-Type PTC Thermistors

Silistor-type PTC thermistors are made from heavily doped silicon. In these devices, resistance increases gradually with temperature because carrier mobility decreases due to increased lattice scattering. They exhibit a nearly linear positive temperature coefficient over a limited temperature range and are mainly used for temperature compensation.

Switching-Type PTC Thermistors

Switching-type PTC thermistors are commonly made from doped barium titanate ($BaTiO_3$), a ferroelectric ceramic material. Below the Curie temperature, the material exhibits relatively low resistance. As the temperature approaches the Curie temperature, the material undergoes a phase transition from ferroelectric to paraelectric state, which increases the grain boundary potential barriers and causes a sharp rise in resistance. This sudden increase in resistance makes switching-type PTC thermistors suitable for protection applications such as overcurrent and overtemperature protection.

Resistance Behaviour in PTC Thermistors

The resistance–temperature characteristic of a switching PTC thermistor typically shows:

- A relatively low resistance region at lower temperatures
- A sharp increase in resistance near Curie temperature
- A high resistance plateau at higher temperatures

This behaviour enables automatic current limiting. When excessive current flows, the device heats up, resistance increases sharply, and current is reduced. Once the device cools down, resistance returns to its original low value.

Applications of PTC Thermistors

PTC thermistors are widely used in:

- Overcurrent protection devices
- Resettable fuses
- Overtemperature protection
- Motor protection circuits
- Battery safety systems
- Self-regulating heaters

Because of their self-limiting behavior, PTC thermistors are highly reliable in safety-critical applications.

Table 4.1: Comparison Between NTC and PTC Thermistors

Feature	NTC Thermistor	PTC Thermistor
Temperature Coefficient	Negative	Positive
Resistance with Temperature	Decreases	Increases
Behavior	Smooth exponential	Gradual or abrupt rise
Typical Materials	Metal oxides	Doped silicon, $BaTiO_3$
Main Use	Temperature sensing	Protection and switching

4.6 CHARACTERISTICS OF THERMISTORS

Heat and electricity traits shape how well a thermistor works in real use. Because they affect how fast it reacts, how stable it stays, plus its overall dependability. Picking the right one means knowing those details, especially when used to sense heat, balance circuits, or guard systems. The main characteristics of thermistors include:

1. Resistance–Temperature Characteristics
2. Temperature Coefficient of Resistance
3. Sensitivity
4. Thermal Time Constant
5. Dissipation Constant
6. Stability and Aging Effects
7. Power Handling Capability

Each of these is discussed in detail below.

Resistance–Temperature Characteristics

The most fundamental characteristic of a thermistor is the variation of resistance with temperature. For an NTC thermistor, the resistance decreases exponentially with increasing temperature, and the relation is given by:

$$R(T) = R_0 \exp[\beta(\frac{1}{T} - 1/T_0)]$$

This exponential nature makes thermistors highly nonlinear devices. The resistance–temperature curve is steep at lower temperatures and gradually flattens at higher temperatures.

If the natural logarithm of resistance is plotted against reciprocal temperature, a straight-line relationship is obtained:

$$ln\, R = ln\, R_0 + \beta(\frac{1}{T} - \frac{1}{T_0})$$

This linearized representation is useful for calibration and parameter extraction.

In PTC thermistors, the resistance–temperature characteristic is different. The resistance increases with temperature, and in switching-type PTC thermistors, the increase is abrupt near the Curie temperature. The curve shows three regions:

- Low-resistance region
- Sharp transition region
- High-resistance region

This behaviour is particularly useful in protection circuits.

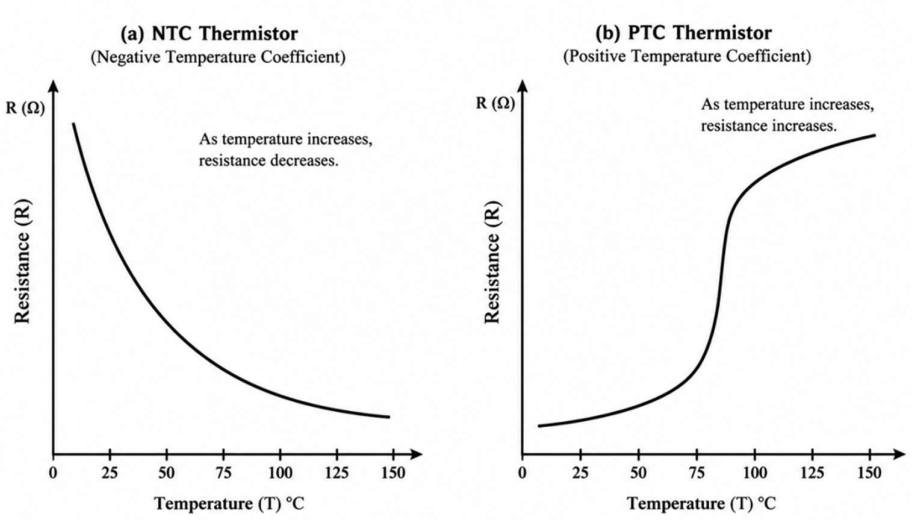


Figure 4.2: Resistance-Temperature curve of NTC and PTC thermistor

Temperature Coefficient of Resistance (TCR)

The temperature coefficient of resistance (TCR) is defined as the fractional change in resistance per unit change in temperature:

$$\alpha = (\frac{1}{R})(\frac{dR}{dT})$$

For NTC thermistors, substituting the exponential resistance relation gives:

$$\alpha = -\frac{\beta}{T^2}$$

This expression shows that:

- The coefficient is negative for NTC thermistors.
- Its magnitude decreases as temperature increases.
- Sensitivity is highest at lower temperatures.

For PTC thermistors, the temperature coefficient is positive and may become very large near the transition temperature.

Compared to metallic resistance temperature detectors (RTDs), thermistors have a much larger magnitude of TCR, which makes them more sensitive within a limited temperature range.

Sensitivity

Sensitivity indicates how effectively a thermistor responds to small temperature changes. It is mathematically defined as:

$$S = \frac{dR}{dT}$$

Because of the exponential relationship between resistance and temperature, the sensitivity of an NTC thermistor is very high at lower temperatures.

High sensitivity is advantageous for precise temperature measurement, but it also contributes to nonlinearity, requiring compensation techniques in practical circuits.

Thermal Time Constant

The thermal time constant determines how quickly a thermistor responds to temperature changes.

It is defined as the time required for the thermistor to reach approximately 63% of the total temperature difference when exposed to a sudden change in ambient temperature.

It is expressed as:

$$\tau = \frac{c}{k}$$

where

T is the thermal time constant,

C is the thermal capacitance (heat capacity) of the thermistor,

K is the dissipation constant.

A small bead thermistor has a small thermal mass and therefore a small time constant, making it suitable for fast-response applications. Larger disc thermistors have larger time constants due to greater thermal inertia.

Dissipation Constant

The dissipation constant indicates the ability of the thermistor to transfer heat to its surroundings.

It is defined as:

$$K = \frac{P}{T - T_a}$$

where

P is the power dissipated,

T is the thermistor temperature,

T_a is the ambient temperature.

The unit of dissipation constant is typically milliwatts per degree Celsius (mW/°C).

A higher dissipation constant means better heat transfer and lower self-heating error.

Stability and Aging

Over long periods of operation, thermistor characteristics may shift slightly due to:

- structural relaxation
- grain boundary modification

- dopant redistribution
- environmental exposure

High-quality thermistors are designed to maintain stability within specified tolerance limits over their operating lifetime.
Aging effects are particularly important in precision measurement systems where calibration accuracy is critical.

Power Rating and Operating Limits
Each thermistor has a specified maximum power rating beyond which permanent damage may occur. The power rating depends on:

- physical size
- material composition
- mounting configuration
- cooling conditions

Exceeding the rated power can lead to thermal runaway in NTC thermistors or irreversible structural changes.

4.7 APPLICATIONS OF THERMISTORS

Thermistors are widely used in scientific, industrial, medical, automotive, and consumer electronic systems because of their high sensitivity, small size, fast response, and low cost. Their applications depend mainly on whether they are NTC or PTC thermistors.

1. Temperature Measurement
NTC thermistors are widely used for accurate temperature measurement. Their resistance changes significantly with small temperature variations, making them suitable for:

- Digital thermometers
- Laboratory temperature probes
- HVAC systems
- Environmental monitoring
- Food processing

2. Temperature Compensation
Thermistors compensate for temperature-induced changes in electronic components and help maintain stable circuit performance.

Applications:

- Amplifier circuits
- Oscillators
- Voltage regulators
- Bridge circuits
- Precision instruments

3. Inrush Current Limiting
NTC thermistors have high resistance at room temperature and limit the initial surge current. As they heat up, their resistance decreases, allowing normal current flow.

Used in:

- Switched-mode power supplies
- AC–DC converters
- LED drivers
- Motor controllers

4. Overcurrent and Overtemperature Protection

PTC thermistors increase their resistance sharply when temperature or current exceeds a critical value, thereby limiting current. They can reset automatically after cooling.

Applications:

- Battery protection
- USB ports
- Consumer electronics
- Transformer protection
- Motor protection

5. Automotive Applications

Thermistors are used for temperature monitoring and control in:

- Engine coolant systems
- Air and oil temperature sensors
- Cabin climate control
- Battery thermal management

6. Biomedical Applications

Due to their small size and high sensitivity, thermistors are suitable for medical temperature measurement.

Applications:

- Clinical thermometers
- Respiratory monitoring
- Skin temperature sensing
- Incubators
- Dialysis equipment

7. Industrial Process Control

Thermistors provide accurate temperature monitoring in:

- Chemical processing
- Food production
- Plastic moulding
- Heating and cooling systems
- Power plants

8. Self-Regulating Heating Systems

PTC thermistors are used in heating systems where increasing temperature automatically increases resistance, reducing current and preventing overheating.

Examples:

- Seat heaters
- Defrosting systems
- Temperature-controlled enclosures
- Dehumidifiers

Advantages

- High sensitivity
- Compact size
- Fast response
- Low cost
- Simple circuitry

Limitations

- Nonlinear response
- Limited temperature range
- Self-heating effects
- Possible aging drift

Despite these limitations, thermistors remain important and widely used temperature-sensing, compensation, protection, and control devices in modern electronic systems.

W CHECK YOUR PROGRESS

Q1: Why the resistance of an NTC thermistor decreases with increasing temperature.
Q2: Derive the resistance–temperature relation of an NTC thermistor starting from thermally activated conduction.
Q3: What is the physical significance of the beta constant in thermistors?
Q4: Distinguish between intrinsic carrier excitation and grain boundary conduction in ceramic thermistors.
Q5: Define the temperature coefficient of resistance and explain why it is negative for NTC thermistors.
Q6: What is the thermal time constant of a thermistor? How does it influence response speed?
Q7: What is dissipation constant?
Q8: How does a switching-type PTC thermistor work?
Q9: How are thermistors used for inrush current limiting?
Q10: Why are thermistors preferred over RTDs in certain applications?

2 4.8 LET US SUM UP

- A thermistor is a temperature-sensitive semiconductor resistor whose resistance changes with temperature and is widely used in sensing and protection circuits.
- Thermistors are generally made from semiconductor metal oxides such as manganese oxide, nickel oxide, cobalt oxide, and copper oxide, whose electrical conductivity depends strongly on temperature.

- The operation of thermistors is based on semiconductor conduction, where the number of charge carriers increases with temperature, leading to an increase in conductivity and corresponding change in resistance.
- In NTC thermistors, resistance decreases with increase in temperature and follows the relation $R(T) = R_0 \exp[\beta (1/T - 1/T_0)]$, where β is a material constant related to activation energy.
- The temperature coefficient of resistance (TCR) of an NTC thermistor is negative and is given by $\alpha = -\beta/T^2$, which indicates high sensitivity to temperature changes.
- Grain boundary effects and hopping conduction play an important role in determining the electrical behaviour of ceramic thermistors.
- Thermistor performance is influenced by practical factors such as self-heating, dissipation constant, thermal time constant, and environmental stability.
- Thermistors are classified into Negative Temperature Coefficient (NTC) and Positive Temperature Coefficient (PTC) types based on the variation of resistance with temperature.
- NTC thermistors are mainly used in temperature measurement, temperature compensation, and inrush current limiting applications.
- PTC thermistors are mainly used in overcurrent protection, overtemperature protection, and self-regulating heating systems.
- Important thermistor characteristics include resistance–temperature behaviour, sensitivity, thermal response, stability, and power handling capability.
- Thermistors have wide applications in automotive systems, biomedical instruments, industrial process control, and electronic protection circuits because of their high sensitivity, compact size, and fast response.

K 4.9 ANSWERS TO CHECK YOUR PROGRESS

1. The resistance of an NTC thermistor decreases because carrier concentration increases exponentially with temperature, increasing conductivity.
2. Starting from thermally activated conduction, the resistance varies as
$$R(T) = R_0 \exp[\beta(\frac{1}{T} - \frac{1}{T_0})]$$
where β is related to activation energy.
3. The beta constant represents the material-dependent parameter that determines the sensitivity of resistance to temperature changes.
4. Intrinsic excitation involves band-to-band carrier generation, whereas grain boundary conduction involves thermally activated transport across potential barriers at grain interfaces.
5. The temperature coefficient of resistance is defined as
$$\alpha = (\frac{1}{R})(\frac{dR}{dT})$$
It is negative in NTC thermistors because resistance decreases with increasing temperature.
6. The thermal time constant is the time required to reach 63% of the final temperature change. It determines response speed.
7. The dissipation constant is the power required to raise thermistor temperature by one degree above ambient. It affects self-heating error.

8. In switching-type PTC thermistors, resistance increases sharply near the Curie temperature due to phase transition and increased grain boundary barriers.
9. In inrush current limiting, the high initial resistance restricts current, and subsequent heating reduces resistance for normal operation.
10. Thermistors offer higher sensitivity and smaller size compared to RTDs, making them suitable for limited-range precision sensing.

O 4.10 FURTHER READING

- S. M. Sze and Kwok K. Ng, *Physics of Semiconductor Devices.*
- B. G. Streetman and S. Banerjee, *Solid State Electronic Devices.*
- R. F. Pierret, *Semiconductor Device Fundamentals.*
- A. S. Sedra and K. C. Smith, *Microelectronic Circuits.*
- D. A. Bell, *Electronic Devices and Circuits.*

M 4.11 MODEL QUESTIONS

Q1. Explain the semiconductor basis of thermistor action with appropriate equations.
Q2. Derive the resistance–temperature relation for an NTC thermistor and discuss the significance of the beta constant.
Q3. Discuss grain boundary conduction mechanisms in ceramic thermistors.
Q4. Compare NTC and PTC thermistors with reference to material composition and applications.
Q5. Explain thermal time constant and dissipation constant in detail.
Q6. Discuss thermistor applications in power electronics and protection systems.
Q7. Describe switching-type PTC thermistors and their role in circuit protection.
Q8. Analyse the advantages and limitations of thermistors as temperature sensors.

*** ***** ***

UNIT 5: OPTOELECTRONIC DEVICES

UNIT STRUCTURE

5.1 Learning Objectives
5.2 Introduction
5.3 Photodiodes
5.3.1 Current and Voltage in an Illuminated Junction
5.4 Solar Cells
5.5 Light Emitting Diodes (LEDs)
5.5.1 Light Emitting Materials
5.6 Laser Diodes
5.7 Let Us Sum Up
5.8 Answers to Check Your Progress
5.9 Further Reading
5.10 Model Questions

5.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- describe the working of photodiodes and the generation of current and voltage in an illuminated junction
- discuss the working principles and characteristics of solar cells and photodetectors
- explain the construction and working of LEDs and the properties of light-emitting materials
- evaluate the working principles and applications of laser diodes.

5.2 INTRODUCTION

In the previous unit, you learned about thermistors and their applications. In this unit, we will study optoelectronic semiconductor devices, which convert electrical energy into light or light into electrical energy. We will begin with photodiodes and examine the current and voltage characteristics of an illuminated junction. Next, we will study solar cells, including their working principle, characteristics, and applications in converting solar energy into electrical energy. We will then discuss Light Emitting Diodes (LEDs), with emphasis on the materials used for light emission and their operating principles. Finally, we will learn about laser diodes, their characteristics, working principle, and applications. These devices are widely used in optical communication, solar energy conversion, lighting, sensing, and display technologies.

5.3 PHOTODIODES

Optoelectronic devices are semiconductor devices that convert electrical energy into optical energy and optical energy into electrical energy. A photodiode is an optoelectronic device that converts incident light into an electrical signal. Examples of photodiodes include silicon PN photodiodes, PIN photodiodes, avalanche photodiodes (APDs), Schottky photodiodes and InGaAs photodiodes. For example, a silicon PIN

photodiode is commonly used in light sensors, while InGaAs photodiodes are widely used in optical-fibre communication systems. The figure 5.1 shows different types of photodiodes used for detecting light and converting it into electrical signals.

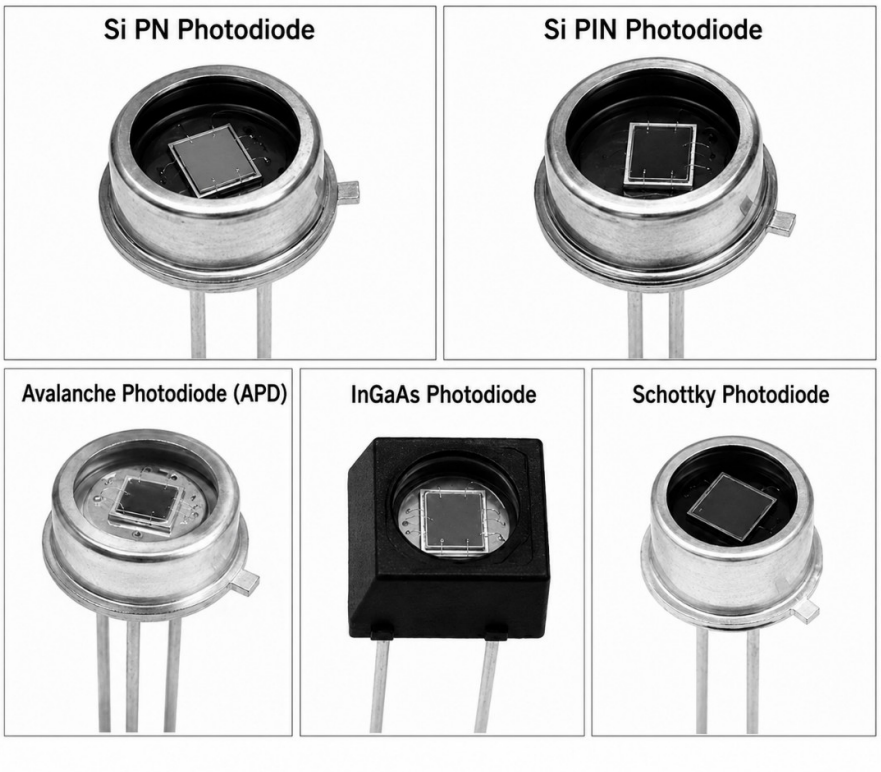
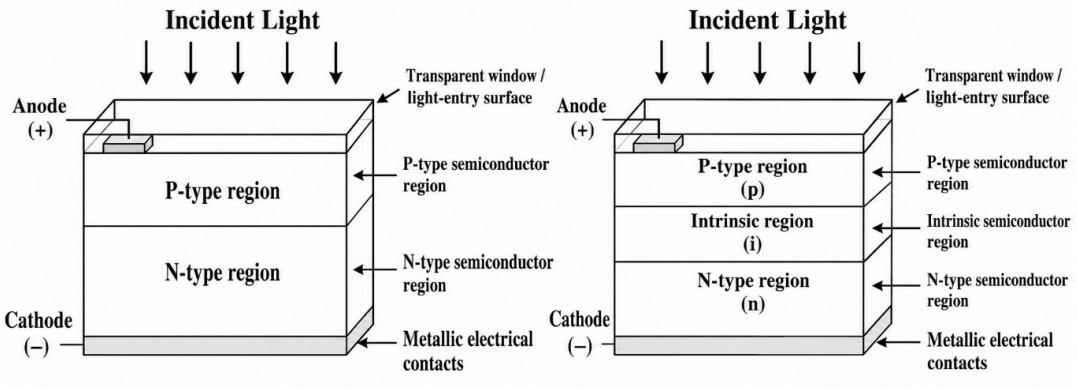


Figure 5.1: Examples of different photodiodes

Basic structure of a photodiode:
A basic photodiode is formed by joining p-type and n-type semiconductor regions, creating a p–n junction. It also contains electrical contacts and a light-entry surface that allows incident light to reach the photosensitive region.
The main structural components are:

- P-type semiconductor region – for example, boron-doped silicon, where holes are the majority carriers.
- N-type semiconductor region – for example, phosphorus-doped silicon, where electrons are the majority carriers.
- Electrical contacts – provide connection to the external circuit.
- Transparent window or light-entry surface – allows light to enter the device.

For a PIN photodiode, an additional intrinsic (I) region, such as nearly pure silicon, is placed between the p-type and n-type regions. Figure 5.2(a) and (b) shows the basic structure of a PN juncrtion photodiode and P-I-N photodiode respectively.



(a) **(b)**

Figure 5.2: Basic structure of (a) PN junction photodiodes and (b) p-i-n photodiode

Basic operational principle of photodiodes:

When a p-type and an n-type semiconductor are joined together, a PN junction is formed. The n-region contains a large number of free electrons, while the p-region contains a large number of holes. After the two regions are joined, electrons from the n-region diffuse toward the p-region, and holes from the p-region diffuse toward the n-region. Near the junction, electrons and holes recombine with each other. As a result, fixed positive donor ions remain on the n-side and fixed negative acceptor ions remain on the p-side. This region becomes almost free of mobile charge carriers and is therefore called the depletion region. The fixed charges in this region produce an internal electric field across the junction. A photodiode is generally operated in reverse bias because reverse bias increases the width of the depletion region and strengthens the electric field. When light with sufficient photon energy falls on the depletion region, the photons are absorbed and generate electron–hole pairs. The strong electric field quickly separates these carriers, driving electrons toward the n-side and holes toward the p-side. This movement of charge carriers produces a photocurrent in the external circuit. Reverse bias also keeps the dark current small, making the photocurrent produced by the incident light easier to detect. The photocurrent is approximately proportional to the intensity of the incident light. Thus, a photodiode converts the incident optical signal into a measurable electrical signal.

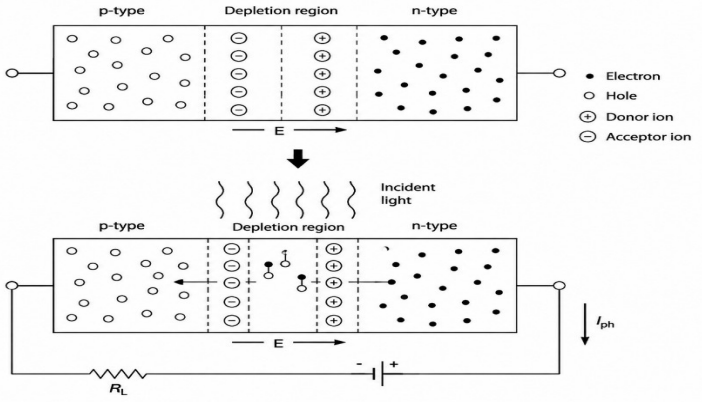


Figure 5.3: Operational principle of a PN junction photodiodes photodiode

A PIN photodiode consists of a p-type layer, an intrinsic (I) layer, and an n-type layer, arranged as P–I–N. Although a conventional PN junction already has a depletion region, an intrinsic layer is introduced in a PIN photodiode to provide a much wider and well-defined active region. The intrinsic layer contains very few free charge carriers and therefore becomes almost completely depleted when the photodiode is operated under reverse bias. As a result, the depletion region extends across most of the intrinsic layer. When light enters the photodiode, photons are absorbed mainly in the wide intrinsic region and generate electron–hole pairs. The strong electric field across the depleted intrinsic region rapidly separates these charge carriers, driving electrons toward the n-side and holes toward the p-side. Their movement produces a photocurrent in the external circuit. Since the intrinsic layer provides a larger active volume for photon absorption, more electron–hole pairs can be generated and efficiently collected compared with a conventional PN photodiode. This produces a larger photocurrent for a given incident optical power, thereby improving the performance of the device. The wide depletion region also reduces the junction capacitance, allowing the photodiode to respond rapidly to changes in the incident light. Thus, a PIN photodiode efficiently converts an optical signal into an electrical signal and is widely used in high-speed optical communication and light detection.

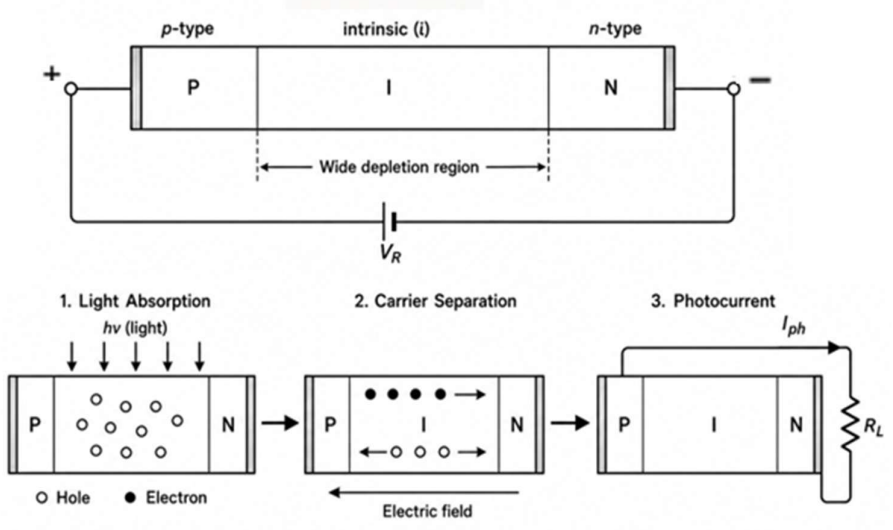


Figure 5.4: Operational Principle of a P-I-N Junction Photodiode

5.3.1 Current and Voltage in an Illuminated Junction

When a photodiode is exposed to light, the total current flowing through the junction is the combination of:

- The diode current (due to applied bias)
- The photocurrent (due to illumination)

The total current can be expressed as:

$$I = I_D - I_{ph}$$

where,

I_D is the normal diode current and
I_{ph} is the light-generated current

The negative sign indicates that the photocurrent flows in the direction opposite to the conventional forward diode current. Under reverse bias, the normal diode current is small, and the output current is approximately equal to the photocurrent. As illumination intensity increases, the photocurrent increases proportionally with the incident light intensity. Under zero bias condition (open-circuit), a voltage known as the photovoltaic voltage develops across the junction, known as the open-circuit voltage or photovoltaic voltage. This principle is utilized in solar cells. Thus, illumination shifts the current–voltage characteristics of the diode downward, by an amount approximately equal to the photocurrent. Consequently, the magnitude of the current in the reverse-bias region increases with increasing light intensity.

Modes of Operation (Photovoltaic and Photoconductive)
Photodiodes operate in two main modes:

1. **Photovoltaic Mode**
   Photovoltaic mode is the operating mode of a photodiode in which no external bias voltage is applied. When light falls on the photodiode, it generates electron–hole pairs and produces a photo voltage across the device. Since no external reverse-bias voltage is applied, this mode produces very low noise. However, it generally has a slower response compared with reverse-biased operation. This mode is commonly used in solar cells and low-noise optical detection systems.
2. **Photoconductive Mode**
   Photoconductive mode is the operating mode of a photodiode in which the device is operated under reverse bias. The reverse bias creates a strong electric field that quickly separates and collects the light-generated charge carriers, producing a photocurrent. It also widens the depletion region, allowing faster collection of the photogenerated carriers and resulting in a faster response time. This mode provides good sensitivity and improved linearity over a wide range of incident light intensity. It is therefore widely used in high-speed optical communication systems and fast optical detection applications.

Characteristics and Performance Parameters
The performance of a photodiode is evaluated using several important parameters:

- **Responsivity**
  It measures the output current per unit incident optical power. Higher responsivity indicates better sensitivity.
- **Quantum Efficiency**
  It represents the fraction of incident photons that successfully generate electron–hole pairs contributing to photocurrent.
- **Dark Current**
  The small current that flows in the absence of light. Lower dark current improves detection accuracy.
- **Response Time**
  The time required for the photodiode to respond to changes in light intensity.
- **Spectral Response**

Indicates sensitivity over different wavelengths. It depends on the band gap of the semiconductor material.

Applications of Photodiodes

Photodiodes are extensively used in:

- Optical fibre communication systems
- Light meters
- Smoke detectors
- Barcode scanners
- Medical imaging instruments
- Infrared detection systems
- Automatic lighting controls

Due to their reliability, compact size, and fast response, photodiodes remain one of the most important optoelectronic detection devices.

5.4 SOLAR CELLS

A solar cell is a device that converts sunlight directly into electrical energy through a phenomenon called the photovoltaic effect. It is basically a semiconductor device that generates electricity when light falls on it. Solar cells are very important for generating clean and renewable energy. They provide electricity without directly producing harmful emissions during operation.

When sunlight falls on a solar cell, it provides energy to the electrons in the semiconductor material, creating electron–hole pairs. The solar cell contains a p–n junction, where the built-in electric field separates these electrons and holes. This separation produces a photovoltage, and when an external circuit is connected, an electric current flows through the circuit.

There are three generations of solar cells, commonly classified as first-generation, second-generation, and third-generation solar cells. The p–n junction solar cell is the basic and most widely used type of first-generation solar cell. In proceeding section, we will discuss the principle and operation of a p–n junction solar cell.

Principle of Photovoltaic Conversion

A p–n junction solar cell works on the principle of photovoltaic conversion, in which light energy is converted into electrical energy through the generation, separation, and collection of charge carriers. When a p-type semiconductor and an n-type semiconductor are brought into contact, electrons diffuse from the n-type region toward the p-type region, while holes diffuse in the opposite direction. This movement of charge carriers produces a depletion region near the junction. Due to the movement of charges in this region, an internal electric field is established. At thermal equilibrium, the drift current produced by this electric field balances the diffusion current near the junction. The internal electric field gives rise to a potential called built-in potential. This built-in potential is very important for photovoltaic conversion because it helps to separate the charge carriers generated by light. When sunlight with energy greater than or equal to the band gap energy of the semiconductors ($h\nu \geq E_g$)falls on the solar cell, the semiconductor absorbs the incident light and generates electron–hole pairs. The absorbed light provides energy to the electrons, allowing them to move to higher-

energy states and leaving holes behind. Thus, electron–hole pairs are generated inside the semiconductor when the junction is illuminated. The photo-generated electron–hole pairs are then separated by the built-in electric field of the junction. The electrons and holes are directed toward different regions of the solar cell. In the p-n junction, the separated these minority carriers move toward the respective semiconductor regions. After separation, the charge carriers move through the semiconductor toward the electrical contacts. Under open-circuit conditions, the photo-generated carriers cannot continuously flow through an external circuit. Therefore, they accumulate in their respective regions of the solar cell. This accumulation increases the carrier density. As a result, an open-circuit voltage V develops across the solar cell. When the solar cell is connected to an external circuit under short-circuit conditions, the photo-generated carriers move toward and are collected by the contact electrodes instead of accumulating in the semiconductor. Their movement through the external circuit produces a short-circuit current. Unlike photodiodes operating in reverse bias, solar cells typically operate in photovoltaic mode (no external bias).

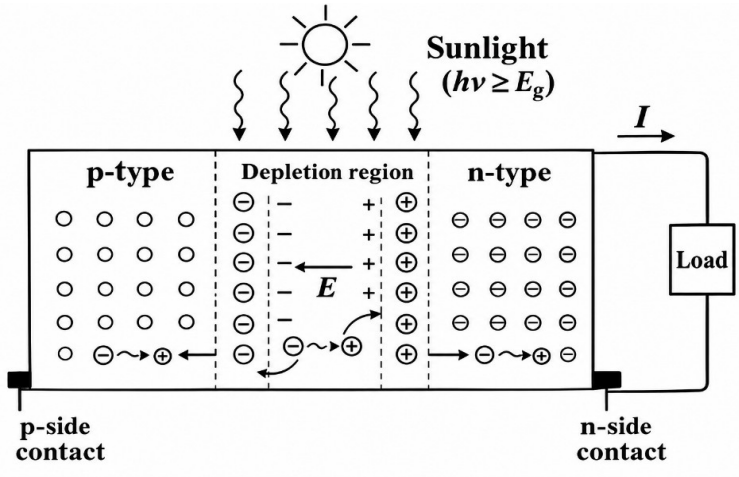


Figure 5.5: Generation of electron hole pairs and photocurrents of a pn junction solar cell

Current–Voltage Characteristics of Solar Cells

The current–voltage (I–V) characteristic is the relationship between the current produced by a solar cell and the voltage across its terminals under a given light intensity. It shows how the current changes when the terminal voltage of the solar cell changes. To obtain the I–V characteristic, the terminal voltage is varied using an external load or measuring instrument, and the corresponding current is measured at each voltage. This gives different pairs of voltage and current values, which are plotted as an I–V curve. The I–V curve allows us to determine important parameters such as short-circuit current (I_{SC}), open-circuit voltage (V_{OC}), and the maximum power point (M_{PP}). Thus, it helps us evaluate the performance and power-generating capability of the solar cell. The illuminated I–V characteristic of a solar cell is generally plotted in the fourth quadrant, where the terminal voltage is positive and the current is negative according to the conventional sign convention. This indicates that the solar cell is delivering electrical power to the external circuit. The power is given by P=VI, and the negative sign represents power supplied by the solar cell. The figure 5.6 IV characteristic of a solar cell.

The total current in a solar cell can be written as:

$$I = I_D - I_L$$

where

I_D is the diode current and

I_L is the light-generated current.

Important operating parameters include:

Short-Circuit Current (Isc)

The current when the output terminals are shorted (voltage = 0). It depends directly on light intensity.

Open-Circuit Voltage (Voc)

The voltage when no external current flows (current = 0).

Maximum Power Point (MPP)

The point on the I–V curve where the product of current and voltage is maximum.

The power output is given by:

$$P = IV$$

The I–V curve of a solar cell lies in the fourth quadrant, indicating power generation

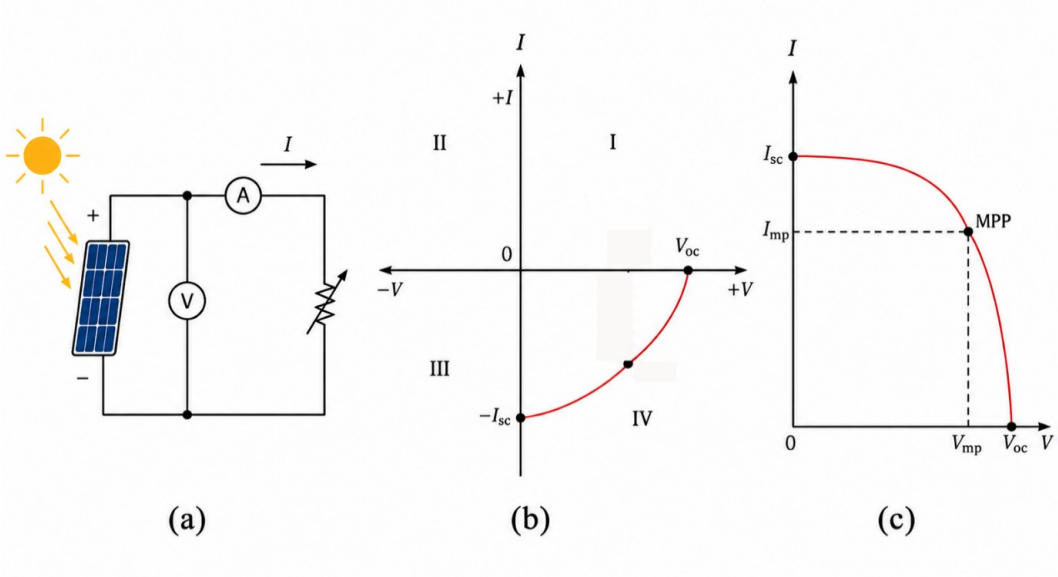


Figure 5.6: (a) I–V measurement circuit; (b) fourth-quadrant I–V characteristic; (c) conventional first-quadrant I–V curve showing I_{SC}, V_{OC}, MPP

The current–voltage (I–V) characteristic of a solar cell under illumination differs from that of a normal diode. Illumination shifts the I–V curve downward due to the generation of photocurrent.

Efficiency and Performance Factors

The efficiency of a solar cell is defined as the ratio of electrical power output to incident solar power.

Several factors influence efficiency:

- Band gap of the semiconductor
- Carrier lifetime
- Recombination losses
- Series and shunt resistance
- Temperature

- Surface reflection losses

Practical losses include:

- Thermalization losses (excess photon energy converted to heat)
- Reflection of incident light
- Incomplete absorption
- Non-radiative recombination

Optimizing these factors improves energy conversion efficiency.

Materials Used in Solar Cells

The choice of semiconductor material is critical for solar cell performance. Materials must have:

- Suitable band gap
- High absorption coefficient
- Good carrier mobility
- Long carrier lifetime

Common materials include:

- **Silicon (Si)**
  Most widely used due to abundance and stability. Available in:
  Monocrystalline
  Polycrystalline
  Amorphous forms
- **Gallium Arsenide (GaAs)**
  High efficiency and suitable for space applications.
- **Cadmium Telluride (CdTe)**
  Used in thin-film solar cells.
- **Copper Indium Gallium Selenide (CIGS)**
  Used in high-efficiency thin-film technologies.

Material engineering and hetero structures are used to enhance performance.

Applications of Solar Cells

Solar cells are widely used in:

- Residential and commercial rooftop systems
- Solar power plants
- Space satellites
- Remote sensing systems
- Calculators and small electronic devices
- Street lighting systems
- Rural electrification projects

With increasing global demand for renewable energy, solar cells have become one of the most important applications of semiconductor physics.

5.5 LIGHT EMITTING DIODES (LEDs)

A Light-Emitting Diode (LED) is a semiconductor device that emits light when an electric current passes through it. This is essentially the opposite of what devices such as photodetectors do, which convert light into electrical signals. LEDs convert electrical energy into light energy through a process called electroluminescence. LEDs are

widely used in display systems, lighting, optical communication, and electronic indicators.

The colour of light emitted by an LED mainly depends on the semiconductor material used to make it and its energy band gap. Different semiconductor materials have different band-gap energies, resulting in the emission of different colours of light.

Principle of Electroluminescence

Electroluminescence is the phenomenon in which light is emitted from a semiconductor when charge carriers recombine under forward bias conditions.

When a p–n junction diode is forward biased:

- Electrons are injected from the n-region into the p-region
- Holes are injected from the p-region into the n-region

In the junction region, electrons and holes recombine. If the recombination process is radiative, energy is released in the form of photons. The energy of the emitted photon is approximately equal to the band gap energy of the material. Therefore, the wavelength (and hence colour) of the emitted light depends on the semiconductor band gap.

Direct band gap semiconductors are preferred for LEDs because they allow efficient radiative recombination. In indirect band gap materials, most recombination energy is lost as heat rather than light.

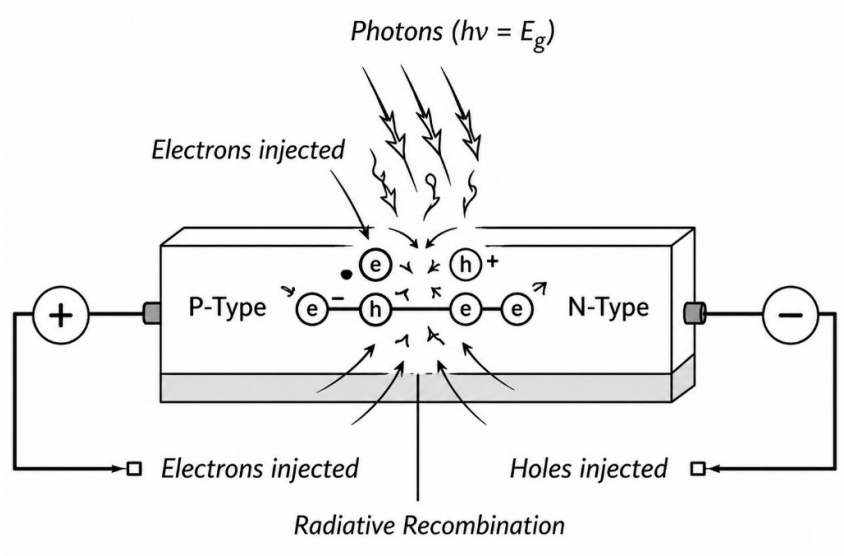


Figure 5.7: Electroluminescence in an LED through radiative recombination of electrons and holes at the p–n junction.

5.5.1 Light Emitting Materials

The choice of semiconductor material determines the wavelength and efficiency of the emitted light. Materials used in LEDs must possess:

- Direct band gap
- High radiative recombination probability
- Good thermal stability
- Suitable electrical conductivity

Common light-emitting materials include:

- Gallium Arsenide (GaAs) – Infrared emission
- Gallium Phosphide (GaP) – Red and green emission
- Gallium Arsenide Phosphide (GaAsP) – Red to yellow emission
- Gallium Nitride (GaN) – Blue emission
- Indium Gallium Nitride (InGaN) – Blue and green emission

White LEDs are typically produced using blue LEDs coated with phosphor materials that convert part of the blue light into longer wavelengths.
Band gap engineering through alloying allows precise control of emission wavelength.

Spectral Characteristics and Efficiency
Spectral Characteristics
The spectral output of an LED is not perfectly monochromatic. Instead, it exhibits a narrow wavelength distribution around a central peak wavelength. The spectral width depends on:

- Material composition
- Temperature
- Carrier concentration

LEDs emit incoherent light with a relatively broad spectral width compared to laser diodes.

Efficiency of LEDs
The efficiency of an LED is determined by several factors:

- Internal quantum efficiency – Fraction of recombination events that produce photons
- External quantum efficiency – Fraction of generated photons that escape the device
- Carrier injection efficiency
- Optical extraction efficiency

Losses occur due to:

- Non-radiative recombination
- Absorption within the semiconductor
- Reflection at semiconductor-air interface
- Heat generation

Improved device design and surface texturing enhance light extraction efficiency.

LED Structures and Fabrication
LEDs are fabricated using compound semiconductor materials through processes such as epitaxial growth and doping.
Typical structural features include:

- p–n junction
- active recombination region
- ohmic contacts
- encapsulation with transparent material

Modern LEDs often use heterojunction structures to improve carrier confinement and recombination efficiency. Double heterostructures are commonly employed to enhance performance.

Encapsulation with a transparent epoxy lens improves light extraction and mechanical protection.

Applications of LEDs

LEDs are widely used due to their:

- Low power consumption
- Long lifetime
- High efficiency
- Compact size
- Fast switching capability

Major applications include:

- Indicator lamps
- Digital displays
- Traffic signals
- Automotive lighting
- LCD backlighting
- Solid-state lighting systems
- Optical communication (short distance)
- Medical devices
- Decorative lighting

With advancements in high-brightness LEDs, solid-state lighting has become a major alternative to conventional incandescent and fluorescent lamps.

5.6 LASER DIODES

A laser diode is a semiconductor device that produces a highly intense, narrow, and directional beam of light through stimulated emission. The emitted light is typically nearly monochromatic, meaning it has a very narrow range of wavelengths. Unlike ordinary light-emitting devices, which generally produce light over a broader range of wavelengths and with wider angular spread, a laser diode produces a more coherent and directional beam. Laser diodes are widely used in optical communication, barcode scanners, optical storage devices, laser printers, medical instruments, and industrial systems. They are widely used because of their small size, high efficiency, reliability, and very fast response, allowing them to be modulated at high speeds.

Principle of Stimulated Emission

The operation of a laser is based on stimulated emission. To understand this process, consider an atom or electron with two energy levels, E_1 and E_2.

When the electron absorbs energy, it moves from the lower energy level E_1 to the higher energy level E_2. This is called excitation. An excited electron can return to the lower energy level in two ways: spontaneous emission and stimulated emission.

In spontaneous emission, the excited electron returns to the lower energy level on its own and emits a photon with energy

$$h\nu = E_2 - E_1$$

where E_2 and E_2 are the higher and lower energy states, respectively. The emitted photons in this process have random directions and phases.

In stimulated emission, an incoming photon with energy equal to the energy difference between the two levels interacts with an excited electron. It stimulates the electron to return to the lower energy level and emit another photon. The emitted photon has the same frequency, phase, direction of propagation, and polarization as the incoming photon.

Thus, one photon produces two coherent photons, and these photons can stimulate further emissions. This leads to amplification of light, which is the fundamental principle of laser operation

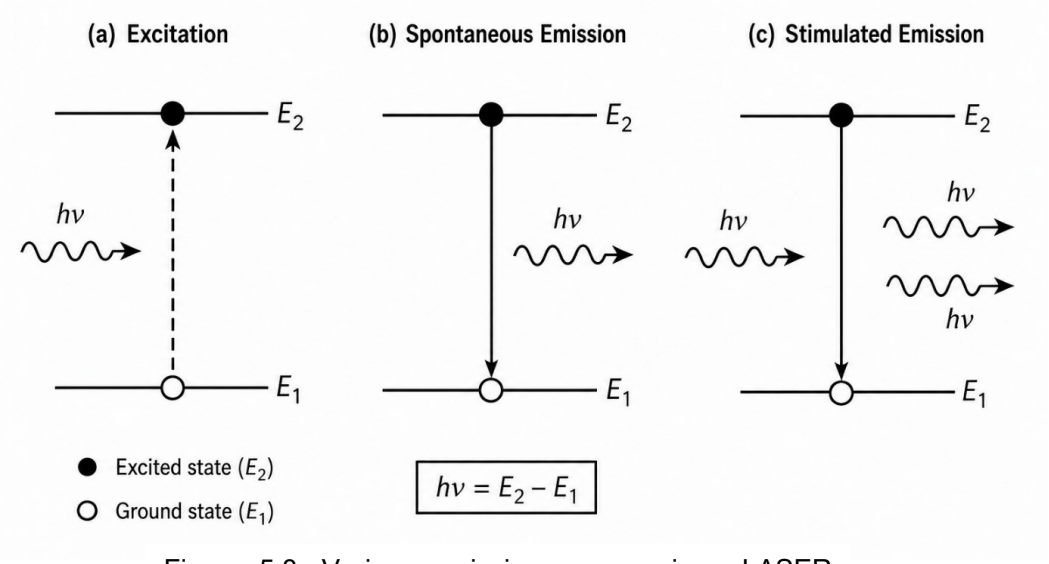


Figure 5.8: Various emission process in a LASER Diode

Population Inversion in Semiconductors

For stimulated emission to dominate over absorption, a condition known as population inversion must be achieved.

Under normal thermal equilibrium:

- Most electrons occupy the lower energy (valence band) states
- Very few electrons occupy higher energy (conduction band) states

For laser action, the number of electrons in the excited state must exceed those in the lower state within the active region.

In semiconductor laser diodes, population inversion is achieved by:

- Heavy forward bias injection
- Carrier confinement within the active region
- Use of heterojunction structures

When carrier injection exceeds a critical level, the probability of stimulated emission becomes greater than absorption.

Structure and Operation of Laser Diodes

A basic laser diode consists of:

- p-type semiconductor layer
- n-type semiconductor layer
- Active region (where recombination occurs)
- Reflective end facets forming an optical cavity

When forward biased:

- Electrons and holes are injected into the active region
- Recombination produces photons
- Photons reflect back and forth between the cleaved end faces
- Optical amplification occurs through stimulated emission

The reflective end faces act as a resonant cavity, allowing only certain wavelengths to be amplified. This leads to narrow spectral emission.

Modern laser diodes often use:

- Double heterostructures
- Quantum well structures

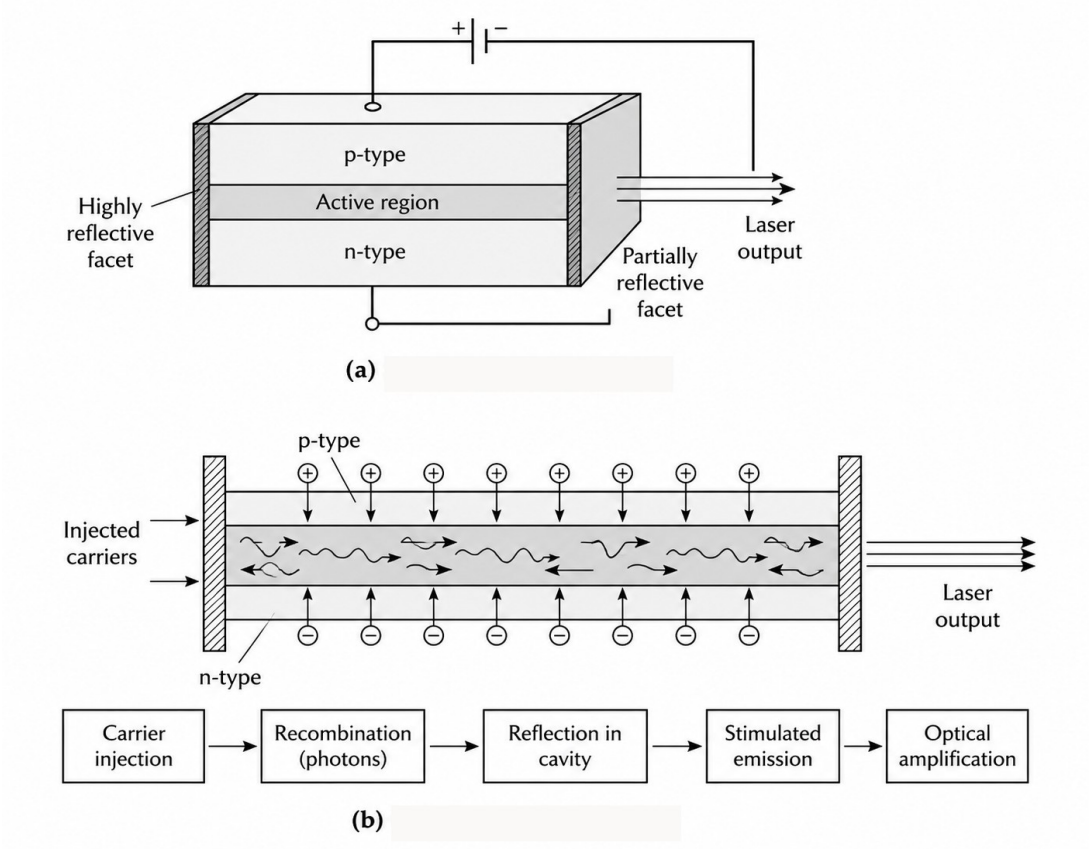


Figure 5.9: LASER Diode (a) Structure and (b) Operation

Threshold Condition and Output Characteristics

Laser action does not occur at all current levels. A minimum current, called the threshold current, must be exceeded for stimulated emission to dominate.

Below threshold:

- Device behaves like an LED
- Emission is spontaneous and incoherent

Above threshold:

- Stimulated emission dominates
- Output intensity increases rapidly
- Emission becomes coherent

The output power increases almost linearly with current above threshold.

Important characteristics include:

- Threshold current
- Slope efficiency
- Spectral width
- Beam divergence
- Modulation speed

Laser diodes exhibit narrow spectral width and highly directional output compared to LEDs.

Applications of Laser Diodes

Laser diodes are used in:

- Optical fiber communication systems
- Barcode scanners
- CD/DVD/Blu-ray players
- Laser printers
- LIDAR systems
- Medical surgery and therapy
- Industrial cutting and alignment systems
- Optical sensing

In optical communication, laser diodes are preferred due to their:

- High modulation speed
- Narrow spectral width
- High output power
- Coherent radiation

They form the backbone of modern fiber-optic transmission systems.

W CHECK YOUR PROGRESS

Q1. What is meant by an optoelectronic device?
Q2. State the condition required for optical absorption in a semiconductor.
Q3. Differentiate between radiative and non-radiative recombination processes.
Q4. What is the working principle of a photodiode under illumination.
Q5. What is the difference between photovoltaic mode and photoconductive mode of photodiode operation?
Q6. Define short-circuit current and open-circuit voltage in a solar cell.
Q7. What factors affect the efficiency of a solar cell?
Q8. Define responsivity of a photodetector.
Q9. Why are direct band gap semiconductors preferred for LEDs?
Q10. What is the principle of electroluminescence.
Q11. What is population inversion and why is it necessary for laser action?
Q12. Define threshold current in a laser diode.
Q13. Compare LED and laser diode in terms of emission characteristics.

2 5.7 LET US SUM UP

- Optoelectronic devices are semiconductor devices that convert energy between optical and electrical forms.
- Optical absorption occurs when photon energy is equal to or greater than the semiconductor band gap, leading to electron–hole pair generation.

- Radiative recombination produces light, while non-radiative recombination dissipates energy as heat.
- Photodiodes convert incident light into electrical current and operate in photovoltaic or photoconductive modes.
- Solar cells function on the photovoltaic effect and convert solar energy into electrical power without external bias.
- The performance of solar cells depends on parameters such as short-circuit current, open-circuit voltage, and maximum power point.
- Photodetectors include PIN and avalanche photodiodes, offering high sensitivity and fast response for optical communication systems.
- Responsivity and quantum efficiency are key performance parameters of photodetectors.
- Light Emitting Diodes (LEDs) operate on the principle of electroluminescence and require direct band gap materials for efficient light emission.
- Laser diodes operate based on stimulated emission and require population inversion and threshold current for coherent radiation.
- Laser diodes produce monochromatic, coherent, and highly directional light and are widely used in optical fiber communication systems.
- Material properties, band gap engineering, and device structure strongly influence the efficiency and performance of optoelectronic devices.

K 5.8 ANSWERS TO CHECK YOUR PROGRESS

Answer to Q No 1. An optoelectronic device converts optical energy into electrical energy or vice versa through carrier generation and recombination processes in semiconductors.

Answer to Q No 2. Optical absorption occurs when photon energy is equal to or greater than the band gap energy of the semiconductor.

Answer to Q No 3. Radiative recombination releases energy in the form of photons, while non-radiative recombination releases energy as heat.

Answer to Q No 4. A photodiode generates photocurrent when light creates electron–hole pairs that are separated by the built-in electric field of the junction.

Answer to Q No 5. In photovoltaic mode, the photodiode operates without external bias. In photoconductive mode, it operates under reverse bias for faster response.

Answer to Q No 6. Short-circuit current is the current when output voltage is zero. Open-circuit voltage is the voltage when output current is zero.

Answer to Q No 7. Solar cell efficiency depends on band gap, carrier lifetime, recombination losses, temperature, and resistive losses.

Answer to Q No 8. Responsivity is the ratio of photocurrent to incident optical power.

Answer to Q No 9. Direct band gap semiconductors allow efficient radiative recombination, making them suitable for LEDs.

Answer to Q No 10. Electroluminescence is light emission due to radiative recombination of carriers under forward bias.

Answer to Q No 11. Population inversion is a condition where more carriers occupy higher energy states than lower ones, necessary for stimulated emission.

Answer to Q No 12. Threshold current is the minimum current required for stimulated emission to dominate and produce laser action.

Answer to Q No 13. LEDs emit incoherent light with broader spectral width, while laser diodes emit coherent, monochromatic, and highly directional light.

O 5.9 FURTHER READING

- Bhattacharya, P. (2006). *Semiconductor optoelectronic devices* (2nd ed.). Pearson Education India.
- Sze, S. M., & Ng, K. K. (2007). *Physics of semiconductor devices* (3rd ed.). Wiley India.
- Streetman, B. G., & Banerjee, S. (2016). *Solid state electronic devices* (7th ed.). Pearson India.
- Singh, J. (2003). *Semiconductor optoelectronics: Physics and technology*. McGraw Hill Education (India).
- Neamen, D. A. (2012). *Semiconductor physics and devices* (4th ed.). McGraw Hill Education (India).
- Tyagi, M. S. (2008). *Introduction to semiconductor materials and devices*. Wiley India.
- Saxena, A. K. (2009). *Optoelectronic devices and circuits*. PHI Learning Pvt. Ltd.

M 5.10 MODEL QUESTIONS

Short Answer Questions

Q1. Define optoelectronic devices.
Q2. What is electroluminescence?
Q3. Define responsivity.
Q4. What is short-circuit current?
Q5. State the condition for stimulated emission.
Q6. What is threshold current in a laser diode?

Descriptive Questions

Q1. Explain the working principle of a photodiode with suitable diagram.
Q2. Describe the photovoltaic effect and operation of solar cells.
Q3. Discuss PIN and avalanche photodiodes with advantages.
Q4. Explain the principle of electroluminescence in LEDs.
Q5. Describe the structure and working of a laser diode.

Analytical / Long Answer Questions

Q1. Discuss the energy conversion mechanisms in optoelectronic devices.
Q2. Compare photodiodes, solar cells, LEDs, and laser diodes based on structure and operation.
Q3. Analyze the current–voltage characteristics of a solar cell under illumination.
Q4. Explain population inversion and threshold condition in semiconductor lasers.
Q5. Discuss efficiency limitations in optoelectronic devices.

*** ***** ***

UNIT 6: BIPOLAR JUNCTION TRANSISTOR

UNIT STRUCTURE

6.1 Learning Objectives
6.2 Introduction
6.3 Introduction to Bipolar Junction Transistor
  6.3.1 Types of Transistor
  6.3.2 Configuration
  6.3.3 Operation Modes
6.4 Working of Transistor
  6.4.1 PNP Transistor
  6.4.2 NPN Transistor
6.5 Transistor Characteristics
6.6 Let Us Sum Up
6.7 Answers to Check Your Progress
6.8 Further Reading
6.9 Model Questions

6.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- define a Bipolar Junction Transistor and list the types of transistors
- explain the different transistor configurations
- differentiate between the operation modes of a transistor
- illustrate the working of PNP and NPN transistors in basic electronic circuits
- analyse the input and output characteristics of a transistor to determine its performance in circuits.

6.2 INTRODUCTION

In the previous unit, we have learned about photoelectronic devices. In this unit, learners will explore the Bipolar Junction Transistor (BJT), a fundamental semiconductor device widely used in electronic circuits for amplification and switching applications. The unit covers the basic concepts and types of transistors, along with different transistor configurations, such as common-base, common-emitter, and common-collector configurations. Learners will also learn about the various operation modes of a transistor and understand the working principles of PNP and NPN transistors**.** Finally, the input and output characteristics of a transistor will be explored at the end of the unit.

6.3 INTRODUCTION TO BIPOLAR JUNCTION TRANSISTOR

A transistor is a semiconductor device widely employed for the amplification and switching of electrical signals in electronic circuits. Among various types, the Bipolar Junction Transistor (BJT) is particularly significant due to its ability to control large output currents using relatively small input signals. A BJT is constructed from a single semiconductor crystal in which two p–n junctions are formed by placing a thin region of

one type of semiconductor between two regions of the opposite type and form either an n–p–n or p–n–p structure.

A transistor is a semiconductor device generally used to amplify or switch electric signals. A transistor is made of a single piece of a semiconductor crystal containing two P-N junctions formed by sandwiching one type of semiconductor material between two layers of other type. A transistor has three regions, namely; emitter, base and collector.

Emitter: It is the left most region of transistor and is more heavily doped than the other regions because it is mainly used to release electrons or holes to the next region i.e. base.

Base: The middle region is called the base. It is very lightly doped and is very thin because it passes most of the emitter injected carriers to the next region i.e. collector.

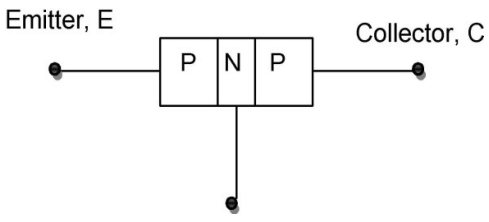
Collector: It is the right most region of the transistor. The main function is to collect the majority carriers through base. It is moderately doped and large in size because it is required to dissipate more heat. The junction between the emitter and base is called the emitter- base junction which is in forward biased and has low resistance. The other junction between the base and collector is called the collector- base junction which is in reverse biased and has high resistance. The Transistor has the capability of transferring a signal from low resistance circuit to high resistance circuit. Hence the name transistor is joining the two terms “transfer” and “resistor”.

6.3.1 Types of Transistor

Transistor is also known as Bipolar Junction Transistor (BJT) because both the carriers i.e. electrons and holes are takes place in current conduction. The BJTs are divided into two types.

- P-N-PTransistor and
- N-P-NTransistor.

P-N-P Transistor: A P-N-P transistor is composed of two P-type semiconductors separated by a thin layer of N-type semiconductor. In this transistor the conduction is by holes.



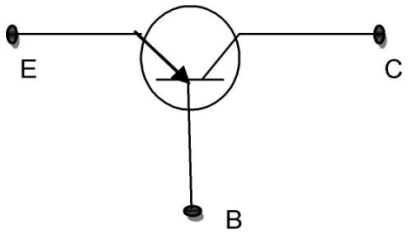


Fig.6.1(a) Structure of P-N-P Transistor Figure 6.1(b): Circuit Symbol of PNP transistor

N-P-N Transistor: An N-P-N transistor is made of two N type material sandwiched between a P type material.

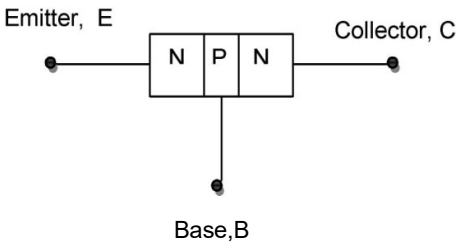


Figure 6.2: (a) Structure of N-P-N Transistor

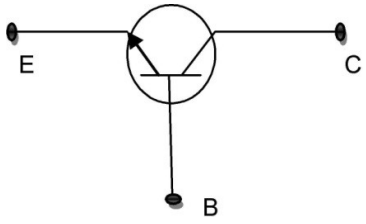


Figure 6.2 (b) Circuit Symbol

6.3.2 Configuration

Transistor Configuration

Generally, there are three types of circuit arrangements for operating a transistor. Theyare–

- Common Base (CB)
- Common Emitter (CE)
- Common Collector (CC).

Common Base: In this configuration the base terminal is common to both the input and output circuit. The input signal is applied between the emitter and base and the output is taken out from the collector and the base terminal.

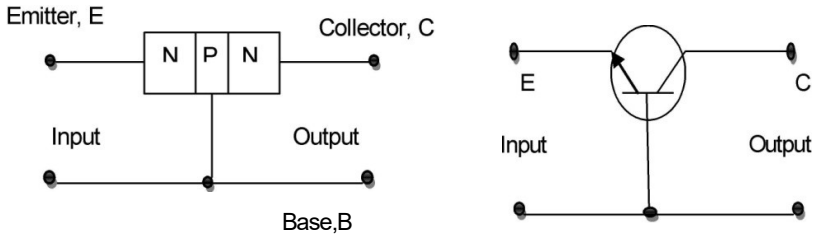


Figure 6.3: Common base configuration of NPN transistor

Common Emitter: Here the emitter terminal is common to both the input and output circuit. The input signal is applied between the base and emitter terminal and the output is taken out from the collector and emitter terminal.

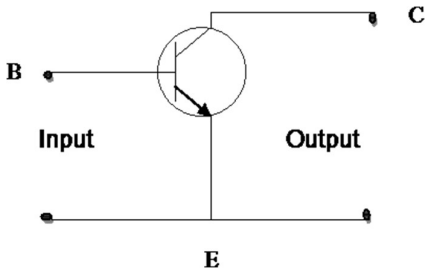


Figure 6.4: Common emitter configuration of NPN transistor

Common Collector: In this configuration the collector terminal is common to both input and output circuit. Here the input signal is applied between the base and collector terminal and the output is obtained from the emitter and collector terminal.

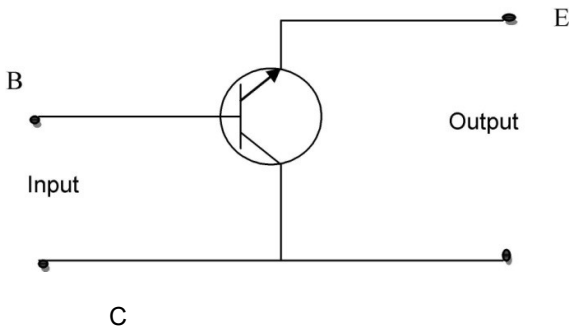


Figure 6.5: Common Collector Configuration of NPN transistor

6.3.3 Operation Modes

Transistor can be operated in three different modes. They are discussed below.

Active Mode: In this mode, the emitter-base junction of the BJT is forward biased and the collector-base junction is reversed biased. This mode of the transistor is used for amplification.

Saturation mode: In this mode, the emitter-base junction and collector-base junction are forward biased and the transistor has very large value of current. This mode is suitable when the BJT is operated as a switch.

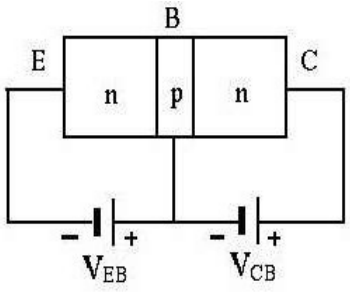
Cut-off mode: In this mode both the emitter base and the collector base junctions are reversed biased. In this case, the current in the transistor is practically zero, because emitter provides no current.

The BJT operates as an open switch when in cut-off mode for switching application of transistor; saturation and cut-off mode are used alternately.

The three operating modes are shown in figure 6.6.

Operating Mode	EBJ	CBJ
Cut-off	Reverse	Reverse
Active	Forward	Reverse
Saturation	Forward	Forward

For p-n-p transistor the polarities of the batteries will be just reversed.



Forward active mode

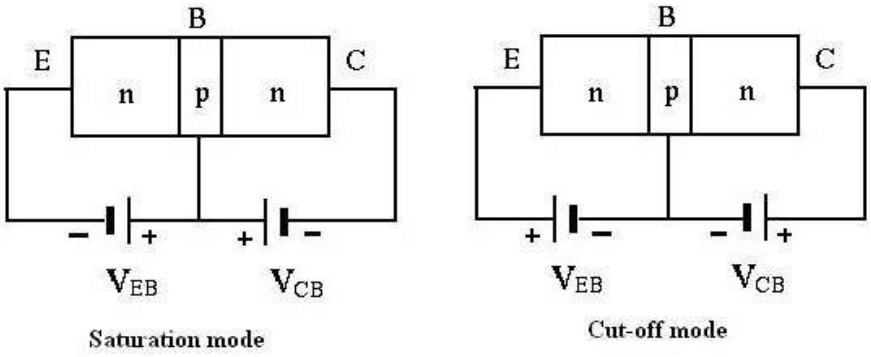


Figure 6.5: Modes of BJT Operation

6.4 WORKING OF TRANSISTOR

When there is no battery connection between the different terminals of a transistor, the transistor is said to be in an unbiased state. The process of applying DC voltages across the different terminals of a transistor is called biasing. In normal operation the EB junction is forward bias and the CB junction is in reverse biased. The forward bias reduces the barrier potential and narrows the depletion region whereas the reverse biased produces a wide depletion region.

6.4.1 PNP Transistor

The working of a PNP transistor which is shown in Figure 6.6. The EB junction of a PNP transistor is forward biased by the battery V_{EB} and the CB junction is reverse biased by a bigger battery V_{CB}. The forward bias causes the holes in the P-region cross over to the N- region or the base region. As the base is comparatively thin and is lightly doped, a majority of holes are cross into the collector region. A small number of holes are lost in the base region due to recombination with electrons. The holes that enter the collector region are attached by the negative terminal of the battery V_{CB}. Each hole is neutralized by the electrons from the negative terminal of the battery.

As each hole is lost by combination with an electron in the base and the collector region, an electron is free by the breaking down of a covalent bond near the emitter and the electrons so enlightened by the positive terminal of the battery V_{EB}. Hole is produced by the breaking of covalent bond that moves towards the emitter base junction and this process goes continuously.

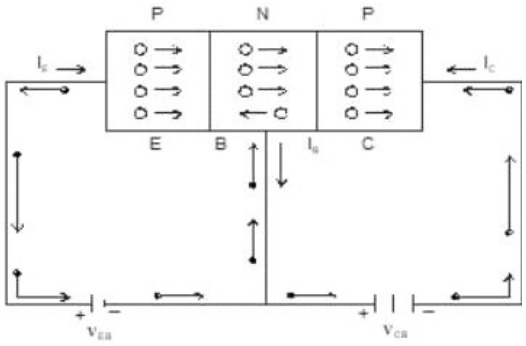


Figure 6.6: PNP transistor

6.4.2 NPN Transistor

The operation of an NPN transistor in its active region is governed by a biasing arrangement in which the emitter base junction is forward biased and the collector–base junction is reverse biased. This configuration establishes the necessary conditions for efficient carrier injection, transport, and collection forming the basis of transistor action and amplification.

In this mode, the heavily doped n-type emitter serves as a rich source of electrons, which are the majority carriers. Due to the forward bias across the emitter base junction, these electrons gain sufficient energy to overcome the potential barrier and are injected into the p-type base region. The base is intentionally designed to be extremely thin and lightly doped, minimizing the probability of recombination. As a result, only a small fraction of the injected electrons recombine with holes in the base, producing a relatively small base current.

The majority of electrons diffuse across the base region and reach the collector base junction. Since this junction is reverse biased, it creates a strong electric field that rapidly sweeps the electrons into the collector region. This drift mechanism ensures that most of the carriers injected by the emitter contribute directly to the collector current.

Thus, the transistor exhibits a key amplification property: a small variation in the base current I_B controls a much larger collector current I_C. The emitter current I_E, is the sum of the base and collector currents, expressed as $\mathbf{I_E=I_B+I_C}$. This relationship highlights the fundamental principle underlying the operation of an NPN transistor as a current-controlled device.

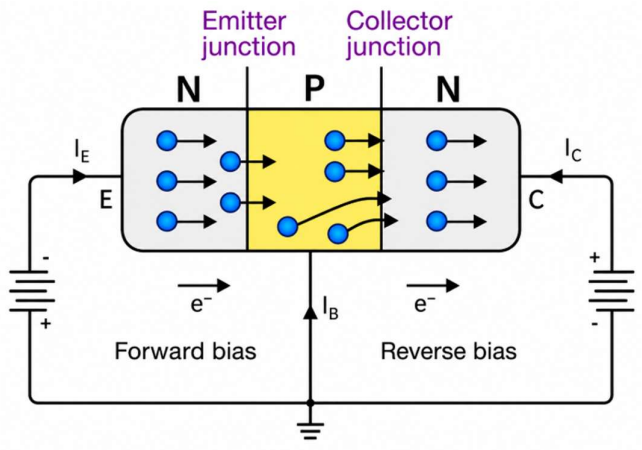


Figure 6.7: NPN transistor

6.5 TRANSISTOR CHARACTERISTICS

To analyse its operation and performance of a transistor, it is necessary to study the relationship between its input and output currents and voltages under different operating conditions. These relationships are known as transistor characteristics. Here we will discuss the characteristics of common-base and common-emitter configurations.

Characteristics of common base (CB) configuration

Input Characteristics: The graph drawn between Base emitter voltage (V_{BE}) vs I_E characteristics is called input characteristics. I_E increases rapidly with V_{EB}. It means input resistance is very small. I_E almost independent of V_{CB}

Output characteristics: of a transistor is the variation of the collector current I_C with the collector-base voltage V_{CB} is called the output characteristic. Collector Current I_C varies linearly with V_{CB}, only when V_{CB} is very small. As, V_{CB} increases, Ic becomes constant.

Input and Output Resistance of common base configuration

Input Resistance: The ratio of change in emitter-base voltage to the change in emitter current is called Input Resistance.

Output Resistance: The ratio of change in collector-base voltage to the change in collector current is called Output Resistance.

Current amplification factor (α): The ratio of change in collector current to the change in emitter current at constant V_{CB} is known as current amplification factor.

$$\alpha = (\frac{\Delta I_C}{\Delta I_E})_{V_{CB}=Constant}$$

Practical value of α is less than unity, but in the range of 0.9 to 0.99
Collector diode is reverse biased, so very few minority carrier passes the collector-base junction which actually constitute leakage current. So, collector current constitute of portion of emitter current αI_E and leakage current I_{CBO}.

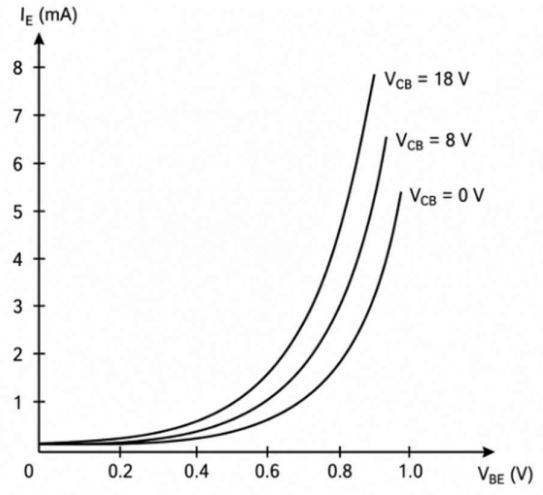


Figure 6.8(a): Input characteristics of CB

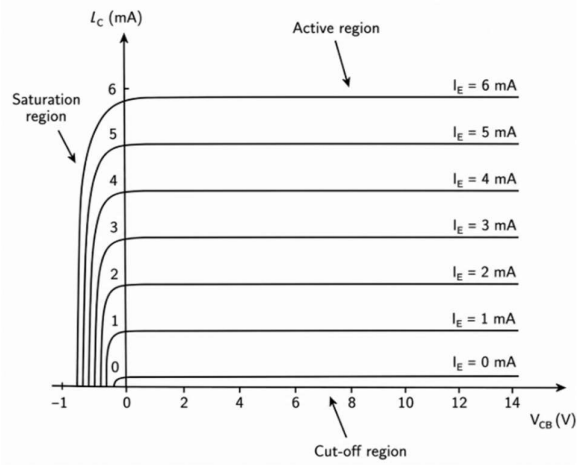


Figure 6.8(a): Output characteristics of CB

Characteristics of common emitter (CE) configuration

Input Characteristics: In a common emitter configuration, the input characteristics are represented by a graph showing the relationship between the base current (I_B) and the base-emitter voltage (V_{BE}) while keeping the collector-emitter voltage (V_{CE}) constant. I_B increases rapidly with V_{BE}. It means input resistance is very small. I_B is of the range of micro amps.

Output Characteristics: The V_{CE} vs I_c characteristics is called output characteristics when I_B is constant. I_C varies linearly with V_{CE}, only when V_{CE} is very small. As, V_{CE} increases, I_C becomes constant.

Base Current amplification factor (β): In common emitter connection input current is base current and output current is collector current. The ratio of change in collector current to the change in base current is known as base current amplification factor.

$$\beta = \left(\frac{\Delta I_C}{\Delta I_E}\right) V_{CE} = Cons\tan t$$

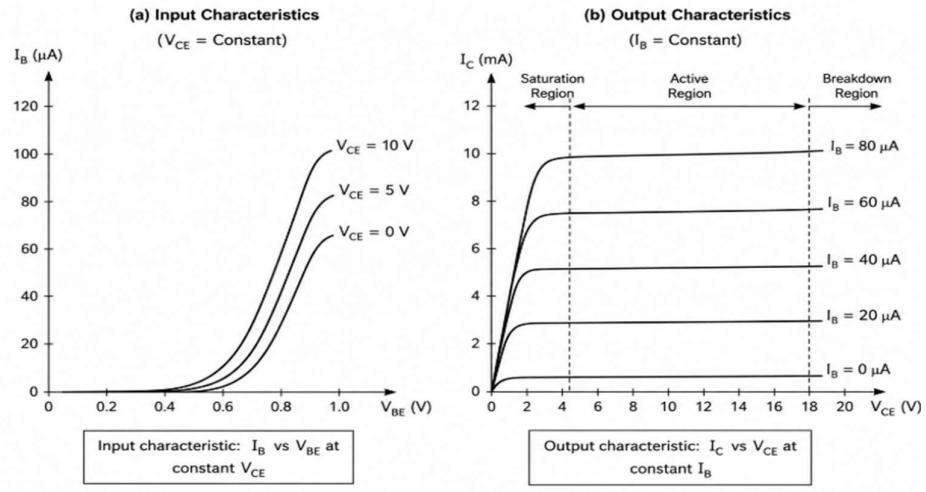


Figure 6.9: Input and output characteristics of a common-emitter configuration of BJT

W CHECK YOUR PROGRESS

Q1. Why BJT is it called bipolar?
Q2. What is meant by transistor configuration?
Q3. Why transistor characteristics are important?
Q4. What are the operation modes of a transistor?

2 6.6 LET US SUM UP

- A Bipolar Junction Transistor (BJT) is a three-layer, three-terminal semiconductor device used for amplification and switching, consisting of emitter, base, and collector regions.
- BJTs are of two types: NPN and PNP, classified based on the arrangement of n-type and p-type semiconductor materials.
- A transistor can be connected in three configurations: common emitter (CE), common base (CB), and common collector (CC), each having distinct input–output characteristics.
- The main modes of operation of a transistor are active, cut off, and saturation, determined by the biasing of the emitter–base and collector–base junctions.
- In the active mode, the transistor acts as an amplifier with the emitter–base junction forward biased and the collector–base junction reverse biased.
- In a PNP transistor, current conduction occurs mainly due to the movement of holes from emitter to collector.
- In an NPN transistor, electrons are the majority carriers, and current flows from emitter to collector due to electron movement.
- Transistor characteristics describe the relationship between current and voltage and are used to analyze transistor performance.

K 6.7 ANSWERS TO CHECK YOUR PROGRESS

Answer to Q1. It is called *bipolar* because both electrons and holes participate in current conduction within the device.

Answer to Q2. Transistor configuration refers to the way a transistor is connected in a circuit. The three configurations—common base (CB), common emitter (CE), and common collector (CC)—differ in gain, input impedance, and output impedance, determining their suitability for amplification or impedance matching.

Answer Q3. They are important for determining operating regions, gain, and performance of the transistor in circuits.

Answer to Q4. A transistor operates in three modes:

- Cutoff (OFF state)
- Active (amplification region)
- Saturation (fully ON state)

O 6.8 FURTHER READING

- Boylestad, R. L., & Nashelsky, L. (2013). *Electronic devices and circuit theory* (11th ed.). Pearson Education.

- Malvino, A. P., & Bates, D. J. (2015). *Electronic principles* (8th ed.). McGraw-Hill Education.
- Sedra, A. S., & Smith, K. C. (2015). *Microelectronic circuits* (7th ed.). Oxford University Press.
- Floyd, T. L. (2012). *Electronic devices* (9th ed.). Pearson Education.
- Sarma, S. (2012). *Electronic principles*. S. K. Kataria & Sons.
- Streetman, B. G., & Banerjee, S. K. (2015). *Solid state electronic devices* (7th ed.). Pearson Education.

M 6.9 MODEL QUESTIONS

Q1. Differentiate between NPN and PNP transistors based on construction and working.

Q2. Explain the different configurations of a transistor and mention their key features.

Q3. Describe the operation modes of a transistor with proper biasing conditions.

Q4. Explain the working of an NPN transistor in the active region with necessary details.

Q5. What are transistor characteristics? Explain input and output characteristics of a transistor in CE configuration

*** ***** ***

UNIT 7: TRANSISTOR APPLICATIONS

UNIT STRUCTURE

7.1 Learning Objectives
7.2 Introduction
7.3 Application of Bipolar Junction Transistor
    7.3.1 BJT as an Amplifier
    7.3.2 BJT as a Switch
7.4 Introduction to Integrated Circuits
7.5 Oscillator Basics
7.6 Let Us Sum Up
7.7 Answers to Check Your Progress
7.8 Further Reading
7.9 Model Questions

7.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- identify the operating regions of a BJT
- explain the role of BJTs in amplification and switching circuits
- differentiate between analog and digital ICs based on their signal-processing functions
- compare integrated and discrete circuits in terms of cost, efficiency, reliability, and size
- analyze the conditions for sustained oscillations using the Barkhausen criterion.

7.2 INTRODUCTION

In the previous unit, you learned about the Bipolar Junction Transistor (BJT). In this unit, you will explore about the applications of BJTs in amplifier and switching circuits. You will examine how a transistor is used as an amplifier to increase the strength of a weak signal and as a switch by operating in the cut-off and saturation regions. The unit will also introduce you to the basics of Integrated Circuits (ICs). Finally, you will learn the basic concepts of oscillators and the conditions required for generating sustained oscillations.

7.3 APPLICATION OF BIPOLAR JUNCTION TRANSISTOR

Bipolar Junction Transistors (BJTs) are one of the most important things that have been made in the last hundred years. Their ability to control the flow of a large current with a very small input signal makes them the "valves" of the electronic world. In short, a BJT can be used in two main ways: as a switch to turn power on and off or as an amplifier to make weak signals stronger.

7.3.1 BJT as an Amplifier

An ideal amplifier has the capacity to increase the amplitude level from input signal. The amplification can be measured by the ratio of output to input signal. The amplification can be denoted as amplification factor. Amplification factor for voltage, current and power denoted as $A_{v,}$ A_i and A_p respectively.

There are two common ways to look at BJT amplifier circuits. One way is to employ hybrid parameters. There are four basic kinds of hybrid parameters. We have already seen one of them, h_{fe}, which is the forward current gain. We just call it Beta. The other three are h_{ie}, which stands for input impedance; h_{oe}, which stands for output admittance; and h_{re}, which stands for reverse voltage gain. The second letter of the subscript (the "e" in h_{fe}) means that it is for the common emitter configuration and the input goes to the base, the output comes from the collector, and the emitter goes to the common ground.

An amplifier's heart is the transistor. We shall discuss about bipolar transistor amplifiers in this chapter. Because they offer a relatively high gain, bipolar transistors have long been utilized in linear amplifier circuits. We will start our conversation with the same bipolar circuit that we discussed about in the prior chapter. As illustrated in the image, the transistor needs to be biased with a DC voltage at a quiescent point (Q-point) so that it is in the forward-active region. This is how the circuit works as an amplifier. If a time-varying signal, like a sine wave, is added to the dc input voltage, $V_{B\,B}$, the output voltage will move along the transfer curve, making the output voltage shift with time. The circuit is a linear amplifier if the output voltage that changes with time is directly proportional to and greater than the input voltage that changes with time. This figure shows that if the transistor is not biased in the active area (either in cutoff or saturation), the output voltage does not change when the input voltage changes.

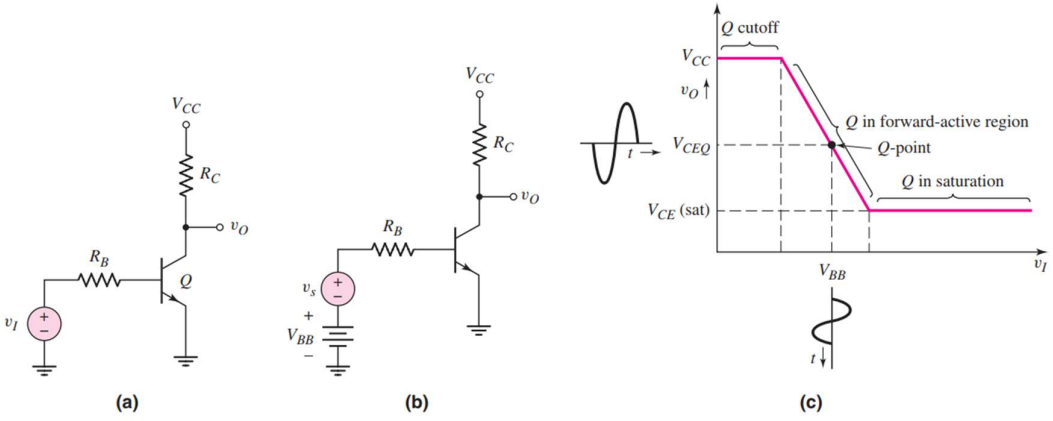


Figure 7.1: (a) BJT Inverter circuit (b) BJT with both input dc bias and output ac signal (c) transistor inverter voltage transfer characteristics

The time-varying output and input signals relations will discuss based on the ac analysis and design of bipolar transistor amplifiers. A graphical method that helps to understand how the circuit works in a simple way. Next, we will make a small-signal equivalent circuit on the ac signals. We will mostly be looking at circuits in a steady state and using a sinusoidal analysis. We shall presume that any time-changing signal

may be expressed as a summation of sinusoidal signals with varying frequencies and amplitudes (Fourier series), thereby rendering sinusoidal analysis suitable.

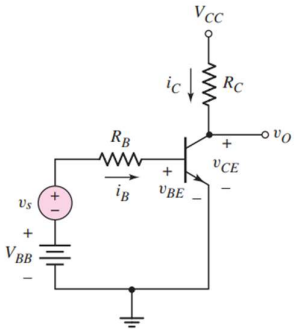


Figure 7.2: The common emitter circuit characteristics

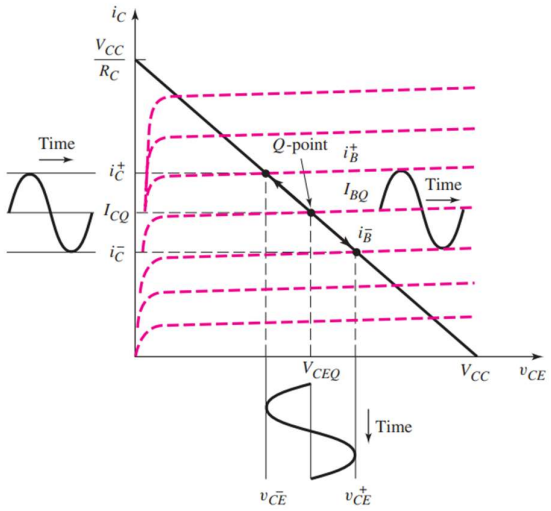


Figure 7.3: Common emitter transfer with time varying signal

A mathematical derivation can relate how sinusoidal changes in currents and voltages in the circuit are related to each other. As was said before, a linear amplifier means that superposition works, so the dc and ac analyses can be done separately. To get a linear amplifier, the ac currents and voltages must be small enough to keep the relationship between the ac signals linear. To attain this goal, it is assumed that the time-varying signals are tiny signals. This means that the amplitudes of the ac signals are small enough to create linear relationships. We shall discuss more about the idea of "small enough," or tiny signal, when we work on the small-signal equivalent circuits.

A time-varying signal source, as seen in the base of the circuit in Figure 7.1(c), creates a time-varying part of the base current. This means that there is also a time-varying part of the base-emitter voltage. The graph in Figure 7.3 shows how base current and base-emitter voltage are related in an exponential way. We can create a linear relationship between the ac base-emitter voltage and the ac base current if the time-varying signals that are added to the dc quiescent point are not too big. This relationship is the same as the curve's slope at the Q-point.

We may now use Figure 7.3 to find one quantitative definition of a small signal. The relationship between base-emitter voltage and base current can be expressed as

$$i_B = \frac{I_s}{\beta}\exp\left(\frac{v_{BE}}{V_T}\right) \qquad (1)$$

If $V_{BE} = V_{BEQ} + v_{be}$

Where V_{BE} is the instantanous value of base emitter value, v_{be} is the phasor value of base emitter voltage and V_{BEQ} is the base emitter turn on voltage

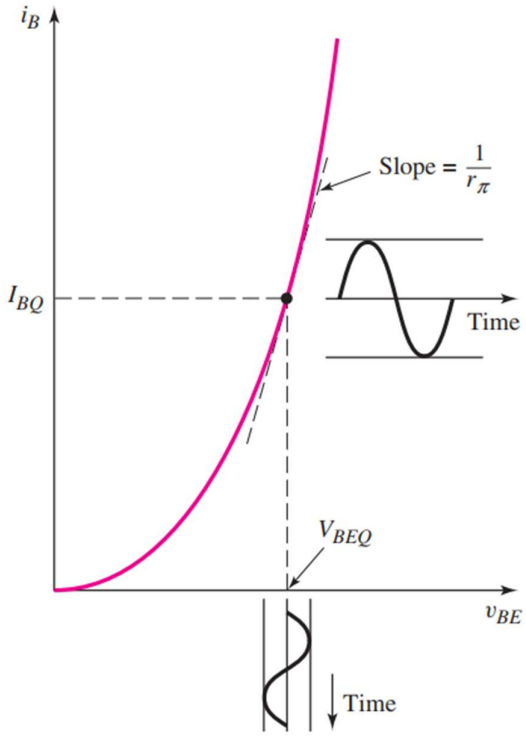


Figure 7.4 Base current verses base emitter voltage characteristics with superimposed signals

Then, $i_B = \frac{I_S}{\beta}\exp\left(\frac{V_{BEQ}+v_{be}}{V_T}\right) = \frac{I_S}{\beta}\exp\frac{V_{BEQ}}{V_T}\exp(\frac{v_{be}}{V_T})$ (2)

The term, $\frac{I_S}{\beta}\exp\frac{V_{BEQ}}{V_T}$ is the quiescent base current

Therefore, we can write

$$i_B = I_{BQ}\exp(\frac{v_{be}}{V_T}) \quad (3)$$

The base current, as presented, is nonlinear and cannot be expressed as an AC current overlaying a DC quiescent value. We can enlarge the exponential term in a Taylor series and only preserve the linear term if $v_{be} \ll V_T$. This is what is meant by a little signal. Then the equation for i_B is

$$i_B \cong I_{BQ}\left(1+\frac{v_{be}}{V_T}\right) = I_{BQ} + \frac{I_{BQ}}{V_T}v_{be} = I_{BQ} + i_b \quad (4)$$

Here, i_B is the time varying sinusoidal current
We can write the equation as

$$i_B = (\frac{I_{BQ}}{V_T})v_{be} \quad (5)$$

The sinusoidal base current, i_b, has a straight-line relationship with the sinusoidal base-emitter voltage, v_{be}. In this situation, "small-signal" means that v_{be} is small enough for the linear relationships between i_b and v_{be} in Equation 5 to hold true. If v_{be} is less than 10 mV, the exponential relationship in Equation 3 and its linear expansion

in Equation 4 are usually within 10 percent of each other. Another good rule of thumb for designing linear bipolar transistor amplifiers is to make sure that v_{be} is less than 10 mV.

If we think the v_{be} signal as a sine wave, with very high strength, the output signal will no longer be a pure sine wave voltage. Instead, it will be warped and have harmonics.

According to the common emitter transistor characteristics (Figure 7.4) all the time varying signals will be linearly related and are superimposed on dc values.

Therefore, we can write,

$$i_B = i_{BQ} + i_b \quad (6)$$

$$i_c = i_{CQ} + i_c \quad (7)$$

$$v_{CE} = v_{CEQ} + v_{ce} \quad (8)$$

$$v_{BE} = v_{BEQ} + v_{be} \quad (9)$$

The base emitter loop equation for zero v_s is

$$V_{BB} = I_{BQ}R_B + V_{BEQ} \quad (10)$$

For a time varying signal, the base emitter loop equation can be expressed as

$$V_{BB} + v_s = i_B R_B + v_{BE} \quad (11)$$

$$V_{BB} + v_s = \left(I_{BQ} + i_b\right)R_B + v_{BEQ} + v_{be} \quad (12)$$

Or

$$V_{BB} - I_{BQ}R_B - v_{BEQ} = i_b R_B + v_{be} - v_s \quad (13)$$

From the equation (11) it is seen that left hand side of equation (13) is zero. Therefore,

Equation (13) can be written as,

$$v_S = i_b R_B + v_{be} \quad (14)$$

This is the base-emitter loop equation with all the DC terms set to zero.

The collector emitter loop: For zero signal source, v_s, the collector emitter loop equation is

$$V_{CC} = I_{CQ}R_C + V_{CEQ} \quad (15)$$

For a time varying signals, the collector emitter loop equation is expressed as,

$$V_{CC} = I_C R_C + v_{CE} = (I_{CQ} + i_c)R_C + V_{CEQ} + v_{ce} \quad (16)$$

Rearranging the equation 16, we can write

$$V_{CC} - I_{CQ}R_C - V_{CEQ} = i_c R_C + v_{ce} \quad (17)$$

From equation 15 and 17,

$$i_c R_C + v_{ce} = 0 \quad (18)$$

In this collector emitter loop equation, all the dc terms are zero.

Equations (14) and (18) show how the ac parameters in the circuit are connected. To get these equations, we can set all the dc voltages and currents to zero. This makes the dc voltage sources short circuits and the dc current sources open circuits. These results are a direct outcome of using superposition on a linear circuit. The BJT circuit

that comes out of this is termed the ac equivalent circuit (Figure 7.5). All of the currents and voltages depicted are signals that change over time.

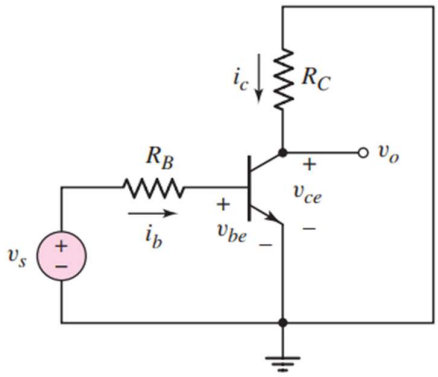


Figure 7.5: AC equivalent circuit of common emitter circuit

Small-Signal Hybrid-π Equivalent Circuit of the Bipolar Transistor:

We now need to make a small-signal equivalent circuit for the transistor. The hybrid-π model is one of these circuits, and it is quite similar to how transistors work. A more sophisticated hybrid-π model that takes into consideration the transistor's frequency response will make this impact more clear.

As indicated in the Figure 7.6, we can think of the bipolar transistor as a two-port network.

The input port is located between the base and the emitter, whereas the output port is situated between the collector and the emitter. The input port is between the base and emitter, while the output port is between the collector and emitter.

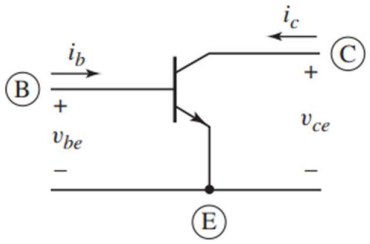


Figure 7.6: BJT two port network as a small signal

One part of the hybrid-π model has already been discussed. In the base current in relation to the base-emitter voltage characteristic figure, featuring minor time-varying signals superimposed at the Q-point. Because the sinusoidal signals are tiny, we can treat the slope at the Q-point as a constant with units of conductance. The small-signal resistance, r_π, is the opposite of this conductance. We can then connect the small-signal input base current to the small-signal input voltage by

$$v_{be} = i_b r_\pi \qquad (19)$$

It is seen that $1/r_\pi$ is the slope of I_B-V_{BE} curve. From equation (2) we can find that

$$\frac{1}{r_\pi} = \frac{\partial i_B}{\partial v_{BE}} = \frac{\partial}{\partial v_{BE}}\left[\frac{I_S}{\beta} exp\, \frac{v_{BE}}{V_T}\right] \qquad (20)$$

Or

$$\frac{1}{r_\pi} = \frac{1}{V_T}\left[\frac{I_S}{\beta} exp\, \frac{v_{BE}}{V_T}\right] = \frac{I_{BQ}}{V_T} \qquad (21)$$

Then

$$\frac{v_{be}}{i_b} = r_\pi = \frac{V_T}{I_{BQ}} = \frac{\beta V_T}{I_{CQ}} \quad (22)$$

The resistance r_π is known as the diffusion resistance or the base-emitter input resistance. We can observe that the Q-point parameters affect r_π. This is the same expression as in Equation 4.

Output Collector-Emitter Port

We can look at the properties of the output terminal of the bipolar transistor. If we start by thinking about the case when the output collector current doesn't depend on the collector-emitter voltage, then the collector current just depends on the base-emitter voltage.

We can then write

$$\Delta i_C = \frac{\partial i_C}{\partial v_{BE}} . v_{BE} \quad (23)$$

$$i_C = \frac{\partial i_C}{\partial v_{BE}} . i_b \quad (24)$$

$$i_c = I_S \left[exp \frac{v_{BE}}{V_T} \right] \quad (25)$$

$$\frac{\partial i_c}{\partial v_{BE}} = \frac{1}{V_T} \left[I_S exp \frac{v_{BE}}{V_T} \right] = \frac{I_{CQ}}{V_T} \quad (26)$$

$\frac{I_{CQ}}{V_T}$ term is called transconductance

$$g_m = \frac{I_{CQ}}{V_T} \quad (27)$$

The small signal collector current is

$$i_c = g_m v_{be} \quad (28)$$

The small-signal transconductance depends on the Q-point characteristics and is directly related to the dc bias current. The change in transconductance with quiescent collector current will be helpful in designing amplifiers. Figure 7.7 shows simplified small signal signal hybrid π equivalent circuit for npn transistor.

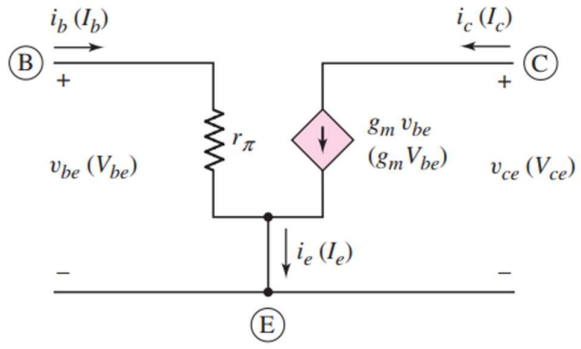


Figure 7.7: A simplified small signal hybrid π equivalent circuit for npn transistor

COMMON EMITTER CURRENT GAIN

The common-emitter current gain in Equation 28 is actually an ac beta and does not take into account dc leakage currents.

Equations 22 and 26 defined the small-signal hybrid-π parameters r_π and g_m. We can find what happens when we multiply r_π and g_m.

$$r_\pi g_m = \frac{\beta V_T}{I_{CQ}} \frac{I_{CQ}}{V_T} = \beta \qquad (29)$$

For a given transistor, we will generally assume that the common-emitter current gain β is always the same. We need to remember, though, that β might be different for different devices and that it does change with collector current.

SMALL SIGNAL VOLTAGE GAIN:

The small-signal voltage gain, $A_v = V_o/V_s$, of the circuit is the ratio of the output signal voltage to the input signal voltage. In the Figure of small-signal equivalent circuit of the common-emitter circuit shows a new variable. V_π, which stands for "control voltage," is the standard way to write the small-signal base-emitter voltage. The dependent current source is $g_m V_\pi$.

The dependent current $g_m V_\pi$ goes through RC, which makes the collector-emitter voltage negative.

$$V_o = V_{ce} = -(g_m V_\pi) R_c \qquad (30)$$

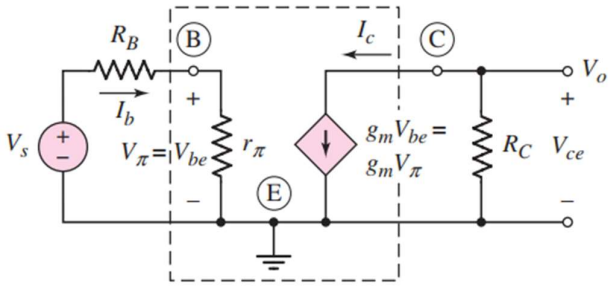


Figure 7.8: The small-signal equivalent circuit of the common-emitter circuit.

In figure 7.8 the small-signal hybrid-π model of the npn bipolar transistor is shown within the dotted lines.

From the input part, we get

$$V_\pi = \frac{r_\pi}{r_\pi + R_B} V_S \qquad (31)$$

The small signal voltage gain is

$$A_V = \frac{V_0}{V_S} = -(g_m R_c) \frac{r_\pi}{r_\pi + R_B} \qquad (32)$$

7.3.2 BJT as a Switch

Conceptualizing a Bipolar Junction Transistor (BJT) as a current-controlled switch helps us understand its switching operation. Unlike a mechanical switch, a BJT is controlled by a small electrical current. For switching, the BJT operates in two extreme regions: cut-off and saturation.

The Two States of the Switch
OFF State:
In the Cut-off region, ideally there is no current flowing into the base, i.e. in this state $I_B = 0$. The path between the Collector and the Emitter is basically broken because there is no "instruction" at the base. The load (such as an LED or motor) connected to the output of the switch remains OFF as a result.

Voltage: The voltage across the collector-emitter v_{CE} is at its highest point, which is the same as the supply voltage.

ON State:

Area of Saturation: When sufficient current is applied to the base terminal, the BJT transitions into its saturation region. In this state, the collector-emitter path achieves maximum conductivity, minimizing resistance and allowing peak current flow, thereby maintaining the device in a fully conducting ("on") state.

Voltage: The voltage between the collector and emitter v_{CE}lowers to almost zero (around 0.2V), which makes it look like a closed wire.

A standard BJT switching circuit, usually using an NPN transistor, has three main parts: the transistor operating as the solid-state control element, a base resistor, and the load device.

- **Base Resistor (R_B):** Placed in series with the base terminal, this component limits the control current delivered from signal sources (such as microcontrollers) to prevent thermal overload and damage to the base-emitter junction.
- **Load Device:** The target actuator or indicator (e.g., DC motor, LED, or relay), typically connected in series with the collector terminal to receive switched supply power.
- **Current Gain (h_{FE}/β) :** Enables small base control currents (e.g., 20 mA) to modulate significantly larger collector load currents (e.g., 500 mA) supplied by an auxiliary power source.
- **Voltage Level Shifting and Isolation:** Facilitates safe switching of higher-voltage loads (such as a 12 V peripheral) using low-voltage logic-level control signals (such as a 3.3 V micro-controller output).
- **High-Frequency Operation:** Supports microsecond-scale switching frequencies, enabling applications in pulse-width modulation (PWM) power delivery and high-speed digital processing architectures.

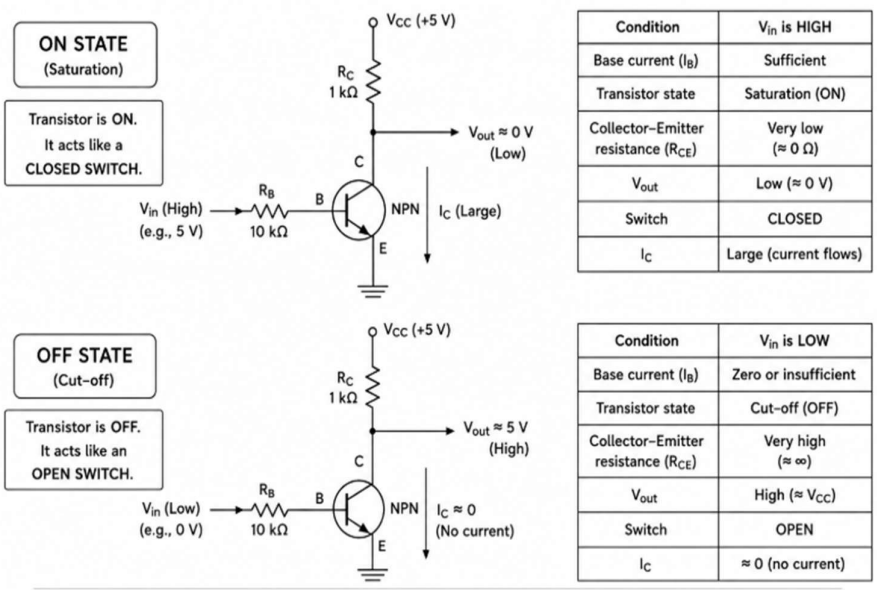


Figure 7.9: BJT Transistor as a Switch

7.4 INTRODUCTION TO INTEGRATED CIRCUITS

Microelectronic chips, also called integrated circuits (ICs), are monolithic semiconductor devices fabricated to carry out a designated electronic function, such as signal amplification and processing. An integrated circuit is the integration of a number of interconnected semiconductor devices on a single semiconductor substrate. Standard integrated circuits are capable of a wide range of fundamental functions, like analog signal processing and continuous power regulation.

Integrated circuits are one of the most important parts of electronics that we utilize every day in many of our products. William B. Shockley and his team at AT&T Bell Laboratories invented transistors in 1947. The scientists learnt to control the flow of electricity (flow of charges) through the crystal by regulating the barrier. They found that under the correct conditions, electrons will form a barrier at the surface of some crystals. Later in 1947, Jack S. Kilby was successful in bonding two BJTs in the same substrate, thereby paving the way to a new world of Integrated Circuits.
An integrated circuit (IC) can have thousands or even millions of:

- Transistors
- Resistors
- Capacitors
- Diodes

Integrated Circuit (IC) is also known as a microchip, micro electronic chip, or chip. It is a set of electronic circuits that are put on a metal plate made of semiconductor materials. An IC is the basic part of all modern electronic circuits that includes active devices like R, C, L, diodes, transistors etc. Their connections are constructed on a thin layer of semiconductor material, usually silicon. An IC can be a microcontroller, microprocessor, timer, counter, logic gate or amplifier.

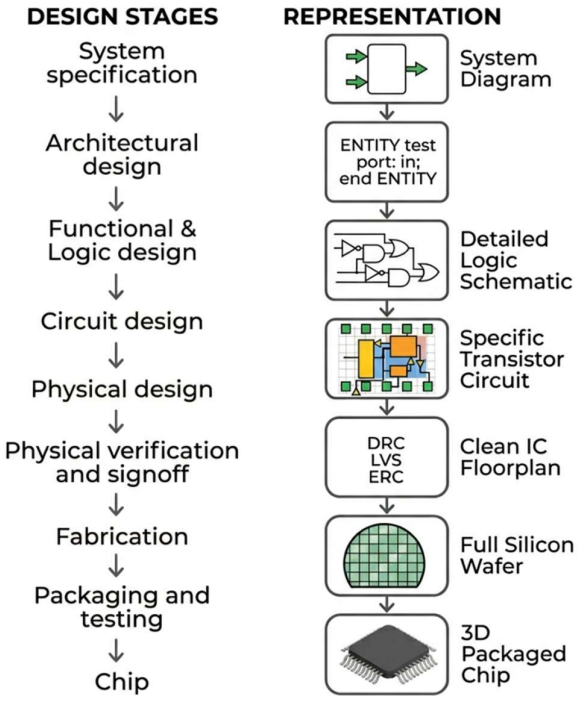


Figure 7.10: Flowchart of IC processing

Based on the characteristics of the signals processed by them and their modes of operation, integrated circuits (ICs) are of three types: digital, analog, and mixed-signal.

- **Digital Integrated Circuits**: Also known as logic or computational circuits , operate on the basis of discrete binary logic states (0 and 1) to represent off-state and on-state conditions. They are the basic building blocks of microprocessors and memory storage devices such as RAM and ROM. Flip-flops and basic logic gates (AND, OR) are examples of sequential components.
- **Analog Integrated Circuits**: Also known as linear ICs, these circuits are designed to handle input signals that change continuously. The input continuum is in one-to-one correspondence with the output response. One of the most common uses is for the amplification of radio-frequency (RF) and audio-frequency signals.
- **Mixed signal integrated circuits** are a combination of analog and digital design topologies. It enables smooth interaction between discrete digital processing platforms and continuous physical signals. The most common implementations are Analog-to-Digital Converters (ADCs), Digital-to-Analog Converters (DACs), and timing/clock generation ICs.

7.5 OSCILLATOR BASICS

A transistor oscillator is a circuit that produces continuous oscillations without requiring an external input signal. It uses a transistor as an amplifying device and a feedback circuit to return a part of the output signal to the input. A tank circuit, usually consisting of an inductor (L) and a capacitor (C), determines the frequency of oscillation by storing and exchanging energy between the inductor and capacitor. When the feedback is

properly connected, the transistor compensates for the energy losses in the tank circuit and maintains the oscillations at the desired frequency.

The condition required to maintain continuous oscillations is known as the Barkhausen criterion. It states that the total phase shift around the feedback loop must be 0° or 360°, so that the feedback signal is in phase with the input signal. In addition, the magnitude of the loop gain must be equal to 1 for sustained oscillations. The loop gain is the product of the amplifier gain and the feedback factor.

The **loop gain** of an oscillator is given by: $A\beta$

where "A" is the voltage gain of the amplifier and "β" is the feedback factor.
For sustained oscillations, according to the Barkhausen criterion:

$$|A\beta| = 1$$

with a total phase shift of **0° or 360°** around the loop.

The schematic diagram of a transistor oscillator circuit is shown in figure 7.11. This circuit makes it easy to show how to use a transistor as an oscillator. This circuit is divided into three components, as shown below.

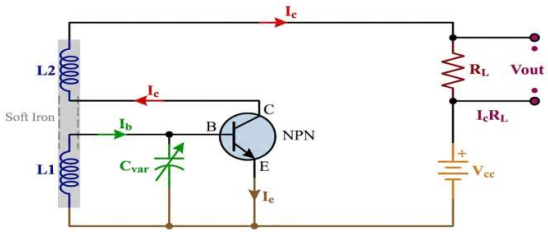


Figure 7.11: Basic transistor oscillator circuit

Tank Circuit
A tank circuit is an LC circuit made of an inductor (L) and a capacitor (C). It stores electrical energy in the inductor and capacitor and produces electrical oscillations at a particular frequency. It also determines the frequency of oscillation in an oscillator circuit, and provides the oscillating signal that can be amplified by the transistor.

Amplifier Circuit
This circuit takes the small sinusoidal oscillations from the base-emitter circuit and makes them bigger, which is what the output is.

Feedback Circuit
The feedback circuit is a very important part of this circuit since it needs some energy to work with the tank circuit to make an amplifier. Using the Mutual induction phenomena, the energy from the collector circuit is sent back to the base circuit. This circuit sends energy back from the output to the input.

Transistor Works as an Oscillator
In the transistor oscillator circuit above, the transistor works as a CE (common emitter) circuit. This means that the emitter is the same for both the base and collector terminals. A tank circuit is linked between the emitter and base input terminals. The inductor and capacitor in the tank circuit are linked in parallel so that the circuit can oscillate.

The voltage and charge oscillations in the tank circuit cause the current flow at the base terminal to shift, which in turn changes the forward biasing of the base current and the collector current.

Because LC oscillations are sinusoidal, both the base and collector currents change in a sinusoidal way. The diagram shows that if the current at the collector terminal changes in a sinusoidal way, the output voltage may be easily represented as Ic RL. This output is thought to be sinusoidal.

If we make a graph of time and output voltage, the curve will be a sine wave. We need some energy to keep the tank circuit oscillating all the time. But there is no battery or DC source in this circuit.

We used a soft iron rod to connect the L1 and L2 inductors in the collector and base circuits. The L2 inductor will be connected to the L1 inductor via this rod because of mutual induction. This means that some of the energy in the collector circuit will flow to the base side of the circuit. So, the oscillation in the tank circuit keeps going and getting stronger.

Conditions for Oscillation

The transistor oscillator circuit needs to follow these rules:

- The loop's phase shift should be between 0 and 360 degrees.
- The gain of the loop has to be more than 1.

Different kinds of transistor oscillators

There are several types of oscillators, each designed to generate oscillations over a specific frequency range. They differ mainly in how the feedback and frequency-selective circuits are arranged to overcome circuit losses and maintain stable oscillations. Transistor oscillators that use LC circuits as tank circuits are commonly used to generate high-frequency signals. The different types of transistor oscillators are discussed below.

Hartley Oscillator

The Hartley oscillator is a type of electronic oscillator that uses a tuned circuit to find the frequency of the oscillation. The most important part of this oscillator is that the tuned circuit has one capacitor connected in parallel to two inductors in series. The feedback signal needed for oscillation comes from the center connection of the two inductors. The Hartley oscillator works well for RF oscillations up to 30 MHz.

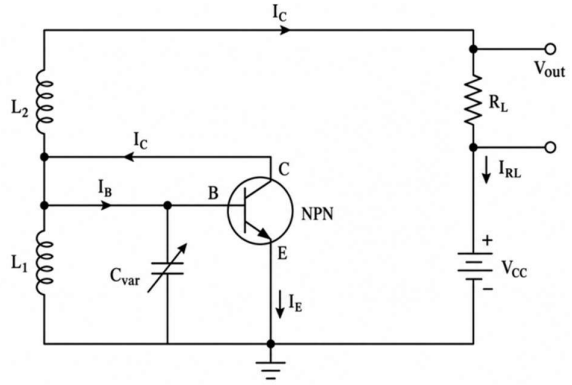


Figure 7.12: Hartley Oscillator

Crystal Oscillator

Transistor crystal oscillators can be used in many fields of electronics, including radio. These kinds of oscillators are very important for making an inexpensive CLK signal that may be used in logic or digital circuits. This oscillator can also be used to make a consistent and accurate RF signal source. So, radio amateurs or "radio hams" often utilize these oscillators in radio transmitter circuits where they work best.

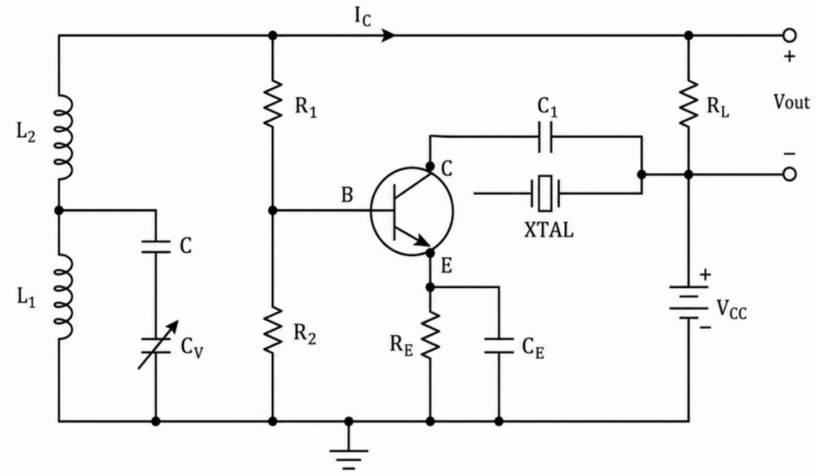


Figure 7.13: Crystal oscillator

Colpitt's Oscillator

The Colpitts oscillator is different from the Hartley oscillator. The only difference is that the inductors and capacitors are switched places in the tank circuit. The fundamental benefit of this type of oscillator is that it has less mutual and self-inductance in the tank circuit, which makes the frequency more stable. This oscillator makes very high frequencies from sinusoidal signals. These oscillators are stable at high frequencies and can handle both low and high temperatures.

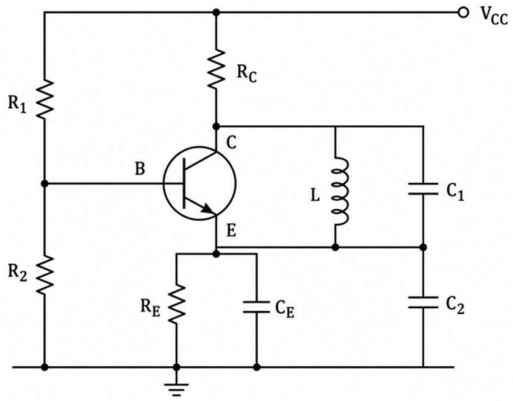


Figure 7.14: Colpitts oscillator

Wien Bridge Oscillator

The Wien bridge oscillator is an audio frequency oscillator that is often utilized because it has several useful qualities. This kind of oscillator doesn't change with the temperature of the circuit or the environment. The best thing about this type of

oscillator is that it can alter the frequency from 10Hz to 1MHz. This oscillator circuit keeps the frequency stable, which is desirable.

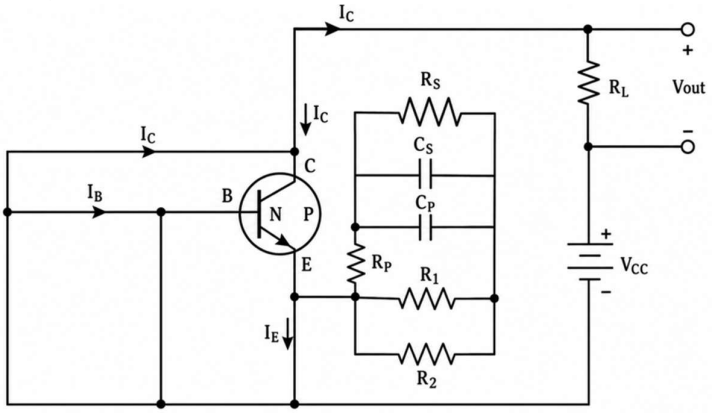


Figure 7.15: Wien bridge Oscillator

Phase shift Oscillator

An RC phase shift oscillator is a type of oscillator that uses a simple RC network to change the phase of the feedback signal. This oscillator, like the Hartley and Colpitts oscillators, uses an LC network to give the necessary positive feedback. This oscillator has great frequency stability and can make pure sine waves over a wide range of loads.

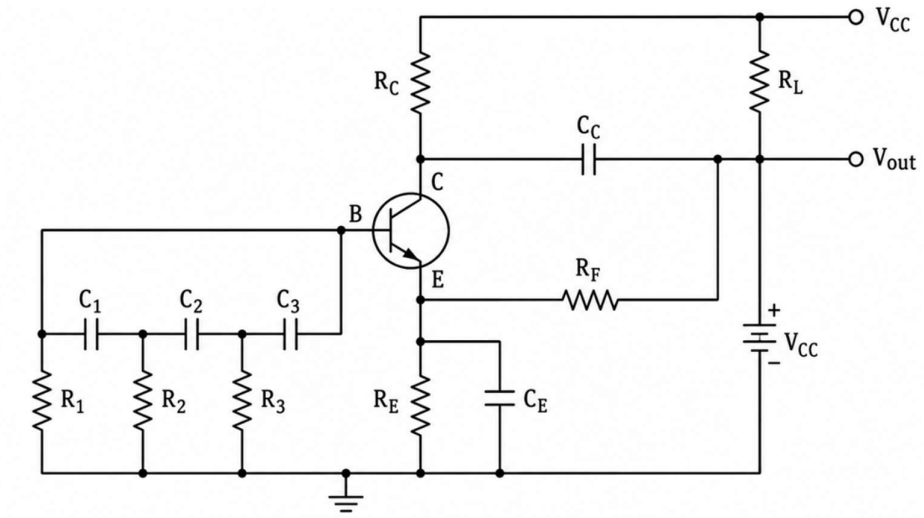


Figure 7.16: Phase shift oscillator

W CHECK YOUR PROGRESS

Q1. How does a BJT function as an amplifier?
Q2. What are the two states of a BJT when used as a switch?
Q3. What is an Integrated Circuit (IC)?
Q4. What is the main purpose of an Oscillator?
Q5. What are the three essential parts of an Oscillator circuit?

2 7.6 LET US SUM UP

- The Bipolar Junction Transistor (BJT) is a key component in modern electronics, mainly used for amplification and switching.
- A BJT amplifies a weak input signal at the base into a stronger output signal at the collector.
- A small base current controls a much larger collector current, enabling signal amplification.
- In digital circuits, a BJT operates in two main regions: Cut-off (OFF state) and Saturation (ON state).
- In Cut-off, the transistor behaves like an open switch; in Saturation, it behaves like a closed switch.
- An Integrated Circuit (IC) is a compact electronic chip that integrates many components on a single silicon wafer.
- ICs are formed by fabricating thousands to billions of components such as transistors, resistors, and capacitors on silicon.
- ICs offer advantages such as miniaturization, low power consumption, high speed, and low cost due to mass production.
- ICs are broadly classified into analog ICs and digital ICs.
- An oscillator is an electronic circuit that converts DC input into an AC output without any external signal.
- The output signal of an oscillator is in the form of a sine wave or square wave.
- Oscillator consists of an amplifier and a positive feedback network.

K 7.7 ANSWERS TO CHECK YOIUR PROGRESS

Answer to Q1: In amplifier mode, the BJT operates in the Active Region. A small input current at the Base controls a much larger current flowing through the Collector. This allows a weak signal (like audio) to be strengthened into a much more powerful version of itself.

Answer to Q2: The two states are Cut-off and Saturation.

Cut-off: The transistor is "OFF" (no current flows), acting like an open switch.

Saturation: The transistor is "Fully ON," acting like a closed switch to allow maximum current to the load.

Answer to Q3: An IC is a single small chip (usually made of Silicon) that contains thousands or millions of tiny interconnected components like transistors, resistors, and capacitors. It replaces large circuits made of individual, bulky parts.

Answer to Q4: An oscillator converts DC (Direct Current) from a power supply into an AC (Alternating Current) signal, such as a sine wave or square wave, without needing an external input signal.

Answer to Q5: The three parts are:

Amplifier: To provide the necessary signal gain.

Feedback Network: To send a portion of the output back to the input.

Frequency Determining Network: Usually an LC or RC circuit that sets the specific frequency (speed) of the vibration.

O 7.8 FURTHER READING

- Boylestad, R. L., & Nashelsky, L. (2012*). Electronic devices and circuit theory* (11th ed.). Pearson.
- Sedra, A. S., Smith, K. C., Carusone, T. C., & Gaudet, V. (2020). *Microelectronic circuits* (8th ed.). Oxford University Press.
- Floyd, T. L. (2017). *Electronic devices: Conventional current version* (10th ed.). Pearson.
- Millman, J., & Halkias, C. C. (2017). *Integrated electronics: Analog and digital circuits and system*s (2nd ed.). McGraw Hill Education.
- Razavi, B. (2013). *Fundamentals of microelectronics* (2nd ed.). Wiley.

M 7.9 MODEL QUESTIONS

Q1. Explain why a BJT is considered a "current-controlled" device when used as an amplifier.

Q2. Contrast the operation of a BJT in the Active Region versus the Saturation Region.

Q3. In a BJT switching circuit, what is the state of the transistor when the base current I_B is zero?

Q4. If a BJT has a current gain β of 100 and a base current of 10μA, what is the resulting collector current I_C in the active region?

Q5. Why is a "pull-down" resistor often used at the base of a BJT when it is being used as a digital switch?

Q6. What is the primary difference between a discrete circuit and an integrated circuit?

Q7. Give two examples of Digital ICs and two examples of Analog ICs commonly found in consumer electronics.

Q8. Why is Silicon the preferred semiconductor material for fabricating integrated circuits?

Q9. Explain what this means in terms of energy input and output.

Q10. What happens if the loop gain is less than 1?

*** ***** ***

UNIT 8: TRANSISTOR BIASING

UNIT STRUCTURE

8.1 Learning Objectives
8.2 Introduction
8.3 Transistor Biasing and Stabilization
  8.3.1 Need for Biasing
  8.3.2 DC Load Line
  8.3.3 Bias Stability
8.4 Two Port Network
  8.4.1 Hybrid Model
  8.4. 2 h-parameter
8.5 Thevenin's and Norton's Theorem
8.6 Let Us Sum Up
8.7 Answers to Check Your Progress
8.8 Further Reading
8.9 Model Questions

8.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- explain the need for DC biasing to establish a stable Q-point for linear amplification
- construct the DC load line and identify the Q-point, cutoff, and saturation regions
- describe the effects of temperature and current-gain variations on transistor stability and explain the stability factor
- apply two-port network theory and h-parameters to analyse transistor circuits
- use Thevenin's and Norton's theorems to simplify and analyse transistor networks.

8.2 INTRODUCTION

In the previous unit, we have learned about applications of Bipolar Junction Transistor (BJT). In this unit we will learn about different biasing configurations of BJT, stabilization, load line and operating point. Moreover, you will learn about two-port networks, hybrid models, and h-parameters. You will also study Thevenin's and Norton's theorems and their applications in circuit analysis at the end of this unit.

8.3 TRANSISTOR BIASING AND STABILIZATION

To truly comprehend and build a transistor amplifier, you need to look at it from two different angles: the AC and DC responses. An amplifier takes power from a DC source and sends it to a weak AC signal. This makes the signal stronger in general. Because of this, the design of an electrical amplifier has two primary parts: the AC portion and the DC part. It is extremely crucial to choose the proper DC levels throughout the

design phase since these choices have a direct impact on how effectively the AC functions.

The biasing circuit uses DC voltages to establish fixed values of current and voltage, which makes this simple setup work. The basic aims of DC biasing are to turn the device "ON" and relocate its operating point, or Q-point, to the region of its characteristic curves that is the most linear. A well configured biasing circuit ensures that the transistor operates within this linear range, often targeting a centered Q-point, sometimes referred to as halfway biasing, to maximize signal oscillation. But bad biasing may put the device in non-linear regions, which can mess up the signal and make the output waveforms either limited or clipped.

There are different configurations of transistor biasing. They can be categorized as

- Fixed Biasing
- Collector Derived Biasing
- Emitter Stabilized Biasing
- Voltage Divider Biasing

Fixed Biasing

Amplifiers are essential in many electronic devices, serving purposes in sound and image reproduction as well as in control systems and communications. The design of amplifiers is to create a circuit that achieves a specified gain throughout a designated frequency range with minimal distortion. The amplifier must be steady and not susceptible to oscillation. Bipolar PNP or NPN transistors, as well as FETs, may be used in several designs based on their specific applications. Let us see the basic bipolar NPN common emitter amplifier seen in Fig. 8.1, which comprises a transistor and two resistors. For optimal operation, the amplifier must provide an amplified output of the input signal without distortion. To do this, the quiescent or no signal (DC) conditions must be accurately established first. Its output can only be unaltered if its input is unaltered.

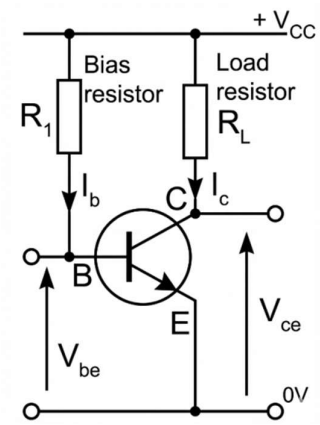


Figure 8.1: A common-emitter fixed bias circuit

A typical input characteristic curve for a small signal amplifier transistor is depicted in Fig. 8.2. This curve plots the changes in base voltage V_b against the resultant changes in base current I_b. If the AC signal voltage (changes in V_b) applied to the base is to result in proportional changes in the AC base current I_b, a DC value of VB must be

used to ensure that positive and negative excursions of the signal voltage occur exclusively on the linear part of the input curve (waveform b in Fig. 8.2). The base bias voltage is the DC voltage (0.7V in Fig. 8.2) that is applied to the base. It is evident from Fig. 8.2 that if the bias voltage is insufficient, the base current waveform a will exhibit severe distortion as only the positive ends of the input voltage waveform generate base current. It is also evident that a quiescent (DC) base current of 40μA is generated by a DC base bias voltage of 0.7V for this transistor. The precise selection of R_B resistance value in Fig. 8.1 establishes these values.

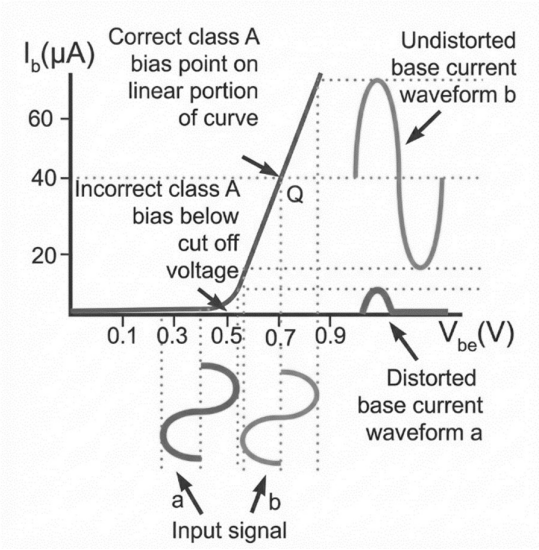


Figure 8.2: Input Characteristics

Additionally, the quiescent output conditions must be taken into account, as the quiescent collector current Ic will be contingent upon the current gain h_{fe} of the transistor and the quiescent base current Ib. Additionally, the potential difference across the load resistor (RL) as a result of Ic flowing through it will yield the value of the transistor collector/emitter voltage (V_{ce}) when the supply voltage (V_{cc}) is subtracted.

Problems in fixed biasing: Although the design depicted in Fig. 8.1 is straightforward and necessitates a minimal number of components, there are certain challenges that must be resolved in order to make it practicable. The bias voltage will also be affected by any variations in the supply voltage or transistor temperature. The collector current will increase as a result of the increased base current, which is induced by an increase in the bias voltage. This will result in a further increase in current and a subsequent rise in junction temperature within the transistor. The transistor will eventually conduct a higher level of current, resulting in an additional increase in temperature, and so forth. The transistor will continue to heat up until it is annihilated, which the ultimate consequence of this process is known as "Thermal Runaway." Although Thermal Runaway is significantly less of a concern in contemporary power transistors, it remains a potential hazard in small signal types that should be mitigated by incorporating some form of bias stabilization into the amplifier design.

Collector Derived Biasing (DC Stabilization)

A straightforward technique of enhancing the temperature stabilization of a common emitter amplifier is illustrated in Fig. 8.3. The bias current is supplied from the collector end of R_L, rather than from V_{cc}. In this configuration, any increase in collector current will result in an increase in the potential difference across R_L. Consequently, the collector voltage V_{ce} at the bottom of R_L must decrease, as the top of R_L is maintained at a constant voltage by V_{cc}. This, in turn, will result in a decrease in V_{be}, which will subsequently decrease the collector current. The bias conditions are largely self-adjusting and are believed to be stabilized by a form of DC feedback.

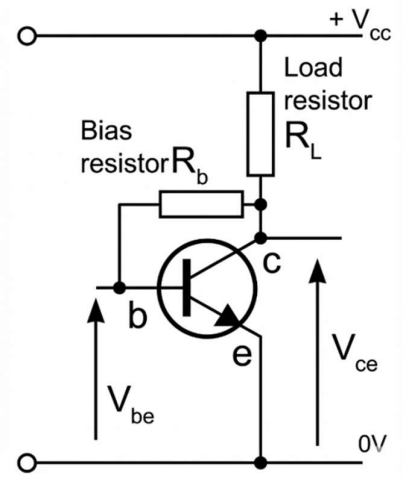


Figure 8.3: Collector Derived Biasing

Emitter Stabilized Biasing

An emitter resistor R_E is added to the fixed bias circuit to improve stability of the circuit. The addition of the emitter resistor to the dc bias of the BJT provides improved stability, that is, the dc bias currents and voltages remain closer to where they were set by the circuit when outside conditions, such as temperature, and transistor beta, change.

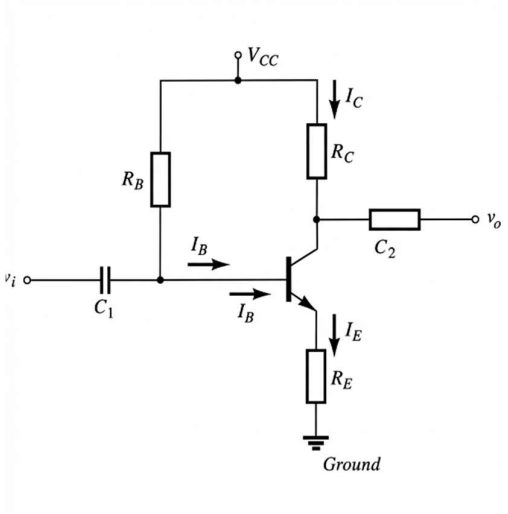


Figure 8.4: Emitter Stabilized Biasing

Voltage Divider Biasing

An alternative bias arrangement that is much more prevalent in commercial circuits is the use of a potential divider that consists of two resistors (R1 and R2 in Fig. 8.5) to maintain a consistent value of V_{be} and an emitter resistor Re to provide stabilization through DC feedback. The emitter current increases in tandem with the collector current in this circuit, resulting in an increase in the emitter voltage V_e. This increase, in contrast to the constant base voltage, results in a decrease in the base-emitter voltage V_{be} and a resultant decrease in collector current. DC feedback, which employs an emitter stabilizing resistor, maintains circuit conditions that are susceptible to fluctuations in other variables, such as temperature or transistor h_{fe}.

Nevertheless, the emitter resistor will also generate unwanted AC feedback. This is due to the fact that the AC waveform that appears on the emitter will be in phase with the base waveform under signal conditions. The two waveforms that are changing simultaneously will tend to reduce the variations in base-emitter voltage, resulting in a substantial reduction in gain. In order to prevent this issue, it is customary to bypass the emitter stabilizing resistor Re by connecting a large value capacitor across RE. This capacitor will create a very low impedance path to any AC signal present, thereby preventing any AC from appearing on the emitter while maintaining the same DC conditions.

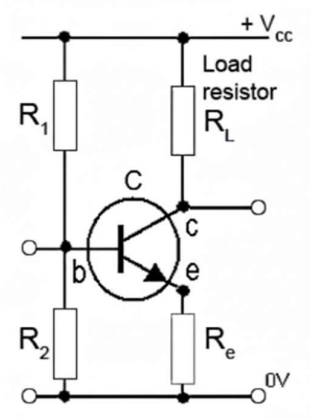


Figure 8.5: Voltage Divider Biasing

8.3.1 Need for Biasing

For a Bipolar Junction Transistor (BJT) to function optimally as an amplifier or switch, it requires appropriate biasing. Biasing denotes the application of designated external DC voltages to the terminals of a transistor to create a stable level of current and voltage, referred to as the Operating Point or Q-point (Quiescent Point).

The Necessity of BJT Biasing

The main aim of biasing is to locate the Q-point inside the active area of the transistor's characteristic curves. This is crucial for many technical reasons:

- **Linear Amplification:** Appropriate biasing guarantees that the transistor stays inside the active zone for the whole of the input signal cycle. Should the Q-point be positioned too near the saturation or cutoff zones, the output signal may experience clipping, resulting in considerable distortion.

- **Thermal Stability**: Transistors exhibit significant sensitivity to variations in temperature. An elevation in temperature results in an augmentation of the collector leakage current (I_{CBO}), subsequently leading to an increase in the collector current (I_C). This may result in a change of the Q-point, possibly culminating in thermal runaway, when the transistor self-destructs owing to excessive heat. Biasing circuits are designed to mitigate these temperature variations.
- **Stabilization against Parameter Variation**: Transistors of the same type may exhibit significant variations in β (current gain) attributable to manufacturing tolerances. A reliable biasing circuit guarantees the stability of the Q-point, even when substituting the transistor with another of same kind.
- **Optimal Signal Swing**: The Q-point is typically established at the center of the DC Load Line to achieve the highest possible undistorted output. This enables the output voltage and current to exhibit the greatest possible symmetrical variation.

The BJT's intended function is guaranteed to be consistently performed across a wide range of environmental conditions and component tolerances by establishing a stable Q-point, which is the cornerstone of reliable circuit design.

8.3.2 DC Load Line

A DC load line in a BJT (Bipolar Junction Transistor) is a straight line drawn on the output characteristic curve (V_{CE} vs. I_C) that shows all the potential DC voltage and current operating points for a circuit. The only things that affect these points are the supply voltage (V_{CC}) and the external resistors. It shows the difference between saturation ($V_{CE}\approx 0$) and cutoff ($I_C\approx 0$) and finds the Quiescent point (Q-point).
To plot the load line let us consider the fixed bias circuit shown in fig. 8.1.
We can find that

$$V_{CE} = V_{CC} - I_C R_C$$

I_C (sat) occurs when the transistor is in saturation, hence

$$I_{C_{sat}} = \left.\frac{V_{CC}}{R_C}\right|_{V_{CE}=0}$$

Again VCE (cut-off) occurs when the transistor is operating in cut-off region, hence

$$V_{CE_{(off)}} = V_{CC} - I_C R_C\big|_{I_C=0}$$

The load line drawn is shown in fig. 8.6

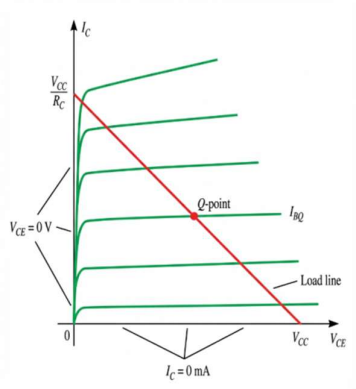


Figure 8.6: The DC load line

8.3.3 Operating point

The DC operating point is one of several intriguing parameters that must be considered when designing with a BJT transistor. The transistor's state when no input current is applied to the component is referred to as a DC operating point, also known as a quiescent or Q point. The DC operating point is the intersection of the DC load line and the base current, as depicted on a chart. In order to ascertain the transistor's DC operating point, it will be necessary to plot the DC load line. The procedure is straightforward, as the relationship between Ic and Vce is delineated by

$$V_{cc} = V_{ce} + I_c R_c$$

Vce is the collector-emitter voltage, I_cR_c is the voltage drop across the base resistor (R_b), and V_{cc} is the supply voltage in this equation. The value at both coordinates can be obtained by solving the equation for V_{ce} =0 and I_c =0. The DC load line can be obtained by connecting a line between the two values, as the equation is linear. The DC operating point can be determined on the chart for a known Ib.

Nevertheless, it is crucial to comprehend the impact of the DC operating point on transistor biasing. The transistor's behavior is subject to change in the active, cut-off, and saturation regions.

The DC operating point should be maintained within a specific range, contingent upon the intended function of the transistor. For instance, the DC operating point should be situated in the center of the active region when constructing a Class A amplifier. The input AC signal will be truncated or distorted if it is near the saturation or cutoff. Simultaneously, there are specific amplifier varieties that are engineered to function in the cutoff or saturation region. The DC operating point of the Class B amplifier is located in the cutoff region, as it amplifies only the positive half of the input signal and eliminates the negative portion.

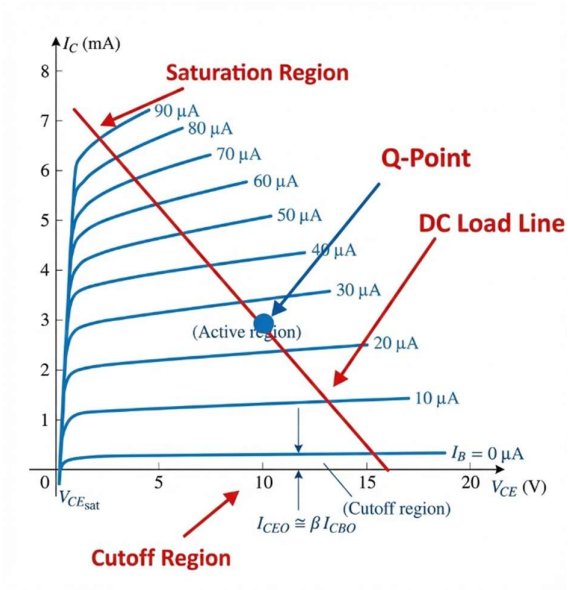


Figure 8.7: DC Load line and Q-point

It is also to be noted that different circuit values impact the operating point. Let us see the impacts of V_{cc}, R_c and I_B . In Fig. 8.8 we can see that the decreasing values of V_{cc}

shifts the load line as well as the operating point. Moreover, changing the value of R_c will also change both the load line and the Q-point which is depicted in fig. 8.9. The change in the base current, IB does not change the load line unlike the previous cases but shifts the Q-points which is presented in fig. 8.10.

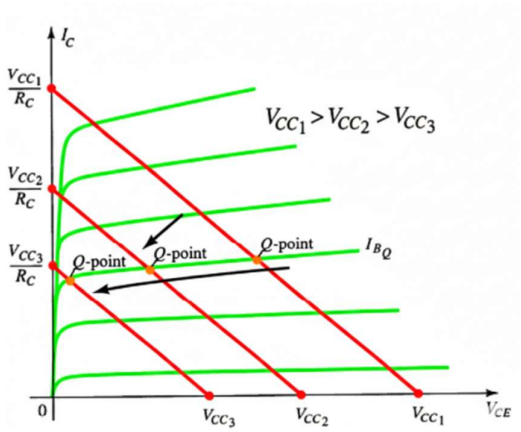


Figure 8.8: Change of Q-point with change in the values of V_{cc}

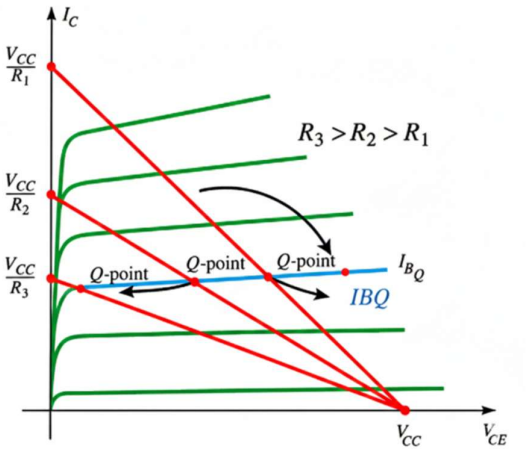


Figure 8.9: Shifting of Q-point with the change in R_c value

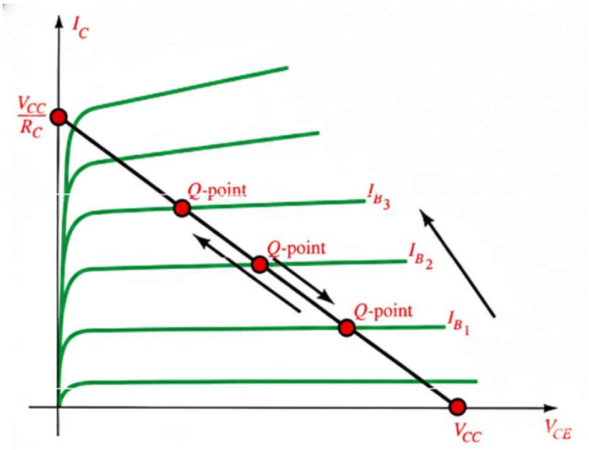


Figure 8.10: Shifting of Q-point with change in base current

The Significance of DC Operating Point Analysis

It is possible that the subsequent assumption will be that biasing a transistor is not a straightforward process. In the end, what could possibly go awry when the DC operating point of a transistor is established? The harsh reality is that the task is not complete with the placement of a few resistors.

Certain biasing circuits are unreliable, contingent upon their configuration. The fixed base bias circuit, as illustrated in the diagram above, is straightforward to implement; however, it is susceptible to instability during active operation. Additionally, fluctuations in temperature may result in fluctuations in the DC operating point. The transistor is configured in collector-feedback bias in the subsequent circuit, resulting in a modest enhancement. The DC operating point is perpetually maintained in the active region as a result of the negative feedback generated by Rb. The voltage divider bias is demonstrated in the circuit below to enhance stability. This configuration guarantees that the DC operating point is entirely unaffected by external circuitry elements and is completely independent of the β value. The β value in this context denotes the DC current gain and illustrates the correlation between large current fluxes and tiny biasing currents.

8.3.4 Bias Stability

Bias stability in Bipolar Junction Transistors Transistor (BJT) circuit design ensures the circuit maintains a stable Operating Point (Q-point) despite variations in temperature or individual transistor characteristics. Maintaining this stability is a significant technological challenge, since the Q-point, determined by the collector current (I_C) and the collector-emitter voltage (V_{CE}), is very sensitive to the operational characteristics of the semiconductor material. In the absence of adequate stability safeguards, a circuit may experience "Q-point drift," leading to signal distortion, reduced gain, or even catastrophic failure. Factors Influencing Bias Stability three very temperature-sensitive variables primarily account for the instability of a BJT's operating point:

Collector Leakage Current (I_{CBO}): This current originates from minority carriers and about doubles with each 10°C rise in temperature. As I_{CBO} increases, the total collector current (I_C) similarly rises, causing the Q-point to ascend towards the saturation region.

Voltage between the base and the emitter (V_{BE}): To maintain a constant current, the required V_{BE} decreases by about 2.5 mV for each degree Celsius increase in temperature. Maintaining a constant base bias voltage, the reduction in V_{BE} results in an increase in base current (I_B), subsequently causing a further increase in collector current (I_C).

Present Gain (β): The gain of a transistor is well recognized as inconsistent. It not only increases with temperature, but it may also vary significantly between two transistors of the same kind due to manufacturing differences. A stable circuit must ensure that I_C remains relatively constant despite variations in the β values.

Consequences and Strategies for Mitigation

If these variables are not regulated, the circuit may enter thermal runaway. This constitutes a positive feedback loop: as the temperature increases, so does I_C, generating more heat inside the transistor, which further elevates I_C, and this process continues until the transistor fails. Engineers use the Stability Factor (S) to assess and

mitigate this issue. It is the derivative of collector current with respect to leakage current. A reduced value of S indicates enhanced stability. The Voltage Divider Bias configuration is often used in due to its utilization of negative feedback via an emitter resistor (R_E). This configuration autonomously compensates for augmentations in I_C by increasing the emitter voltage. This reduces the effective V_{BE} and reverts the current to its designated quiescent condition.

8.4 TWO PORT NETWORK

A two-port network (Fig. 8.11) has four port variables: V_1, I_1, V_2, and I_2. If the two-port networkis linear, we can use two of the variables as excitation variables and the other two as responsevariables. For instance, the network can be excited by a voltage V_1 at port 1 and a voltage V_2at port 2, and the two currents, I_1 and I_2, can be measured to represent the network response.

In this case, V_1 and V_2 are independent variables and I_1 and I_2 are dependent variables, andthe network operation can be described by the two equations

$$I_1 = y_{11}V_1 + y_{12}V_2$$

$$I_2 = y_{21}V_1 + y_{22}V_2$$

Here, the four parameters y_{11}, y_{12}, y_{21} and y_{22} are admittances, and their values completely characterize the linear two-port network. Depending on which two of the four port variables are used to represent the network excitation, a different set of equations (and a correspondingly different set of parameters) is obtained for characterizing the network. Based on the parameters used in a two port network there are three types of models: Admittance model, Impedance model and Hybrid model.

8.4.1 Hybrid Model

The hybrid (or h-parameter) characterization of two-port networks is based on exciting the network by I1 and V2, as shown in Fig. 8.11

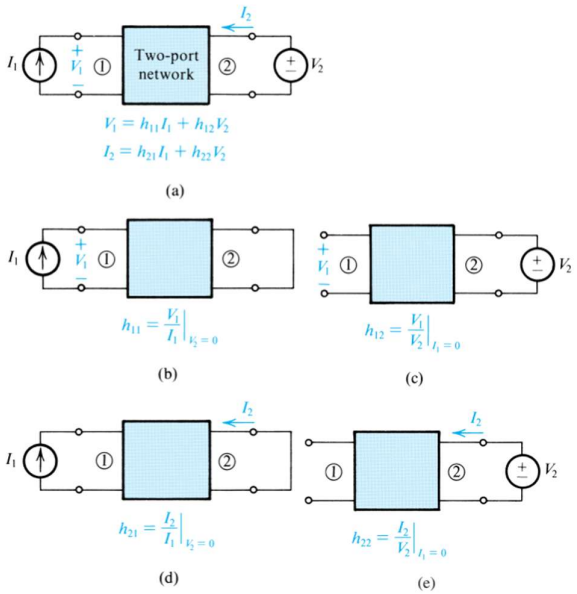


Figure 8.11: Depiction of conceptual measurement circuits of hybrid parameters

8.4.2 h-Parameter

The describing equations are

$$V_1 = h_{11}I_1 + h_{12}V_2$$

$$I_2 = h_{21}I_1 + h_{22}V_2$$

from which the definition of the four h parameters can be obtained as

$$h_{11} = \left.\frac{V_1}{I_1}\right|_{V_2=0} \qquad h_{21} = \left.\frac{I_2}{I_1}\right|_{V_2=0}$$

$$h_{12} = \left.\frac{V_1}{V_2}\right|_{I_0=0} \qquad h_{22} = \left.\frac{I_2}{V_2}\right|_{I_1=0}$$

Thus, h_{11} is the input impedance at port 1 with port 2 short-circuited. The parameter h_{12} represents the reverse or feedback voltage ratio of the network, measured with the input port open-circuited. The forward-transmission parameter h_{21} represents the current gain of the network with the output port short-circuited; for this reason, h_{21} is called the short-circuit current gain. Finally, h_{22} is the output admittance with the input port open-circuited. The definitions and conceptual measuring setups of the h parameters are given in Fig. 8.11.

8.5 THEVENIN'S AND NORTON'S THEOREM

Thevenin's theorem states that a linear two-terminal circuit can be replaced by an equivalent circuit consisting of a voltage source V_{Th} in series with a resistor R_{Th}, where V_{Th} is the open-circuit voltage at the terminals and R_{Th} is the input or equivalent resistance at the terminals when the independent sources are turned off.

The linear circuit in fig. 8.12.(a) can be substituted by the one in fig.8.12.(b) in accordance with Thevenin's theorem. The load in the figure may be a solitary resistor or another circuit. The Thevenin equivalent circuit is the circuit located to the left of the terminals in fig. 8.12.(b).

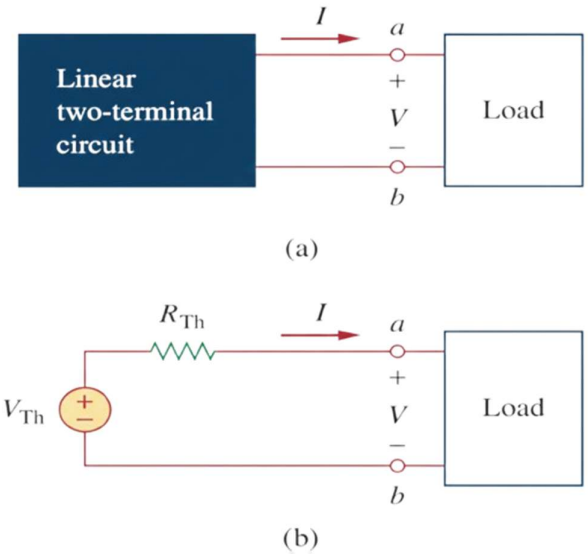


Figure 8.12: A two terminal network and it's Thevenin's equivalent circuit

A primary use of Thévenin's theorem is to substitute a substantial portion of a circuit, sometimes a complex and unremarkable segment, with a straightforward counterpart. The new, streamlined circuit allows for quick estimates of the voltage, current, and power that the original circuit can provide to a load. It also assists us in selecting the optimal value of this load resistance.

How to find the Thevenin equivalent voltage (V_{Th}) and resistance (R_{Th})

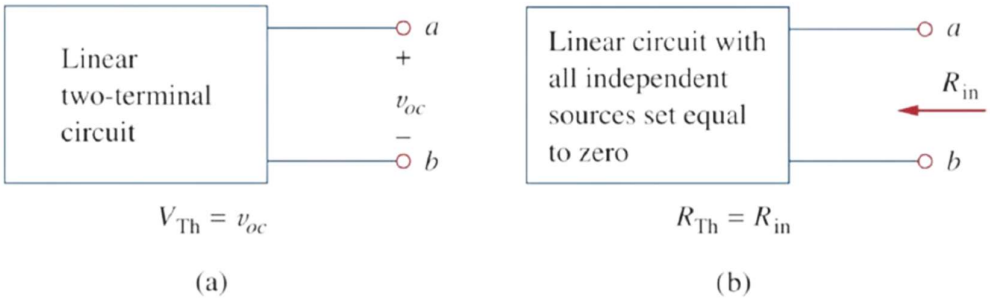


Figure 8.13: Finding the Thevenin's equivalent resistance

- Assume that the two circuits shown in figures 8.13.(a) and 8.13.(b) are equivalent. Two circuits are considered comparable if they exhibit identical voltage-current relationships at their terminals.
- Terminals a-b are rendered open-circuited (by detaching the load), resulting in no current flow; hence, the open-circuit voltage across the terminals in figure 8.13.(a) must be equivalent to the voltage source in figure 8.13.(b). Therefore, V_{Th} represents the open-circuit voltage between the terminals, namely $V_{Th} = V_{oc}$.
- Deactivate all independent sources. The input resistance, or equivalent resistance, of the inactive circuit at terminals a-b in figure 8.13.(a) must equal RTh in figure 8.13.(b) due to the equivalence of the two circuits. Consequently, R_{Th} represents the input resistance at the terminals when the independent sources are off. Thus, $R_{Th} = R_{in}$.

To find the Thevenin's resistance, two cases are to be considered:

- **Case 1**: In the absence of dependent sources inside the network, deactivate all independent sources. R_{Th} is the input resistance of the network seen between terminals a-b.
- **Case 2**: In the presence of dependent sources inside the network, all independent sources are to be deactivated. Dependent sources must remain operational since they are governed by circuit variables.

Norton's theorem states that a linear two-terminal circuit can be replaced by an equivalent circuit consisting of a current source i_N in parallel with a resistor R_N, where i_N is the short circuit current through the terminals and R_N is the input or equivalent resistance at the terminals when the independent sources are turned off.

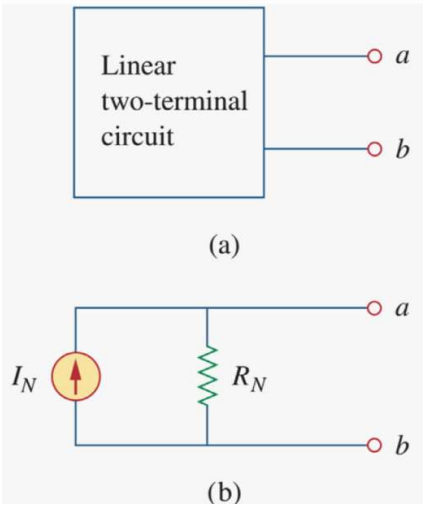


Figure 8.14: A two terminal network and Norton's equivalent circuit

- According to Norton theorem, the circuit figure 8.14.(a) can be replaced by the one in figure 8.14.(b).
- We determine R_N in the same manner that we ascertain R_{Th}. Thevenin and Norton resistances are equivalent; specifically, $R_N = R_{Th}$.
- To ascertain the Norton current, we calculate the short-circuit current traversing from terminal a to b in figure 8.14.(a); so, $i_N = i_{sc}$. So, $i_N = v_{TH}/R_{Th}$. This scenario can be visualized from fig.8.15

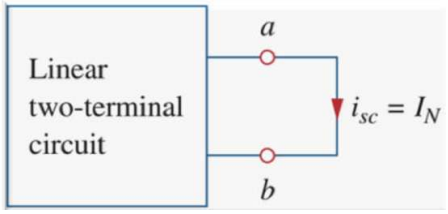


Figure 8.15: To find the short circuit current

W CHECK YOUR PROGRESS

Q1. What is the primary purpose of biasing a transistor?
Q2. What does the "Q-point" represent on a DC load line?
Q3. Why is thermal stability a concern in BJT biasing?
Q4. What are the four standard h-parameters for a transistor?
Q.5. What is the main difference between Thevenin's and Norton's theorems?

2 8.6 LET'S SUM UP

- The most important part of biasing is setting up a steady DC operating point, or Q-point, in the transistor's active area.
- This prevents the signal from going into the saturation or cutoff zones, which would cause non-linear distortion.
- The DC Load Line is a graph that shows all the possible combinations of I_C and V_{CE} for a certain circuit.

- The Q-point should be in the middle for the best signal swing.
- Bias stability is affected by temperature variations in transistor parameters and temperature changes may have a big effect on transistor properties.
- Higher temperatures increase the collector leakage current (I_{CBO}) and the current gain (β), but lower the base-emitter voltage (V_{BE}).
- Without stabilization, which is commonly done by negative feedback, these changes may cause thermal runaway, which is when heat builds up and kills the device.
- Transistors are commonly represented as Two-Port Networks to make AC analysis easier.
- The Hybrid (h-parameter) model works very well since it uses both voltage and current to explain how the device works.
- Transistors can be modeled as two-port networks using four h-parameters: input resistance (h_i), reverse voltage gain (h_r), forward current gain (h_f), and output conductance (h_o).
- Thevenin's and Norton's Theorems provide us the math we need to make complicated linear networks easier to understand.
- Thevenin's theorem says that a network may be simplified to a single voltage source in series with a resistor.
- Norton's theorem says that a network can be simplified to a parallel current source.
- These equivalents may be used in place of each other and are important for finding the best load for power transfer and stability in BJT biasing networks.

K 8.7 ANSWERS TO CHECK YOUR PROGRESS

Answer to Q 1: The primary purpose of biasing is to establish a stable DC operating point (Q-point) in the active region of the transistor's characteristic curves. This ensures that the transistor can amplify an input signal faithfully without clipping or distortion.

Answer to Q 2: The Q-point, or Quiescent point, represents the specific DC values of collector-emitter voltage (V_{CE}) and collector current (I_C) when no input signal is applied. It is the intersection point between the transistor's characteristic curve and the DC load line.

Answer to Q 3: BJTs are sensitive to temperature; as temperature rises, the collector leakage current (I_{CBO}) increases, which in turn increases the total collector current (I_C). This can cause the Q-point to shift toward the saturation region, potentially leading to thermal runaway and damage to the device.

Answer to Q 4: The four h-parameters are:

- **h_{11} (h_i):** Input resistance (measured in Ohms).
- **h_{12} (h_r):** Reverse voltage gain (unitless).
- **h_{21} (h_f):** Forward current gain (unitless).
- **h_{22} (h_o):** Output conductance (measured in Siemens or mhos).

Answer to Q 5:

- **Thevenin's Theorem** reduces a circuit to a single voltage source (V_{Th}) in series with a single resistor (R_{Th}).

- **Norton's Theorem** reduces a circuit to a single current source (I_N) in parallel with a single resistor (R_N).

O 8.8 FURTHER READING

- Boylestad, R. L., & Nashelsky, L. (2012). *Electronic devices and circuit theory* (11th ed.). Pearson.
- Alexander, C. K., & Sadiku, M. N. O. (2020). *Fundamentals of electric circuits* (7th ed.). McGraw-Hill Education.
- Sedra, A. S., Smith, K. C., Carusone, T. C., & Gaudet, V. (2019). *Microelectronic circuits* (8th ed.). Oxford University Press.
- Floyd, T. L. (2017).*Electronic devices* (10th ed.). Pearson.
- Hayt, W. H., Kemmerly, J. E., Phillips, S. M., & Durbin, S. M. (2018). *Engineering circuit analysis* (9th ed.). McGraw-Hill Education.

M 8.9 MODEL QUESTIONS

Q.1. Explain why a BJT requires a DC bias to function as a linear amplifier. What happens to the output signal if the bias point is too close to the saturation region?

Q.2. For a fixed-bias circuit with V_{CC} = 15V and R_C = 3kΩ, calculate the saturation current ($I_{C,sat}$) and the cutoff voltage ($V_{CE,off}$). How does changing R_C affect the slope of the DC load line?

Q.3. Define the Stability Factor (S) with respect to I_{CBO}. Why is a lower value of S preferred in circuit design, and which biasing configuration typically offers the best stability?

Q.4. Describe the positive feedback mechanism of thermal runaway in a BJT. How does the inclusion of an emitter resistor (R_E) help mitigate this effect?

Q.5. If a transistor with β = 100 is replaced by one with β = 200 in a circuit without stabilization, how will the collector current (I_C) be affected? Compare this to the behaviour in a voltage-divider bias circuit.

Q.6. Why is the "Hybrid" model named as such? Explain how it uses a mix of open-circuit and short-circuit conditions to define its parameters.

Q.7. In a common-emitter configuration, identify the physical meaning of h_{fe} and h_{ie}. What are their respective units?

Q.8. Given a set of h-parameters, derive the expression for the current gain (A_i) of a transistor amplifier in terms of h_{fe}, h_{oe}, and the load resistance (R_L).

Q.9. State Thevenin's Theorem and explain the step-by-step process to find R_{Th} in a circuit containing dependent sources. How is I_N (Norton current) related to V_{Th} and R_{Th}?

Q.10. In a voltage-divider bias circuit, the base circuit can be simplified using Thevenin's Theorem. Determine the Thevenin voltage (V_{Th}) and resistance (R_{Th}) as seen from the base of the transistor.

*** ***** ***

UNIT 9: FIELD EFFECT TRANSISTOR-PART I

UNIT STRUCTURE

9.1 Learning Objectives
9.2 Introduction
9.3 Introduction to FET
  9.3.1 Comparison between BJT and FET
  9.3.2 Advantages and Applications of FETs
9.4 Classification of FETs
  9.4.1 Construction of n-channel and p-channel JFET
9.5 Principle of Operation
  9.5.1 Drain Characteristics
  9.5.2 Transfer Characteristics
9.6 JFET Parameters
  9.6.1 Transconductance
  9.6.2 Drain Resistance
  9.6.3 Amplification Factor
9.7 Let Us Sum Up
9.8 Answers to Check Your Progress
9.9 Further Reading
9.10 Model Questions

9.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- discuss basic concept of FETs and comparing them with Bipolar Junction Transistors (BJTs)
- identify the various types of FET with their construction
- explain the principle of operation of JFETs and their characteristics
- analyze important JFET parameters used in circuit design.

9.2 INTRODUCTION

In the previous unit we have discussed about the Bipolar Junction Transistor (BJT) biasing. In this unit, we will discuss about the other type of transistor called Field Effect Transistor (FET). You will then study the classification of FETs and the construction of n-channel and p-channel Junction Field-Effect Transistors (JFETs). The principle of operation of JFETs will also be explained through their drain and transfer characteristics. Finally, the unit will introduce important JFET parameters, such as transconductance, drain resistance, and amplification factor.

9.3 INTRODUCTION TO FET

A Field Effect Transistor (FET) is a three-terminal semiconductor device. Its operation is based on a controlled input voltage. FETs are voltage controlled devices i.e the input voltage/field controls the output current. FET is a type of semiconductor device which is having a semiconductor-based channel connected to two electrodes. The two

electrodes named the drain and the source connected to either end. It has a third electrode named the gate which controls the current flow between the given two electrodes named the drain and the source. The FET is classified into two types based on its mode of operation, specifically the enhancement mode and depletion mode FETs, depending upon whether the voltage applied at the gate terminal then increases or decreases the current flow through the channel. The idea of the field effect transistor depends on the possibility that a charge on a nearby thing can draw in charges inside a semiconductor channel. It essentially works by using an electric field effect, and this is how its name started.

Working of FET Transistor

It has 3 terminals: gate, drain, and source. The working principle is as per the following: To allow electrons or holes to flow between the source and drain terminals, they are doped with material of the n-type or p-type variety.

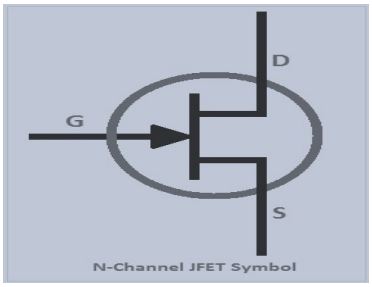


Figure 9.1: Symbol of JFET

An electric field is created in the channel area between the source and drain when a voltage is applied to the gate terminal. This electric field modulates the conductivity of the channel, behaving like a "valve" that controls the current flow between the source and drain.

9.3.1 Comparison between BJT and FET

The FET has the following advantages over a conventional transistor:

i. FET is a unipolar device, depending upon only majority carriers (either electrons or holes) whereas BJT is a bipolar device.
ii. FET has high input impedance due to reverse bias input whereas BJT has low input impedance due to forward bias.
iii. FET is a voltage controlled device whereas BJT is a current controlled device.
iv. FET is less noisy than BJT.
v. FET has better thermal stability and high efficiency.
vi. FET has very high power gain

9.3.2 Advantages and Applications of FETs

The high input impedance and low noise level is a great advantage of FET. Some of the circuit applications of the FET are:

i. It can act as a buffer amplifier

ii. It minimises the loading effect in Phase shift oscillators
iii. FETs can act as the input stage in voltmeters.
iv. FETs are used in RF amplifiers, FM tuners and communication equipment for the low noise level.
v. It is used as voltage variable resistors (VVRs) in operational amplifiers and tone controllers as it is a voltage controlled device.
vi. FETs are used in mixer circuits in FM and TV receivers, and communication equipment because of their low intermodulation distortion.
vii. FETs are used in low frequency amplifiers in hearing aids and inductive transducers because of the small coupling capacitors.
viii. It is used in digital circuits in computers, LSI and memory circuits because of very small size.
ix. FETs can be used as Cascode and Chopper amplifiers.
x. It can act as current limiter circuit for protection against excessive current.

9.4 CLASSIFICATION OF FETS

Field Effect Transistor scan be broadly classified in to two categories. The classification is shown below:

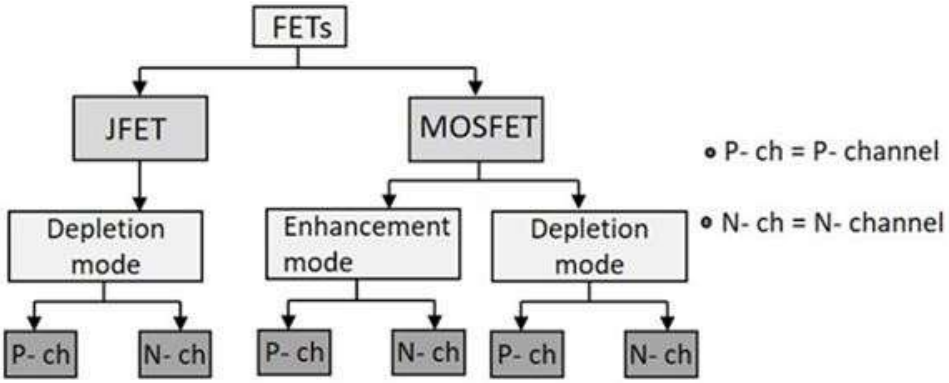


Figure 9.2: Classification of FET

9.4.1 Construction of n-channel and p-channel JFET

The junction field effect transistor is a fundamental component, operating on the principle of electric field control. A Junction Field-Effect Transistor (JFET) is a three-terminal semiconductor device constructed from a central doped bar (channel) of N-type or P-type material, with opposite-type impurities diffused into the sides to form two P-N junctions (the Gate). It consists of a Drain (D), Source (S), and Gate (G), acting as a voltage-controlled, majority-carrier device. There are two types of JFETs: N-channel JFET and P-channel JFET. The N channel JFET controls the flow of current through electrons, while the P channel JFET depends on the movement of the holes where the electrons are absent. The gate terminal plays a major role in conducting current and controls the electron flow from source to drain.

Some of the FET notations to be remembered:

- **Source**: The source S is a terminal through which the majority carriers enter the bar.
- **Drain**: The drain D is a terminal through which the majority carriers leave the bar.
- **Gate**: These are two internally connected heavily doped impurity regions which form two P-N junctions.
- **Channel**: The space between two gates through which majority carriers pass.
- **Depletion Region:** The area around the P-N junction where free carriers are absent, formed when the junction is reverse-biased.

N Channel JFET

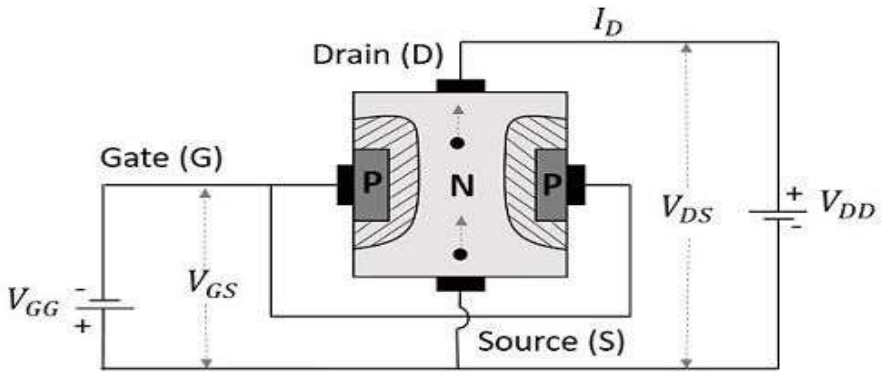


Figure 9.3: Biasing of Nchannel JFET

The n-channel junction field effect transistor is a semiconductor devicethat has three terminals: gate, source, and drain. It is made up of an N-type semiconductor thin bar material. The majority of charge carriers for n-typeare electrons. This forms a channel between source and drain. When a negative voltage is applied to the gate terminal, it creates anelectric field across the insulating material between the gate and the channel. Whenever a negative voltage in the gate increases, the electric field also increases, narrowing the channel and decreasing the current flow between the source and drain. To control the flow of electrons and to be used in applications such as voltage regulator, impedance switches, etc., the high input of this n channel JFET can be used.

The working of JFET can be explained by discussing about how to turn on N-channelJFET and how to turn off N-channel JFET. For turning ON a N-channel JFET, positive voltage of VDD has to be applied to the drain terminal of the transistor w.r.t (with respect to) source terminal such that the drain terminal must be appropriately more positive than the source terminal. Thus, current flow is allowed through the drain to source channel. If the voltage at the gate terminal, VGG is 0V, then there will be maximum current at the drain terminal and N-channel JFET is said to be in ON condition. For turning off the N- channel JFET, the positive bias voltage can be turned off or a negative voltage can be applied to the gate terminal. Thus, by changing the polarity of the gate voltage the drain current can be reduced and then N-channel JFET is said to be in OFF condition.

P Channel JFET

The p-channel junction field effect transistor is a semiconductor device that has three terminals: gate, source, and drain. It is made up of a semiconductor p-type thin bar material. The majority of carriers are holes in the p type. This forms a channel between source and drain. When a positive voltage is applied to the gate terminal, it creates an electric field across the insulating material between the gate and the channel. Whenever a positive voltage in the gate increases, the electric field also increases, narrowing the channel and decreasing the current flow between the source and drain. In applications such as amplifiers, switches, and signal processing, the high input of this p channel impedance JFET can be used to control the flow of holes.

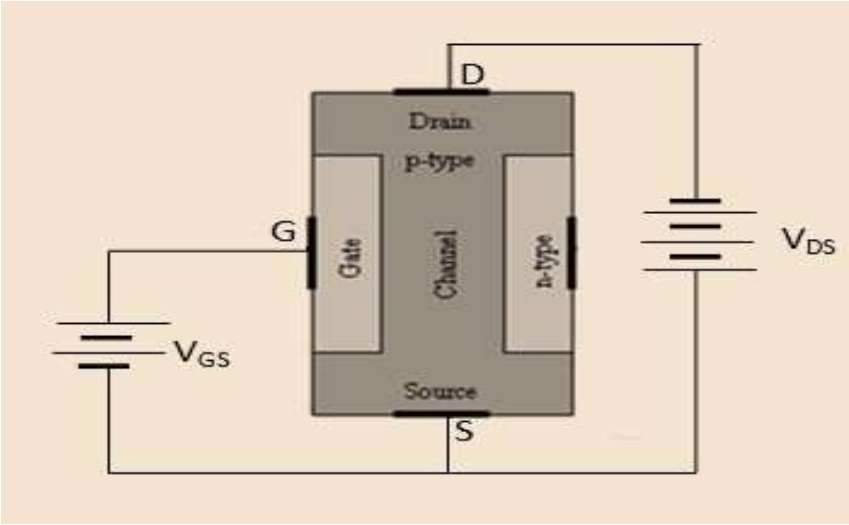


Figure 9.4: Biasing of Pchannel JFET

For turning ON P-channel JFET, negative voltage can be applied across the drain terminal of the transistor w.r.t source terminal such that the drain terminal must be appropriately more negative than the source terminal. Thus, the current flow is allowed through the drain to source channel. If the voltage at the gate terminal, VGG is 0V, then there will be maximum current at the drain terminal and the P-channel JFET is said to be in ON condition. For turning OFF the P-channel JFET, the negative bias voltage can be turned off or positive voltage can be applied to the gate terminal. If the gate terminal is given positive voltage, then the drain currents starts reducing (until cut-off) and thus the P-channel JFET is said to be in OFF condition.

9.5 PRINCIPLE OF OPERATION

The characteristics of Junction Field Effect Transistor (JFET) are as follows:

- **Temperature sensitivity**: The JFETs are sensitive to temperature variations, which allow them to adapt their properties with changing temperatures. In certain applications, proper thermal management may be necessary.
- **Low Noise**: JFETs are known for their low noise, which is valuable in applications such as amplifier and sensors where signal integrity is essential.
- **High input impedance**: The high input impedance of JFETs remains relatively

stable over a range of temperatures and voltages, making them suitable for precision applications.

- **Variants**: JFETs, which allow circuit design flexibility and enable applications that require both sourcing and sinking current, are available in two types of channels: N channel and P Channel.

9.5.1 Drain Characteristics

The output characteristics of the JFET were plots between drain current (Id) and drain source voltage (Vds). These characteristics find the variation of drain current concerning the voltage applied at the drain-source terminals while keeping the gate-source voltage constant.

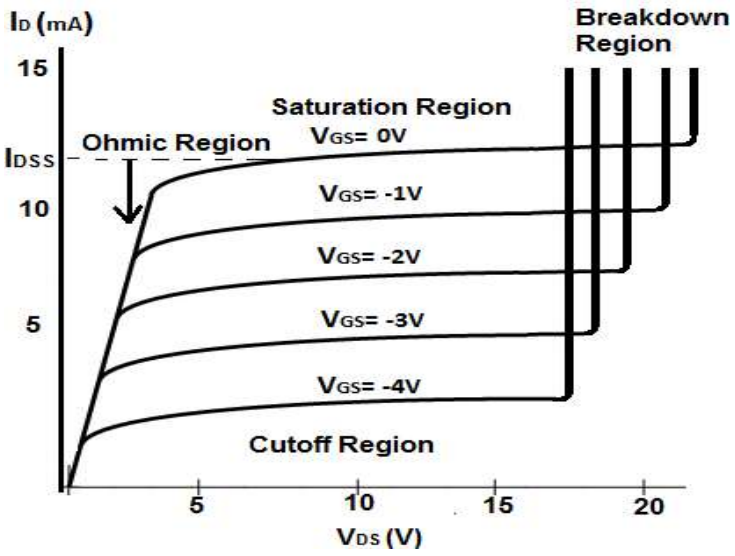


Figure 9.5: Static/Drain characteristics curve of N-channel

- **Saturation region:** In this region, the JFET operates with low V_{ds}, and the current is constant. The current is controlled by the gate source voltage (Vgs).
- **Ohmic region:** JFET becomes an ohmic region with increasing V_{DS}. The drain current increases with Vds.
- **Pinched off region:** By increasing the V_{DS}, it can lead to a pinch of region; in this region, the channel narrows and the drain current becomes independent of the V_{DS}.
- **Breakdown Region:** As the voltage of drain source becomes very high then the JFET channels gets breakdown and current flows in an uncontrolled manner.

The gate-to-source voltage creates an electric field that JFETs use to function. JFET regulates the conductivity of the channel between the source and drain:

A) Depletion Mode Operation:

- Depletion mode refers to the JFET's initial state, when no voltage is provided to the gate. Between the source and the drain in this mode, there is a conducting channel. to permit unrestricted stream flow.

- When the gate receives an applied negative voltage (for n-channel JFETs) or positive voltage (for p-channel JFETs). The channel will be used by the carrier. There is a decrease in the channel width. and hence, the current flow.

B) Voltage Control of Current Flow:

- The current flow from the source to the drain (I_D) is controlled by the voltage applied to the gate (V_{GS}). A more negative V_{GS} (for n-channel JFETs) narrows the channel, reducing I_D. Conversely, a less negative V_{GS} or a positive V_{GS} widens the channel, increasing I_D.
- There is a threshold voltage (V_{GS} (th)) at which the channel becomes completely pinched off, and current flow stops. This is called the pinch-off voltage.

When V_{GS} = 0 V, the channel between the gate junction is fully open. When V_{DS} = 0, then no electric field is applied and the majority carrier (electrons) are not attracted from the source to the drain. Hence the drain current I_D = 0. When the value of V_{DS} increases from zero, then the electric field is set up between drain and source.

Now the current carriers electrons flow from source to the drain and hence drain current increases linearly. As the value of I_D increases, the voltage drop takes place in the bar. This voltage drop, reverse biases the gate and hence the width of the channel decreases.

The narrowing of the channel is not uniform as the voltage drop takes place along the length of the channel. The width of the channel goes on decreasing as we move away from the source. Ultimately a critical value of V_{DS} is reached when the channel is pinched off as shown in the figure. This value of voltage V_{DS} is known as pinch-off voltage V_P. At this voltage, the value of the gate current I_D also becomes constant. However, the channel is not completely blocked. If the channel is completely blocked at the Pinch off voltage, then I_D would have been reduced to zero. As the value of V_{DS} increases, the reverse voltage between the channel and the gate becomes, hence the breakdown of the gate junction takes place, resulting into a sharp increase in the drain current. When V_{GS} is negative, the reverse bias between the channel and the gate increases. As a result of this, the resistance of the channel increases, the drain current decreases. The Pinch of takes place at a lower value of V_{DS}. Similarly, the breakdown of the gate junction also takes place at lower value of V_{DS}.

Since that gate current is controlled by the gate bias beyond Pinch off, so the amplification of the signal can be achieved by changing the gate bias.

9.5.2 Transfer Characteristics

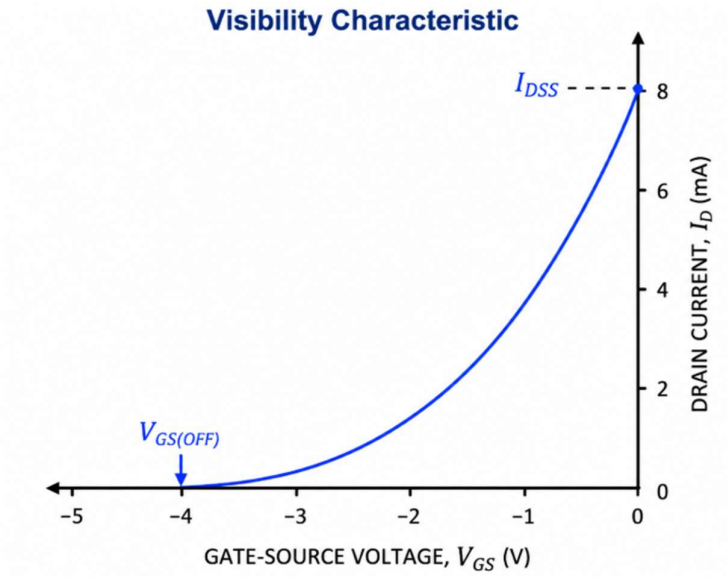


Figure 9.6: Transfer characteristics curve

The transfer characteristics of a JFET plotted between the drain current (I_d) and drain source voltage (V_{ds}). It can be determined by keeping the V_{ds} constant and drain current can be observed by changing the gate source voltage. So we can observe that when the gate source voltage V_{gs} is increased, the drain current I_d decreases. When the drain source voltage is constant, it can be observed that the value of the drain current varies inversely with respect to the gate source voltage. The above transfer characteristics curve of JFET is described below; it can be observed that the value of drain current varies inversely with respect to gate-source voltage (v_{gs}) when the drain-source voltage is constant. The transfer characteristics of JFET is a plot of output (drain) current versus input controlling quantity (gate-source voltage) and is used extensively in JFET amplifiers. In contrast to linear input-output relationship of BJT ($I_C = \beta I_B$), the input-output relationship of JFET is not linear. The relationship between I_D and V_{GS} is defined by Shockley's equation:

$$I_D=I_{DSS}(1-Vgs/Vp)^2$$

where, I_D is the drain current, I_{DSS} is the saturation drain source current, V_{gs} is the gate source voltage and V_p is the pinch off voltage.

The squared term on the right-hand side of the equation suggests that the relationship of I_D vs V_{GS} is nonlinear and exponential in nature. The transfer characteristics defined by Shockley's equation are unaffected by the network in which the device is employed. The transfer curve can be obtained using Shockley's equation or from the o/p characteristics.

9.6 JFET PARAMETERS

JFET has certain parameters which determine the performance. Such parameters of JFET are (i) ac drain resistance, (ii) transconductance, (iii) amplification factor, (iv) dc drain resistance, (v) Gate Cut Off Voltage and (vi) Shorted Gate Drain Current

i) **AC Drain Resistance (r_d):** It is defined as the ratio of change in drain-source voltage to change in drain current at constant gate-source voltage and is denoted by r_d. It is also called the **dynamic drain resistance**.

$$r_d=\Delta V_{DS}/\Delta I_D \text{ at constant } V_{GS}$$

ii) **Transconductance (g_m):** The control that the gate-source voltage has over the drain current, I_D is measured by transconductance. It is denoted by g_{fs}. It may be defined as the ratio of change indrain current to the change in gate-source voltage at constant drain-source voltage. It is also called the forward transconductance (g_{fs}) or forward transadmittance (Y_{fs}). It is measured in mA/volt or millisiemens.

$$g_m=\Delta I_D/\Delta V_{GS} \text{ at constant } V_{DS}$$

iii) **Amplification Factor (μ):** It is defined as the ratio of change in drain-source voltage to the change in gate-source voltage at constant drain current and is denoted by μ.

$$\mu=\Delta V_{DS}/\Delta V_{GS} \text{ at constant } I_D$$

Also, $\mu=\Delta V_{DS}/\Delta I_D X \Delta V_{GS}/\Delta I_D \mu = r_d \text{ X } g_m$

amplification factor = ac drain resistance X transconductance Amplification factor μ of a FET may be as high as 100

iv) DC Drain Resistance (R_{DS}): It is also called the static or ohmic resistance of thechanneland isdefinedasratio ofdrain-sourcevoltageanddraincurrent. It is denoted by R_{DS}.

$$R_{DS}=V_{DS}/I_D$$

v) Gate Cut Off Voltage (V_{GS} (off)): At a fixed drain voltage, the drain current (I_D) of a JFET depends on the gate to source voltage (V_{GS}). If the gate to source voltage decreases from zero in n channel JFET, the drain current also gets decreased accordingly. The relation between V_{GS} and I_D is as follows:

$$I_D = I_{DSS}\left[1 - \frac{V_{GS}}{V_{GS(off)}}\right]^2$$

After a certain gate to source voltage (V_{GS}), the drain current I_D becomes zero. This voltage is known as Cut Off Gate Voltage (V_{GS} (off)). This voltage numerically equals the pinch-off drain to source voltage (Vp). In the case of p channel JFET if we increase the gate terminal voltage from zero the drain current gets decreased and after a certain gate to source voltage, the drain current becomes zero. This voltage is the cut off gate voltage for p channel

JFET. It is the gate cut off voltage for p channel JFET.

vi) **Shorted Gate Drain Current (I_{DSS}):** When the gate terminal is grounded (V_{GS}= 0) and the drain-source voltage (V_{DS}) is slowly increased in an n-channel JFET, the drain current increases linearly. After the pinch-off voltage (Vp), the drain current stays constant, reaching its maximum value. This maximum current, called the Shorted Gate Drain Current (I_{DSS}), is fixed for each JFET.

W CHECK YOUR PROGRESS

Q1. Why is a JFET called a "unipolar" device?

Q2. What is the primary function of the Gate terminal?

Q3. What are the main advantages of JFETs over BJTs?

Q4. What is the input impedance of a JFET like?

Q5. How is a JFET biased in a simple amplifier circuit?

Q6. What does "pinch-off voltage" mean?

Q7. What are the primary disadvantages of a JFET?

Q8. Why is a JFET considered a voltage-controlled device?

Q9. How is the gate-source junction normally biased?

Q10. What happens to the channel width as VGS becomes more negative (for n-channel)?

Q11. What is the input impedance of a JFET?

Q12. Does a JFET operate in enhancement or depletion mode?

Q13. What is IDSS?

Q14. Why is JFET preferred in RF amplifiers?

2 9.7 LET US SUM UP

- A Field Effect Transistor (FET) is a voltage-controlled device where the output current is controlled by the input voltage.
- FET has high input impedance.
- FETs are unipolar devices which means they operate using only one type of charge carrier i.e. either electrons or holes.
- FETs are classified into Junction FET (JFET) and Metal Oxide Semiconductor FET (MOSFET) based on their structure.
- Each type is further divided into n-channel and p-channel depending on the type of charge carrier involved.
- JFETs are widely used in amplifiers and switching circuits due to their low noise, high efficiency, and thermal stability.
- High reliability and low noise make JFETs suitable for audio, scientific, and space applications.
- High input impedance minimizes signal distortion and loading effects.
- Drain characteristics shows the relationship between drain current and drain-

source voltage for different gate-source voltages.

- Transfer characteristics shows the relationship between drain current and gate-source voltage, indicating how the input voltage controls the output current.
- Three parameters of JEFT are drain characteristic, transconductance and amplification factor
- Drain resistance (r_n or r_d) represents the opposition offered by the channel to the flow of drain current.
- Transconductance (g_m) indicates how effectively the gate-source voltage controls the drain current.
- Amplification factor (μ) defines the overall voltage gain capability of the JFET.

K 9.8 ANSWERS TO CHECK YOIUR PROGRESS

Answer to Q1: Current conduction relies on only one type of majority charge carrier (either electrons in N-channel or holes in P-channel), unlike BJT which is bipolar.

Answer to Q2: To control the current flow between the source and drain by changing the channel width using reverse bias.

Answer to Q3: Very high input impedance, lower noise, and better thermal stability.

Answer to Q4: JFETs have a very high input impedance because the gate-source junction is reverse-biased, resulting in negligible gate current.

Answer to Q5: It is typically self-biased using a source resistor (R_S) which creates a voltage drop, ensuring the gate-source junction remains reverse-biased.

Answer to Q6: It is the value of the drain-source voltage (VDS) at which the drain current (I_D) becomes constant and stops increasing, regardless of further increases in V_{DS}.

Answer to Q7: JFETs have a low gain-bandwidth product and lower voltage gain compared to conventional BJTs.

Answer to Q8: The drain current (I_D) is controlled by the input voltage applied between the gate and source (V_{GS}), rather than by an input current.

Answer to Q9: The gate-source junction is always reverse-biased to ensure high input impedance and control the depletion region.

Answer to Q10: As V_{GS} becomes more negative, the depletion region widens, narrowing the conducting channel and reducing the drain current.

Answer to Q11: It is extremely high (often in the $M\Omega$ to $G\Omega$ range) because the gate-source junction is reverse-biased.

Answer to Q12: Only in depletion mode (normally ON)

Answer to Q13: The maximum Drain-to-Source current that flows when the Gate-to-Source voltage (V_{GS}) is zero.

Answer to Q14: JFETs have a very low noise level, making them ideal for amplifying weak signals from antennas.

O 9.9 FURTHER READING

- Vančo Litovski, “Junction Field Effect Transistor—JFET” Lecture Notes in Electrical Engineering, DOI:10.1007/978-981-19-9868-3_5, In book: Lecture Notes in Analog Electronics (pp.143-163)

- Physics of Semiconductors: Grundmann, Marius (2010). *The Physics of Semiconductors*. Springer-Verlag.
- Circuit Theory: Horowitz, Paul; Hill, Winfield (1989). *The Art of Electronics* (2nd ed.). Cambridge University Press
- M. Citterio, "Radiation effects on Si-JFET devices for front-end electronics." *IEEE* Transactions on Nuclear Science (1996).
- Shockley, W. (1952), "A Unipolar 'Field-Effect' Transistor." *Proceedings of the IRE*, 40(11), 1365–1376.

M 9.10 MODEL QUESTIONS

Q1. Explain the basic structure and working principle of a JFET.
Q2. Compare JFET and BJT based on operation, input impedance, and applications.
Q3. Describe the construction and operation of n-channel and p-channel JFETs.
Q4. Draw and explain the drain characteristics of a JFET.
Q5. Draw and explain the transfer characteristics of a JFET.
Q6. Define JFET parameters and explain their significance in circuit performance

*** ***** ***

UNIT 10: FIELD EFFECT TRANSISTOR-PART II

UNIT STRUCTURE

10.1 Learning Objectives
10.2 Introduction
10.3 Introduction to MOSFET
10.4 Classification of MOSFETs
10.4.1 Depletion type MOSFET
10.4.2 Enhancement type MOSFET
10.5 Operation of Depletion type MOSFET
10.6 Operation of Enhancement type MOSFET
10.7 Characteristics of MOSFET
10.7.1 Threshold voltage
10.7.2 Body effect
10.8 Let Us Sum Up
10.9 Answers to Check Your Progress
10.10 Further Reading
10.11 Model Questions

10.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- discuss the basic structure and operating principle of MOSFET
- Identify different types of MOSFETs based on their mode of operation
- explain the working of depletion-type MOSFET and its behavior under different bias conditions
- describe the operation of enhancement-type MOSFET and its switching characteristics
- analyse the electrical characteristics of MOSFET for effective application in electronic circuits.

10.2 INTRODUCTION

In the previous unit we have discussed about the construction and biasing of Junction Field Effect Transistor (JFET). In this unit, we will discuss about the other type of Field Effect Transistor (FET) called as Metal Oxide Semiconductor Field Effect Transistor (MOSFET). This unit will first focus on the basic concept of MOSFET. It will also provide the classification of MOSFETs. Operation of depletion and enhancement type MOSFET will also be discussed in this unit. Finally, the characteristics of MOSFET will be explored.

10.3 INTRODUCTION TO MOSFET

MOSFET referred to as MOS tube, has become the most widely used semiconductor device at present, and is widely used in various electronic equipment, including

power supplies, computers, TVs, etc. It is also referred to as an Insulated-Gate Field-Effect Transistor (IGFET) because the gate terminal is electrically isolated from the channel by a thin insulating oxide layer. MOSFETs contain an intrinsic body diode (also known as a parasitic diode) connected between the drain and source. The primary role of this parasitic diode is protection and freewheeling. In inductive load applications (like motor drives), it provides a path for back-EMF current when the transistor switches off, preventing high voltage spikes from exceeding the breakdown voltage (V_{DSS}) and burning out the MOSFET. MOSFETs include extremely high input impedance, fast switching speeds, and good thermal stability. They are voltage-controlled devices, widely used as electronic switches in high-frequency power supplies and amplifiers.

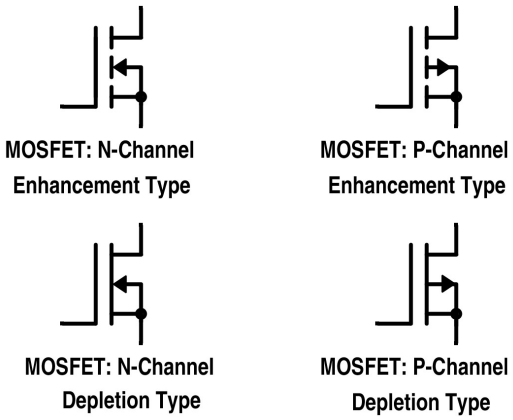


Figure 10.1: Symbol of Enhancement and Depletion type MOSFET

MOSFET basically consists of three parts: active region (source), drain region (drain) and gate region (gate). In an N-channel MOSFET, on a P-type substrate (substrate), N-type deposits form the source and drain regions, and an insulating material (usually silicon oxide) is deposited between them to form the gate. By changing the voltage of the gate, the carrier concentration in the channel is changed, thereby changing the conductance between the source and the drain.

Comparison between JFET and MOSFET:

The following is the comparison table between JFET and MOSFET:

JFET	MOSFET
JFET stands for Junction Field-Effect Transistor	MOSFET stands for Metal Oxide Semiconductor Field Effect Transistor
It only operates in the depletion mode	It operates in both depletion mode and enhancement mode
It has less input impedance than a MOSFET (about $10^9\Omega$)	It offers higher input impedance than a JFET (about $10^{14}\Omega$)
Manufacturing process is simple which makes them cheaper	Manufacturing process is complex and the additional metal oxide layer adds to the high cost
It is less susceptible to damage because of the high input capacitance	The metal oxide insulator reduces the input capacitance making it more susceptible to damage
It is mainly used in low noise applications	It is mainly used in high noise applications

Applications of MOSFETs:

The Applications of MOSFETs:

- **Switch Mode Power Supplies (SMPS):** MOSFETs dominate power management in devices like laptops, smartphone chargers, and adapters due to high efficiency and rapid switching speeds.
- **Motor Control & Drive Systems:** They are used in industrial automation, drones, and electric vehicles (EVs) for pulse-width modulation (PWM) to control speed and torque.
- **Digital Logic Circuits:** Millions to billions of MOSFETs are utilized in microprocessors, RAM, and CMOS technology, acting as switches for logic gates and data processing.
- **Automotive Electronics:** Essential for battery management systems (BMS), electric power steering, and ignition systems due to their robustness and thermal efficiency.
- **Renewable Energy Conversion:** Critical in solar inverters, solar charge controllers, and wind energy systems to minimize losses during DC-AC or DC-DC conversion.
- **Amplifiers and Signal Processing:** Used in audio, radio frequency (RF), and operational amplifiers (op-amps) because of their high input impedance, allowing precise amplification.

Lighting Systems: Employed in LED lighting for precise current control to improve efficiency.

10.4 CLASSIFICATION OF MOSFETS

Depending upon the type of materials used in the construction, and the type of operation, the MOSFETs are classified as follows:

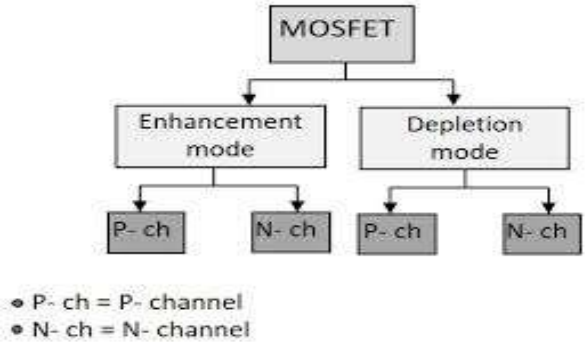


Figure 10.2: Classification of MOSFET (Source: Tutorials point)

10.4.1 Depletion type MOSFET

A Depletion-type MOSFET (D-MOSFET) is a type of Metal-Oxide-Semiconductor Field Effect Transistor that is “normally ON” at zero gate-to-source voltage (V_{GS}). This is because the channel forms during manufacturing, so current flows between the drain and source even without any external gate voltage. Because of this behavior, D-MOSFETs are also called “normally ON” transistors. D-MOSFETs can operate in both depletion mode (reducing current) and enhancement mode (increasing current).

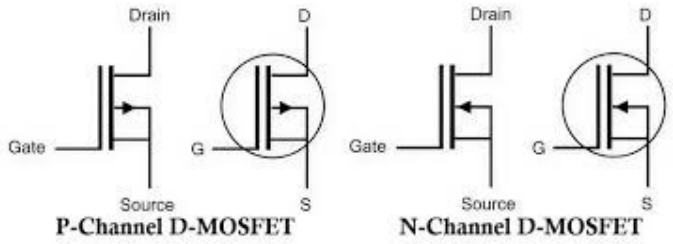


Figure 10.3: Symbol of P-channel and N-channel D-MOSFET (Source: Utmel)

A D-MOSFET can be constructed as either an **n-channel** or **p-channel** device, depending on the type of doped semiconductor used to create the channel.

N-Channel D-MOSFET

- **Substrate:** Lightly doped p-type.
- **Channel:** n-type region diffused between source and drain; present at fabrication.
- **Source & Drain:** Heavily doped n+ regions at either end of the channel.
- **Gate:** A metal or polysilicon electrode placed above the channel, insulated by a thin SiO_2 (silicon dioxide) layer.
  The channel contains electrons as majority carriers, and these electrons allow current flow when a voltage is applied between drain and source.
  P-Channel D-MOSFET
- **Substrate:** Lightly doped n-type.
- **Channel:** p-type region between source and drain; present at fabrication.
- **Source & Drain:** Heavily doped p+ regions at either end of the channel.
- **Gate:** A metal or polysilicon electrode placed above the channel, insulated by a thin SiO_2 (silicon dioxide) layer.

The channel contains holes as majority carriers; they allow current to flow when a voltage is applied between drain and source.

10.4.2 Enhancement type MOSFET

An Enhancement-type MOSFET (E-MOSFET) is a type of MOSFET that is normally OFF when the gate-to-source voltage (V_{GS}) is zero. Unlike D-MOSFETs, the channel is not pre-formed during manufacturing. Instead, the conducting channel is induced by applying a suitable gate voltage. E-MOSFETs operate only in enhancement mode, i.e., they require a gate voltage to create a channel and allow current flow between the drain and source.

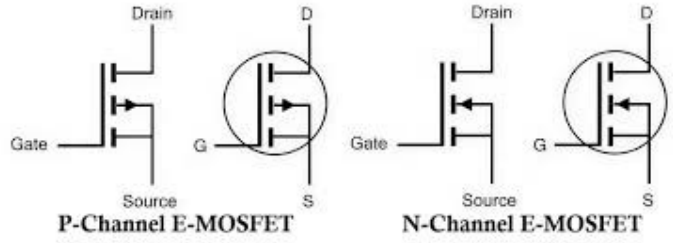


Figure 10.4: Symbol of P-channel and N-channel E-MOSFET (Source: Utmel)

E-MOSFETs are fabricated in n-channel and p-channel configurations.

N-Channel E-MOSFET

- **Substrate:** Lightly doped p-type semiconductor.
- **Source & Drain:** Heavily doped n+ regions diffused into the p-type substrate.
- **Gate:** A metal or polysilicon gate is deposited above the substrate.
- **Insulating Layer:** A thin layer of SiO_2 (silicon dioxide) insulates the gate from the substrate.
- **No physical channel** exists between source and drain initially.

P-Channel E-MOSFET

- **Substrate:** Lightly doped **n-type** semiconductor.
- **Source & Drain:** Heavily doped **p+ regions** diffused into the substrate.
- **Gate:** A metal or polysilicon gate is present above the substrate.
- **Insulating Layer:** A thin layer of **SiO_2 (silicon dioxide)** insulates the gate from the substrate.
- **No physical p-type channel** exists at zero gate bias.

10.5 OPERATION OF DEPLETION TYPE MOSFET

A depletion-mode MOSFET is naturally conducting in the absence of an external gate voltage (V_{GS} = 0V). In this state, a conducting channel formed by free carriers (usually electrons) exists between the source and drain, allowing current to flow freely.

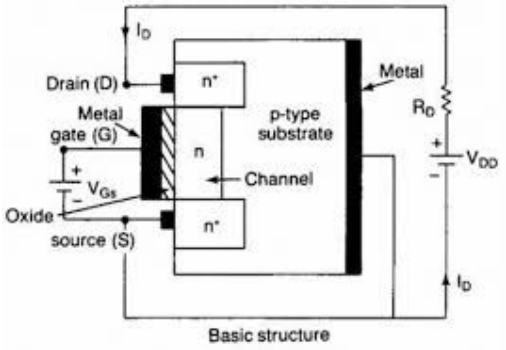


Figure 10.5: Biasing of Depletion type MOSFET (Source: Electronics for you)

When a negative voltage is applied to the gate of an N-channel depletion-mode MOSFET, the operation changes subtly. The negative voltage repels electrons in the gate region toward the source, reducing the free carrier density in the channel near the gate. As the negative voltage increases, the free carriers in the channel decrease until the channel are completely "depleted," shutting off the flow of current. Conversely, applying a positive voltage to the gate attracts more electrons into the channel, increasing the density of free carriers. This enhances the conductivity of the channel, increasing the current flow between the source and drain. This process allows the MOSFET channel to be precisely controlled by adjusting the polarity and magnitude of the gate voltage, thereby regulating the current.

Drain Characteristics

Drain Characteristics of N-Channel Depletion-Mode MOSFET

The drain characteristic of an N-channel depletion-mode MOSFET enables effective use. As the drain-source voltage (V_{DS}) rises, the drain current (I_D) of this MOSFET increases until it reaches a point called the saturation current (I_{DSS}), typically when the gate-source voltage (V_{GS}) is 0V.

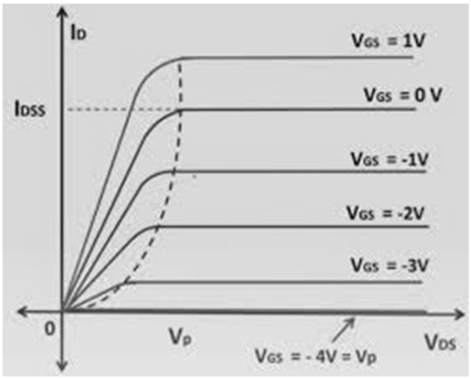


Figure 10.6: Drain Characteristics of N-Channel Depletion-Mode MOSFET
(Source: Electronics for you)

As V_{DS} increases, the drain current initially rises, reflecting the freer movement of electrons in the channel. However, once the current reaches saturation, the current remains constant even if V_{DS} increases further because the electron velocity in the channel is close to the maximum value. When a negative voltage is applied to the gate, it repels the electrons in the channel to the P-type substrate, causing them to recombine with holes. This electron-hole recombination reduces the number of free electrons, thus reducing the current. As negative V_{GS} increases, the recombination rate accelerates, further depleting the channel of electrons until the current reaches zero at the pinch-off voltage. Conversely, applying a positive V_{GS} attracts more electrons into the N channel, increasing the electron density and channel conductivity, thereby raising the current I_D above I_{DSS}. This allows precise control of the current by adjusting V_{GS}, suitable for applications that require fine current regulation.

Drain Characteristics of P-Channel Depletion MOSFET

The drain characteristics of a P-channel depletion-mode MOSFET are determined by the drain-source voltage (V_{DS}) and the gate-source voltage (V_{GS}). When V_{DS} is negative, it means that the voltage is applied from source to drain, and the setting is opposite to that of an N-channel MOSFET. If V_{GS} is positive, the drain current (I_D) decreases as V_{GS} increases. This is because positive voltage reduces the number of holes (carriers in a P-channel MOSFET) in the channel, thereby weakening its conductivity.

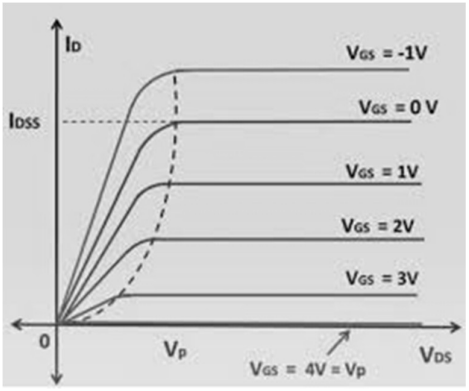


Figure 10.7: Drain Characteristics of P-Channel Depletion MOSFET
(Source: Electronics for you)

When V_{GS} increases to a certain point, the MOSFET enters a pinch-off state, pushing all holes out of the channel, and causing the drain current I_D to drop to zero, enabling precise current control. Conversely, applying a negative voltage to V_{GS} increases the number of holes in the channel, causing the drain current I_D to rise above the drain saturation current (I_{DSS}).

Transfer Characteristics

Transfer Characteristics of N-Channel Depletion MOSFET

The transfer characteristics of an N-Channel depletion-mode MOSFET describe how the drain current (I_D) varies with the gate-source voltage (V_{GS}) at a fixed drain-source voltage (V_{DS}).

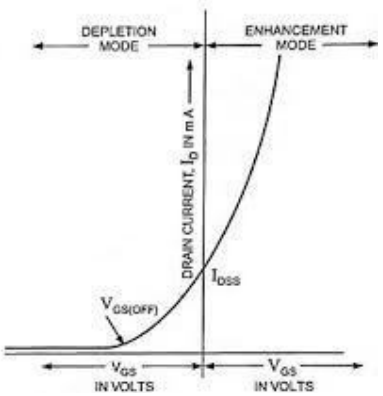


Figure 10.8: Transfer characteristics curve (Source: eeeguide)

When V_{GS} is positive, I_D increases as V_{GS} increases. This is because positive V_{GS} attracts more electrons into the channel, which reduces resistance and increases current. This occurs in the "enhancement region," where more electrons make the channel more conductive. Conversely, when V_{GS} is negative, the MOSFET enters the "depletion region." Here, negative V_{GS} repels electrons from the channel, which increases resistance and reduces current. By adjusting the negative V_{GS}, engineers can precisely control the reduction in current until it is completely cut off. This shows that I_D depends on the ratio between V_{GS} and the pinch-off voltage (V_P), which is the voltage at which the MOSFET is fully turned off. This equation helps engineers predict the current behavior at different V_{GS} values, which can optimize MOSFET applications.

Transfer Characteristics of P-Channel Depletion MOSFET

The transfer characteristics of a P-Channel depletion-mode MOSFET are very different from those of an N-Channel depletion-mode MOSFET. In actual operation, these characteristics show how the gate-source voltage (V_{GS}) regulates the drain-source current (I_D).

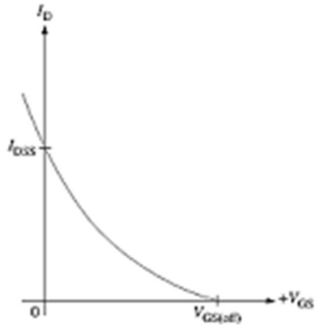


Figure 10.9: Transfer characteristics curve (Source: eee guide)

When V_{GS} is positive, I_D gradually increases from zero. As V_{GS} continues to rise, the current reaches I_{DSS}, the maximum saturation current, reflecting the gradual increase in channel conductance. When V_{GS} becomes negative, the P-Channel MOSFET allows the current to continue to increase. This unique characteristic of the depletion-mode MOSFET means that increasing the negative gate-source voltage reduces the hole repulsion within the channel, further enhancing the current flow. This behaviour is opposite to that of an N-Channel MOSFET, where increasing negative V_{GS} reduces the current.

10.6 OPERATION OF ENHANCEMENT TYPE MOSFET

The working of E-MOSFET is when VGS is 0V then there is no channel that will connect the source & drain. The p-type substrate has only a small number of thermally produced minority charge carriers like free electrons thus the drain current is zero. Because of this reason, this MOSFET will be normally OFF.

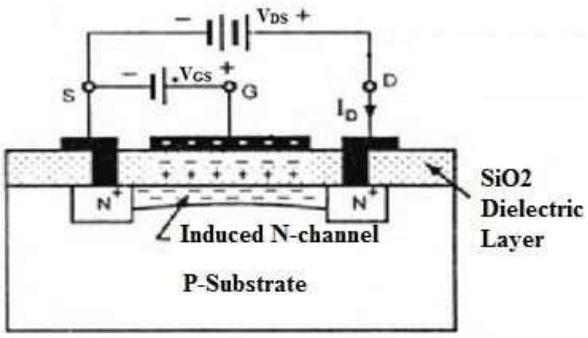


Figure 10.10: Biasing of Enhancement type MOSFET (Source: Electronics for you)

Once the gate (G) is positive (+ve), then it attracts minority charge carriers like electrons from p–substrate where these charge carriers will combine through the holes under the layer of SiO2. Further VGS is increased then the electrons will have enough potential to overcome and bonding and more charge carriers i.e. electrons get deposited in the channel. Here, the dielectric is used to prevent the electron's movement across the silicon dioxide layer. This accumulation will result in the n-

channel formation between Drain and Source terminals. So this can lead to the generated drain current flow throughout the channel. This drain current is simply proportional to the channel's resistance which depends further on the charge carriers attracted to the +ve terminal of the gate. Now on increasing V_{GS} further, a high electric field is developed forcing atoms inside the P substrate to break. The free electrons generated will fill the holes near the gate region. This way holes are pushed away from the gate terminal increasing N-type behavior near the gate terminal. A time will come when an N-channel is created between the two N wells. The VGS voltage at which the channel is created is called the threshold voltage or V_T. We can conclude from this discussion when $\mathbf{V_{GS} > V_T}$ an N channel is induced near the gate terminal. A channel is created still we are not getting any current. Apply a voltage source between the drain and the source keeping the drain at a higher potential. This voltage is called V_{DS}. On applying this voltage current will start flowing from drain to source. This current is called drain current or I_D. Thus, when $\mathbf{V_{GS} > V_T}$ and $\mathbf{V_{DS} > 0}$, the current I_D flows from drain to source. On increasing the positive voltage at the drain terminal a reverse bias is formed at the PN junction near the drain terminal. This will result in a thick depletion region near the PN junction. Hence on increasing VDS further, you will see the channel near the drain terminal is becoming narrow. The drain current will face more resistance near the drain terminal. A situation will reach when the drain current becomes constant and will not increase further. This situation is called the pinch-off situation and the drain current is called the saturation current. The voltage at which we will get saturation current is called saturation voltage. We can conclude from this discussion, that pinch-off is reached when $\mathbf{V_{GS} > 0}$ (constant) and $\mathbf{V_{DS} = V_{DS}(SAT)}$, $\mathbf{I_D = I_D(SAT)}$.

Drain characteristics

Drain Characteristics of N-Channel Enhancement-Mode MOSFET

The N channel enhancement MOSFET drain characteristics are shown below. In these characteristics, we can observe the drain characteristics plotted between the Id and Vds for different Vgs values as shown in the diagram. When the Vgs value is increased, then the current 'Id' will also be increased. The parabolic curve on the characteristics will show the locus of V_{DS} where the Id (drain current) will get saturated. In this graph, the linear or ohmic region is shown. In this region, the MOSFET can function as a voltage-controlled resistor. So, for the fixed Vds value, once we change the Vgs voltage value, then the channel width will be changed or we can say that the resistance of the channel will change.

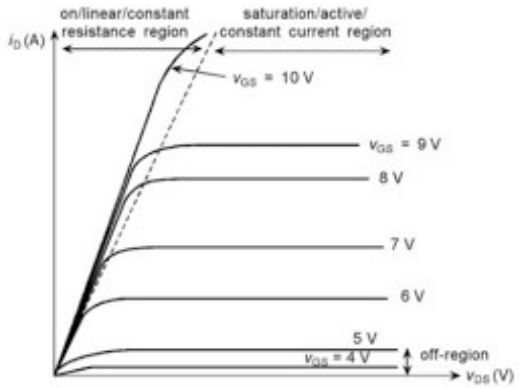


Figure 10.11: Drain Characteristics of N-Channel Enhancement-Mode MOSFET (Source: tutorialpoint)

The ohmic region is a region where the current 'I_{DS}' raises with an increase in the V_{DS} value. Once MOSFETs are designed to work in the ohmic region, then they can be utilized as amplifiers**.** The gate voltage at which point the transistor turns ON & starts flowing current throughout the channel is known as threshold voltage (V_T or V_{TH}). For N-channel, this threshold voltage value ranges from 0.5V – 0.7V whereas for P-channel devices it ranges from -0.5V to -0.8V. Whenever the Vds<Vgs-Vt & Vgs > Vt then, in this case, the MOSFET will operate in a linear region. So in this region, it can function as a voltage-controlled resistor. In the cut-off region, when the voltage Vgs <VT then the current throughout the MOSFET is zero otherwise we can say that the MOSFET will stay in the OFF condition. Whenever the mosfet is operated on the right side of the locus then we can say that it is operated in a saturation region. So, mathematically, whenever the Vgs voltage is > or = Vgs-Vt then it is operating in a saturation region.

Drain Characteristics of P-Channel Enhancement-Mode MOSFET

The P channel enhancement MOSFET drain characteristics are shown below. Here, the Vds and Vgs will be negative. The drain current 'Id' will supply from the source to the drain terminal. As we can notice from this graph, when Vgs become more negative then the drain current 'Id' will also increase.

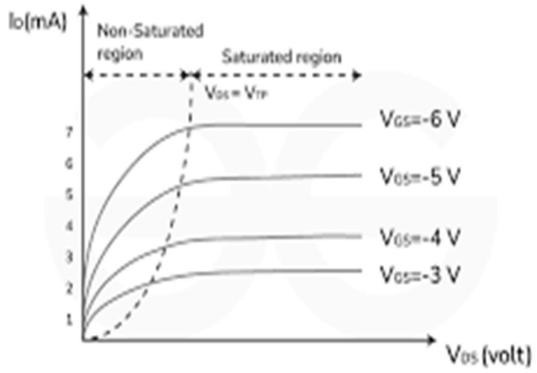


Figure 10.12: Drain Characteristics of P-Channel Enhancement-Mode MOSFET (Source: tutorial point)

Transfer Characteristics

Transfer Characteristics of N-Channel Enhancement MOSFET

The transfer characteristics of the N channel enhancement MOSFET is shown below. The transfer characteristics show the relationship between the input voltage 'Vgs' and output drain current 'I_D'. These characteristics basically show how the 'I_D' changes when Vgs values change. So from these characteristics, we can observe that the drain current 'I_D' is zero upto the threshold voltage. After that, when we increase the Vgs value, then the 'I_D' will increase. The relationship between the current 'I_D' and Vgs can be given as $I_D = k(Vgs-Vt)^2$. Here, the 'K' is the device constant which depends on the device's physical parameters. So by using this expression, we can find out the drain current value for the fixed Vgs value.

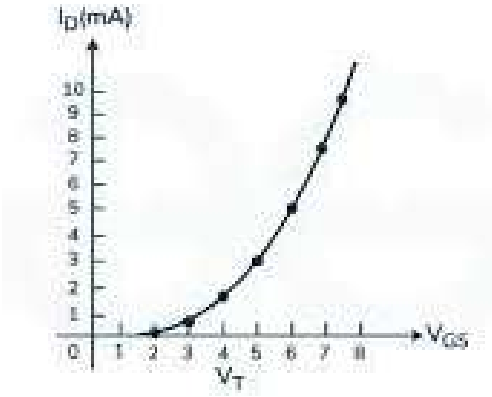


Figure 10.13: Transfer characteristics curve (Source: eeeguide)

Transfer Characteristics of P-Channel Enhancement MOSFET

When the Vgs >V_T, then this MOSFET will operate in the cut-off region. Similarly, if you observe the transfer characteristics of this MOSFET then it will be a mirror image of the N-channel.

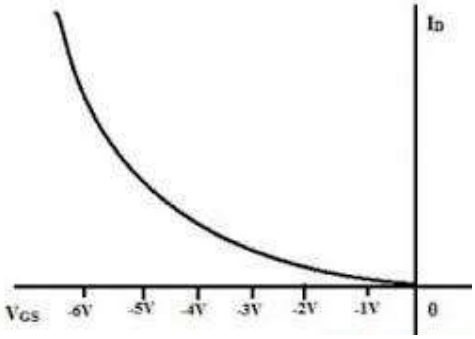


Figure 10.14: Transfer characteristics curve (Source: eeeguide)

10.7 CHARACTERISTICS OF MOSFET

Threshold voltage and body effect are crucial as they determine how transistors turn on and off, impacting performance and power consumption.

10.7.1 Threshold Voltage

Threshold voltage is the minimum voltage required at the gate terminal of the FET to create a conducting path (or channel) between the source and drain terminals. The existence of this channel allows current to flow through the transistor.

For an N-channel MOSFET it is as follows:

- When the voltage difference between the gate and the source (Vgs) is below the threshold voltage, no current flows from the drain to the source. This is because there's no conductive channel formed. The MOSFET is in the off state, or cutoff region.
- When Vgs equals the threshold voltage, the MOSFET begins to turn on, and a channel just begins to form. This is the onset of the active region (or saturation region).

- When Vgs is greater than the threshold voltage, the MOSFET is fully on, the channel is formed, and current can flow freely from drain to source. The MOSFET operates in the active region, or saturation region.

For a P-channel MOSFET, the principles are the same, but the voltage polarities are reversed.

The exact value of the threshold voltage depends on the specific design and fabrication of the transistor and is typically provided in the device's datasheet.

10.7.2 Body Effect

Whenever there is a voltage (potential) difference between the source and substrate (body), this leads to an increase or decrease in the threshold voltage of the transistor. This is called a “body effect.”

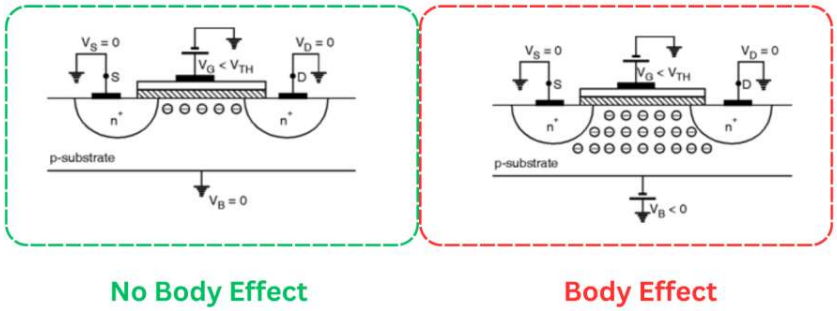


Figure 10.15: MOSFET with and without Body Effect

To understand this effect suppose $V_S = 0$ and $V_D = 0$ and V_G is somewhat less than V_{TH} so that depletion region is formed under the gate but inversion channel does not exist. As V_B becomes more negative (i.e. $V_B < V_S$ where $V_S = 0$) more holes are attracted to the substrate connection leaving a larger negatively charged ions behind i.e. the depletion region becomes wider. As we know that the threshold voltage is a function of the total charge in the depletion region (i.e. Q_{dep}). Thus as the body voltage V_B drops then depletion charge (Q_{dep}) increases which increases the threshold voltage (V_{TH}). This effect is called as the body effect or back gate effect. When the body bias voltage V_{SB} is applied between source and body the surface potential required for strong inversion is increased.

W CHECK YOUR PROGRESS

Q1: What does MOSFET stand for?
Q2: How many terminals does a MOSFET have?
Q3: What is the Threshold Voltage (Vth)?
Q4: How does a MOSFET compare to a BJT?
Q5: Is a MOSFET current-controlled or voltage-controlled?
Q6: What is the function of the Gate terminal?
Q7: Why are MOSFETs preferred in integrated circuits (ICs)?
Q8: What happens to the threshold voltage if substrate doping increases?
Q9: Why is a resistor sometimes used on the gate?
Q10: What is the main application of a Power MOSFET?

2 10.8 LET US SUM UP

- MOSFET (Metal Oxide Semiconductor Field Effect Transistor) is a voltage-controlled device widely used for amplification and switching applications.
- It has very high input impedance and low power consumption, making it suitable for modern electronic circuits.
- MOSFETs are classified into depletion and enhancement types, each available in n-channel and p-channel configurations.
- Depletion MOSFET is normally ON at zero gate-source voltage, and applying a reverse gate voltage reduces the channel conductivity and decreases the drain current.
- Enhancement MOSFET normally OFF at zero gate-source voltage, and applying a suitable gate voltage induces a channel, allowing current to flow between drain and source.
- MOSFET characteristics are strongly influenced by the threshold voltage (V_t), which determines the minimum gate-source voltage required to form a conducting channel and control the drain current.
- The body effect modifies the threshold voltage when there is a change in the substrate-to-source voltage, leading to variations in drain current and overall device characteristics.

K 10.9 ANSWERS TO CHECK YOUR PROGRESS

Answer to Q1: Metal-Oxide-Semiconductor Field-Effect Transistor.

Answer to Q2: Three: Gate (G), Drain (D), and Source (S). (Note: The substrate/body is a fourth terminal, often internally connected to the source).

Answer to Q3: The minimum gate-to-source voltage (VGS) required to create a conducting channel in an enhancement-mode MOSFET.

Answer to Q4: MOSFETs have higher input impedance, faster switching speeds, and lower power consumption.

Answer to Q5: It is a voltage-controlled device. The gate-to-source voltage (VGS) controls the drain current (I_D).

Answer to Q6: It controls the conductivity of the channel between the drain and source by producing an electric field.

Answer to Q7: Due to their small size, high packing density, and low power dissipation.

Answer to Q8: The threshold voltage increases.

Answer to Q9: To protect the driving circuit from high gate capacitance, commonly 100-200 ohms.

Answer to Q10: High-speed switching in converters and power supplies.

O 10.10 FURTHER READING

- Agajanian A. H., (Springer, 1980) *"MOSFET Technologies: A Comprehensive Bibliography"*
- Rashid A.B.M.H. (Wiley, 2009), *"Compact MOSFET Models for VLSI Design"*
- Liou J. J., Ortiz-Conde A., and . Garcia-Sanchez F (Springer, 1998) *"Analysis and Design of MOSFETs: Modeling, Simulation, and Parameter Extraction"*

- Bhattacharyya A.B., Compact Mosfet Models for VLSI Design
- Bipolar Transistor and MOSFET Device Models by Kunihiro Suzuki
- *The Physics and Modeling of Mosfets Surface-Potential Model HiSIM* by Mitiko Miura-Mattausch (Hiroshima University, Japan), Hans Jürgen Mattausch (Hiroshima University, Japan) and Tatsuya Ezaki (Hiroshima University, Japan)
- *MOSFET - Developments and Trends* by Yuxiang Tu

M 10.11 MODEL QUESTIONS

Q1. Explain the basic structure and working principle of a MOSFET.
Q2. Classify MOSFETs and describe their types.
Q3. Explain the operation of a depletion-type MOSFET with characteristics.
Q4. Describe the working of an enhancement-type MOSFET.
Q5. Draw and explain the drain characteristics of a MOSFET.
Q6. Compare depletion-type and enhancement-type MOSFETs.

.

*** ***** ***

UNIT 11: MICROWAVE DIODES

UNIT STRUCTURE

11.1 Learning Objectives
11.2 Introduction
11.3 Metal Semiconductor and Heterojunctions
11.3.1 Schottky Barrier Diode
11.3.2 Ohmic Contact
11.4 Tunnel Diode
11.5 Varactor Diode
11.6 PIN Diode
11.7 Gunn Diode
11.8 IMPATT Diode
11.9 Let Us Sum Up
11.10 Answers to Check Your Progress
11.11 Further Reading
11.12 Model Questions

11.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- explain the basic concepts of metal–semiconductor junctions and heterojunctions
- discuss the working principle and applications of tunnel diodes
- describe the construction, operation, and characteristics of varactor and PIN diodes
- illustrate the working principles, characteristics, and applications of Gunn and IMPATT diodes and distinguish their key features.

11.2 INTRODUCTION

In the previous unit, we have learned about the field effect transistors which are mainly voltage driving devices, but to deal with signals of microwave range frequencies, which lie from 300MHz to 300GHz, some special type of diodes needs to be studied and constructed. In this unit we shall learn about different microwave diodes that are generally used for generation and detection of high frequency (microwave) signals. This content describes the construction and working principle of such microwave diodes along with their possible band diagrams required for understanding their operations.

11.3 METAL SEMICONDUCTOR AND HETEROJUNCTIONS

The first solid-state device was reported in 1874 and consisted of a fine wire pressed against a lead sulfide crystal. This simple metal–semiconductor interface, known as a whisker contact rectifier, marked the beginning of solid-state electronics. Although whisker contact rectifiers are now largely obsolete, metal–semiconductor junctions

remain fundamental components in microwave integrated circuits. In solid-state physics, a metal–semiconductor (M–S) junction refers to the electrical interface formed when a metal is brought into close contact with a semiconductor. It represents the earliest practical semiconductor device and can exhibit either rectifying or non-rectifying behavior. A rectifying M–S junction forms a Schottky barrier and functions as a Schottky diode, whereas a non-rectifying junction is referred to as an ohmic contact.
Further, a heterojunction is an interface between two layers or regions of dissimilar semiconductors. Heterojunction semiconductors are mainly fabricated by engineering band structures at the interface of two different materials. Unlike the homojunction, heterojunction semiconductor device provides an improved performance in terms of efficiency, power consumption, and wide absorption spectra along with better carrier confinement. Moreover, the heterojunctions enable precise control over band alignment (Type I, II, or III), allowing the deliberate formation of Ohmic or Schottky contacts that are crucial for advanced electronic device engineering.

11.3.1 Schottky Barrier Diode

A rectifying metal–semiconductor junction forms a Schottky barrier that enables rectification and is used in Schottky diodes. The key parameter of any Schottky junction is the Schottky barrier height, which depends on the specific metal–semiconductor combination. As compared to the non rectifying (ohmic) contacts, rectifying (non-ohmic) contacts exhibit a high barrier height.
Now let us have an insight on the band alignment, construction and working principle of Schottky Barrier diode.

Construction of Schottky barrier diode:
Before moving on the construction let us see what is mean by barrier height for a rectifying M-S junction referred to as the Schottky barrier. The figure below depicts a metal-semiconductor junction along with the formation of the barrier height.

Φ_B is the Schottky barrier height which is shown in the figure 11.1. When the Schottky barrier height Φ_B is sufficiently large—specifically, when it is much greater than the thermal energy k_T, the semiconductor becomes depleted in the region adjacent to the metal, and the junction exhibits Schottky barrier behavior. For silicon, this condition typically corresponds to a barrier height in the range of approximately 0.4eV to 0.7eV.

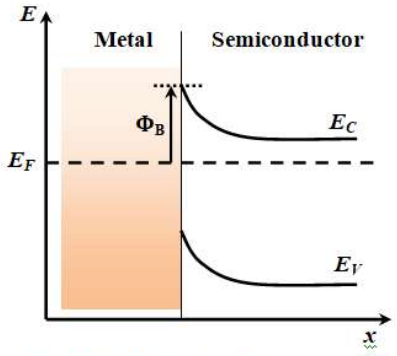
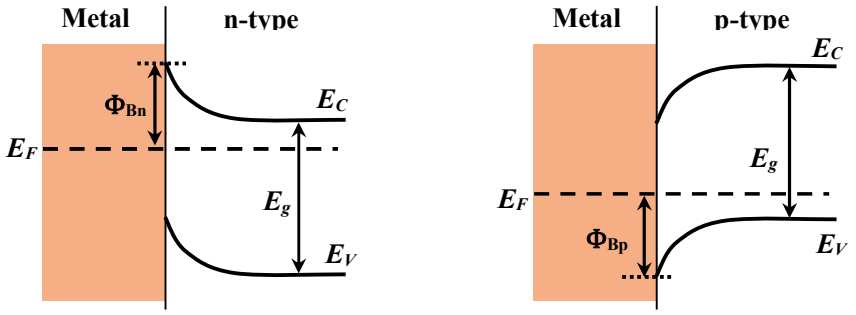


Figure 11.1: Band diagram for metal-semiconductor junction at zero bias (equilibrium)

The definition of the Schottky barrier height differs for n-type and p-type semiconductors, being referenced to the conduction band edge in n-type materials and to the valence band edge in p-type materials which is depicted in figure 11.2 (a and b).



(a) (b)

Figure 11.2: Band diagram of (a) metal and n-type semiconductor (b) metal and p-type semiconductor

For an n-type semiconductor, the Schottky barrier height Φ_{Bn} is given by

$$\Phi_{Bn} = \Phi_M - \chi$$

where Φ_M is the metal work function and χ is the electron affinity of the semiconductor.

For a p-type semiconductor, the Schottky barrier height Φ_{Bp} is expressed as

$$\Phi_{Bp} = E_g - \Phi_{Bn}$$

where E_g is the bandgap of the semiconductor. But in practice, mostly metal and n-type semiconductor junction is used for the construction of Schottky diode.

Till here, we came to know about how the Schottky barrier height is formed when a metal and a semiconductor are joined together to form a junction. Now, figure 11.3 (a and b) shows the schematic diagram of a Schottky barrier diode and with its circuit symbol.

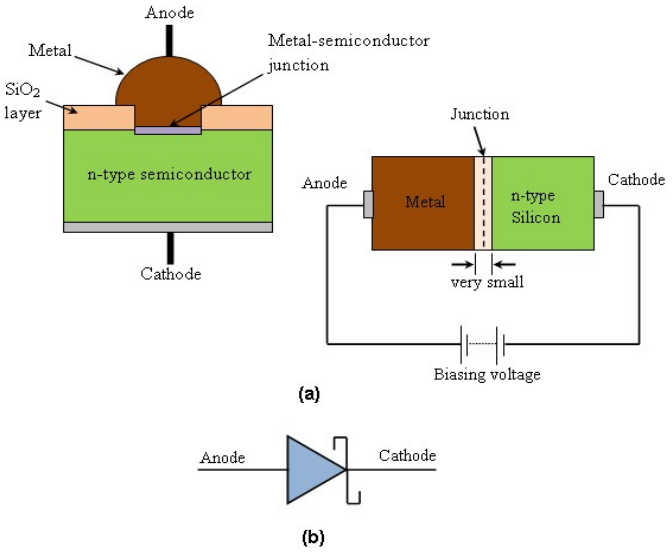


Figure 11.3 (a) Schematic diagram of a Schottky barrier diode

(b) Circuit symbol of a Schottky barrier diode

Working principle of Schottky diode:

Now let us discuss about the working principle of a Schottky diode. It is quite different than that of the conventional PN junction diode as Schottky diode has electron as majority carrier in the metallic region as well as in the semiconductor region. As we know that Schottky diode is made of metal and N-type semiconductor combination. Although metals contain a high density of free electrons, they are electrically neutral overall. In contrast, an n-type semiconductor has an excess of electrons due to donor impurities. When the metal and n-type semiconductor are brought into contact, electrons flow from the semiconductor into the metal until thermal equilibrium is established. As a result of this electron transfer, negative charge accumulates on the metal side of the junction, while a region of positive space charge, due to ionized donor atoms, forms on the semiconductor side. This region of charge separation is known as the depletion region. Figure 11.4 shows an unbiased Schottky diode with the formation of the depletion layer.

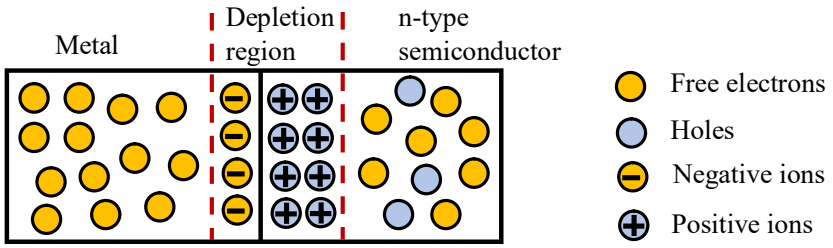


Figure 11.4: Formation of depletion region in an unbiased Schottky diode

The depletion region creates a potential energy barrier that opposes further electron flow across the junction. This barrier is characterized by a built-in potential, which represents the minimum voltage required to initiate significant conduction. The magnitude of the built-in potential depends on the width of the depletion region. Since the metal contains a large concentration of free carriers, the depletion region extends primarily into the n-type semiconductor, with a negligible width on the metal side. Consequently, the built-in potential of a Schottky junction is relatively small. To achieve conduction across the junction, an externally applied forward bias must overcome this built-in potential, resulting in the low forward voltage drop characteristic of Schottky diodes.

Forward bias: In the forward bias condition the metal of the Schottky diode is connected to the positive terminal of the battery whereas the negative terminal of the battery is connected to the semiconductor region. Upon application of an external voltage, charge carriers are supplied in the semiconductor region; however, electrons are unable to cross the potential barrier unless the applied voltage exceeds the built-in potential. For most Schottky diodes, this barrier potential is approximately 0.2V. When the applied voltage exceeds this value, electrons acquire sufficient energy to surmount the barrier and conduction begins. With further increase in the applied voltage, the width of the depletion region decreases and eventually becomes negligible. Figure 11.5 shows the forward biased condition of Schottky diode.

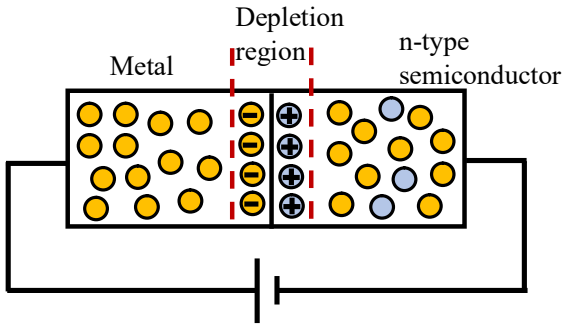


Figure 11.5: Forward biased condition of Schottky diode

Reverse bias: Under reverse-bias conditions, the negative terminal of the external supply is connected to the metal contact of the Schottky diode, while the positive terminal is connected to the n-type semiconductor. This reverse bias increases the height and width of the potential barrier, thereby widening the depletion region. Figure 11.6 shows the reverse bias condition of a Schottky diode.

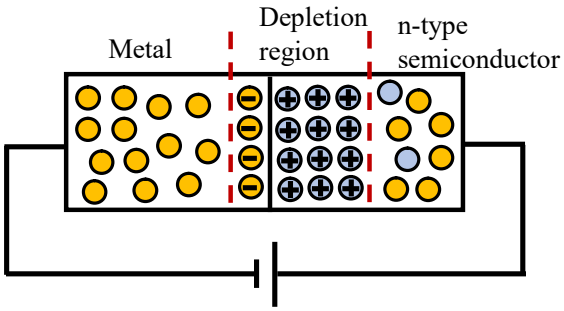


Figure 11.6: Reverse biased condition of Schottky diode

As a result, current conduction is effectively suppressed, although a small reverse leakage current persists. This leakage current increases gradually with increasing reverse bias voltage but remains limited up to a critical value. When the applied reverse voltage exceeds this critical limit, breakdown of the barrier occurs, leading to a sudden increase in reverse current. Such reverse breakdown results in permanent damage to the Schottky diode. A Schottky diode exhibits a very short reverse recovery time compared to a PN junction diode because conduction occurs solely through majority charge carriers (electrons). As a result, there is no charge storage effect, allowing the diode to cease conduction rapidly when the applied voltage is reversed.

I-V characteristics of Schottky diode: You can see that the current–voltage (I-V) characteristics of Schottky and p–n junction diodes are generally similar, differing primarily in the knee (turn-on) voltage. The knee voltage of a Schottky diode is significantly lower than that of a p–n diode; consequently, the forward voltage drop and the associated power dissipation are much smaller.

As indicated by the I-V characteristic shown in figure 11.7, a Schottky diode exhibits negligible conduction until the applied forward voltage reaches the knee point. A small forward leakage current may still be present. Once the applied voltage exceeds the knee voltage, which is typically in the range of 0.2 to 0.4 V for most Schottky diodes the forward current increases linearly after crossing the exponential region.

Under reverse-bias conditions, current flow is limited to a small reverse leakage current, which appears in the third quadrant of the I-V curve. This reverse leakage current increases slowly and remains relatively small with increasing reverse voltage until reverse breakdown occurs. Beyond the breakdown voltage, the diode conducts in

the reverse direction, leading to permanent device failure. Compared to PN junction diodes, Schottky diodes exhibit higher reverse leakage currents and lower reverse breakdown voltages due to their lower barrier height and reduced forward voltage drop.

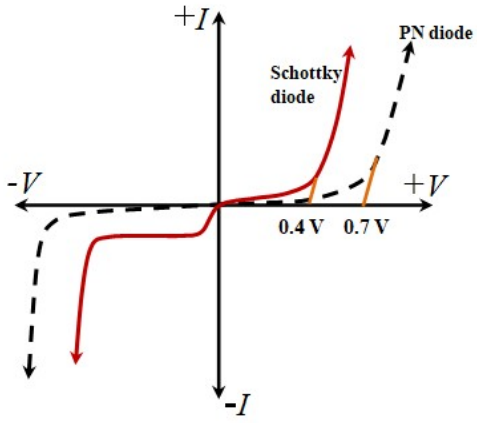


Figure 11.7: I-V characteristic curve of Schottky diode in comparison to PN diode

Applications of Schottky diode:
Schottky diodes are widely used in a variety of electronic applications due to their low forward voltage drop and fast switching characteristics. Typical applications include:

- High-speed, high-power rectification
- High-frequency RF applications, such as RF mixers and detectors
- Signal detection in FM demodulator circuits
- TTL logic circuits in digital systems, owing to low power loss and high switching speed
- Voltage clamping circuits
- Photovoltaic (solar cell) systems, where they are used to prevent reverse current flow from batteries or connected loads
- Switched-mode power supplies (SMPS)
- Battery protection circuits in multi-battery systems, where they prevent reverse current flow between batteries and avoid potential damage

11.3.2 Ohmic contact

Let us now discuss about the ohmic contact. In non-rectifying Schottky contacts, the barrier height is insufficient to produce a significant depletion region, resulting in a negligible depletion layer and this is referred to as ohmic contact. The ohmic contact is formed in an M-S junction where the work function of the employed semiconductor is higher that of the metal. Prior to contact, the Fermi levels of the metal and the semiconductor are at different energy positions as shown in figure 11.8(a) whereas after the contact the Fermi level of the metal and the semiconductor lined up as shown in figure 11.8(b).

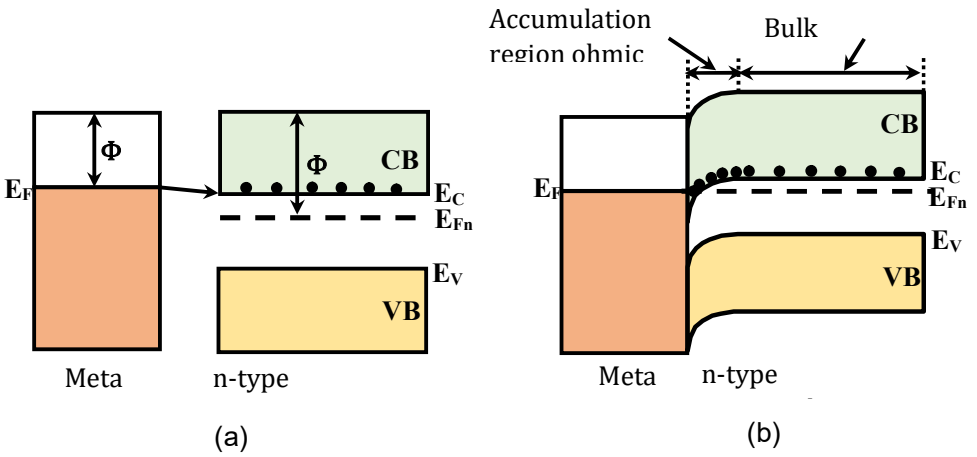


Figure 11.8: Ohmic contact (a) before contact (b) after contact

At thermal equilibrium, electrons transfer from the metal into the available states in the conduction band of the semiconductor until the Fermi levels align. This results in the formation of an accumulation region on the semiconductor side of the interface. The accumulation region contains a higher concentration of electrons than the bulk semiconductor and therefore exhibits increased conductivity. Consequently, an ohmic contact behaves as a low-resistance element, allowing current to flow linearly in both forward and reverse bias. The I-V characteristic curve of an ohmic contact is shown in figure 11.9. The overall resistance of the contact is primarily determined by the bulk resistivity of the semiconductor.

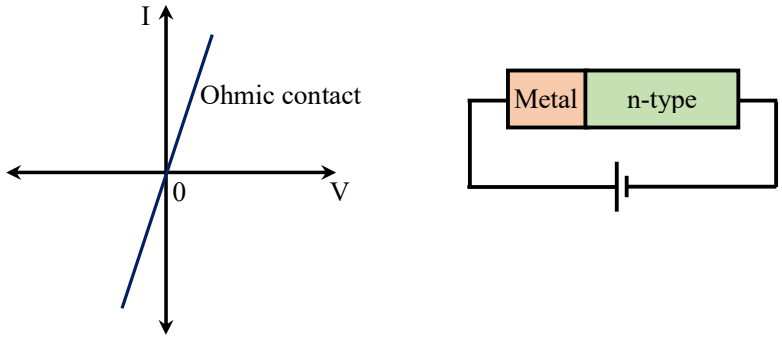


Figure 11.9: I-V characteristic curve of an ohmic contact

Low-resistance ohmic contacts are employed to ensure efficient bidirectional current flow without rectifying behavior and without significant power loss arising from voltage barriers. Ohmic contacts are commonly used to connect semiconductor wafers and chips with external connecting pins or circuitry of a system.

11.4 TUNNEL DIODE

In this next section we shall discuss about the tunnel diode which exhibits an interesting phenomena of negative resistance in its forward bias condition. Let us first see, what is negative resistance? In any conventional resistor when a voltage is applied it restricts the flow of current also the current increases linearly with linear increase in the applied voltage. Moreover, power is dissipated in the resistor given by I^2R, where I is the current flowing through the resistor and R is the value of the resistor used. This in general is referred to as positive resistance. But, if with the increase in the applied voltage the current flowing through the specimen decreases or with the decrease in voltage the current increases, then it can be said that the specimen is showing a negative resistance property. So if a positive resistance can be associated with the dissipation power than the negative resistance can be associated with a release of power. Now let us discuss about the tunnel diode along with its construction and working principle.

A tunnel diode, also known as an Esaki diode, is a heavily doped semiconductor diode that exhibits negative differential resistance as a result of the quantum mechanical tunneling effect. It was invented in August 1957 by Leo Esaki and Yuriko Kurose during their work at Tokyo Tsushin Kogyo, a company now known as Sony.

Construction of Tunnel diode:

A tunnel diode requires extremely heavy doping of the semiconductor material. In an intrinsic semiconductor, the concentrations of electrons and holes are equal, resulting in electrical neutrality. By introducing donor or acceptor impurities, the carrier concentration is significantly increased to produce n-type or p-type material. The doping concentration in a tunnel diode is typically of several orders higher than that of a conventional p–n junction diode. This heavy doping drastically reduces the width of the depletion region, enabling quantum mechanical tunneling. Germanium is commonly used in the fabrication of tunnel diodes; however, gallium arsenide (GaAs) is often preferred because it provides a higher peak-to-valley current ratio. Silicon is generally unsuitable due to its relatively low peak-to-valley ratio. In one construction method, a small tin dot (approximately 50 μm in diameter) is alloyed onto a heavily doped semiconductor wafer of small diameter (about 0.5 mm) to form the junction. Figure 11.10(a) shows the structure of a tunnel diode and figure 11.10(b) shows the circuit symbol of a tunnel diode.

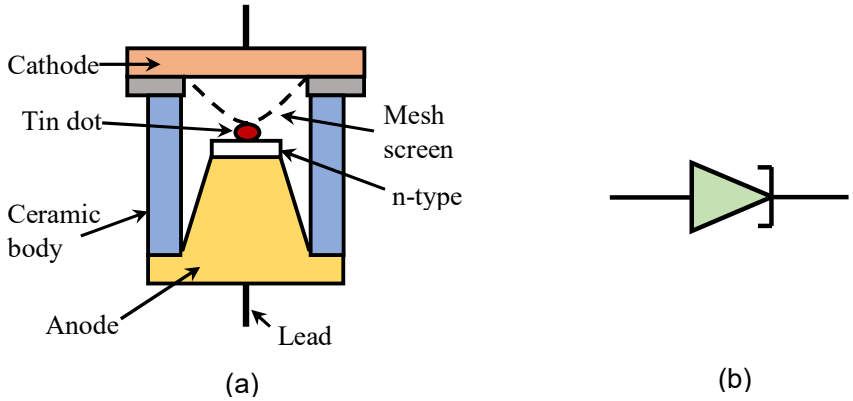
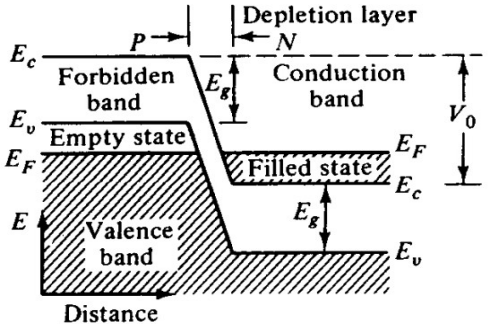


Figure 11.10: (a) Construction of Tunnel diode (b) circuit symbol of Tunnel diode

Working principle of Tunnel diode:

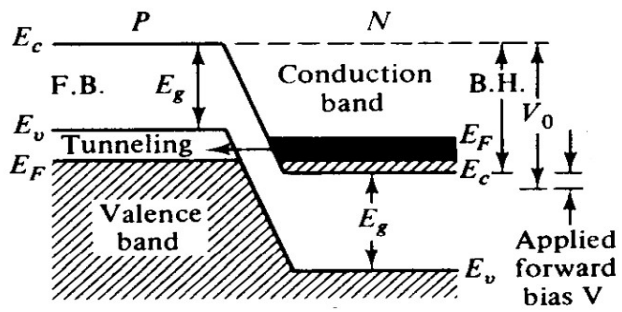
To understand the working principle of Tunnel diode we need to know about the band diagram in various biasing conditions. As we know that the tunnel diode is a heavily doped one upto 1000 times more than that of a conventional PN diode this results a different alignment of the energy bands which in contrast provides the tunnelling effect. The energy band diagram of a tunnel diode in the unbiased condition is shown in figure 11.11(a).



E_F is the Fermi level
V_O is the potential barrier of the junction
E_g is the energy required to break the covalent bond
F.B. is the forbidden band
B.H. represents the barrier height

Figure 11.11(a): Tunnel diode in the unbiased condition
(*source: Microwave Devices and Circuits by S. Y. Liao*)

Under open-circuit conditions, or at zero-bias equilibrium, the electron energy distributions in the p-type and n-type regions align such that their Fermi levels coincide. Since there are no occupied energy states on one side of the junction that correspond to available empty states at the same energy on the other side, carrier transfer across the junction cannot occur. Consequently, there is no net charge flow in either direction, and the current is zero initially. In conventional p–n junction diodes, the Fermi level lies within the forbidden energy gap. However, in a tunnel diode, the extremely heavy doping shifts the Fermi level into the valence band on the p-type side and into the conduction band on the n-type side. When a forward bias voltage say $0<V<V_P$ (V_P is the peak valley voltage [refer to the I-V characteristic curve of the tunnel diode shown in figure 12]) is applied across the tunnel diode the potential barrier is reduced by an amount equal to the applied voltage, resulting in a separation of the Fermi levels on the two sides of the junction. In this condition, occupied states in the conduction band of the n-type region align energetically with empty states in the valence band of the p-type region. Consequently, electrons tunnel through the narrowed barrier from the n-type side to the p-type side, producing a forward tunneling current across the junction. This situation can be understood via the diagram shown in figure 11.11(b).



*V*is the applied forward bias voltage

Figure 11.11(b): Forward bias of tunnel diode with voltage $0<V<V_P$
(*source: Microwave Devices and Circuits by S. Y. Liao*)

As the forward-bias voltage increases to V_p, the alignment between the filled states in the n-type region and the empty states in the p-type region becomes maximum, allowing the largest number of electrons to tunnel through the barrier. This results in the peak tunneling current, denoted by I_p [for I_p please refer to figure 12]. This situation is shown in figure 11.11(c).

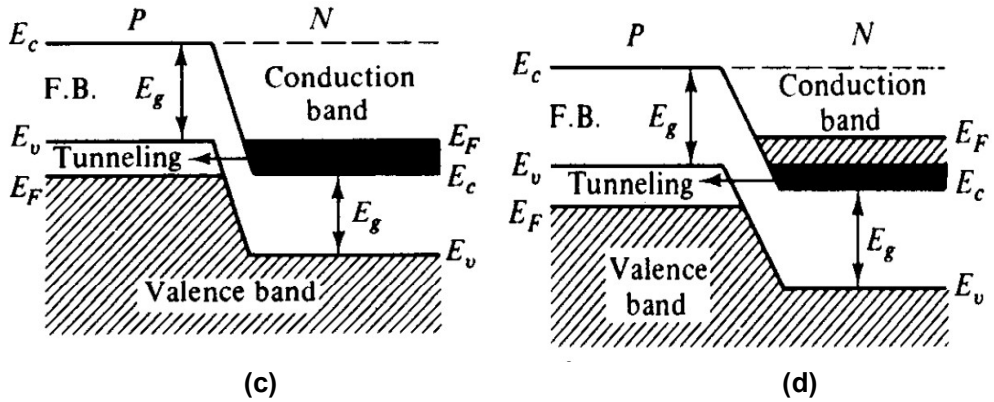


Figure 11.11: Tunnel diode with forward bias voltage (c)$V = V_p$(d)$V_p < V < V_v$
(*source: Microwave Devices and Circuits by S. Y. Liao*)

When the biasing voltage is further increased exceeding the peak voltage V_p till valley voltage V_v the tunneling current slowly decreases as the number of allowed empty states decreases due to the alignment of the bands as shown in figure 11.11(d). Further when the biasing voltage reached the valley point (V_v), there are no available empty states in the p-type region that correspond in energy to the filled states in the n-type region. As a result, electron tunneling across the junction is no longer possible, and the tunneling current decreases to zero. This situation can be realized from the figure 11.11(e) in the next page. The I-V characteristic curve of a tunnel diode is shown in figure 11.12.

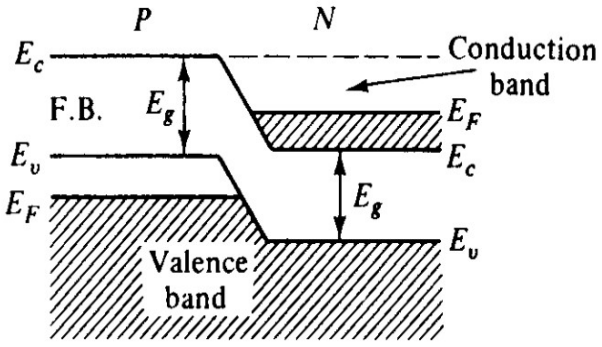


Figure 11.11(e): Tunnel diode with forward bias voltage $V_v < V < \infty$
(*source: Microwave Devices and Circuits by S. Y. Liao*)

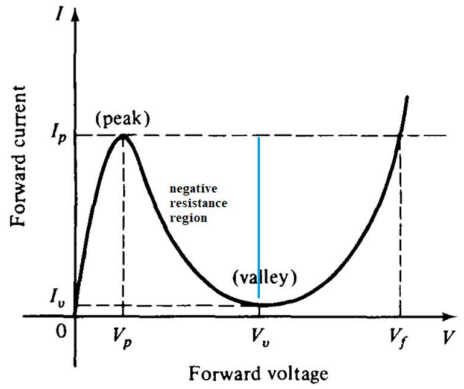


Figure 11.12: I-V characteristic curve of a tunnel diode
(*source: Microwave Devices and Circuits by S. Y. Liao*)

When the forward-bias voltage V exceeds the valley voltage V_v, the conventional carrier injection current of the PN junction begins to dominate. This injection current increases exponentially with further increases in the applied forward voltage. The total current through the tunnel diode, being the sum of the tunneling current and the conventional injection current, gives rise to its characteristic I-V behavior. The ratio of peak current to valley current, I_p/I_v, can theoretically range from 50 to 100; however, in practical devices, this ratio is typically around 15.

Advantages of Tunnel diode:

- Long operational lifetime
- High-speed performance
- Low noise characteristics
- Low power consumption

Disadvantages of Tunnel diode:

- Tunnel diodes are difficult to manufacture in large quantities.
- As a two-terminal device, the input and output are not electrically isolated

Applications of Tunnel diode:

- Microwave Oscillators: Utilized for generating high-frequency, low-power signals. They can produce oscillations up to 100 GHz.
- Reflection Amplifiers: Tunnel diodes act as low-noise amplifiers in microwave receivers, often employing a circulator-coupled design to achieve gain at frequencies such as 6000 MHz.
- Detectors/Mixers: They act as frequency converters in radar and communication receivers, converting high-frequency signals.
- High-Speed Switching: Their ultrafast switching capability makes them suitable for pulse generators, analog-to-digital converters (ADCs).

11.5 VARACTOR DIODE

A varactor diode, also called a Varicap or volt-cap, is a type of p–n junction diode that is primarily operated in reverse bias. Its distinguishing feature is a capacitance that varies with the applied reverse-bias voltage, hence the name “Varicap,” derived from

"variable" and "capacitor". Unlike standard diodes, which are used for current conduction, varactor diodes are designed to exploit their voltage-dependent capacitance. This property makes them particularly useful in oscillator circuits, voltage-controlled tuning, and other applications requiring precise capacitance control. Although their basic structure is similar to that of ordinary PN diodes, their nonlinear reactance under reverse bias is the key characteristic that enables these specialized applications.

The basic concept lies within the modulation of the width of the depletion region formed by the PN junction which is called the transition capacitance. This depletion region can be related to as like the dielectric within the plates of a capacitor whose width gets varied in accordance to the voltage applied specially in the reverse bias connection. The circuit symbol of varactor is shown in figure 11.13 (a) whereas the non-linear response of the capacitance and applied voltage is shown in figure 11.13(b).

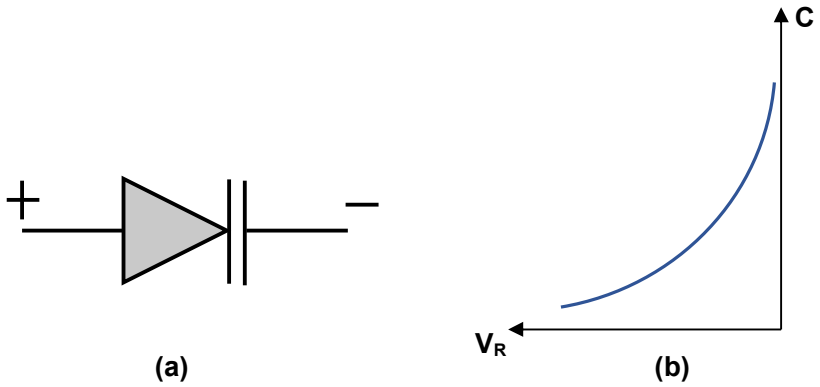


Figure 11.13(a): Circuit symbol of varactor diode (b) varactor diode capacitance v/s applied voltage

The transition capacitance $\mathbf{C_T}$ is given by,

$$C_T = \varepsilon A / W$$

where ε is the dielectric constant, A is the plate area of the capacitor and W is the depletion layer width.

The above relation indicates that the transition capacitance is inversely proportional to the width of the depletion region. Therefore, to obtain a large capacitance, the depletion width must be small. Since the depletion width decreases with lower reverse-bias voltage, a small reverse voltage results in a larger transition capacitance. Thus it can be simplified to obtain:

$$C_T(V_R) = \frac{C_o}{\left(1 + V_R / V_{bi}\right)^n} \qquad or \qquad C_T = \frac{K}{(V_R + V_{bi})^m}$$

Where,

$\boldsymbol{C_o}$ is the junction Capacitance at zero bias
$\boldsymbol{V_R}$ is the applied reverse bias voltage

V_{bi} is the built-in potential/barrier voltage
n or ***m*** is the gradient coefficient (or doping profile index)

The quality factor of the varactor diode is given by,

$$Q = F/f$$

Where ***Q*** is the quality factor, ***F*** is the maximum operating frequency, ***f*** is the operating frequency.

Applications of varactor diode:

- **Electronic tuning:** Used in place of mechanical variable capacitors in radio and television receivers for frequency selection.
- **Voltage-controlled oscillators (VCOs):** Essential for generating variable frequencies in communication systems, phase-locked loops (PLLs), and radar equipment.
- **RF filters and resonant (tank) circuits:** Employed in resonant networks and tracking filters to adjust the operating frequency or bandwidth through a control voltage.
- **Frequency modulation:** Utilized in transmitters to achieve frequency modulation by varying capacitance in response to the modulating signal.
- **Frequency multiplication:** Applied in microwave, radar, and sensing systems to produce higher-order harmonics from a lower-frequency input signal.
- **Parametric amplification:** Used in low-noise, high-frequency amplification systems where voltage-dependent capacitance enables signal gain.

11.6 PIN DIODE

Now let us discuss another kind of diode which is called as the PIN diode, as the name said P-I-N, it has a wide, intrinsic semiconductor layer positioned between the p-type and n-type regions. The p-type and n-type regions are usually heavily doped to ensure low-resistance (ohmic) contacts. Some of the features of PIN diode are listed below:

- The capacitance of a PIN diode is largely independent of the applied bias voltage because the intrinsic layer contains very little net charge.
- It has a very short reverse recovery time.
- For low-frequency signals, the diode follows the standard diode equation.
- It behaves more like a linear resistor than a typical non-linear device, resulting in minimal distortion or rectification.

Construction of PIN diode:
The structure of a PIN diode is similar to a conventional p–n diode, with the key difference being the inclusion of an intrinsic layer between the p-type and n-type regions. This intrinsic layer is usually made from high-purity silicon or gallium arsenide. In a PIN diode, the p-type and n-type regions are heavily doped, while the intrinsic region is lightly doped or nearly undoped. This configuration results in a wide depletion region and a large potential barrier between the p-type and n-type regions. Figure 11.14 shows the construction of a PIN diode along with the circuit symbol.

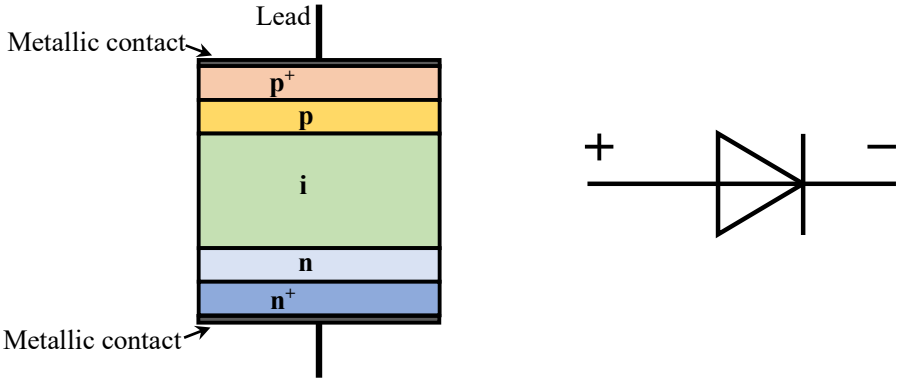


Figure 11.14: Construction and circuit symbol of PIN diode

Working of PIN diode:

The working principle of PIN diode can be understood from the following diagram as shown in figure 11.15. The operating principle of a PIN diode is based on its three-layer semiconductor structure, consisting of a p-type region, an intrinsic (undoped or lightly doped) region, and an n-type region. The intrinsic layer plays a fundamental role in the device's operation. Under the influence of an applied voltage, this layer exhibits variable resistance characteristics, thereby controlling the current flow through the diode. Under forward-bias conditions, a positive potential is applied to the p-type region while a negative potential is applied to the n-type region. This biasing causes the injection of charge carriers (holes) from the p-type region and electrons from the n-type region into the intrinsic layer. The resulting increase in carrier concentration within the intrinsic region significantly reduces its resistivity and narrows the effective depletion width.

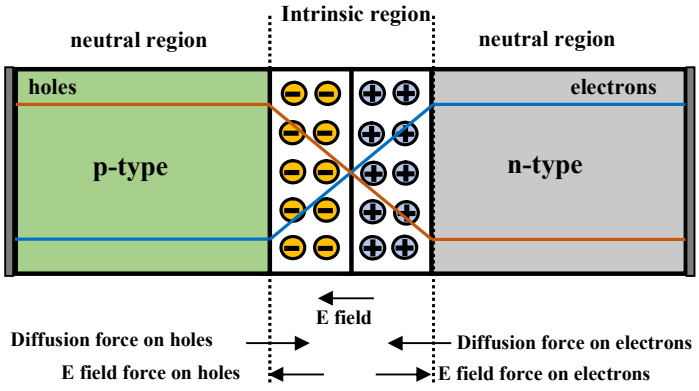


Figure 11.15: Diffusion of electrons and hole into the intrinsic region of PIN diode

Consequently, the PIN diode exhibits low series resistance and functions effectively as a low-resistance switch.

Under reverse-bias conditions, a positive potential is applied to the n-type region and a negative potential to the p-type region. This biasing increases the width of the depletion region, particularly across the intrinsic layer, thereby significantly reducing the concentration of free charge carriers. As a result, the conductivity of the device decreases, and the PIN diode behaves as a high-resistance element, effectively restricting current flow. The ability of the PIN diode to transition between low-resistance

(forward-biased) and high-resistance (reverse-biased) states under varying bias conditions is fundamental to its operation. By modulating the conductivity of the intrinsic region through the application of an external voltage, the device enables precise control of resistance. This characteristic makes the PIN diode particularly suitable for applications requiring dynamic impedance control, such as RF switching and variable attenuation in high-frequency electronic systems.

Applications of PIN diode:

The following are few applications of PIN diode—

- **RF switching:** PIN diodes are widely used in radio-frequency (RF) switching applications. Their fast switching speed, low insertion loss, and high-frequency capability make them well suited for communication systems, radar equipment, and other wireless technologies where efficient and precise signal control is essential.
- **Voltage controlled attenuators:** Due to their voltage-dependent resistance characteristics, PIN diodes are commonly employed in voltage-controlled attenuators. In communication systems, they enable dynamic control of signal amplitude by varying the applied bias voltage, allowing accurate adjustment and modulation of signal strength.
- **Switched-Mode Power Supplies (SMPS):** PIN diodes are utilized in switched-mode power supplies to facilitate efficient high-frequency switching in converters and inverters. Their ability to operate at high frequencies with rapid response times contributes to improved performance and overall system efficiency.
- **RF Phase Shifters:** In RF phase-shifting circuits, PIN diodes are used to control the phase of transmitted signals. This function is critical in radar systems, phased-array antennas, and communication networks, where precise phase adjustment is required for proper signal transmission and reception.
- **Variable RF Attenuators:** PIN diodes are also used in variable RF attenuators, where controlled variation of signal power is necessary. Such attenuators are important in test and measurement equipment, communication systems, and satellite communication applications, where accurate signal level management is required.

Advantages of PIN diode:

- Fast Switching Speed: Enables quick response in high-frequency and power electronic applications.
- Low Reverse Recovery Time: Minimizes power loss during switching transitions, improving efficiency.
- Variable Resistance: Allows dynamic control of resistance through applied voltage.
- High Power Handling Capability: Suitable for high-power applications such as amplifiers.
- Wide Range of Applications: Used in RF switching, attenuators, SMPS, and phase shifters, making it highly versatile.

Disadvantages of PIN diode:

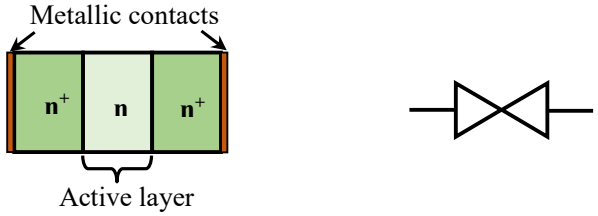
- High Reverse Leakage Current: May exhibit higher leakage current under reverse bias, leading to power losses.

- Limited Extremely High-Frequency Performance: Performance can degrade at very high frequencies.
- Temperature Sensitivity: Characteristics may vary with temperature, requiring thermal management.
- Manufacturing Complexity: Multi-layer structure and precise doping increase fabrication complexity and cost.
- Less Suitable for Low-Frequency Applications: Not as efficient in low-frequency operations compared to other diode types.

11.7 GUNN DIODE

In a simple way a Gunn diode can be defined as a transferred electron device (TED), which is a two-terminal semiconductor device that exhibits negative differential resistance. It is primarily used in high-frequency and microwave electronic applications. The operation of the Gunn diode is based on the Gunn effect, which was discovered in 1962 by physicist J. B. Gunn. Unlike the PN diode the Gunn diode consist of only n type material which consist of multiple, initially-empty, closely-spaced energy valleys in their conduction band like Gallium Arsenide (GaAs), Indium Phosphide (InP), Gallium Nitride (GaN), Cadmium Telluride (CdTe), Cadmium Sulfide (CdS), Indium Arsenide (InAs), Indium Antimonide (InSb) and Zinc Selenide (ZnSe).

Construction and working principle of Gunn diode:
The general fabrication process of a Gunn diode involves the growth of an epitaxial layer on a highly doped (degenerate) n^+ substrate, resulting in a three-layer n-type semiconductor structure. In this configuration, the outer regions are heavily doped to provide low-resistance contacts, while the intermediate layer is lightly doped and serves as the active region responsible for carrier transport and microwave generation. Ohmic metal contacts are formed at both ends of the structure to enable proper biasing of the device. The internal structure of Gunn diode and circuit symbol is shown in figure 11.16. The circuit symbol of the Gunn diode differs from that of a conventional p–n junction diode, reflecting the absence of a p–n junction in its structure.



(a) (b)

Figure 11.16: (a) internal structure of Gunn diode (b) circuit symbol of Gunn diode

I-V characteristics curve of a Gunn diode showing different region is shown in figure 11.17. Up to the applied threshold voltage (V_{th}), the Gunn diode exhibits a linear, ohmic response, characterized by a direct proportionality between current and voltage. As the applied voltage exceeds this threshold, the device enters a regime in which the current decreases with increasing voltage, reaching a minimum at the valley point (V_V). This region of the current–voltage characteristic corresponds to the manifestation of negative differential resistance.

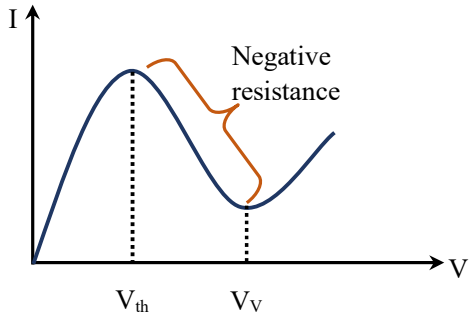
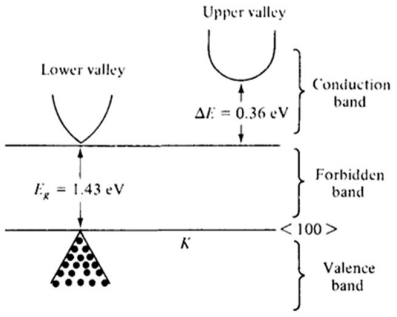


Figure 11.17: I-V characteristics of Gunn diode

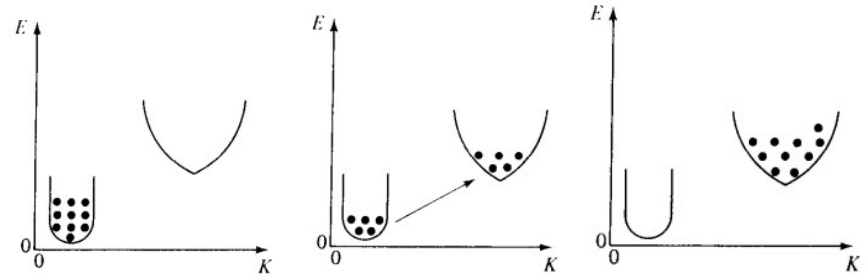
Upon further increase in the applied voltage beyond the valley region, the device resumes normal ohmic behavior, with current once again increasing proportionally with voltage. This phenomenon of achieving negative resistance in Gunn diode can be understood properly by the two valley model theory-

According to the energy band theory of n-type GaAs, the conduction band comprises a high-mobility lower valley that is energetically separated by approximately 0.36 eV from a higher-energy, low-mobility upper valley as shown in figure 11.18(a). When the applied electric field is less than the critical field associated with the lower valley ($E < E_l$), inter-valley transfer does not occur, and electrons remain confined to the lower valley as shown in figure 11.18(b). However, when the applied electric field exceeds the lower valley threshold but remains below that of the upper valley ($E_l < E < E_u$), electrons begin to transfer from the lower valley to the upper valley as shown in figure 11.18(c). And when the applied electric field is higher than that of the upper valley ($E_u < E$), all electrons will transfer to the upper valley as shown in figure 11.18(d).



(a)

Figure 11.18(a): Two valley model of electron energy versus wave number of n-type GaAs
(*source: Microwave devices and circuits by S. Y. Liao*)



(b)$E<E_l$ (c)$E_l<E<E_U$ (d)$E_U<E$

Figure 11.18(b, c & d): Electron transfer from lower to higher valley at different applied field (*source: Microwave devices and circuits by S. Y. Liao*)

The conductivity of the n-type GaAs with electron densities nl and nu in the lower valley and upper valley is given by,

$$\sigma = e\left(\mu_l n_l + \mu_u n_u\right) \qquad (i)$$

Where,

e = charge of an electron
μ = electron mobility
n = $n_l + n_u$ is the electron density

When a sufficiently high field ***E*** is applied to the specimen, electrons are accelerated and their effective temperature rises above the lattice temperature. Furthermore, the lattice temperature also increases. Thus electron density ***n*** and mobility ***μ*** are both functions of electric field ***E***. Differentiating equation (***i***) with respect to ***E*** we have,

$$\frac{d\sigma}{dE} = e\left(\mu_l \frac{dn_l}{dE} + \mu_u \frac{dn_u}{dE}\right) + e\left(n_l \frac{d\mu_l}{dE} + n_u \frac{d\mu_u}{dE}\right) \qquad (ii)$$

If the total electron density is given by $n = n_l + n_u$ nad it is assumed that μ_l and μ_u are proportional to E^P, where p is a constant, then

$$\frac{d}{dE}(n_l + n_u) = \frac{dn}{dE} = 0$$

$$\Rightarrow \frac{dn_l}{dE} = -\frac{dn_u}{dE} \qquad (iii)$$

and

$$\frac{d\mu}{dE} \propto \frac{dE^p}{dE} = pE^{p-1} = p\frac{E^p}{E} \propto p\frac{\mu}{E} = \mu\frac{p}{E} \qquad (iv)$$

Substituting (***iii***) and (***iv***) in equation (***ii***) we have,

$$\frac{d\sigma}{dE} = e\left(\mu_l - \mu_u\right)\frac{dn_l}{dE} + e\left(n_l\mu_l + n_u\mu_u\right)\frac{p}{E} \qquad (v)$$

Now, the differentiation of Ohm's law $J = \sigma E$ with respect to ***E*** will give us,

$$\frac{dJ}{dE} = \sigma + \frac{d\sigma}{dE}E$$

$$\Rightarrow \frac{1}{\sigma}\frac{dJ}{dE} = 1 + \frac{d\sigma/dE}{\sigma/E} \qquad (vi)$$

It is clear that for negative resistance, the current density ***J*** must decrease with the increasing filed ***E*** or the ratio of ***dJ/dE*** must be negative. This would be the case only when the right hand term of equation (***vi***) is less than zero, i.e.

$$-\frac{d\sigma/dE}{\sigma/E} > 1 \qquad (vii)$$

Substituting equation (***i***) and (***v***) with $g = n_u/n_l$ in (***vii***) we will have,

$$\left[\left(\frac{\mu_l - \mu_u}{\mu_l + \mu_u g}\right)\left(-\frac{E}{n_l}\frac{dn_l}{dE}\right) - p\right] > 1 \qquad (viii)$$

When lattice scattering is dominant, however, ***p*** is negative and must depend on the lattice and carrier temperature. The first bracket in equation (***viii***) must be positive in order to satisfy unequality. This means that $\mu_l > \mu_u$. Electrons must begin in a low-mass valley and transfer to a high-mass valley when they are heated by the electric field. The maximum value of this term is unity, that is, $\mu_l >> \mu_u$. The factor dn_l/dE in the second bracket must be negative. This quantity represents the rate of the carrier density with field at which electrons transfer to the upper valley; this rate depends on differences between electron densities, electron temperature, and gap energies in the two valleys.

*The frequency (***f***) of a Gunn diode, operating in the transit-time mode, is primarily determined by the velocity of the charge carriers (v_d) and the length of the drift region (***L***). The fundamental frequency is given by $f = v_d / L$; where vd is approximately 10^7 cm/s and L is the device length.

Applications of Gunn diode:

Due to the advantages of the Gunn diode like high frequency stability, low noise, compact and robustness, it can be used in different applications that are listed below:

- **Microwave Oscillators and Sources:** Used as Transfer Electron Oscillators (TEO) for generating microwave signals in radar, radio communication systems, and signal generators.
- **Radar Technology:** Found in CW Doppler radar systems, police radar guns, traffic monitoring sensors, and airborne collision avoidance systems.
- **Sensors and Security:** Used for automatic door openers, security systems, and intrusion detectors due to their ability to detect movement.
- **Communication and Data Links:** Employed in microwave transmitters and satellite communication links for transmitting data.
- **Industrial and Measuring Devices:** Utilized in tachometers for speed measurement, moisture content monitors, and remote vibration detectors.
- **Amplifiers:** Act as Transfer Electron Amplifiers (TEA) in microwave receivers.

11.8 IMPATT DIODE

This IMPATT diode comes under the classification of Avalanche Transit Time devices. It stands for Impact ionization Avalanche Transit Time diode. It is a high-power semiconductor diode designed for microwave applications operating at frequencies ranging from a few gigahertz to several hundred gigahertz (3 to 100 GHz and above). Because it exhibits negative resistance, the IMPATT diode is inherently suited for use as an oscillator in high-frequency signal generation. A major advantage of IMPATT diodes compared to other microwave diodes is their relatively high output power capability. Historically, the IMPATT diode is also known as the “Read diode,” named after W. T. Read, who in 1958 first proposed the p^+–n–i–n^+ structure based on carrier injection through a forward-biased PN junction.

Physical structure of IMPATT diode:

The physical structure of an IMPATT diode (based on Read diode) has been shown in figure 11.19 along with its doping concentration profile. The configuration of the diode

can be either n^+-p-i-p^+ or p^+-n-i-n^+. Its basic physical mechanism is the interaction of the impact ionization avalanche and the transit time of charge carriers.

The construction of an IMPATT diode can be summarised as:

- **Semiconductor Material:** IMPATT diodes are typically fabricated using silicon (Si) or gallium arsenide (GaAs), owing to their favorable electronic properties such as high breakdown field strength, suitable carrier mobility, and well-characterized avalanche behavior.
- **P–N Junction:** The fundamental structural element of the IMPATT diode is the p–n junction, formed by adjoining p-type and n-type semiconductor regions. Under reverse-bias conditions, this junction constitutes the active region in which avalanche breakdown occurs, initiating carrier multiplication necessary for microwave generation.

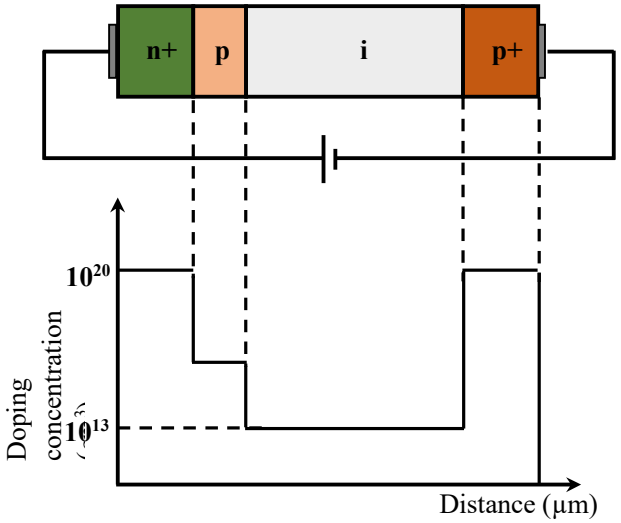


Figure 11.19 Physical structure of IMPATT diode along with doping profile

- **Depletion Region:** Application of a reverse-bias voltage across the p–n junction gives rise to a depletion region devoid of mobile charge carriers. This space-charge region sustains a high electric field, which suppresses normal conduction and establishes the conditions required for avalanche ionization.
- **Transit-Time Region:** Adjacent to the avalanche region lies the drift or transit-time region, which is engineered to control the carrier transit time. The physical dimensions of this region are precisely designed so that the generated charge carriers (primarily electrons) require a specific time interval to traverse the depletion layer. This transit-time delay, in conjunction with avalanche multiplication, produces the phase shift responsible for negative resistance at microwave frequencies.
- **Metal Contacts:** Ohmic metal contacts are formed on the p-type and n-type regions to permit the application of reverse-bias voltage and the extraction of the generated microwave power. These contacts are typically fabricated from highly conductive metals such as aluminium or gold to minimize contact resistance and power loss.
- **Packaging:** The device is encapsulated within a protective package that shields it from environmental influences such as moisture, mechanical stress,

and contamination. The packaging is also designed to facilitate impedance matching and integration into microwave circuits.

- **Heat Dissipation:** Given the substantial power dissipation associated with avalanche processes, thermal management is a critical design consideration. IMPATT diodes are often mounted on heat sinks or thermally conductive substrates to ensure efficient heat removal and maintain stable operation under high-power conditions.

Operation of IMPATT diode:

An extremely high electric field of the order of 400 kV/cm, is applied across the IMPATT diode, giving rise to a correspondingly large current density. Under such conditions, a conventional p–n junction diode would undergo irreversible breakdown. However, the IMPATT diode is specifically engineered to sustain repeated operation in the avalanche breakdown regime without degradation. A simple operational concept of an IMPATT diode is shown in figure 11.20. The intense reverse-bias potential gradient across the junction initiates avalanche multiplication, resulting in a substantial flow of minority carriers across the depletion region. Assuming that oscillations are present, let us consider the effect of a positive RF voltage swing superimposed on the large DC reverse bias. The instantaneous increase in voltage further strengthens the electric field across the junction. Under these conditions, the carrier velocities become sufficiently high that electrons and holes acquire enough kinetic energy to liberate additional carriers from the crystal lattice. This process, known as impact ionization, results in the generation of new electron–hole pairs and sustains the avalanche mechanism. The operation can be understood in two steps:

Firstly, these additional carriers continue the process at the junction and the voltage will be exceeded during the whole of the +ve RF cycle. The avalanche current multiplication will be taking place during this entire time. Since avalanche is a multiplication process, it is not instantaneous or we can say it is a cumulative process. This process takes time such that the current pulse maximum, at the junction, occurs at the instant when the RF voltage across the diode is zero and going negative. A 90^0 phase difference between voltage and current has been obtained. Because avalanche multiplication is inherently cumulative rather than instantaneous, it requires a finite buildup time. As a result of this delay, the peak of the avalanche current pulse at the junction does not coincide with the peak of the RF voltage. Instead, the maximum current occurs at the instant when the RF voltage crosses zero and begins to swing in the negative direction as can be seen from figure 11.20. This temporal displacement gives rise to an approximate 90° phase difference between the RF voltage and the resulting current.

Secondly, the generated current pulse initially appears in the avalanche region at the p–n junction. Owing to the applied reverse bias, this pulse does not remain localized but drifts toward the cathode under the influence of the strong DC electric field.

The drift velocity of the carriers is determined by the magnitude of this electric field and the transport properties of the semiconductor material. The time required for the current pulse to reach the cathode depends on both the carrier drift velocity and the thickness of the drift region, typically the lightly doped or intrinsic layer adjacent to the junction. The device is designed such that the transit time of the carriers through this drift space

introduces an additional 90° phase delay. Consequently, when combined with the avalanche buildup delay, the total phase shift between the RF voltage and the terminal current approaches 180°, thereby establishing the negative resistance characteristic essential for microwave oscillation.

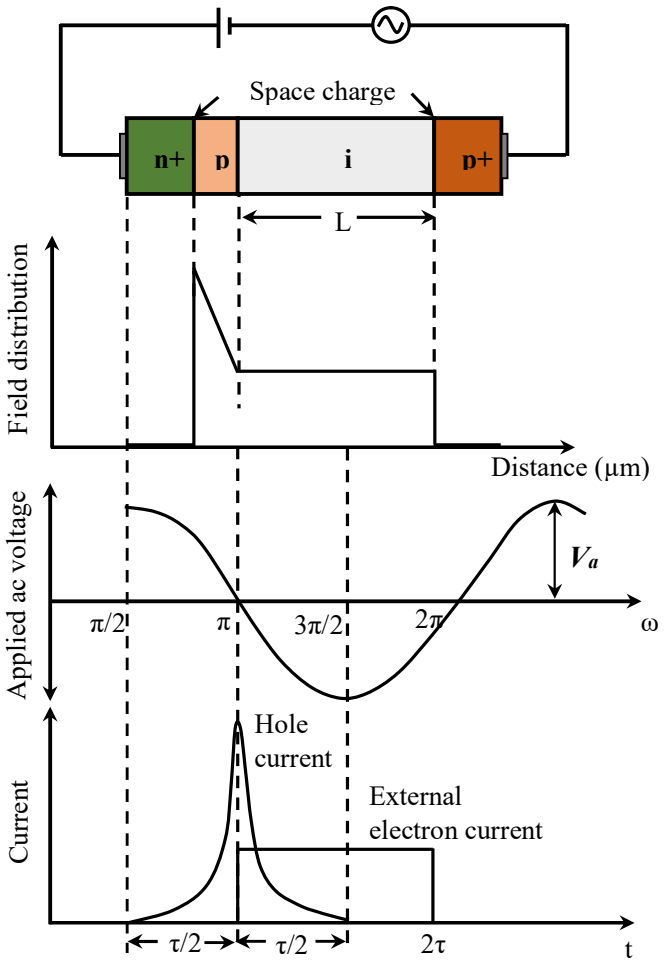


Figure 11.20**:** Field voltage and current of an IMPATT diode

In summary, negative resistance phenomenon is taken into account by using

- The impact multiplication avalanche effect, which causes the minority current to lag the microwave output voltage by 90°. phase shift
- The effect of transit time through the drift region, this result in the external current lagging the microwave voltage by a further 90° phase shift.

However, the resonant frequency at which the IMPATT diode work is given by:

$$2\pi f = \frac{\pi}{\tau} \qquad (i)$$

Then,

$$f = \frac{1}{2\tau} = \frac{v_d}{2L} \qquad (ii)$$

Where v_d is the drift velocity of holes and L is the length of the drift '*i*' region.

Applications of IMPATT diode:

- **Microwave Oscillators and Generators:** Provide high-power microwave signals in the GHz range for compact and reliable source applications.

- **Radar Systems (CW Doppler and Pulsed):** Used in speed detection, target tracking, and imaging systems due to their high power and robustness.
- **Telecommunication Systems:** Serve as microwave sources and amplifiers in transmitters and receivers for high-capacity communication links.
- **Electronic Warfare (EW):** Applied in jamming and electronic countermeasure equipment to interfere with hostile radar and communication systems.
- **Parametric Amplifiers:** Function as negative-resistance devices in microwave amplification circuits.
- **Intruder Alarm Systems:** Employed in motion sensors and security detection devices.
- **Scientific and Industrial Applications:** Operate as stable microwave sources in EPR spectroscopy, terahertz imaging, and material analysis.

Limitations of IMPATT diode:

- **High Noise Generation:** The fundamental physical process of avalanche multiplication produces substantial AM/FM noise. They are considered much noisier than Gunn or Klystron diodes.
- **Low Efficiency:** IMPATT diodes typically offer low power-added efficiency (often less than 10-15% for CW), which is generally lower than TRAPATT diodes.
- **Limited Tuning Range:** They offer a narrow frequency tuning range compared to other devices like Gunn diodes.
- **High Operating Voltage/Sensitive Sensitivity:** They require high DC supply voltages and are extremely sensitive to temperature and operating conditions.
- **Thermal Management Issues:** Due to high current density and high power dissipation, improper thermal management can cause the diode to fail.
- **Spurious Oscillations:** These devices tend to produce unwanted harmonics (spurious noise) at higher frequencies.

W CHECK YOUR PROGRESS

Answer the following multiple-choice questions

Q1. The junction capacitance of a Schottky diode is generally smaller than a p–n junction diode because:

A) It has a narrower depletion region due to high doping
B) It has no minority carrier charge storage
C) It operates under forward bias only
D) It has higher series resistance

Q2. Tunnel diodes are suitable for microwave oscillators because:

A) They have very low forward voltage drop
B) They exhibit fast response due to minority carrier injection
C) They have extremely fast tunneling-based negative resistance with negligible charge storage
D) They operate only at low frequencies

Q 3. The negative differential resistance in a Gunn diode arises due to:

A) Avalanche multiplication in a p–n junction
B) Quantum tunneling across a narrow depletion region

C) Transfer of electrons to a higher-energy satellite valley with lower mobility
D) Minority carrier injection across a junction

Q4. In a reverse-biased PIN diode used as an RF switch, the device primarily behaves as:
A) A low-resistance path
B) A high-resistance path with small junction capacitance
C) A voltage regulator
D) A frequency multiplier

Q5. Compared to Gunn diodes, IMPATT diodes generally exhibit:
A) Lower noise levels
B) Higher efficiency and lower power
C) Higher power output but higher noise
D) No negative resistance

11.9 LET US SUM UP

- Metal–semiconductor junctions form Schottky (rectifying) or ohmic (non-rectifying) contacts, while heterojunctions use two different semiconductors for improved device performance.
- Schottky barrier diode is a metal–semiconductor diode with majority carrier conduction, giving fast switching and low forward voltage drop.
- Ohmic contact is a metal–semiconductor junction that provides a linear, bidirectional current–voltage relationship with negligible barrier.
- Tunnel, varactor, PIN, and Gunn diodes are special devices used for high-speed switching, voltage tuning, RF applications, and microwave generation.
- Gunn and IMPATT diodes are microwave devices used for high-frequency oscillators based on transferred electron effect and avalanche breakdown respectively.

11.10 ANSWERS TO CHECK YOUR PROGRESS

Answer to Q 1. B (*Explanation:* Schottky diodes are majority-carrier devices, so there is negligible charge storage, giving very low junction capacitance and fast switching speed.)

Answer to Q 2. C (*Explanation:* Tunnel diodes are majority-carrier devices with tunneling current, so they can switch extremely fast, making them ideal for microwave frequency oscillators.)

Answer to Q 3. C (*Explanation:* Under a high electric field, electrons transfer from the central valley to a satellite valley of the conduction band where mobility is lower, causing a decrease in current with increasing voltage (negative differential resistance).)

Answer to Q 4. B (*Explanation:* Under reverse bias, the intrinsic region is depleted, resulting in high resistance and low capacitance, making it suitable for RF isolation.)

Answer to Q 5. C (*Explanation:* IMPATT diodes can generate higher microwave power but suffer from higher noise due to the stochastic nature of avalanche multiplication.)

O 11.11 FURTHER READING

- Liao S. Y., "*Microwave Devices and Circuits",* Publisher-Prentice Hall.
- George D. Vendelin, Anthony M. Pavio & Ulrich L. Rohde, *"Microwave Semiconductor Devices",* Publisher-Wiley-Interscience.
- Raymond S. Pengelly, "*Microwave Field-Effect Transistors: Theory, design and applications (3rd Edition)"*, Publisher- IET.
- Bahl J. Inder & Bhartia Prakash, *"Microwave Solid State Circuit Design",* Publisher-Wiley-Interscience.

M 11.12 MODEL QUESTIONS

Q1. How is a Schottky diode constructed?
Q2. Explain the construction and working principle of a PIN diode. Discuss the role of the intrinsic layer in determining its high-frequency characteristics.
Q3. Explain the structure and working principle of a varactor diode. How does the reverse bias voltage affect the junction capacitance?
Q4. How does a PIN diode behave under forward bias at high frequencies?
Q5. Why varactor diodes are used in voltage-controlled oscillators (VCOs)?
Q6. Discuss the factors affecting the oscillation frequency of a Gunn diode, including active region length, material properties, and applied electric field.
Q7. Explain the mechanism of negative resistance in an IMPATT diode.
Q8. Why are tunnel diodes suitable for microwave circuits?

*** ***** ***

UNIT 12: POWER SEMICONDUCTOR DEVICES

UNIT STRUCTURE

12.1 Learning Objectives
12.2 Introduction
12.3 Thyristors: Construction, Characteristics, Applications
12.4 TRIAC and DIAC: Construction, Characteristics, Applications
12.5 Power MOSFET and IGBT: Construction, characteristics, Applications.
12.6 Let Us Sum Up
12.7 Answers to Check Your Progress
12.8 Further Reading
12.9 Model Questions

12.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- identify different types of power electronic devices with the respective symbols
- describe the construction and operating principles of thyristors, TRIACs, DIACs, Power MOSFETs, and IGBTs
- discuss the characteristics of devices such as MOSFET, and IGBT
- explain the practical applications of thyristors, TRIACs, DIACs, Power MOSFETs, and IGBTs in power electronic circuits.

12.2 INTRODUCTION

In the previous unit, you learned about various microwave devices used for high-frequency applications. In this unit, you will focus on power semiconductor devices, which form the backbone of modern power electronic systems. Unlike conventional semiconductor devices used primarily for signal processing, power semiconductor devices are designed to handle high voltages, high currents, and large power levels with improved efficiency and reliability. The unit begins with the construction, working principles, characteristics, and applications of thyristors. You will then study TRIACs and DIACs, with emphasis on their construction, characteristics, and applications in AC power-control circuits. The unit concludes with advanced switching devices such as Power MOSFETs and Insulated Gate Bipolar Transistors (IGBTs). These devices offer advantages such as high switching speed, low power loss, compact size, and efficient thermal performance, making them essential components of modern, energy-efficient power electronic systems.

12.3 THYRISTORS: CONSTRUCTION, CHARACTERISTICS, APPLICATIONS

Thyristors are a class of semiconductor devices that primarily operate on the principle of internal regeneration, which enables them to remain in the conducting state once triggered. Before the development of high-power transistors, thyristors were widely used in power conversion applications due to their excellent power-handling capability

and reliability. The thyristor family consists of several devices designed for different power control applications. Some commonly used members of the thyristor family include:

- Silicon Controlled Rectifier (SCR)
- Gate Turn-Off Thyristor (GTO)
- Reverse Conducting Thyristor (RCT)
- Light Activated Silicon Controlled Rectifier (LASCR)
- MOS-Controlled Thyristor (MCT)
- TRIAC

Among these devices, the Silicon Controlled Rectifier (SCR) is the most widely used thyristor. It is extensively employed in various power converters, particularly in controlled rectifiers (AC–DC converters), as indicated by its name. Although thyristors have relatively lower switching speeds compared to power transistors, they offer very high voltage and current handling capabilities. Additionally, thyristors require comparatively simple gate drive circuits, making them highly suitable for high-power industrial applications.

CONSTRUCTION OF SCR:

The Silicon Controlled Rectifier (SCR) is a four-layer, three-junction semiconductor device. Figure 12.1 shows the symbolic representation of an SCR. It consists of three terminals: Anode (A), Cathode (K), and Gate (G). The SCR is turned ON by applying a small positive voltage between the gate and cathode terminals.

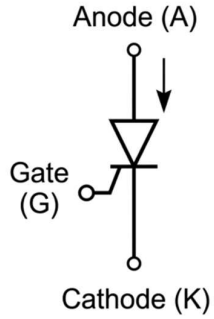


Figure 12.1: SCR Symbol

Structurally, the SCR is composed of alternating p-type and n-type semiconductor layers (p–n–p–n) as shown in Figure 12.2. The outer p^+ layer, connected to the anode, is heavily doped with a concentration of approximately $\mathbf{10^{19}}$ $\mathbf{cm^{-3}}$. The adjacent n^- layer is lightly doped, typically around $\mathbf{10^{14}}$ $\mathbf{cm^{-3}}$, with a thickness ranging from 50 to 1000 μm. This lightly doped region plays an important role in supporting the depletion region of junction $\boldsymbol{J_2}$ and determines the forward blocking capability of the SCR. The next p-layer, having a doping concentration of about $\mathbf{10^{17}}$ $\mathbf{cm^{-3}}$ and thickness between 30 to 100 μm, forms the gate region. Finally, the heavily doped n^+ layer ($\mathbf{10^{19}}$ $\mathbf{cm^{-3}}$) constitutes the cathode terminal.

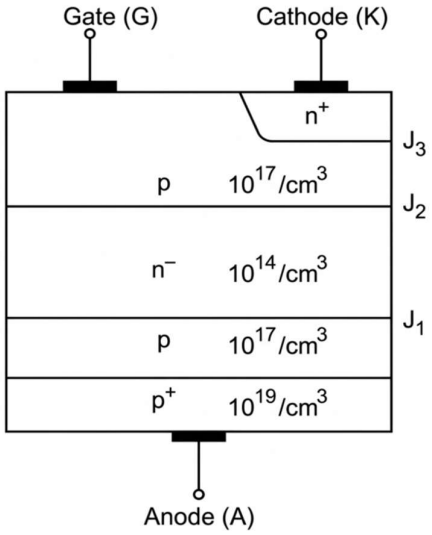


Figure 12.2: Detailed structure of SCR

When the SCR is forward biased (i.e., anode positive with respect to cathode), junctions J_1 and J_3 become forward biased, while junction J_2 remains reverse biased, preventing current flow. Upon application of a gate signal, the gate–cathode junction (J_3) becomes forward biased, allowing charge carriers to flow. These carriers reduce the depletion region at junction J_2, initiating internal regenerative action. As a result, the SCR switches to the conducting state and remains ON even after removal of the gate signal, provided the anode current exceeds the holding current.

SCR CHARACTERISTICS AND MODES OF OPERATION:

The operation of an SCR can be explained under three distinct modes: reverse blocking mode, forward blocking mode, and forward conduction mode. Figure 12.3 shows the V–I characteristics of the SCR, commonly referred to as its static characteristics.

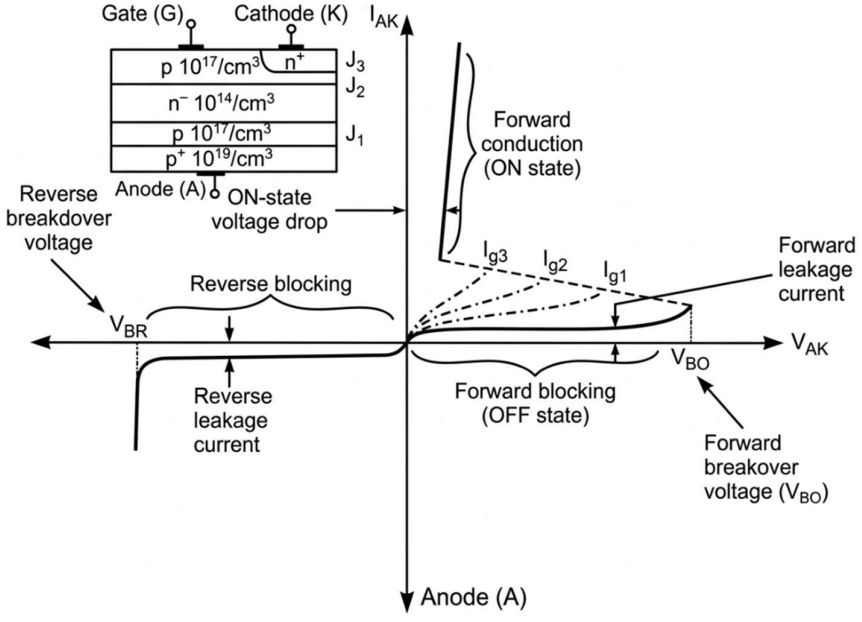


Figure 12.3: Static V-I characteristics of a SCR

In these characteristics, the anode-to-cathode current ($\boldsymbol{I_{AK}}$) is plotted against the anode-to-cathode voltage ($\boldsymbol{V_{AK}}$). The parameter $\boldsymbol{V_{BO}}$ represents the forward breakover voltage, which is the minimum forward voltage at which the SCR turns ON without gate triggering. Similarly, $\boldsymbol{V_{BR}}$ denotes the reverse breakdown voltage, beyond which the device conducts heavily under reverse bias conditions. The curves corresponding to $\boldsymbol{I_{g1}}$, $\boldsymbol{I_{g2}}$, and $\boldsymbol{I_{g3}}$ indicate different levels of gate current applied to the SCR. An increase in gate current reduces the forward breakover voltage, thereby enabling the SCR to turn ON at a lower anode-to-cathode voltage.

Reverse Blocking Mode

The following Figure 12.4 illustrates the condition under which the thyristor operates in the reverse blocking mode. In this mode, the anode (A) is made negative with respect to the cathode (K), while the gate terminal is kept open. An SCR consists of three PN junctions, namely J_1, J_2, and J_3. Under reverse bias conditions, junctions J_1 and J_3 become reverse biased, whereas junction J_2remains forward biased. Due to the overall reverse voltage applied across the device, the SCR does not conduct. Only a very small current flows from cathode to anode, known as the reverse leakage current. Since the device blocks the applied reverse voltage without significant conduction, this operating condition is referred to as the reverse blocking mode of the SCR.

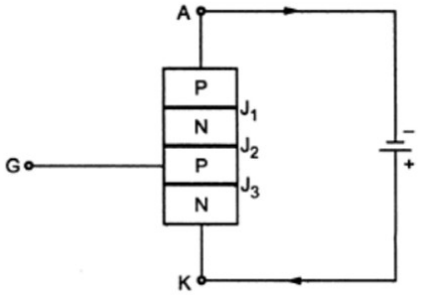


Figure 12.4: Reverse blocking mode

Forward Blocking Mode

The SCR is said to be forward biased when the anode is made positive with respect to the cathode, as shown in the following Figure 12.5. Under this condition, junctions $\boldsymbol{J_1}$ and $\boldsymbol{J_3}$ become forward biased, while junction $\boldsymbol{J_2}$ remains reverse biased. Consequently, the applied forward voltage is mainly supported by junction $\boldsymbol{J_2}$. In this state, only a small current flows from anode to cathode, known as the forward leakage current, which is typically of the order of a few milliamperes. Although the SCR is forward biased, it does not conduct significantly; therefore, this operating condition is called the forward blocking mode. During the forward blocking mode, the anode-to-cathode voltage ($\boldsymbol{V_{AK}}$) can be increased up to the forward breakover voltage ($\boldsymbol{V_{BO}}$), as illustrated in the V–I characteristics of the SCR. When the applied voltage reaches $\boldsymbol{V_{BO}}$, the SCR switches from the forward blocking mode to the forward conduction mode. In practical applications, however, the SCR is normally turned ON by applying a suitable gate signal, allowing conduction at a voltage lower than $\boldsymbol{V_{BO}}$. If the anode-to-cathode voltage exceeds $\boldsymbol{V_{BO}}$, the SCR turns ON evenwithout gate triggering. This condition does not damage the device; instead, it initiates conduction. The forward breakover voltage arises due to the blocking capability of junction $\boldsymbol{J_2}$, whereas the reverse breakover voltage is determined by the combined blocking action of junctions $\boldsymbol{J_1}$ and $\boldsymbol{J_3}$.

Since the combined blocking capability of $\boldsymbol{J_1}$ and $\boldsymbol{J_3}$ is higher than that of $\boldsymbol{J_2}$, the reverse blocking voltage of an SCR is generally greater than its forward blocking voltage.

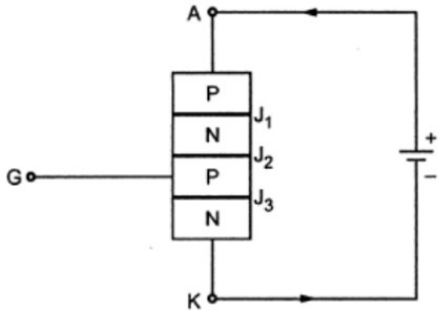


Figure 12.5: Forward blocking mode

Forward Conduction Mode (Thyristor Turn-ON / Triggering Methods)

When the SCR is forward biased, it can be turned ON (triggered) by any of the following methods:

- Forward voltage triggering ($\boldsymbol{V_{AK} > V_{BO}}$)
- Gate triggering (application of gate current)
- $\boldsymbol{dv/dt}$ triggering (rapid rate of rise of anode voltage)
- Light triggering (illumination of the gate–cathode junction)

It should be noted that the SCR can enter the forward conduction mode only if it was previously operating in the forward blocking mode.

Forward Voltage Triggering ($\boldsymbol{V_{AK} > V_{BO}}$)

The SCR enters the forward conduction mode when the anode-to-cathode voltage (V_{AK}) exceeds the forward breakover voltage (V_{BO}). At this point, the SCR is said to be turned ON. During the forward blocking mode ($V_{AK} < V_{BO}$), junction J_2remains reverse biased and prevents current flow. However, when V_{AK} exceeds V_{BO}, avalanche breakdown occurs at junction J_2, even if no gate signal is applied. As a result, a large anode current begins to flow through the device. Once conduction starts, the anode-to-cathode voltage rapidly drops to a very small value, typically less than 2 V, while a substantial current flows through the SCR. In this condition, the SCR operates in the forward conduction mode, indicating that the device has successfully turned ON.

Merits and Demerits of SCR

Category	Points
Merits of SCR	Very small gate drive power is required since SCR is a regenerative device
	SCRs are available with very high voltage and current ratings
	On-state power losses in SCR are comparatively low
Demerits of SCR	Gate has no control once the SCR is turned ON
	External commutation circuits are required to turn OFF the SCR
	Suitable only for low operating frequencies due to slow switching speed
	Snubber (RC) circuits are required for protection against high (dv/dt).

Applications of SCR:

- SCRs are widely used in controlled rectifier circuits for efficient AC–DC power conversion.
- Employed in AC voltage regulators, as well as in lighting control and industrial heating applications.
- Used in DC motor drive systems, high-power regulated power supplies, and electronic circuit breakers for reliable power control and protection.

12.4 TRIAC AND DIAC: CONSTRUCTION, CHARACTERISTICS, APPLICATIONS

TRIAC:

A TRIAC (Triode for Alternating Current) is a three-terminal AC switching device that differs from conventional silicon controlled rectifiers (SCRs) in its ability to conduct current in both directions. The device can be triggered into conduction by either a positive or negative gate signal, making it suitable for bidi rectional current flow. Due to this bidirectional conduction capability, the TRIAC is widely used as an electronic switch in AC power control applications. It is a three-terminal, four-layer, bidirectional semiconductor device designed specifically for controlling AC power. TRIACs are commercially available in a wide range of power ratings, with devices capable of handling loads up to approximately 16 kW, making them suitable for medium- and high-power AC control applications such as lighting dimmers, fan regulators, and heating systems.

Construction:

A TRIAC can be considered equivalent to two SCRs connected in inverse parallel configuration with a common gate terminal. The gate terminal is connected to both the P-type and N-type regions, allowing the device to be triggered by gate signals of either polarity (Figure 12.6). Since the TRIAC conducts current in both directions, it does not have distinct anode and cathode terminals like an SCR. Instead, it operates as a bilateral device, capable of controlling current flow during both positive and negative half cycles of an AC supply. The TRIAC consists of three terminals, namely:

- Main Terminal 1 (MT1)
- Main Terminal 2 (MT2)
- Gate Terminal (G)

This construction enables effective bidirectional switching and control of AC power using a single semiconductor device.

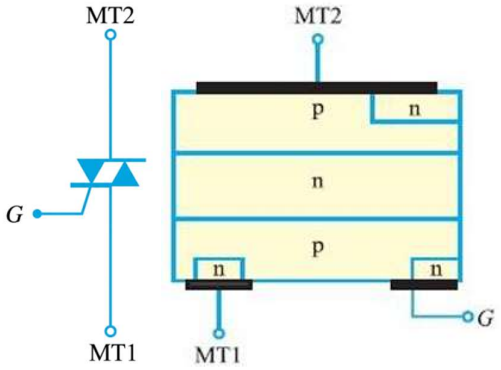


Figure 12.6: TRIAC symbol and its basic structure

Operation of TRIAC:

The TRIAC can be turned ON when the applied voltage exceeds its breakover voltage. However, in practical applications, the device is normally triggered using a gate pulse (typically about 35 μs), allowing conduction at voltages lower than the breakover value. This method is known as gate triggering, which is the most commonly used technique for TRIAC operation.

The TRIAC operates in four different triggering modes, depending on the polarity of Main Terminal 2 (MT2) and the Gate (G) with respect to Main Terminal 1 (MT1).

Mode 1: MT2 Positive and Gate Positive with respect to MT1

In this condition, current flows through the path P_1–N_1–P_2–N_2. Junctions P_1–N_1 and P_2–N_2 are forward biased, while N_1–P_2 remains reverse biased initially. A positive gate signal forward biases the gate junction, causing breakdown and turning the TRIAC ON. This mode provides high triggering sensitivity.

Mode 2: MT2 Positive and Gate Negative with respect to MT1

Here, the main current still flows through the P_1–N_1–P_2–N_2 path. The negative gate signal injects charge carriers into the P_2 region, initiating conduction. The triggering sensitivity in this mode is lower compared to Mode 1.

Mode 3: MT2 Negative and Gate Negative with respect to MT1

In this mode, current flows through the path P_2–N_1–P_1–N_4. Junctions P_2–N_1 and P_1–N_4 become forward biased, while junction N_1–P_1 remains reverse biased initially. The TRIAC operates in the negative bias region and turns ON due to carrier injection from the negative gate signal.

Mode 4: MT2 Negative and Gate Positive with respect to MT1

Under this condition, junction P_2–N_2 becomes forward biased, and injected charge carriers trigger conduction. This mode is lesspreferred, especially in circuits with high $\boldsymbol{di/dt}$, due to reduced triggering stability.

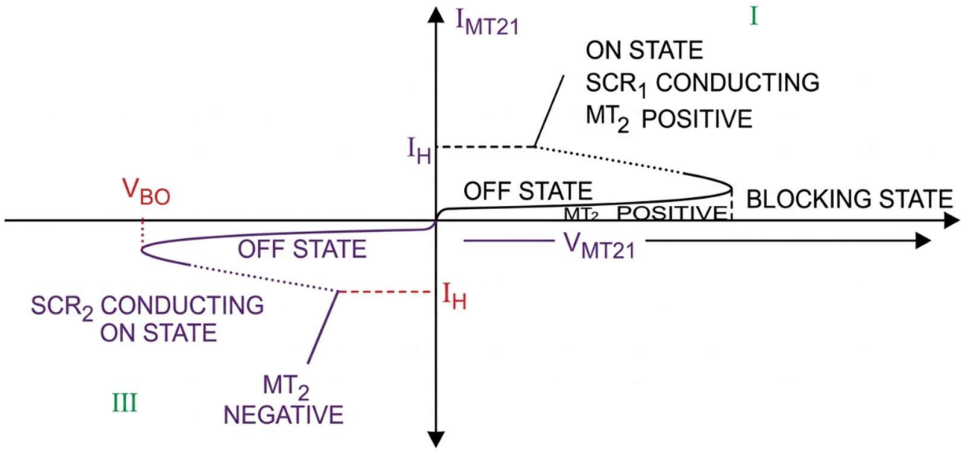


Figure 12.6: TRIAC Characteristics

Characteristics of a TRIAC:

The V–I characteristics of a TRIAC are similar to those of an SCR; however, unlike an SCR, a TRIAC operates in both positive and negative voltage regions due to its bidirectional conduction capability. Therefore, TRIAC operation is described in four quadrants based on the polarity of MT2 voltage and gate voltage with respect to MT1.

The operation of a TRIAC can be summarized as follows:

First Quadrant Operation (Mode I)

- Voltage at MT2 is positive with respect to MT1.
- Gate voltage is also positive with respect to MT1.
- The TRIAC conducts in the positive direction and exhibits high triggering sensitivity.

Second Quadrant Operation (Mode II)

- Voltage at MT2 is positive with respect to MT1.
- Gate voltage is negative with respect to MT1.
- Conduction occurs in the positive half cycle with comparatively lower triggering sensitivity.

Third Quadrant Operation (Mode III)

- Voltage at MT2 is negative with respect to MT1.
- Gate voltage is also negative with respect to MT1.
- The TRIAC conducts during the negative half cycle of the AC supply.

Advantages and Disadvantages of TRIAC:

Category	Description
Advantages of TRIAC	Can be triggered using either positive or negative gate pulses
	Requires only a single heat sink (slightly larger), whereas two SCRs require separate heat sinks
	Only one fuse is needed for circuit protection
	Provides safe bidirectional breakdown capability without the need for an additional parallel diode, unlike SCR-based circuits
Disadvantages of TRIAC	Less reliable compared to SCR in high-power applications
	Has a lower dv/dtrating than SCR, making it more sensitive to voltage transients
	Available with comparatively lower voltage and current ratings than SCR
	Triggering circuit design requires careful consideration since the device can be triggered in either direction.

Applications of TRIAC:

- Widely used in control circuits for regulating AC power.
- Employed in high-power lamp and lighting control systems, such as dimmers.
- Used in AC power control applications, including fan speed regulators, heating control, and motor speed control systems.

DIAC:

A DIAC (Diode for Alternating Current) is a two-terminal semiconductor device belonging to the thyristor family. It is primarily used for triggering TRIACs and other thyristor devices in AC power control circuits. One of the main advantages of a DIAC is that it can be turned ON or OFF simply by varying the applied voltage with respect to its avalanche breakdown voltage. The device conducts current in both directions,

making it suitable for bidirectional operation in AC circuits. The DIAC remains in the non-conducting state until the applied voltage reaches its breakdown level, after which avalanche breakdown occurs, allowing current to flow through the device.

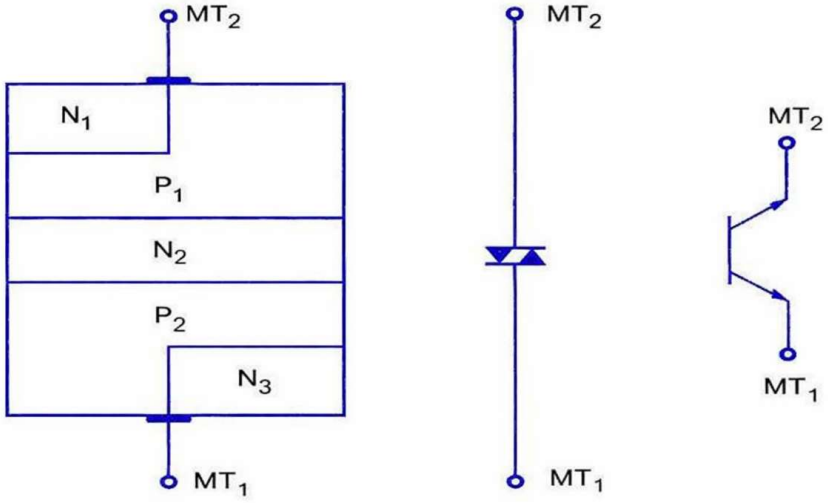


Figure 12.7: DIAC structure and symbol

Construction:

A DIAC is a semiconductor device consisting of four layers and two terminals. Its construction is somewhat similar to that of a transistor; however, it differs in several important aspects. Unlike a transistor, the DIAC does not have a base terminal, as it operates as a bidirectional switching device. The semiconductor regions in a DIAC are nearly equally doped, which enables uniform operation in both directions. Due to this symmetrical structure, the DIAC exhibits symmetrical switching characteristics for both positive and negative applied voltages, making it highly suitable for AC triggering applications.

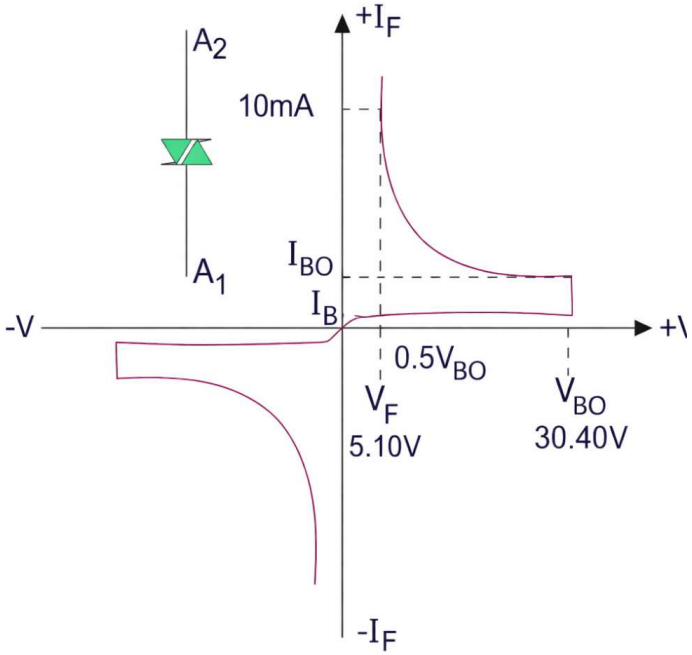


Figure 12.8: V-I characteristics of DIAC

Operation:
From the figure, it can be observed that the DIAC structure consists of two P-type regions and three N-type regions, and it does not contain any gate terminal as shown in Figure 12.7. The DIAC is capable of conducting for both polarities of the applied voltage, making it a bidirectional device. When terminal A_2 becomes positive with respect to A_1, current conduction takes place through the path P_2–N_2–P_1–N_1. Similarly, when A_1 becomes positive with respect to A_2, the current flows through the path P_1–N_2–P_2–N_3. Thus, the device allows symmetrical conduction in either direction. The construction of a DIAC resembles two diodes connected in series in opposite directions. When the applied voltage is small, irrespective of polarity, only a very small current flows due to the drift of charge carriers within the depletion region. This small current is known as the leakage current. At low applied voltages, the current is insufficient to cause avalanche breakdown, and therefore the DIAC remains in the non-conducting state. However, when the applied voltage in either direction exceeds the breakdown voltage, avalanche breakdown occurs, resulting in a rapid increase in current and the DIAC begins to conduct according to its V–I characteristics shown in Figure 12.8.

Applications:
The DIAC is primarily used in triggering circuits of TRIACs to ensure reliable and symmetrical switching in AC power control applications. It is generally connected to the gate terminal of the TRIAC, where it helps provide uniform triggering during both positive and negative half cycles of the AC supply. The DIAC conducts only when the applied voltage exceeds its avalanche breakdown voltage, thereby supplying a triggering pulse to the TRIAC. When the voltage falls below this breakdown level, the DIAC turns OFF, causing the TRIAC to remain in the non-conducting state. Major applications of DIAC include:

- Lamp dimmer circuits for smooth brightness control
- Heating control circuits in domestic and industrial systems
- Speed control of universal motors, such as fans and portable tools
- TRIAC triggering circuits, where the DIAC improves switching stability and reduces harmonics by enabling symmetrical firing angles

12.5 POWER MOSFET AND IGBT: CONSTRUCTION, CHARACTERISTICS, APPLICATIONS

Power MOSFET:
Metal–Oxide–Semiconductor Field Effect Transistors (MOSFETs), possessing significant on-state current carrying capability along with high off-state voltage blocking capacity, have been widely available for power electronic applications since the early 1980s. Over time, MOSFETs have achieved usage comparable to power Bipolar Junction Transistors (BJTs) and are increasingly replacing BJTs in many applications, particularly those requiring high switching speeds and improved efficiency. MOSFETs operate based on physical principles that differ fundamentally from those of BJTs. Therefore, a clear understanding of these operational differences is essential for the effective selection and utilization of both device types in power electronic systems. This

chapter discusses the fundamental physical mechanisms governing MOSFET operation, the parameters that determine their current and voltage handling capabilities, and the potential failure mechanisms that may occur when these operational limits are exceeded.

Construction:

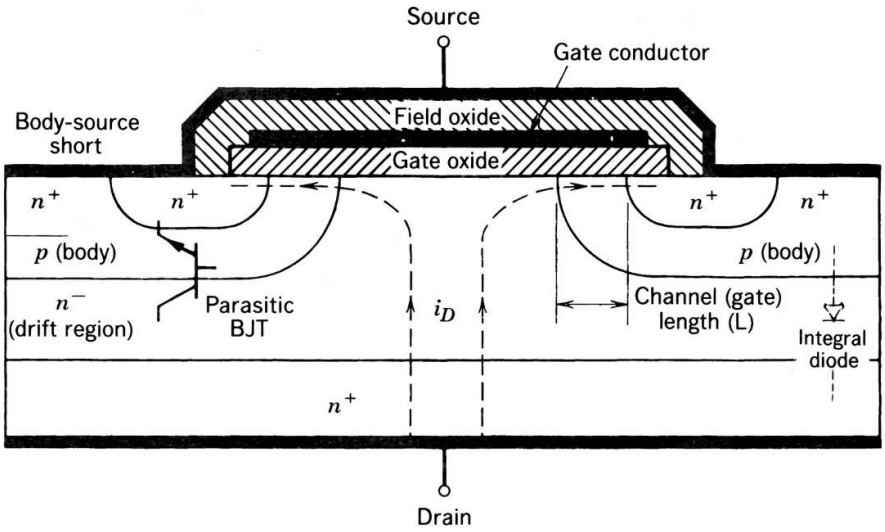


Figure 12.9: Vertical cross section of a power MOSFET

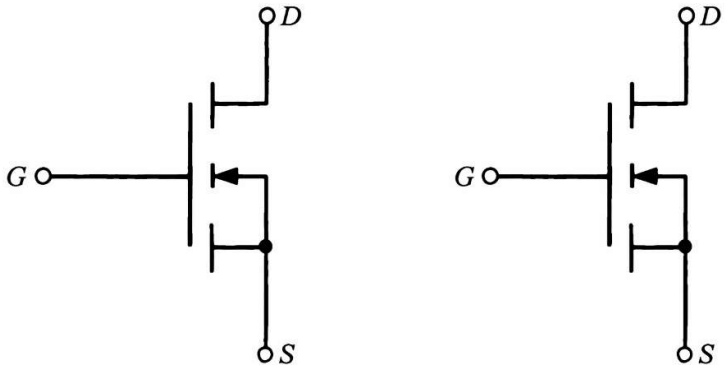


Figure 12.10: a) n-channel MOSFET, b) p-channel MOSFET

A power MOSFET consists of a vertically structured four-layer semiconductor arrangement with alternating p-type and n-type regions as shown in Figure 12.9. The circuit symbol for an n-channel and p-channel MOSFET is shown in Figure 12.10. The commonly used configuration is the enhancement-mode n-channel MOSFET, having an $\boldsymbol{n^+}$–$\boldsymbol{p}$–$\boldsymbol{n^-}$–$\boldsymbol{n^+}$ structure, while the opposite doping profile forms a p-channel MOSFET. The heavily doped $\boldsymbol{n^+}$ regions act as the source and drain, whereas the p-type body region forms the channel area between them. The lightly doped $\boldsymbol{n^-}$drift region primarily determines the device's breakdown voltage capability. Under normal conditions, current cannot flow between drain and source because one of the p-n junctions remains reverse biased. Since the gate is insulated from the body by a thin silicon dioxide (SiO_2) layer, no direct current flows through the gate. However, applying

a positive gate-to-source voltage induces an inversion layer beneath the gate oxide, creating an n-type conductive channel that enables current flow between source and drain. This structure, known as Vertical Diffused MOSFET (VDMOS), is fabricated through epitaxial growth of the drift region followed by masked diffusion of body and source regions. Practical devices consist of thousands of parallel cells to increase the effective gate width-to-length ratio, thereby reducing on-state resistance and improving current capability. A parasitic n-p-n transistor inherently exists within the structure; therefore, the body region is shorted to the source to prevent unintended activation. This configuration also forms an intrinsic body diode between drain and source, widely utilized in power converter circuits.

Additionally, gate overlap over the drift region improves conductivity and electric-field distribution, enhancing device performance and breakdown voltage reliability.

Characteristics:

The MOSFET is a three-terminal device in which the gate voltage controls the current flow between the drain and source terminals. Similar to a BJT, it is widely used for switching and power control applications. The source terminal is commonly shared between the input and output circuits. The output characteristics represent the drain current (i_D) as a function of drain–source voltage (V_{DS}), with gate–source voltage (V_{GS}) as a controlling parameter. The characteristics of n-channel device is shown in Figure 12.11. For a p-channel MOSFET, the characteristics are identical to an n-channel device except that voltage and current polarities are reversed.

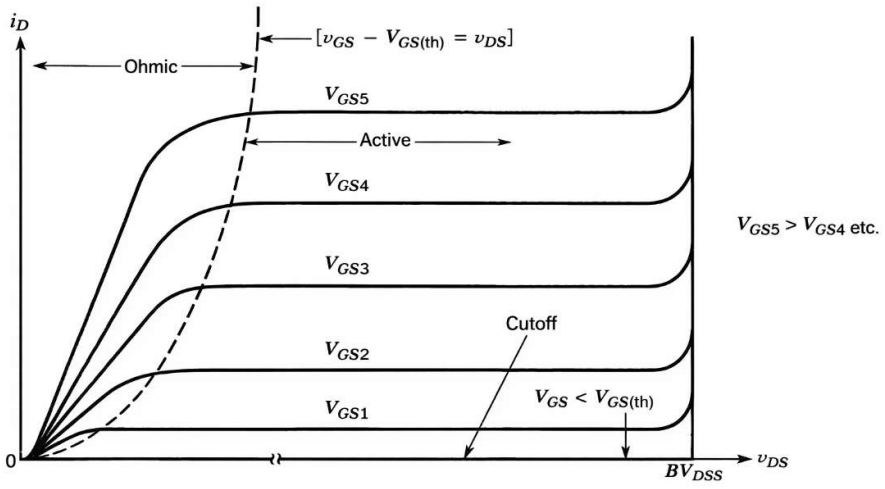


Figure 12.11: V-I characteristics of n-channel MOSFET

In power electronics, MOSFETs operate mainly as switches and move through three operating regions:

1. **Cutoff Region:**
   When V_{GS} is below the threshold voltage $V_{GS(th)}$, the device remains OFF and behaves as an open circuit. The MOSFET must withstand the applied drain voltage without breakdown.
2. **Ohmic (Linear) Region:**
   When V_{GS} is sufficiently high and V_{DS} is small, the MOSFET conducts with low resistance, minimizing $V_{DS(on)}$. The device behaves like a controlled resistor.

3. **Active Region:**
   Here, drain current becomes almost independent of V_{DS} and depends mainly on V_{GS}. The drain current is approximately given by

$$i_D = K(V_{GS} - V_{GS(th)})^2$$

where Kdepends on device geometry.

Power MOSFET transfer characteristics are generally more linear compared to logic-level MOSFETs, making them suitable for high-power switching applications.

Applications:

- Used in Switched Mode Power Supplies (SMPS) for efficient power conversion.
- DC–DC converters for voltage regulation applications.
- Motor drive circuits for speed and torque control.
- Inverters and UPS systems for DC to AC conversion.
- Widely used in electric vehicle motor controllers and BMS.
- Used in switching voltage regulators in electronic devices.
- LED driver circuits for controlled power delivery.
- Class-D audio amplifiers for high-efficiency amplification.
- Solar and renewable energy systems for power management.
- Applied in electronic switching and protection circuits.

IGBT:

The Insulated Gate Bipolar Transistor (IGBT) was developed by integrating the advantages of MOSFET and BJT technologies at the device cell level to achieve improved switching performance and power handling capability. Early developments such as the IGT introduced by the GE Research Laboratory and the COMFET developed by RCA Research Laboratory led to continuous technological improvements, resulting in the modern IGBT. The IGBT exhibits nearly ideal characteristics for high-voltage (greater than 100 V) and medium-frequency (below 20 kHz) applications. Due to its high efficiency, easy gate control, and superior power handling capability, the IGBT, together with MOSFETs for low-voltage high-frequency applications, has largely replaced conventional BJTs in modern power electronic systems.

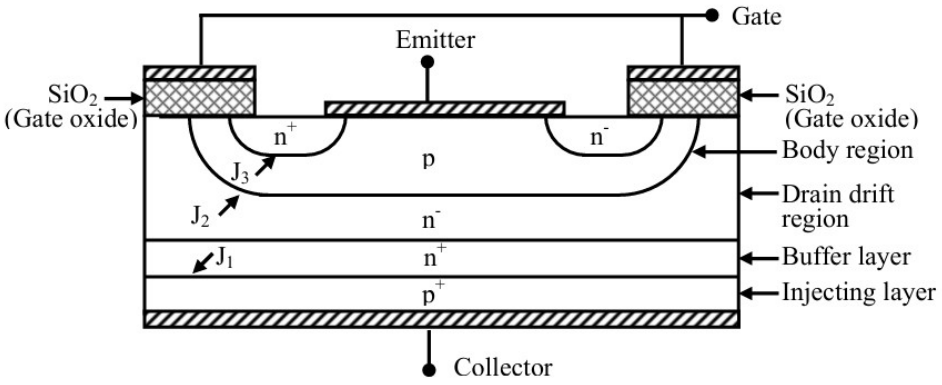


Figure 12.12: Vertical cross section of an IGBT

Construction:

The vertical cross-section of an n-channel IGBT cell consists of a structure similar to a power MOSFET with the addition of a p^+ injecting layer, which forms a p-n junction with the drain region and injects minority carriers shown in Figure 12.12. This improves

conductivity through conductivity modulation, resulting in a lower on-state voltage drop. The lightly doped n^- drift region determines the forward blocking voltage of the device. IGBTs are commonly fabricated using Punch Through (PT) and Non-Punch Through (NPT) designs. The PT structure provides lower conduction loss but comparatively lower reverse breakdown capability, while the NPT structure offers improved voltage blocking performance. The remaining construction, including the insulated gate and shorted body–emitter structure, resembles that of a vertical MOSFET, though the body region is specially designed to prevent latch-up. An inherent parasitic p–n–p–n thyristor structure exists within the IGBT cell. Excess current mayactivate this structure, leading to latch-up and loss of gate control. Modern IGBT designs overcome this issue through optimized doping and geometry, making present devices practically latch-up free. Multiple such cells are connected in parallel on a silicon wafer to form a complete IGBT device.

Operating Principle of IGBT:

The operating principle of an IGBT combines the characteristics of both MOSFET and BJT devices. When the gate–emitter voltage is below the threshold value, no inversion channel is formed in the p-type body region, and the device remains in the OFF state, allowing only a small leakage current to flow while the applied voltage is sustained across the collector–emitter junction. When the gate–emitter voltage exceeds the threshold, an inversion layer (channel) forms, enabling electrons to flow from the emitter to the drift region. This electron flowtriggers hole injection from the p^+ collector, resulting in conductivity modulation of the drift region and significantly reducing conduction resistance. The drift region acts as the base of an internal p–n–p transistor, whose gain is intentionally kept low to prevent latch-up and maintain gate control. The total on-state voltage drop of an IGBT consists of the junction voltage drop, reduced drift-region resistance due to conductivity modulation, and channel resistance similar to that of a MOSFET, making the IGBT more efficient than an equivalent MOSFET in high-voltage applications.

Characteristics of an IGBT:

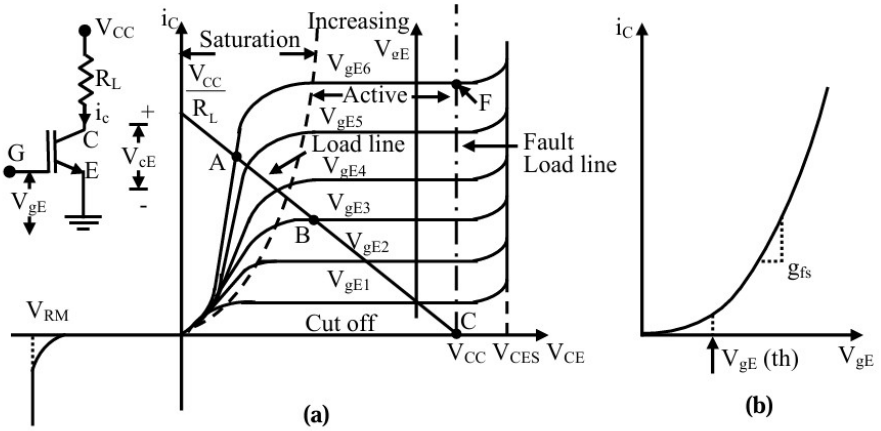


Figure 12.13 (a) Output characteristics (b) Transfer characteristics

The I–V characteristics of an n-channel IGBT are similar to those of a BJT, but the device is controlled by gate–emitter voltage instead of base current. Figure 12.13

shows output and transfer characteristics of IGBT. When the gate–emitter voltage is below the threshold, only a small leakage current flows and the device operates in the cut-off region, sustaining the applied collector–emitter voltage. The maximum blocking voltage is determined by the junction breakdown voltage. Non-Punch Through (NPT) IGBTs can block both forward and reverse voltages, while Punch Through (PT) devices have limited reverse blocking capability. As the gate voltage exceeds the threshold, the IGBT enters the active region, where collector current increases with gate voltage according to the transfer characteristics. The ratio of collector current to effective gate voltage defines the forward transconductance, an important design parameter. With further increase in gate voltage, the device moves into the saturation region, where collector–emitter voltage becomes nearly constant and conduction loss is minimized. In practical power electronic applications, IGBTs operate mainly in cut-off and saturation modes. IGBTs exhibit a positive temperature coefficient, preventing second breakdown failure and allowing safe parallel operation of devices.

Applications:

- Inverter circuits for DC to AC power conversion.
- Variable speed motor drives for industrial motor control.
- Electric vehicle (EV) traction systems and motor controllers.
- Uninterruptible Power Supply (UPS) systems.
- Renewable energy systems such as solar and wind inverters.
- Applied in switch mode power supplies (SMPS) for high-power applications.
- Induction heating and welding equipment.
- Applied in electric traction systems like trains and metro drives.

W CHECK YOUR PROGRESS

Q1. Name different types of power electronic devices.
Q2. What are the operating modes of an SCR?
Q3. Why is DIAC used with TRIAC?
Q4. Name different operating regions of MOSFET.
Q5. State advantages of IGBT over BJT.

2 12.6 LET US SUM UP

- Thyristor (SCR) is a PNPN four-layer device used for controlled power switching with latching action in rectifier and control circuits.
- TRIAC and DIAC are AC control devices where TRIAC handles bidirectional current switching and DIAC provides triggering for phase control.
- Thyristors exhibit low conduction loss and holding/latching characteristics, making them suitable for power regulation and motor control.
- Power MOSFET is a voltage-controlled, fast-switching device with high input impedance used in power electronics and switching applications.
- IGBT combines MOSFET input and BJT output features, offering high efficiency and current capacity for inverters and industrial drives.

K 12.7 ANSWERS TO CHECK YOIUR PROGRESS

Answer to Q 1: Power electronic devices include:

- Power Diode
- Silicon Controlled Rectifier (SCR)
- TRIAC
- DIAC
- Power MOSFET
- IGBT (Insulated Gate Bipolar Transistor)
- GTO (Gate Turn-Off Thyristor)

Answer to Q 2: The operating modes of an SCR are:

- Forward Blocking Mode
- Forward Conduction Mode
- Reverse Blocking Mode

Answer to Q 3: A DIAC is used with a TRIAC to provide symmetrical triggering. It ensures that the TRIAC turns on at a specific voltage in both directions, improving stability and reducing distortion in AC control circuits.

Answer to Q 4: The operating regions of a MOSFET are:

- Cut-off Region
- Linear (Ohmic/Triode) Region
- Saturation (Active) Region

Answer to Q 5: Advantages of IGBT over BJT:

- High input impedance (voltage-controlled device)
- Faster switching speed
- Lower gate drive power requirement
- Better thermal stability
- Easier drive circuitry

O 12.8 FURTHER READING

- Baliga, B. J. (2019). *Fundamentals of power semiconductor devices* (2nd ed.). Springer.
- Baliga, B. J. (1996). *Power semiconductor devices*. PWS Publishing Company.
- Rashid, M. H. (2011). *Power electronics handbook* (3rd ed.). Butterworth-Heinemann.
- Mohan, N., Undeland, T. M., & Robbins, W. P. (2003). *Power electronics: Converters, applications, and design* (3rd ed.). John Wiley & Sons.
- Erickson, R. W., & Maksimović, D. (2020). *Fundamentals of power electronics* (3rd ed.). Springer.

M 12.9 MODEL QUESTIONS

Q1. Explain the need and importance of power semiconductor devices in modern power electronic systems.

Q2. Describe the construction and working principle of a Silicon Controlled Rectifier (SCR) with neat diagram.

Q3. Explain the V–I characteristics and operating modes of an SCR.

Q4. Discuss the construction, operation, and applications of TRIAC and DIAC devices.
Q5. Explain the construction and operating principle of a Power MOSFET.
Q6. Describe the output characteristics and operating regions of a Power MOSFET.
Q7. Explain the construction and operating principle of an Insulated Gate Bipolar Transistor (IGBT).
Q8.Discuss the output and transfer characteristics of an IGBT.
Q9. Compare Power MOSFET and IGBT based on operation, characteristics, and applications.
Q10. Select suitable power semiconductor devices for the following applications and justify your choice:

- Motor drive
- Inverter system
- High-frequency converter
- AC power control circuit

*** ***** ***

UNIT 13: FABRICATION OF P–N JUNCTIONS – PART I

UNIT STRUCTURE

13.1 Learning Objectives
13.2 Introduction
13.3 Wafer Preparation and Doping Techniques
13.4 Thermal Oxidation
13.5 Diffusion Process
13.6 Rapid Thermal Processing (RTP)
13.7 Ion Implantation
13.8 Let Us Sum Up
13.9 Answers to Check Your Progress
13.10 Further Reading
13.11 Model Questions

13.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- acquire an overview of the basic process steps for fabricating a PN junction
- familiarize with the Si wafer preparation techniques and different doping methods
- explain the different setups and systems used for different fabrication steps
- discuss the chemical reactions and mathematical functions associated with different fabrication steps.

13.2 INTRODUCTION

In the previous unit, we studied power semiconductor devices and their applications. In this unit, we shift our focus to wafer preparation and doping techniques, which are essential steps in semiconductor device fabrication. This unit also covers thermal oxidation for growing oxide layers on silicon. It includes diffusion processes for introducing dopants into the wafer. It further discusses rapid thermal processing (RTP) for precise and controlled heat treatment, and ion implantation for accurately introducing dopant atoms into specific regions of the semiconductor.

13.3 WAFER PREPARATION AND DOPING TECHNIQUES

Modern silicon-based integrated circuits are fabricated through a sequence of carefully controlled processing steps. These steps typically include oxide layer formation, selective material removal, and the introduction as well as redistribution of dopant atoms within the silicon substrate. A PN junction, which is the most fundamental semiconductor structure, is formed by creating adjacent regions of p-type and n-type material. In practical fabrication, this is achieved by modifying the surface region of a uniformly doped wafer rather than physically joining two different materials.

The fabrication process starts with a single-crystal Si wafer, often doped with P-type dopant. The wafer must be properly cleaned to remove all particulates, organic films,

and adsorbed metal from the semiconductor surface. Then, a thermal oxide is grown (step 1), followed by lithography (step 2) and etching processes to create an opening in the oxide (step 3). The wafer is introduced with appropriate amount of N-type dopant (such as Arsenic) through this opening using ion implantation (step 4). This is followed by annealing and diffusion (step 5), and eventually, the PN junction is formed.

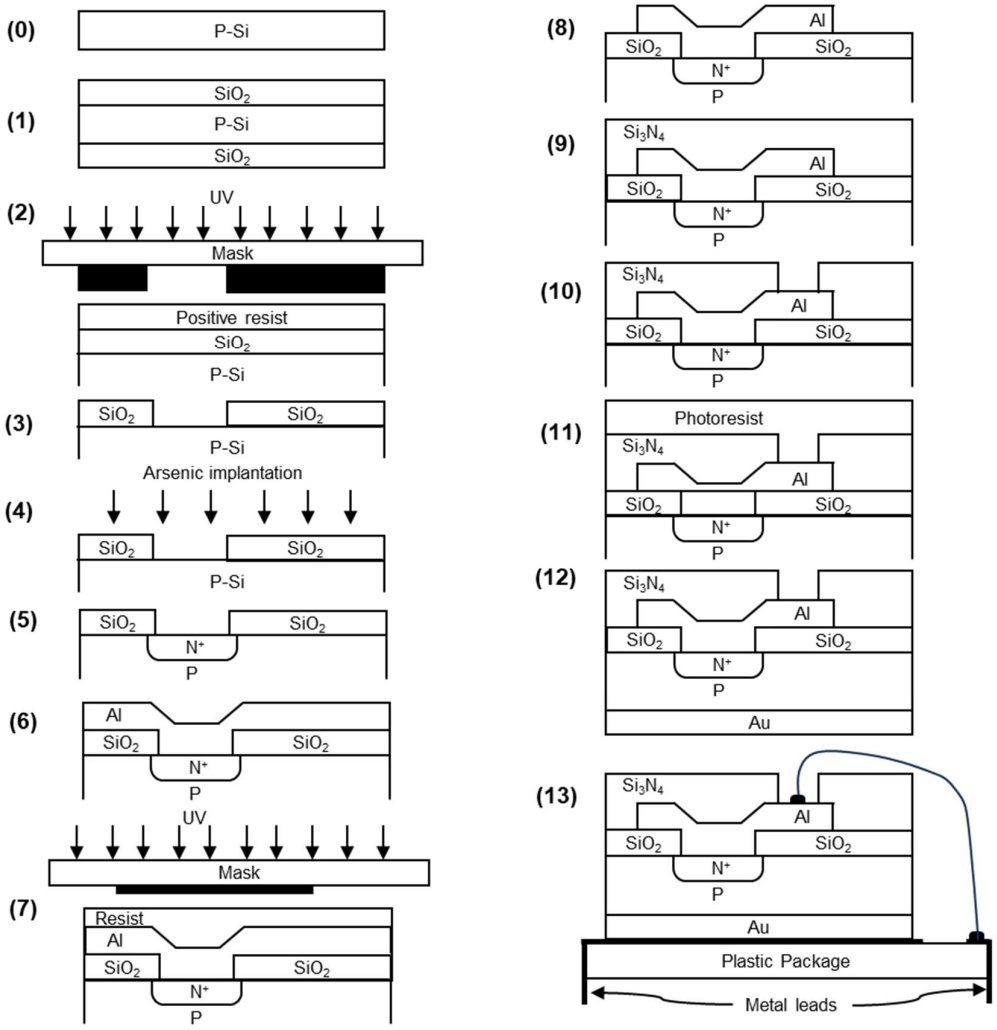


Figure 13.1 Process flow for PN junction diode fabrication.

(0) Pre-cleaned P-Silicon; (1) oxidation; (2) lithography; (3) oxide etching; (4) As implantation; (5) annealing and diffusion; (6) sputtering Al; (7) lithography; (8) metal etching; (9) CVD nitride deposition; (10) lithography and bonding window etching; (11) removal of oxide from back side of wafer; (12) deposition of Au on back side; and (13) dicing and packaging.

Next step is metallization, where a thin metal film (mostly Al) is deposited over the entire surface of the wafer with the help of sputtering (step 6). Then another lithography process (step 7) is carried out to pattern the metal, followed by a low-temperature (≤450 °C) anneal process (step 8) to produce a low-resistance contact between the metal and Si. SiO_2 and Si_3N_4 protective encapsulation layers are deposited over the device surface to guard the wafer against moisture and external contaminants (step 9).

Openings are then created in these layers to expose the aluminium regions, allowing wire bonding to establish the top electrical contact (step 10). To form a proper electrical connection with the p-type substrate, the oxide layer present on the backside of the wafer is removed, while the front surface is protected using a photoresist coating (step 11). After this, a metal layer, commonly gold, is deposited on the backside to serve as the bottom contact (step 12). In the final stage, the wafer is separated into individual diode chips, and each chip is mounted into an appropriate package (step 13). The backside metal contact is directly connected to one terminal of the package, whereas the top aluminium contact is linked to another terminal through a bonding wire.

Single-crystal silicon with minimal defects is grown using the Czochralski (CZ) process, in which high-purity polycrystalline silicon is melted in a quartz crucible at around 1420°C, and a seed crystal is dipped into the melt, slowly rotated, and pulled upward to form a large single-crystal ingot. Dopants like boron (p-type) or phosphorus (n-type) are added to the melt for controlled resistivity. This method allows precise control of growth rate, wafer diameter, and doping concentration, which is why it is widely used in the semiconductor industry to grow large-diameter wafers. Alternatively, another method called the Float Zone (FZ) technique is also used to produce single-crystal silicon, where a molten zone is moved along a polysilicon rod via RF heating for ultra-high purity. This method allows the growth of low impurity crystals with high resistivity, making it suitable to fabricate high-power semiconductor devices. Schematic illustration of the CZ and FZ processes are shown in figure 13.2(a) and (b), respectively.

After crystal growth by any suitable method, the silicon ingot is cut into thin wafers, typically 200–300 μm thick, using a diamond saw. The slicing is done at a controlled orientation to reduce crystal defects. These wafers then undergo several mechanical and chemical treatments to make them suitable for integrated circuit fabrication. Initially, a material removal process such as lapping is performed to achieve uniform thickness and smoother surfaces. This is followed by double-sided polishing using colloidal silica to obtain a highly smooth, mirror-like finish under cleanroom conditions. Finally, the wafers are subjected to RCA cleaning, which effectively removes organic residues, ionic contaminants, metallic impurities, and any native oxide layer present on the surface. Organic residues and particles are removed using a heated mixture of ammonium hydroxide, hydrogen peroxide, and deionized water. This step leaves a thin oxide layer, which can be removed by dipping the wafer in dilute hydrofluoric acid. To remove remaining metallic (ionic) contaminants, a heated mixture of hydrochloric acid, hydrogen peroxide, and deionized water is used.

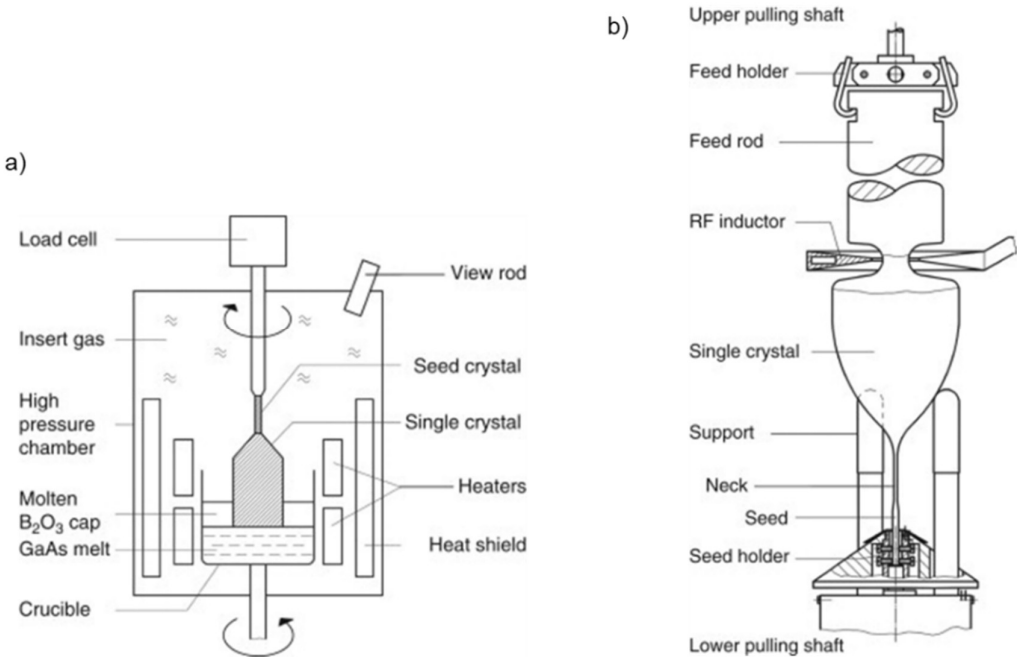


Figure 13.2: Schematic of the single crystal growth techniques; (a) Czochralski (CZ) process, and (b) Float Zone (FZ) process.

Dopants can be incorporated during crystal growth (in-situ doping) or through post-growth processes such as diffusion or ion implantation. In-situ doping ensures uniform incorporation of dopants into the crystal lattice as it solidifies from the melt, creating bulk-doped wafers ideal for PN junction fabrication. Whereas post-growth doping is suitable to create precise n-type or p-type regions on undoped or lightly doped substrates. The spatial distribution of dopant atoms within silicon, commonly referred to as the dopant profile, is typically established through a two-stage process. In the first stage, impurities are introduced at or close to the wafer surface using techniques such as gas-phase doping, solid-source diffusion, or ion implantation. This is followed by a second stage in which the dopant atoms migrate deeper into the silicon substrate through diffusion, which may occur either deliberately or as a result of subsequent thermal processing. Diffusion involves exposing wafers to dopant gases (e.g., $POCl_3$ for phosphorus) at high temperatures (900-1200°C), allowing atoms to migrate into the lattice and forming graded junctions but risking lateral spread and contamination. In solid-source diffusion, a thin dopant-containing layer (e.g., SiGe) is first deposited on the silicon surface, and the dopants are then diffused into the substrate. The remaining film is later removed by wet etching. In ion implantation, dopant ions are accelerated toward the silicon using electric fields, allowing precise and shallow doping profiles. This is followed by annealing (around 1000 °C) to repair lattice damage and activate the dopants electrically. These processes are explained in detail in the following subsections.

13.4 THERMAL OXIDATION

In silicon IC fabrication, silicon dioxide (SiO_2) serves several essential roles, including acting as a masking layer to block dopant entry, providing electrical insulation between devices, and functioning as the gate dielectric in MOSFETs. To obtain oxide layers of

controlled thickness, SiO_2 is grown by reacting silicon with oxidizing agents at elevated temperatures. Oxidation using oxygen gas is known as dry oxidation, while the use of water vapour is termed wet oxidation. In both cases, the oxidizing species (O_2 or H_2O) diffuses through the existing oxide layer and reacts at the silicon surface, leading to oxide growth at the interface. The overall reactions can be written as,

$$Si + O_2 \rightarrow SiO_2 \quad (13.1)$$

$$Si + 2H_2O \rightarrow SiO_2 + 2H_2 \quad (13.2)$$

Dry oxidation proceeds at a slower rate, allowing better control over oxide thickness and making it suitable for forming thin, high-quality films such as MOSFET gate oxides. In contrast, wet oxidation is faster because water vapour diffuses more readily through SiO_2 than oxygen. As a result, it is commonly used to grow thicker oxide layers for field isolation and insulation purposes.

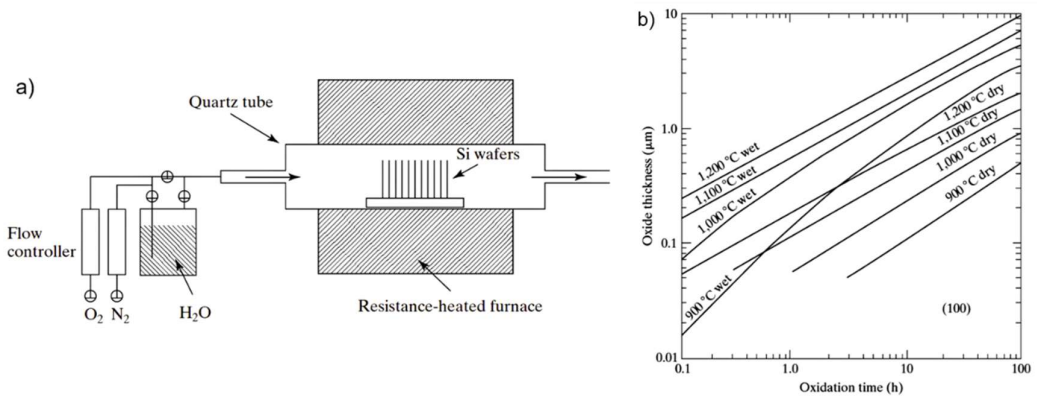


Figure 13.3: (a) Schematic drawing of an oxidation system, and (b) the SiO_2 thickness formed on (100) silicon surfaces as a function of time [1].

Figure 13.3(a) illustrates a typical oxidation furnace, which may be designed in either horizontal or vertical configuration. The central tube, commonly made of fused quartz, silicon carbide (SiC), or polycrystalline silicon, holds the silicon wafers placed on a quartz boat. These wafers are inserted into the furnace, where temperatures between 700°C and 1200°C are maintained using electrical heating elements. During dry oxidation, oxygen gas is directly introduced into the furnace tube. In contrast, wet oxidation involves passing inert gases such as argon (Ar) or nitrogen (N_2) through heated water to generate water vapour, which then reacts with the silicon surface, and water vapour reacts with Si wafers inside the tube. An additional source line to the furnace may allow gaseous Hydrochloric Acid (HCL) or trichloroethane (TCA), which is used to clean the furnace prior to oxidation or as a source of Cl to add a small percentage as an additive. Cl reacts with many unwanted metal ions and produces gas-phase byproducts that may be exhausted from the furnace. The thickness of the oxide grown depends on several factors such as the furnace temperature, the oxidation time, the ambient gas, and the Si surface orientation. The oxide growth curves for wet and dry oxidation at different temperatures are shown in figure 13.3(b). At 1000°C, wet oxidation rates are on the order of 0.1 nm/s, approximately doubling for every 100°C rise in temperature. In the case of dry oxidation, oxide growth is very fast during the initial phase, which makes it very difficult to grow oxides with a thickness of less than 200 Å. Therefore, to achieve ultrathin oxides (<50 Å), specialized dry oxidation techniques are used. Such techniques include dry oxidation with HCL, sequential

oxidation using different temperatures and ambients, reduced pressure oxidation, low temperature oxidation, etc. Anodic plasma oxidation enables the growth of high-quality oxides at temperatures below 600°C.

13.5 DIFFUSION PROCESS

After dopants are introduced through implantation or gaseous methods, they are often driven deeper into the silicon using diffusion. In some cases, diffusion refers to both the initial dopant deposition and the subsequent redistribution process. The stage during which the dopant source is present is known as predeposition, while the later stage when the source is removed and dopants spread further into the material is called drive-in diffusion. Additionally, unintended diffusion may also occur during post-implantation annealing.

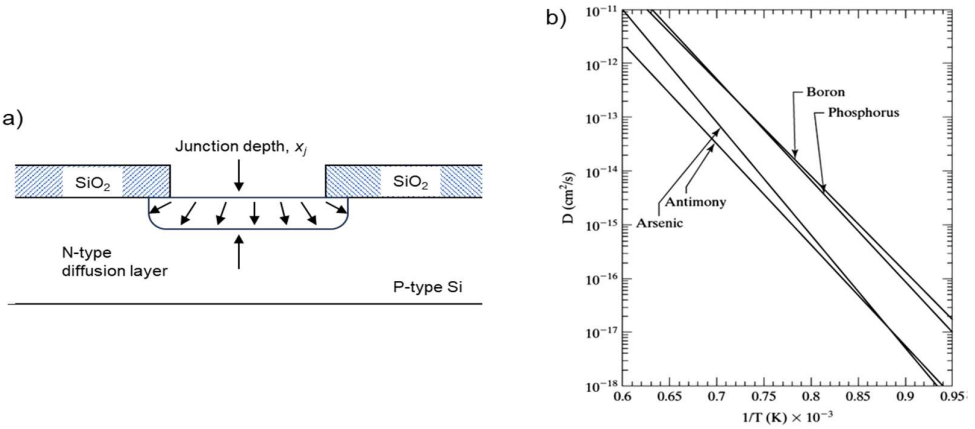


Figure 13.4: (a) The diagrammatic representation of the basic diffusion process, and (b) Diffusivity versus 1/T for As, Sb, P, and B, in silicon [2].

Figure 13.4(a) illustrates the basic diffusion process, while Figure 13.4(b) presents the diffusivity of common dopants such as arsenic (As), antimony (Sb), phosphorus (P), and boron (B) in silicon as a function of the reciprocal temperature (1/T). It is observed that the diffusion rate increases with rising temperature. This process is typically carried out in an open-tube furnace system similar to that used for oxidation, with operating temperatures generally ranging from 900 °C to 1200 °C.

For PN junction formation, the introduced dopant is of opposite conductivity type to that of the substrate, resulting in a junction at the point where the acceptor and donor concentrations become equal (Na = Nd). The distance from the surface to this point is defined as the junction depth. The dopant concentration within the semiconductor, represented as N(x,t), varies with depth x and time t, and follows a Gaussian distribution, which can be expressed mathematically as:

$$N(x,t) = \frac{N_0}{\sqrt{\pi Dt}} e^{-x^2/4Dt} \tag{13.3}$$

Here, N_0 denotes the total number of dopant atoms per unit area (cm²), determined during the dopant introduction stage, while D represents the diffusivity of the impurity, which is highly dependent on the furnace temperature. Depending on the application, either deep or very shallow junctions may be required, making precise control of

diffusion essential. To achieve shallow junctions, the time–temperature product (Dt), known as the thermal budget, must be kept low to restrict dopant penetration.

However, for future generation devices scaled down to tens of nm, the approach of reducing the thermal budget, i.e. the furnace temperature may not work anymore. The new generation diffusion systems are designed with fast-ramp capability and vertical orientation to minimize sample loading, unloading, and ramp times, using ramp rates of 100°C/min. Rapid thermal technique is becoming a key process in diffusion annealing, which is discussed in the following subsection.

13.6 RAPID THERMAL PROCESSING (RTP)

After doping, the wafer typically undergoes thermal annealing to repair lattice damage and activate dopants by placing them at proper lattice sites so they function as donors or acceptors. Conventional furnace annealing can cause significant dopant diffusion, which is carried out at about 900 °C for nearly 30 minutes. This is undesirable for shallow junction devices. For high-performance applications, annealing at higher temperatures for much shorter durations such as around 1050 °C for about 20 seconds helps minimize diffusion.

To achieve such rapid heating and cooling, a specialized method known as rapid thermal annealing (RTA) is used. In this technique, the wafer is heated within seconds using high-intensity lamps, such as tungsten–halogen lamps. A schematic of a typical RTA system is shown in Figure 13.5. The wafers are placed on sharp pins or low-thermal mass holders, and temperature measurements are taken using thermocouples and pyrometers to ensure that the samples are heated to the required temperature while minimizing the risk of overheating or underheating. Temperatures can range from 200°C to over 1300°C, with ramp rates typically ranging from 20°C/sec to 200°C/sec. Cooling is also rapid due to the low thermal mass of the system. Similar setups are employed for processes such as rapid thermal oxidation and rapid thermal chemical vapour deposition (CVD), which are used to alter the crystallographic properties of materials. Collectively, these techniques are referred to as rapid thermal processing (RTP).

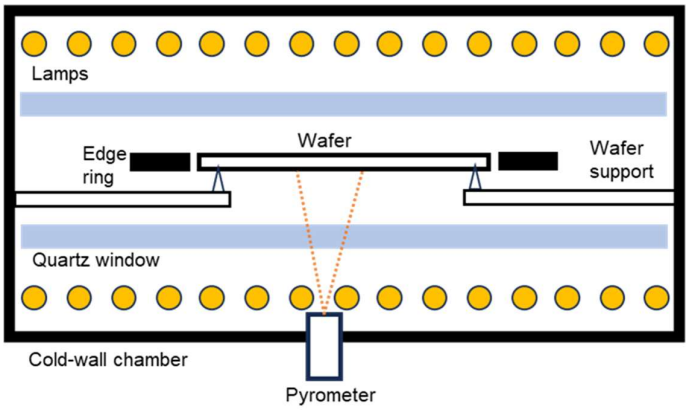


Figure 13.5: Diagrammatic representation of a rapid thermal annealing (RTA) system.

By reducing the annealing time in RTA to about 0.1 s, even shallower junctions can be achieved; this approach is known as flash annealing. For extremely short heating

durations (below a microsecond), laser pulses are used to heat the silicon wafer, a technique referred to as laser annealing, which may sometimes cause melting of a very thin surface layer. In addition to pulsed laser annealing, state-of-the-art RTA techniques include pulsed electron and ion beams, scanned electron beams, continuous wave (scanned) lasers, high beam current implants, and broad-band spectral sources (high-intensity lamps) with programming times down to a few picoseconds. Damage to the crystal lattice caused by ion implantation can temporarily increase dopant diffusivity at lower temperatures beyond the normal values. This phenomenon is known as transient enhanced diffusion (TED). The term transient indicates that this increased diffusion is short-lived and diminishes as the lattice damage is removed during annealing. RTP technology enables the seamless execution of advanced multistage processes, consolidating efficiency and precision in a single processing recipe.

13.7 ION IMPLANTATION

Ion implantation is the most important doping method because of the precise control it provides. In this process, an impurity is introduced into the semiconductor by creating ions of the impurity, accelerating the ions to high energies ranging from sub-keV to MeV, and then literally shooting the ions onto the semiconductor surface. The ions that impact the silicon wafer slow down by a combination of nuclear and electronic interactions with the lattice atoms. This is shown schematically in figure 13.6(a) below.

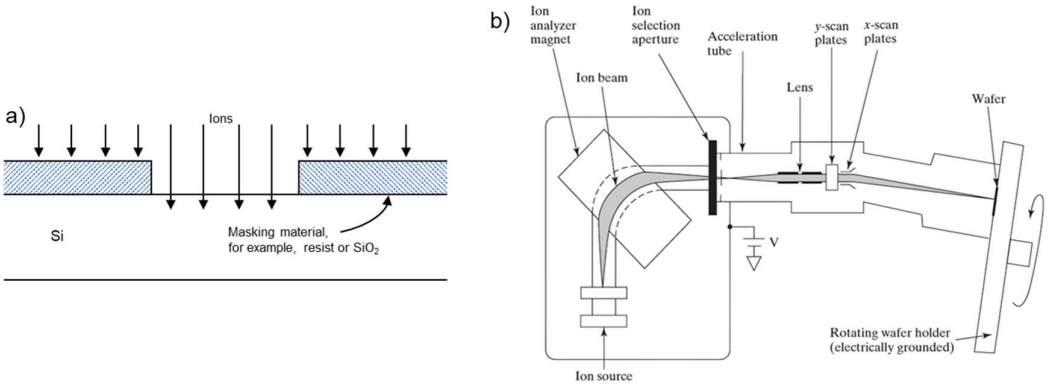


Figure 13.6: Diagrammatic representation of (a) the ion implantation process, and (b) an ion implantation system.

Figure 13.6(b) shows the schematic of an ion implantation system. The dopant material is supplied either from a vaporized solid source or a gas source. This material is ionized using energetic electrons or a plasma discharge to generate ions of the desired impurity. These ions are then accelerated and passed through a mass analyzer, which removes unwanted species and allows only selected ions to pass through an aperture. The resulting ion beam is further accelerated to the required implantation energy and directed onto the wafer surface, which is mounted on a metal holder. The beam is scanned across the wafer either electrostatically, by mechanical movement, or by a combination of both methods. This process can lead to charge buildup on the wafer surface, potentially causing damage. To prevent this, an electrical connection to the wafer supplies electrons to neutralize any charge build up during implantation.

Ion Implantation allows precise control over the number of implanted ions per unit area, known as the implantation dose (N_i), which is obtained by integrating the beam current over time. The resulting dopant distribution typically follows a Gaussian profile, characterized by the peak depth or implantation range (R) and the spread (ΔR), referred to as straggle.

$$N(x) = \frac{N_i}{\sqrt{2\pi}(\Delta R)} e^{-(x-R)^2/2\Delta R^2} \quad (13.4)$$

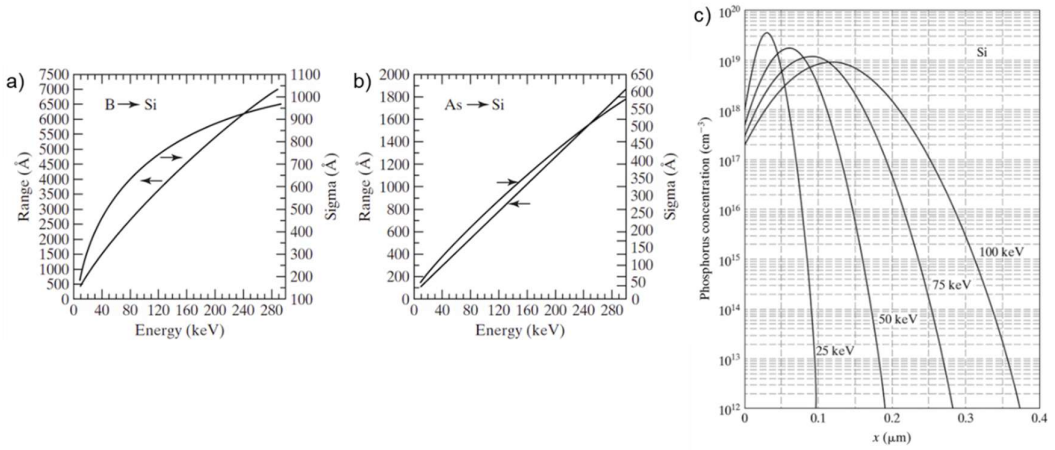


Figure 13.7: (a,b) Implantation range (R) and straggle (ΔR) of implantation of (a) Boron and (b) Arsenic in silicon, versus energy [2], and (c) Implantation profiles of phosphorus for a constant dose of 10^{14}cm^{-2} [3].

These parameters depend on the type of implanted ion and the substrate material, and they generally increase with ion energy. Heavy ions like antimony do not travel as far in the crystal as light ions like boron. The implantation range and straggle for Boron and Arsenic when doped into Si are shown for ion beam energies in the range of 0-300 keV in figures 13.7(a) and (b), respectively. The implantation profile for phosphorus implanted into Si with a constant dose of 10^{14} /cm^3 is shown in figure 13.7(b), at various energies.

2 13.7 LET US SUM UP

- PN junction fabrication begins with single-crystal Si wafer preparation via Czochralski or Float Zone methods, followed by cleaning and polishing.
- Wafer cleaning ensures a defect-free start of the fabrication process.
- Oxidation grows SiO_2 on Si, which masks doping areas.
- Diffusion/implantation create junctions; RTP minimizes deepen.
- Key processes include thermal oxidation (dry/wet) for SiO_2 masks, diffusion for dopant drive-in described by Gaussian profiles

  $$N(x,t) = \frac{N_0}{\sqrt{\pi Dt}} e^{-x^2/4Dt}$$

  and ion implantation for precise doping with profiles

  $$N(x) = \frac{N_i}{\sqrt{2\pi}(\Delta R)} e^{-(x-R)^2/2\Delta R^2}$$

- Rapid thermal processing (RTP) enables short anneals to activate dopants and repair damage while minimizing junction deepening.

W CHECK YOUR PROGRESS

Q1: Outline key fabrication steps for a PN diode.
Q2. Differentiate the two Silicon crystal growth techniques in terms of their applications.
Q3: Compare dry vs. wet oxidation rates/applications.
Q4: What are the advantages of the RTA over furnace annealing?
Q5: Define implantation range and straggle.

K 13.8 ANSWERS TO CHECK YOUR PROGRESS

Answer to Q1: Key steps of PN diode fabrication are: wafer cleaning; oxidation; lithography; oxide etching; ion implantation; annealing and diffusion; sputtering Al; lithography; metal etching; CVD nitride deposition; lithography and bonding window etching; removal of oxide from back side of wafer; deposition of Au on back side; and dicing and packaging.

Answer to Q2: Czochralski (CZ) process melts polycrystalline Si at ~1420°C, dips a seed crystal, and pulls an ingot with controlled dopants for wafers. It allows to grow large-diameter wafers with precise control of growth rate and doping concentration. Float Zone (FZ) technique is used to produce single-crystal silicon, where a molten zone is moved along a polysilicon rod via RF heating for ultra-high purity. It is suitable to fabricate high-power semiconductor devices with low impurity crystals.

Answer to Q3: Dry oxidation ($Si + O_2 \rightarrow SiO_2$) is slower for thin gate oxides; wet oxidation ($Si + 2H_2O \rightarrow SiO_2 + 2H_2$) is faster for thick field oxides due to higher H_2O diffusivity.

Answer to Q3: RTA enables rapid heating (up to 100°C/s via lamps) and short cycles (seconds vs. 30+ minutes), minimizing dopant diffusion for shallower junctions. It also provides precise temperature control (±1%) and uniformity, ideal for post-implant activation and damage repair without excessive thermal budget.

Answer to Q5: Ion implantation accelerates dopant ions to keV-MeV energies, creating Gaussian profiles defined by range (R) and straggle (ΔR), followed by RTP anneal for activation.

O 13.9 FURTHER READING

- Jaeger, R. C. Introduction to Microelectronic Fabrication, Vol. 5, in The Modular Series on Solid State Devices, 2nd ed., G. W. Neudeck and R. F. Pierret. Reading, MA: Addison-Wesley, 2002, pp. 49, 24.
- Sze, S. M. VLSI *Technology*, 2nd ed. New York: McGraw-Hill Book Company, 1988.
- Pierret, R. F. *Semiconductor Device Fundamentals*. Reading MA: Addison-Wesley Publishing Company Inc., 1996, 158 and 167.

Further Reading

- S.M. Sze and Y. Li, *Semiconductor Devices: Physics and Technology*, 4th ed., Wiley, 2020.
- J.D. Plummer et al., *Silicon VLSI Technology*, Prentice Hall, 2009.
- Chenming Hu, *Modern Semiconductor Devices for Integrated Circuits*. 1st ed., Pearson/Prentice Hall, 2010.
- Wolf, S., and R. N. Tauker. *Silicon Processing for the VLSI Era*, 2nd ed. Sunset Beach, CA: Lattice Press, 2000.

M 13.10 MODEL QUESTIONS

Q1. Compare Czochralski (CZ) and Float Zone (FZ) wafer growth methods.
Q2. What are the steps of RCA cleaning?
Q3. Explain predeposition vs. drive-in in diffusion doping.
Q4. Describe RTP advantages over furnace annealing for ion implant repair.
Q5. Differentiate between dry and wet oxidation along with the relevant chemical formula.
Q6. Write the ion implantation dose formula along with the significance of all the parameters.
Q7. What is TED in annealing?

****

UNIT 14: FABRICATION OF P–N JUNCTIONS – PART II

UNIT STRUCTURE

14.1 Learning Objectives
14.2 Introduction
14.3 Thin Film Deposition and Patterning Processes
14.4 Chemical Vapour Deposition (CVD)
14.5 Photolithography
14.6 Etching Techniques
14.7 Metallization
14.8 Let Us Sum Up
14.9 Answers to Check Your Progress
14.10 Further Reading
14.11 Model Questions

14.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- discuss about the thin film deposition techniques
- describe the photolithography steps and advanced variants
- explain wet and dry etching processes and their differences
- outline metallization for IC interconnections.

14.2 INTRODUCTION

In the previous unit, we discussed the basic process involved in the fabrication of a P-N junction. These steps include wafer preparation, thermal oxidation, diffusion/ion implantation for doping, and rapid thermal processing (RTP) for annealing. To accomplish these process steps, multiple thin films need to be deposited and certain portions of them selectively removed. The deposition of thin films on the wafer substrate is done using techniques such as chemical and physical vapour deposition. The removal of a specific portion of the deposited films is achieved using the lithography technique. In this unit, we will discuss about the thin film deposition and patterning processes used in semiconductor device fabrication. The unit covers the principles and processes of Chemical Vapour Deposition (CVD) and photolithography. Finally, the unit will introduce you to etching techniques, which are used at various stages of semiconductor fabrication, including both the front-end and post-junction (back-end) processes.

14.3 THIN FILM DEPOSITION AND PATTERNING PROCESSES

In IC fabrication, various materials such as silicon dioxide, silicon nitride, and metal films are deposited onto silicon wafers. These deposited layers are generally non-crystalline in nature. Thin film deposition methods are broadly divided into two main categories: physical vapour deposition (PVD) and chemical vapour deposition (CVD).

In PVD processes, the source material in solid form is first converted into vapour by physical means and then deposited onto the substrate as a thin film. It produces durable, precise coatings atom by atom without chemical reactions. Different PVD techniques are briefly discussed below, while CVD techniques are discussed in the subsequent subsection.

Thermal Evaporation

In the thermal evaporation method, the source material is heated using resistive heating element until it vaporises in a vacuum environment, which is ideal for pure metallic coatings like optics. The process is carried out in a high-vacuum chamber (typically 10^{-6} to 10^{-9} Torr), and a heating source, such as resistive filament, electron beam, etc., is used to heat the depositing solid material. The source raises the temperature of the solid source material (mostly metals or alloys) to its evaporation point (often 500–2500°C), creating a vapour plume. This vapour travels collision-free to the cooler substrate above, where it condenses, nucleates, and builds into a uniform thin film. The substrate is rotated using rotating substrate holders to ensure even coverage.

In case of the resistive filament (or boat) heating, an electric current is passed through a tungsten coil or crucible holding the material, and the material is evaporated due to the generated heat. This technique is simple and cost-effective and suitable for metals with low melting points (e.g. aluminum). In electron beam evaporation, a focused electron beam is directed onto the source material, causing it to vaporise. This method is particularly effective for depositing high-melting-point materials with good precision, while also reducing contamination. Figure 14.1(a) and (b) illustrate the schematic diagrams of thermal evaporation and electron beam evaporation systems, respectively. These techniques excel in producing pure, optical-quality thin films at low cost but limit film uniformity on complex shapes and suit lower-melting materials only.

Sputtering

In sputtering, the source material and the substrate holding the silicon wafer act as two electrodes connected to a high-voltage supply. The process begins by evacuating the chamber, followed by the introduction of a low-pressure inert gas, typically argon (Ar). When a voltage is applied, the gas becomes ionized, forming a plasma between the electrodes. The target is kept at a negative potential, which accelerates Ar ions toward it. These ions strike the target surface, causing atoms to be ejected and subsequently deposited onto the wafer as a thin film. A schematic of this process is shown in Figure 14.1(c).

Direct current (DC) power is commonly used for metal deposition, while radio frequency (RF) power is required for insulating materials. In reactive sputtering, chemical reactions occur during deposition; for instance, sputtering titanium in a nitrogen environment results in the formation of a titanium nitride (TiN) film on the wafer.

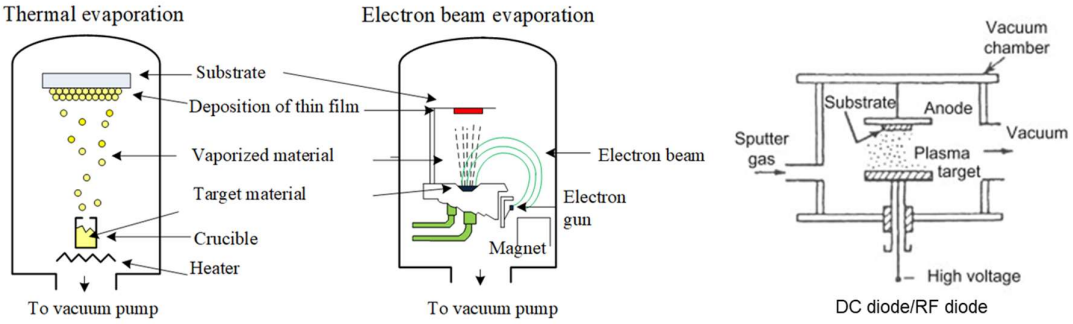


Figure 14.1: Schematic diagrams of different physical vapour deposition (PVD) systems; (a) thermal evapouration process [1], (b) e-beam evapouration process [1], and (c) sputter deposition process [2].

Epitaxy

Like conventional deposition techniques, which generally yield amorphous or polycrystalline films, epitaxy produces a crystalline layer that extends the lattice structure of the underlying substrate. This technique is especially useful for forming junctions by growing a lightly doped silicon layer on a heavily doped base. Epitaxial growth can also involve depositing a different material on the substrate, provided their lattice constants are closely matched. Such interfaces between dissimilar semiconductors are known as heterojunctions (e.g., SiGe grown on Si) and are widely used in devices such as light-emitting diodes and laser diodes.

14.4 CHEMICAL VAPOUR DEPOSITION (CVD)

Physical vapour deposition (PVD) methods are relatively simple and effective for coating flat surfaces; however, they are highly directional and often fail to provide uniform coverage on sidewalls, trenches, or stepped features—a limitation known as the step coverage problem. In contrast, chemical vapour deposition (CVD) produces more conformal films, enabling nearly uniform thickness over both horizontal and vertical surfaces. In CVD, the film is formed through gas-phase processes, where precursor gases either decompose or chemically react at the wafer surface to create the desired layer. The CVD process is routinely used to deposit films of SiO_2, Si_3N_4 (a dielectric with excellent chemical and electrical stability), and polycrystalline silicon (poly-Si). The associated chemical reactions for these materials in the CVD deposition process are given in equations (14.1) to (14.4).

For, poly-Si, SiH_4 (silane) $\rightarrow Si + 2H_2$ (14.1)

For Si_3N_4, $3SiH_2Cl_2$ (dichlorosilane) $+ 4NH_3 \rightarrow Si_3N_4 + 6HCl + 6H_2$ (14.2)

For SiO_2: $SiH_4 + O2 \rightarrow SiO_2 + 2H_2$ (14.3)

For high-temperature SiO_2, $SiH_2Cl_2 + 2H_2O \rightarrow SiO_2 + 2HCl + 2H_2$ (14.4)

High-temperature oxide (HTO) deposition provides excellent conformal coverage, as elevated temperatures enhance surface mobility of deposited species, leading to uniform coating even along sidewalls. Common CVD techniques include low-pressure

chemical vapour deposition (LPCVD) and plasma-enhanced chemical vapour deposition (PECVD). LPCVD operates at reduced pressure, resulting in improved thickness uniformity and more efficient gas usage. In PECVD, plasma-generated electrons transfer energy to the reactant gases, promoting chemical reactions and enabling film deposition at lower temperatures. Schematic representations of LPCVD and PECVD systems are shown in Figure 14.2(a) and (b), respectively. Dopants can also be incorporated during the CVD process itself, a technique known as in-situ doping, which is used to achieve highly doped silicon films.

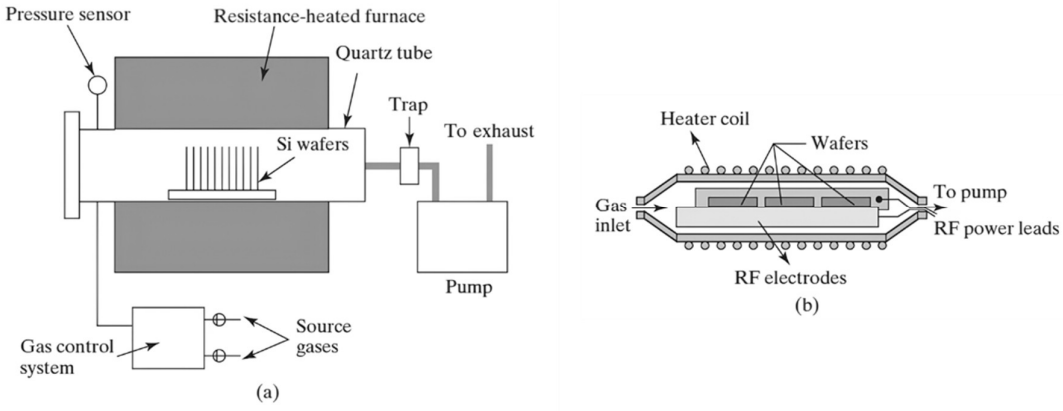


Figure 14.2: Diagrammatic representation of (a) an LPCVD system and (b) a PECVD reactor chamber with plasma generated radio-frequency power.

14.5 PHOTOLITHOGRAPHY

Selective doping or thin-film deposition on specific regions of a wafer requires prior removal of surface layers, such as thermal oxide, using photolithography (optical lithography). The main stages of this process are illustrated in Figure 14.3 for the patterning of an SiO_2 layer. First, a uniform layer of photoresist, which is sensitive to ultraviolet (UV) radiation, is applied to the wafer by spin coating and then soft-baked to enhance adhesion and remove solvents (step (a), Fig. 14.3). The coated wafer is subsequently exposed to UV light through a patterned photomask using a high-resolution optical system (step (b), Fig. 14.3). The photomask, typically fabricated from quartz, carries the desired pattern that is transferred onto the photoresist layer. The opaque regions on the mask block the UV light, while regions of the photoresist exposed to the light undergo different chemical reactions depending on the type of resist used. In negative resists, the areas where the light strikes become polymerized and hard to dissolve in solvents. When placed in a developer (solvent), the polymerized regions remain, while the unexposed regions dissolve and wash away (see right-hand side of figure 14.3(c)). Positive resists contain a stabilizer, which breaks down when exposed to light, leading to the preferential removal of the exposed regions (see left-hand side of figure 14.3(c)). Although photolithography includes only these steps, an additional step of oxide removal is essential to achieve the desired formation on the wafer (see figure 14.3(d)). The removal of exposed regions of the film is carried out through a process known as etching, which is discussed in detail in the following subsection. For oxide layers, buffered hydrofluoric acid (HF) is commonly used to selectively dissolve the unprotected areas. After etching, the remaining photoresist is

eliminated in a step called resist stripping, which can be performed using chemical solutions or by oxidizing the resist in an oxygen plasma or a UV–ozone environment.

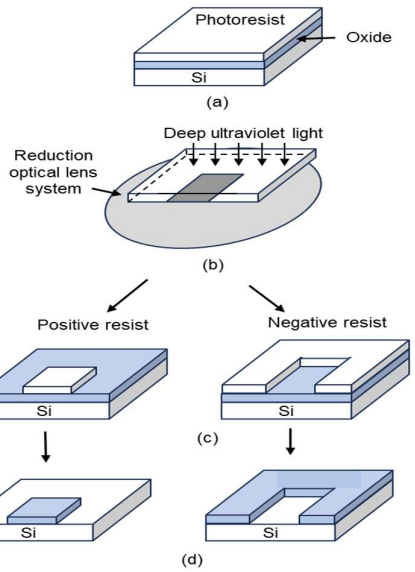


Figure 14.3: Schematic of the key lithography steps: (a) photoresist coating; (b) exposure through a mask using an optical reduction system; (c) development of the resist; and (d) oxide etching followed by resist removal.

Lithography is considered one of the most complex and costly steps in IC fabrication. In a typical process flow, it is repeated multiple times—often more than 20 cycles—each requiring a different photomask. The smallest feature that can be patterned on the wafer, known as the lithographic resolution, is fundamentally limited by optical diffraction. This resolution is directly proportional to the wavelength of the light used during exposure, meaning shorter wavelengths enable finer feature sizes.

$$\text{Lithography Resolution} = k\lambda \qquad (14.5)$$

The factor k depends on the lens system and the photomask technology. A straightforward way to extend the resolution limit is to use UV light of shorter and shorter wavelengths that correspondingly reduce the resolution limit. Laser light sources of 248 and 193 nm (deep UV) are widely used. Further reduction of the wavelength is difficult due to the lack of suitable transparent materials for lenses and mask plates for such small wavelengths.

Wet Lithography

The limitation of conventional optical lithography at wavelengths below 193 nm arises from the difficulty in developing suitable lens and mask materials. To overcome this constraint without reducing the exposure wavelength, immersion (wet) lithography is employed. In this technique, the gap between the projection lens and the wafer is filled with a liquid, usually water. As light enters the liquid medium, its effective wavelength decreases in proportion to the refractive index, resulting in improved lithographic resolution, as indicated by Eq. (14.5). The resolution can be further enhanced by using liquids with refractive indices higher than that of water.

Extreme UV Lithography

An advanced form of optical lithography is extreme ultraviolet lithography (EUVL), which operates at a wavelength of about 13 nm. This drastic reduction in wavelength enables significantly higher patterning resolution. However, EUV radiation is strongly absorbed by most materials, so the system relies entirely on reflective optics rather than conventional lenses. Both the projection system and the photomask are mirror-based, functioning through reflection instead of transmission. To maintain accuracy, the optical components must possess smooth and flat surfaces, with roughness on the order of 0.25 nm, comparable to atomic dimensions. EUV radiation is typically generated by directing high-energy laser pulses onto a stream of xenon (Xe) gas, producing the required short-wavelength light.

Electron Lithography

Electron-beam lithography (EBL) is a high-resolution alternative to optical lithography, capable of defining extremely fine features. In this technique, a focused beam of electrons is directed onto an electron-sensitive resist, where it locally modifies the material to form the desired pattern. Unlike optical lithography, EBL is a maskless process in which the pattern is generated directly on the wafer by scanning the beam under computer. Because the pattern is written sequentially, EBL suffers from low throughput compared to optical methods. Although the use of multi-beam systems can improve exposure rates, the process remains relatively slow. Despite this limitation, EBL is widely used for fabricating high-precision photomasks for conventional optical lithography as well as for extreme ultraviolet lithography (EUVL).

14.6 ETCHING TECHNIQUES

After the pattern is formed in the resist by lithography, the resist pattern is often transferred to the underlying film, and needs to be removed subsequently. This process is called etching. Depending upon the type of interaction between the wafer surface and etchant (the components used for etching), the etching process can be classified broadly into two types; wet and dry etching. In wet etching, the exposed wafer is typically immersed in or sprayed with the etchant, which is typically a liquid chemical solution (etchants like acids or bases). It allows isotropic etching that proceeds equally in all directions due to chemical reactions. Therefore, the etched features are generally larger than the dimensions of the resist patterns. Figure 14.4(a) shows isotropic etching of SiO_2 during wet etching using HF. Dry etching, commonly referred to as plasma etching or reactive-ion etching (RIE), involves exposing a patterned wafer to a plasma environment. This plasma consists of a quasi-neutral mixture of ions, electrons, and reactive species generated using a radio-frequency (RF) electric field. The reactive species chemically interact with the exposed material, while energetic ions are accelerated toward the wafer surface, producing directional bombardment that physically removes material. This combined chemical and physical action enables controlled and anisotropic etching. The ion bombardment in dry etching is highly directional and lead to preferential material removal along the vertical direction. As a result, dry etching is typically anisotropic, making it well suited for fabricating fine features. By optimizing reactor design and selecting appropriate etching chemistries, nearly vertical sidewalls can be achieved, as shown in Figure 14.4(b). Operating at low pressures with a strongly directed electric field further enhances anisotropy. However,

by adjusting process conditions, dry etching can also be tailored to produce isotropic or partially anisotropic profiles when required.

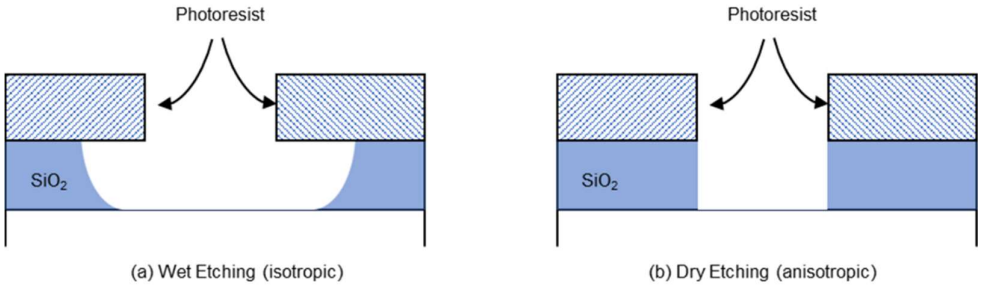


Figure 14.4: Schematic of the etching process of SiO_2; (a) wet (isotropic) etching and (b) dry (anisotropic) etching.

Appropriate gases are introduced into the etching chamber depending on the target material. For instance, fluorine-based plasmas etch silicon and its compounds, while chlorine plasmas handle aluminum. Dry etching generally offers lower material selectivity compared to wet etching. Both the target layer and the substrate beneath it (such as SiO_2 over silicon) can suffer notable damage during the process. Thus, the etching must stop precisely once the intended layer is cleared. There is often a trade-off between selectivity and anisotropy, for example, bromine delivers superior Si-to-SiO_2 selectivity but reduced directional control versus chlorine. Plasma-based processing risks device damage on the wafer, termed plasma-induced or charging damage. This primarily stems from ions charging conductive elements, creating excessive voltage across thin oxides and triggering breakdowns.

14.7 METALLIZATION

During integrated circuit fabrication, electrical connections to diffused regions and interconnections between devices on a wafer are established by depositing patterned metal layers, a process known as metallization. For forming an N-type contact in a PN junction, the silicon dioxide (SiO_2) layer is first selectively etched to expose the required contact regions. A thin metal film, typically aluminium, is then deposited over the wafer surface, commonly using sputtering. Finally, lithography combined with dry etching is used to remove excess metal, leaving behind well-defined contact patterns.

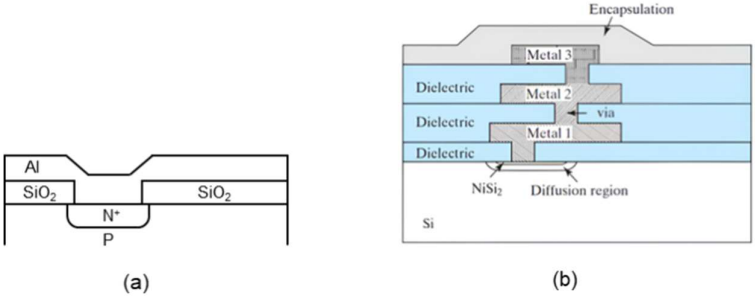


Figure 14.5: Diagrammatic representation of metallization: (a) basic metallization for silicon contact and (b) multilevel metallization structure.

Multilevel metallization is essential in modern IC fabrication for achieving high-density and complex circuit designs. Multiple metal layers—sometimes up to ten—are stacked, with thicknesses ranging from submicron levels to several microns. Thin layers are used for signal routing, while thicker layers handle power distribution. These layers are separated by intermetal dielectrics, and vertical connections are formed using vias. To reduce contact resistance between metal lines and heavily doped (N^+/P^+) regions, silicides such as $NiSi_2$ are commonly introduced. A key reliability concern with aluminium interconnects is electromigration, where prolonged current flow causes atomic displacement, leading to voids and possible circuit failure. To overcome this, copper has become the preferred interconnect material due to its lower resistivity and superior resistance to electromigration. Copper is deposited using electroplating or chemical vapour deposition, and excess material is removed through chemical–mechanical polishing (CMP) to ensure a planar surface. In earlier technologies, silicon dioxide (SiO_2) served as the primary interlayer dielectric. However, it has largely been replaced or supplemented by low-k materials, often incorporating carbon or fluorine, to reduce parasitic capacitance. Lower capacitance improves circuit speed, decreases power consumption, and minimizes crosstalk between interconnects. Due to the complexity and number of processing steps involved, interconnect fabrication constitutes a significant portion of the total IC manufacturing cost.

W CHECK YOUR PROGRESS

Q1: Compare PVD and CVD deposition techniques.
Q2: State photolithography steps for patterning SiO_2.
Q3: Differentiate wet vs. dry etching.
Q4: What do you mean by metallization for PN junction contacts.
Q5: Why use advanced lithography like EUVL?

2 14.8 LET US SUM UP

- Thin films are deposited via PVD (evaporation, sputtering) or CVD (LPCVD, PECVD) for conformal coverage.
- Photolithography patterns wafers using photoresist, masks, and UV/exotic light sources like EUV.
- Etching transfers patterns: wet (isotropic, chemical) vs. dry (anisotropic, plasma-based).
- Metallization creates multilevel interconnects with Cu/low-k dielectrics, addressing electromigration via CMP.

K 14.9 ANSWERS TO CHECK YOUR PROGRESS

Answer to Q1: PVD (e.g., evapouration, sputtering) converts solid sources to vapour physically for line-of-sight deposition on flat surfaces but suffers step coverage issues. CVD uses gas-phase reactions for conformal films on complex topographies, ideal for SiO_2, Si_3N_4, poly-Si via reactions like $SiH_4 \rightarrow Si + 2H_2$.

Answer to Q2: Spin-coat photoresist, soft-bake, expose through mask (positive/ negative resist), develop, etch oxide (e.g., HF), strip resist. Resolution = kλ limits features by wavelength λ.

Answer to Q3: Wet etching is isotropic (chemical, equal all directions, undercuts mask). Dry etching (plasma/RIE) is anisotropic (vertical via ion bombardment), enabling finer features; uses F for Si/SiO_2, Cl for Al.

Answer to Q4: Remove oxide via lithography/etch, sputter Al/Cu, pattern/etch metal, anneal for ohmic contact. Multilevel uses vias, low-k dielectrics, silicides (e.g., $NiSi_2$) to cut resistance; Cu replaces Al for better electromigration reliability.

Answer to Q5: EUVL (13 nm) overcomes optical limits with reflective optics for sub-10 nm features; essential beyond 193 nm immersion lithography.

O 14.10 FURTHER READING

- Physical Vapour Deposition. (2018, July 3). PV Manufacturing. https://pv-manufacturing.org/solar-cell-manufacturing/physical-vapour-deposition/
- Adachi, H., & Wasa, K. (2012). *Thin films and nanomaterials. In Handbook of sputtering technology* (2nd ed.). William Andrew Publishing.
- Hu, C. (2010). *Modern semiconductor devices for integrated circuits* (1st ed.). Pearson/Prentice Hall.
- Plummer, J. D., Deal, M. D., & Griffin, P. B. (2009). *Silicon VLSI technology*. Prentice Hall.
- Sze, S. M., & Li, Y. (2020). *Semiconductor devices: Physics and technology* (4th ed.). Wiley.
- Wolf, S., & Tauber, R. N. (2000). *Silicon processing for the VLSI era* (2nd ed.). Lattice Press.

M 14.11 MODEL QUESTIONS

Q1. Compare PVD and CVD.
Q2. Describe photolithography steps.
Q3. Differentiate wet vs. dry etching.
Q4. Explain EUV lithography challenges.
Q5. What is electromigration in metallization?
Q6. Explain the step coverage problem in PVD and how CVD addresses it.
Q7. Describe reactive sputtering with an example.
Q8. What is chemical-mechanical polishing (CMP) and its role in copper interconnects?
Q9. Discuss plasma process-induced damage mechanisms in dry etching.
Q10. Why are low-k dielectrics used in modern IC metallization?

****