Chapter I: Semiconductor Diodes

Overview

This chapter introduces the fundamental concepts of semiconductor physics, including the nature of semiconducting materials, doping mechanisms, and the formation of *n*-type and *p*-type regions. A quantitative analysis of the PN junction is developed in order to explain its operating principles. The effect of an externally applied voltage on the junction is examined through the derivation and interpretation of the current–voltage characteristic. The response of the junction under time-varying signals is also discussed. The chapter concludes with an overview of essential electronic functions implemented using diode-based circuits.

1. Fundamental Concepts

1. Concepts of Energy Band Theory

1.1 Definition:

If several isolated atoms are brought together to form a crystalline structure, their energy levels interact, leading to a series of closely spaced levels organized into allowed energy bands and forbidden bands.

1.2 Energy Levels in Solids:

Within the crystal lattice, the energy levels can be categorized, in order of increasing energy, as follows:

a) Valence Band:

This band can be occupied by valence electrons when they are in their lowest energy states.

b) Forbidden Band (Band Gap):

This band contains no electrons. The width of the forbidden band, known as the band gap, plays a crucial role in determining the electrical properties of materials.

c) Conduction Band:

This band can be occupied by electrons that have gained sufficient energy to overcome the attraction of the nucleus.

- If this band is empty, the material is an insulator, with a band gap of several electronvolts (eV).
- If the conduction band is partially filled, the material is called a semiconductor, characterized by a small band gap.
- If the band gap is very small, the material behaves as a conductor.

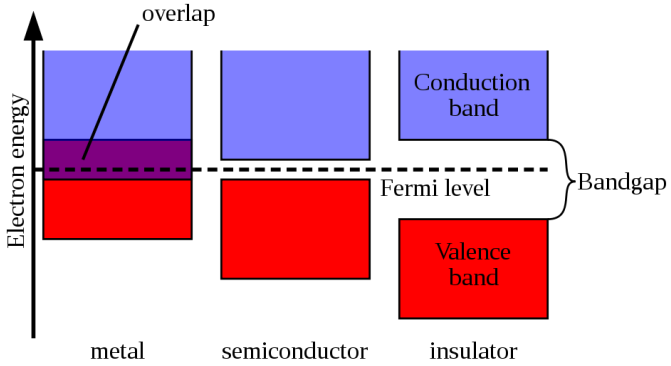


Figure I.1. Energy Levels in Solids.

1.3 Semiconducting Materials

Semiconductors are materials whose electrical conductivity falls between that of conductors and insulators. While charge carriers are not as abundant as in metals, their concentration is sufficient to permit controlled electrical conduction under appropriate conditions. This intermediate behavior makes semiconductors particularly suitable for electronic applications.

The electrical properties of semiconductors are explained using energy band theory, which describes the allowed and forbidden energy levels of electrons within a solid. In semiconductors, the energy gap separating the valence band from the conduction band is relatively narrow. As a result, electrons may acquire enough energy to move into the conduction band, thereby contributing to electrical current. When an electric field is applied, current arises from both free electrons and the effective positive charges known as *holes*.

Among all semiconducting materials, silicon is the most extensively used due to its stability, favorable electronic properties, and wide availability. Nevertheless, alternative materials such as germanium, gallium arsenide, and silicon carbide are employed in specific applications requiring higher frequencies, temperatures, or power levels.

1.4 Intrinsic Semiconductors

An intrinsic semiconductor refers to a material that is free of intentionally introduced impurities. In this idealized case, the electrical behavior is governed exclusively by the crystal lattice and thermal effects. Although absolute purity cannot be achieved in practice, highly refined monocrystalline silicon provides a close approximation to intrinsic behavior.

In intrinsic semiconductors, charge carriers originate primarily from thermal excitation, which generates electron–hole pairs. The concentrations of electrons and holes are equal, resulting in very low conductivity under normal conditions. Significant electrical conduction occurs only at elevated temperatures, where thermal energy increases the number of available charge carriers.

1.4.1 Doping

The existence of forbidden energy bands arises from the periodic and orderly arrangement of atoms within the crystal lattice. Any modification of this regular structure may introduce additional energy levels within the band gap, thereby altering the electronic behavior of the material. **Doping** is the intentional introduction of carefully selected impurity atoms into an intrinsic semiconductor with the purpose of tailoring its electrical properties.

Through doping, the concentration of charge carriers within the semiconductor is significantly modified. When the process results in an increase in free electrons, the material is referred to as **n-type**. Conversely, when the concentration of holes is increased, the material is classified as **p-type**. Semiconductors modified in this manner are known as **extrinsic semiconductors**.

The controlled addition of dopant atoms allows precise regulation of electrical conductivity by creating either an excess of electrons or a deficiency of electrons (holes). By bringing regions with different types of doping into contact, **semiconductor junctions** are formed. These junctions enable control over both the direction and magnitude of electric current, forming the fundamental operating principle of modern electronic devices such as **diodes, transistors, and integrated circuits**.

1.4.2 N-Type Doping

N-type doping is achieved by increasing the population of free electrons within the semiconductor material. This is accomplished by incorporating atoms that possess more valence electrons than the host semiconductor atoms.

In the case of silicon, each silicon atom has four valence electrons that participate in covalent bonding with neighboring atoms, resulting in a tetrahedral crystal structure. When silicon is doped with elements from **Group V** of the periodic table—such as **phosphorus (P)**, **arsenic (As)**, or **antimony (Sb)**—the dopant atom forms four covalent bonds while retaining one additional electron.

This extra electron is weakly bound to the dopant atom and requires only a small amount of thermal energy to be promoted into the conduction band. At room temperature, the majority of these electrons become free charge carriers. Since this process does not generate corresponding holes, electrons largely outnumber holes in the material. As a result, **electrons act as majority carriers**, while **holes are minority carriers**. Because these dopant atoms contribute excess electrons to the crystal lattice, they are referred to as **donor atoms**.

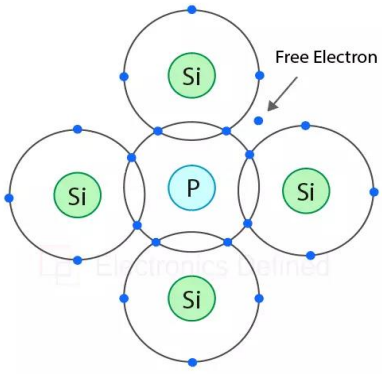


Figure I.2. Illustration of donor impurity doping in a semiconductor lattice (phosphorus atom)

1.4.3 P-Type Doping

P-type doping is achieved by increasing the concentration of **holes** within a semiconductor material. This is accomplished by introducing impurity atoms that possess fewer valence electrons than the atoms of the host lattice, thereby creating electron deficiencies.

In silicon-based semiconductors, this process typically involves the incorporation of **trivalent elements** from Group III of the periodic table, such as **boron** or **indium**. Since these atoms have only three valence electrons, they are able to form only three covalent bonds with neighboring silicon atoms. The incomplete bonding results in the creation of an empty electronic state, known as a **hole**, within the crystal structure.

An electron from a nearby silicon atom may occupy this vacant state, leaving behind a new hole in its original position. This mechanism allows holes to propagate through the lattice and contribute to electrical conduction. When the concentration of acceptor impurities is sufficiently high, holes become the dominant charge carriers, while electrons are present only in small numbers.

In such materials, **holes are the majority carriers**, whereas **electrons act as minority carriers**. Dopant atoms that create holes by accepting electrons are referred to as **acceptor atoms**.

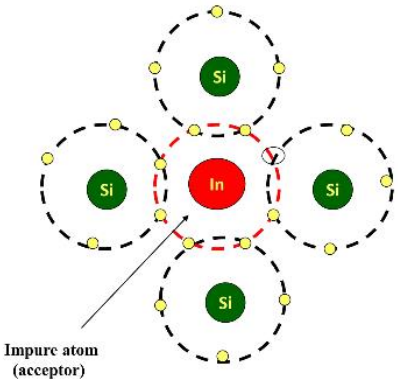


Figure I.3. Schematic illustration of acceptor impurity doping in a semiconductor lattice (indium atom).

2. PN Junction

A **PN junction** is formed when two regions of the same semiconductor crystal, doped respectively with donor (n-type) and acceptor (p-type) impurities, are brought into direct contact. This junction represents the fundamental building block of many semiconductor devices and enables controlled charge transport across the interface.

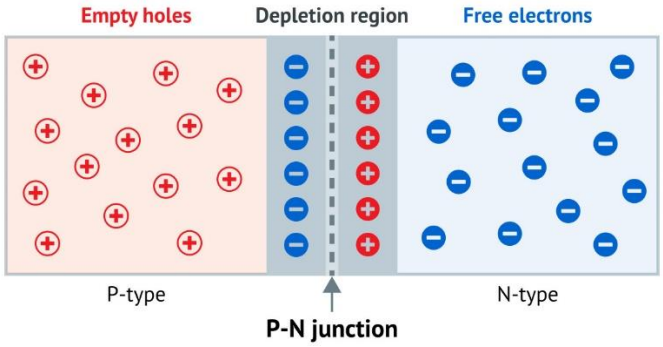


Figure I.4. PN junction

When a p-type semiconductor region is brought into contact with an n-type region, a **PN junction** is formed at their interface. Due to the concentration gradient

of charge carriers, electrons from the n-type region diffuse toward the p-type region, while holes from the p-type region diffuse toward the n-type region. As electrons enter the p-type region, they recombine with holes, and similarly, holes entering the n-type region recombine with electrons.

This diffusion process progressively removes mobile charge carriers from the vicinity of the junction. As a consequence, a region depleted of free carriers develops near the interface. In this zone, immobile ionized donor atoms remain on the n-type side, while ionized acceptor atoms remain on the p-type side. This region, commonly referred to as the **depletion region** or **space-charge region**, is no longer electrically neutral.

The separation of fixed positive and negative charges across the junction creates an internal electric field analogous to that found between the plates of a capacitor. This internal field is directed from the n-type region toward the p-type region and acts to oppose further diffusion of charge carriers. Any electron or hole entering the depletion region experiences a force that drives it back toward its region of origin.

As carrier diffusion continues, the density of fixed charges increases, strengthening the internal electric field. This process persists until the electric field becomes sufficiently strong to counterbalance carrier diffusion. At this point, a **potential barrier** is established, preventing further net movement of charge carriers across the junction. The system then reaches a state of **thermal equilibrium**, in which the internal electric field and the space-charge distribution remain constant.

2.1 Biasing of the PN Junction

2.1.1 Forward Bias

A PN junction operates under forward bias when the p-type region is connected to the positive terminal and the n-type region to the negative terminal of an external voltage source. In this configuration, the externally applied electric field acts in opposition to the built-in electric field established across the depletion region.

At low applied voltages, the internal potential barrier remains dominant, preventing charge carriers from crossing the junction. As the applied voltage increases, the height of this barrier is progressively reduced. Once the applied voltage exceeds a characteristic threshold, the depletion region narrows sufficiently to allow a substantial flow of charge carriers.

Under these conditions, electrons from the n-type region are able to move into the p-type region, while holes from the p-type region migrate toward the n-type region. This bidirectional carrier motion results in a significant electric current through the junction, and the diode is said to be in its conducting state. The approximate forward threshold voltage is material-dependent and is typically around 0.2 V for germanium and 0.6–0.7 V for silicon.

2.1.2 Reverse Bias

A PN junction is said to be reverse biased when the n-type region is connected to the positive terminal and the p-type region to the negative terminal of an external voltage source. In this arrangement, the electric field produced by the external voltage reinforces the internal electric field of the junction.

As a consequence, majority charge carriers are driven away from the junction interface, leading to an increase in the width of the depletion region. This expansion of the space-charge region further inhibits carrier transport across the junction.

In the reverse-biased condition, no appreciable current flows through the junction. Only a very small current, known as the reverse saturation current, persists due to the motion of minority carriers, whose drift is assisted by the strengthened electric field. This current is typically negligible, allowing the reverse-biased PN junction to be modeled as a very high resistance or, in idealized analyses, as an open circuit.

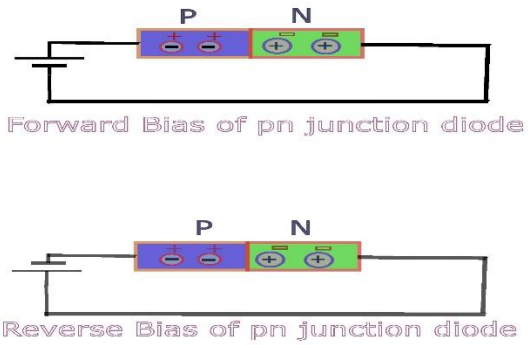


Figure I.5. Forward and Reverse biased junction.

3. The junction diode

3.1 Structure and Physical Principle

A **junction diode** is a semiconductor device formed by a single PN junction. It constitutes one of the most fundamental components in electronic systems and serves as the basis for many more complex semiconductor devices.

The diode is created by joining p-type and n-type regions within the same semiconductor crystal, resulting in a junction that permits current flow primarily in one direction. This directional behavior arises from the internal electric field established at the PN interface. A schematic representation of the diode's internal

semiconductor structure, together with its standardized circuit symbol, is shown in Figure I.6.

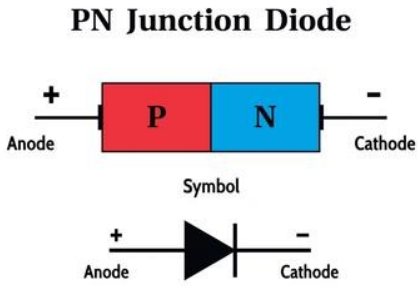


Figure I.6. Semiconductor structure of a junction diode and its corresponding circuit symbol.

3.2 Current–Voltage Characteristic of a Diode

The electrical behavior of a diode is described by the relationship between the voltage applied across its terminals and the resulting current. This relationship is nonlinear and is commonly modeled by an exponential law, expressed as:

$$I = I_S\left(e^{\frac{eV_d}{KT}} - 1\right) \tag{I.1}$$

where $\boldsymbol{I}$ is the diode current, $\boldsymbol{V_d}$ is the voltage across the junction, $\boldsymbol{I_S}$ is the reverse saturation current, $\boldsymbol{e}$ is the elementary charge, $\boldsymbol{k}$ is Boltzmann's constant, and $\boldsymbol{T}$ is the absolute temperature.

Under **forward bias**, the current increases rapidly once the applied voltage exceeds a certain minimum value known as the **threshold voltage** $\boldsymbol{V_0}$. This behavior is illustrated by the characteristic curve shown in Figure I.8. In contrast, under **reverse bias**, the diode conducts only a very small current approximately equal to

the saturation current I_S , until the reverse voltage reaches the **breakdown voltage** V_B. At this point, physical mechanisms such as avalanche breakdown lead to a sudden increase in current.

3.3 Practical Models of the Diode

For circuit analysis and practical calculations, the exact exponential behavior of the diode is often replaced by simplified equivalent models.

- In **forward-biased operation**, the diode may be modeled as an ideal voltage source of value V_0 in series with a small resistance R_d , representing the diode's dynamic resistance.
- In **reverse-biased operation**, the diode is approximated by a very large resistance R_i , reflecting its blocking behavior.

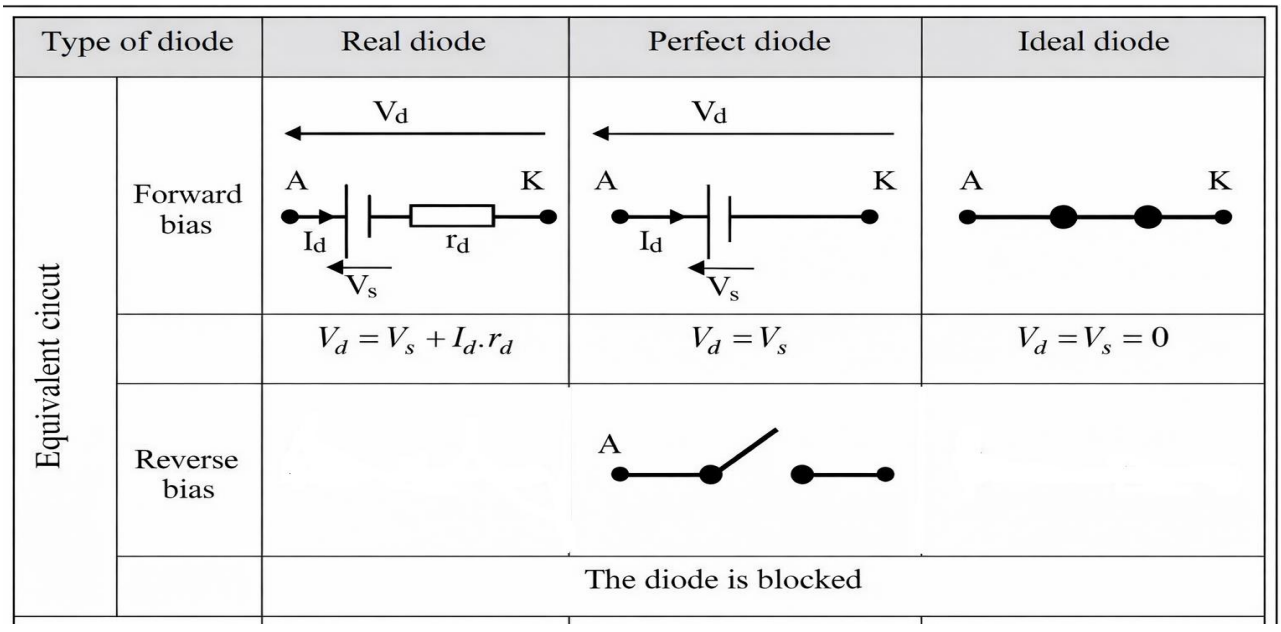


Figure I.7. Types of diode.

Typical values used in practice are:

- $V_0 \approx 0.7$ **V** for silicon diodes
- $V_0 \approx 0.3$ **V** for germanium diodes
- R_d : low forward resistance
- R_i : very high reverse resistance

These simplified representations allow efficient analysis of diode-based circuits without resorting to the full nonlinear model.

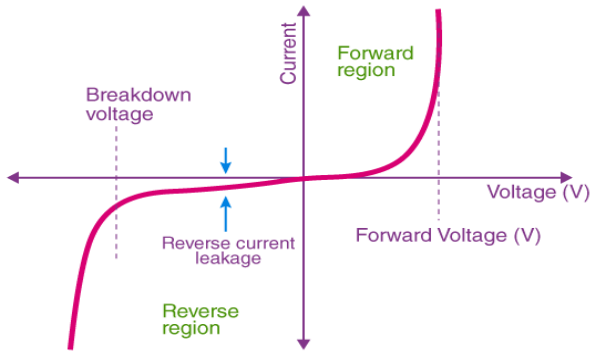


Figure I.8. Current–voltage characteristics of a junction diode in forward-conduction and reverse-blocking regimes.

In practice, and particularly when the threshold voltage V_0 is small compared with the electrical quantities of the circuit under study, diodes can be considered ideal. This assumption simplifies the analysis of diode circuits by modeling a conducting diode as a short circuit and a non-conducting diode as an open circuit.

The characteristic of the diode can be divided into three distinct operating regions:

- **Region 1 (Forward Bias):**

  In this region, the diode is forward biased. It conducts current only when the voltage across its terminals exceeds the threshold (cut-in) voltage. If the

applied voltage satisfies $0 < V_{AK} < V_{threshold}$, the diode remains in the non-conducting state. Conduction begins once the threshold voltage is surpassed.

- **Region 2 (Reverse Bias – Blocking State):**

  Here, the diode is reverse biased, meaning that $V_{AK} < 0$. Under these conditions, the device blocks current flow and operates in the off state.

- **Region 3 (Reverse Breakdown):**

  When the reverse voltage exceeds the maximum value that the diode can withstand, breakdown occurs at the PN junction. This excessive reverse bias leads to irreversible damage of the junction, rendering the diode permanently nonfunctional.

For analytical convenience, simplified characteristic models are commonly adopted to represent real, ideal, and perfect diodes.

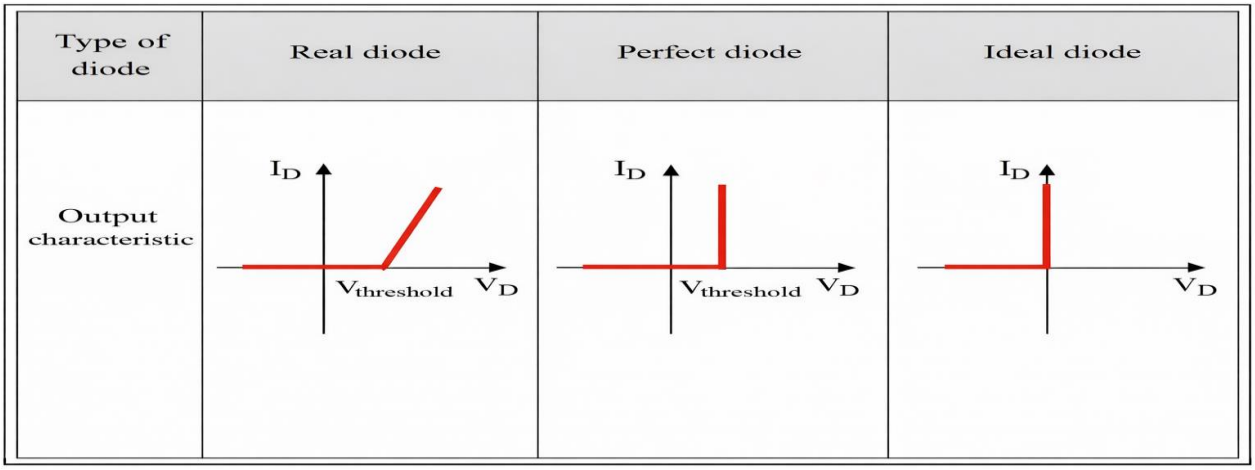


Figure I.9. Real, ideal, and perfect diodes.

4. Diode Applications

Diodes are widely used in electronic systems for various purposes. One of the most fundamental applications is rectification.

4.1 Rectification

Rectification refers to the process by which an alternating voltage is transformed into a unidirectional (DC) voltage. Two principal types of rectifiers are commonly distinguished: half-wave rectifiers and full-wave rectifiers.

(a) Half-Wave Rectification

Half-wave rectification represents the simplest rectifying configuration. In this arrangement, the circuit allows the positive half-cycles of the input signal to pass while suppressing the negative half-cycles. This function can be achieved by placing a single diode in series with the load resistor.

Assumption: The diode D is considered ideal.

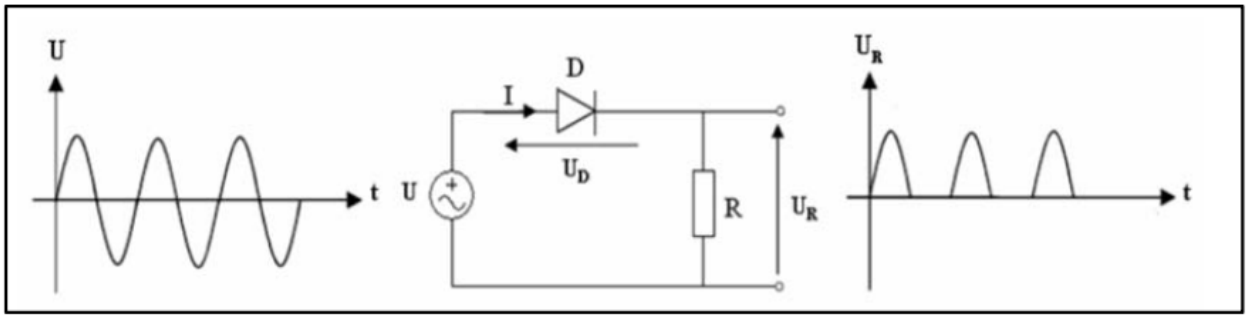


Figure I.10. Real, ideal, and perfect diodes.

The input voltage is assumed to be sinusoidal and can be expressed as:

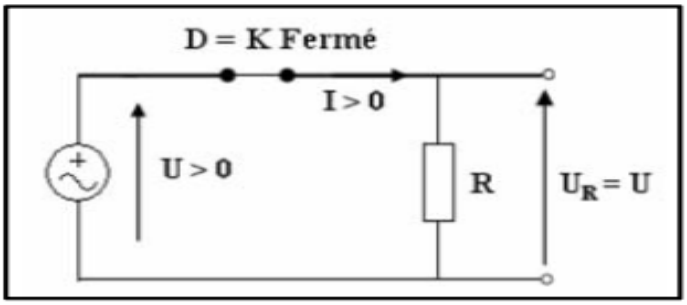
$$U = U_{\max} \sin(\omega t), \quad \text{with} \quad \omega = 2\pi f$$

During the positive half-cycle of the input voltage (U>0):

The diode is forward biased and therefore conducts. Under the ideal diode assumption, the current through the circuit is positive ($I > 0$) and the voltage drop across the diode is zero ($U_D = 0$).

Consequently, the voltage across the load resistor is:

$$U_R = U - U_D = U$$

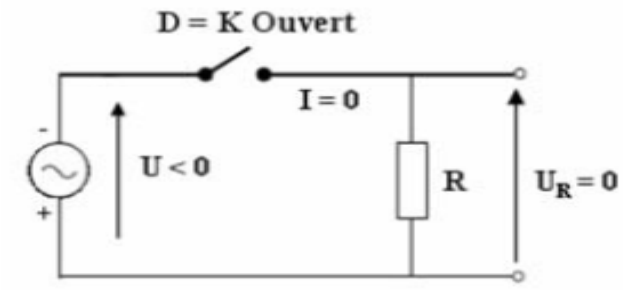


Thus, during the positive alternation, the load voltage follows the input voltage exactly.

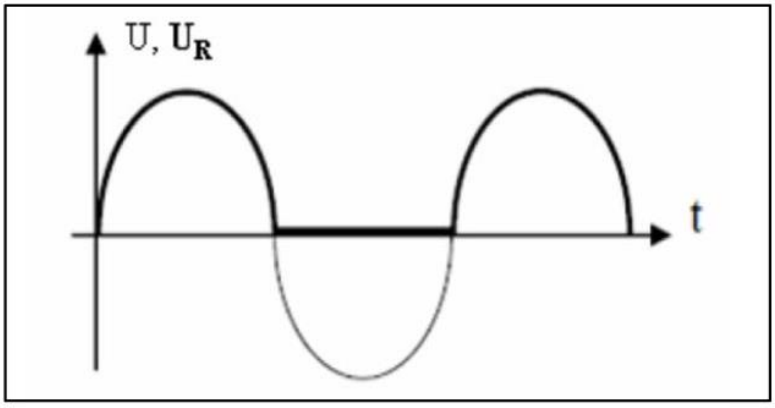
During the negative half-cycle of the input voltage ($U < 0$):

Under this condition, diode D is reverse biased and therefore operates in the blocking state. The circuit current is zero ($I = 0$), and the diode voltage is negative ($U_D < 0$). As a result, the voltage across the load resistor is:

$$U_R = 0$$



Thus, the diode conducts throughout the positive half-cycle and remains non-conducting during the negative half-cycle. Consequently, the circuit suppresses the negative portions of the input waveform, as illustrated in the figure. The resulting output waveform is referred to as a **half-wave rectified signal**. The load current is therefore unidirectional, meaning that it flows in only one direction.



(b) Full-Wave Rectification

In contrast to half-wave rectification, full-wave rectification makes use of both positive and negative half-cycles of the input AC signal Ve to produce a DC output voltage.

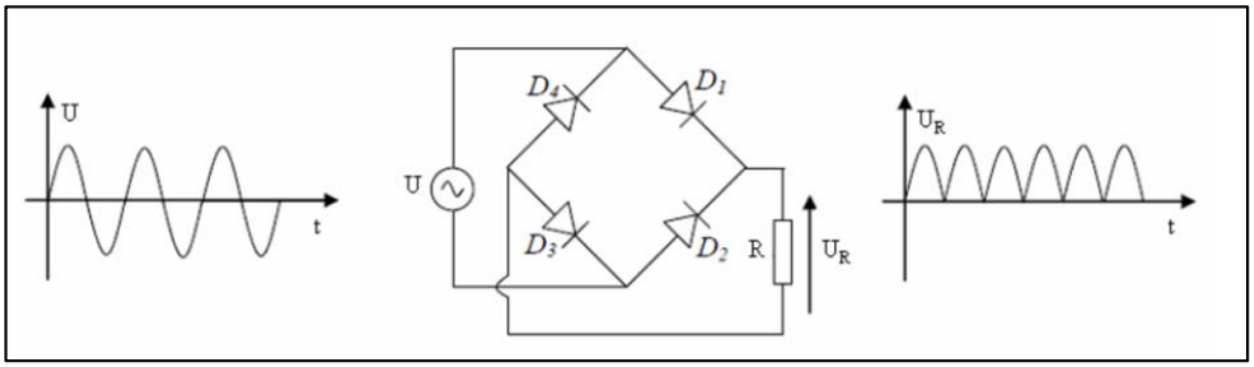
One of the most commonly employed circuits for full-wave rectification is the four-diode bridge configuration, widely known as the Graetz bridge.

The Graetz bridge consists of four diodes arranged in a bridge topology. Diodes D_1 and D_2 share a common cathode connection, while diodes D_3 and D_4 share a common anode connection.

The bridge is supplied by a sinusoidal alternating voltage defined by:

$$U = U_{\max} \sin(\omega t)$$

During the positive half-cycle, diodes D_1 and D_3 are forward biased and conduct. During the negative half-cycle, conduction shifts to diodes D_2 and D_4. As a consequence, the rectified current is present during both half-cycles of the input signal.



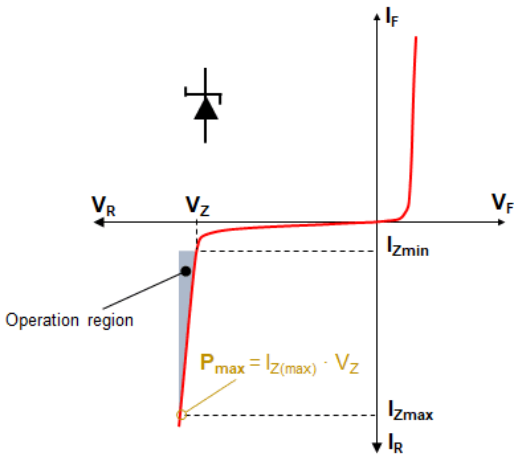
5. Zener Diode

A Zener diode is a PN junction device specifically designed to operate under reverse-bias conditions. When a PN junction is reverse biased, the internal electric field across the depletion region increases in magnitude. Beyond a certain critical value, this electric field becomes sufficiently strong to break covalent bonds within the crystal lattice, thereby generating electron–hole pairs. This mechanism is known as the **Zener effect**.

As the electric field intensity continues to rise, the generated charge carriers (electrons and holes) are rapidly accelerated. Through collisions with lattice atoms, these carriers can release additional charge carriers, leading to a chain reaction process referred to as the **avalanche effect**.

The fundamental consequence of both mechanisms is a sudden and significant increase in the reverse current once the breakdown voltage is reached. To prevent permanent damage to the PN junction and to make practical use of the Zener

breakdown phenomenon, the reverse current must be carefully limited and maintained within specified bounds.



Zener Diode Operation in Reverse Bias

The Zener diode is specifically intended to operate under reverse-bias conditions. The lower current limit, denoted $I_{Z\min}$, defines the minimum reverse current required for the diode to maintain the nominal Zener voltage V_Z across its terminals. The upper limit, $I_{Z\max}$, specifies the maximum permissible reverse current that must not be exceeded to avoid device damage.

From the characteristic curve (Figure III.21), it can be observed that under forward bias the Zener diode behaves like a conventional rectifier diode. In contrast, when reverse biased and when the current remains within the interval

$$I_{Z\min} < I_Z < I_{Z\max},$$

the voltage between the cathode and anode remains approximately constant and equal to V_Z. This voltage stabilization property constitutes the principal function of the Zener diode.

As previously discussed, two physical mechanisms are associated with reverse breakdown in diodes: the Zener effect, which appears first at lower breakdown voltages, and the avalanche effect, which dominates at higher voltages. Accordingly, Zener diodes are generally classified into two major categories:

- **Low-voltage Zener diodes** ($V_Z < 6\ \mathrm{V}$)
  In this range, the Zener effect is predominant. An increase in temperature facilitates the breaking of covalent bonds, thereby reducing the breakdown voltage. Consequently, such diodes exhibit a **negative temperature coefficient**.
- **High-voltage Zener diodes** ($V_Z > 10\ \mathrm{V}$)
  In this case, the avalanche mechanism is dominant. A rise in temperature intensifies lattice vibrations, which reduces carrier mobility and leads to an increase in the breakdown voltage. These diodes therefore possess a **positive temperature coefficient**.

When temperature sensitivity is a critical concern—particularly for relatively high Zener voltages—it is preferable to connect two Zener diodes in series: one with a positive temperature coefficient and the other with a negative coefficient. In this way, the opposing temperature effects compensate each other, improving voltage stability.

6.1 Principle of Voltage Regulation Using a Zener Diode

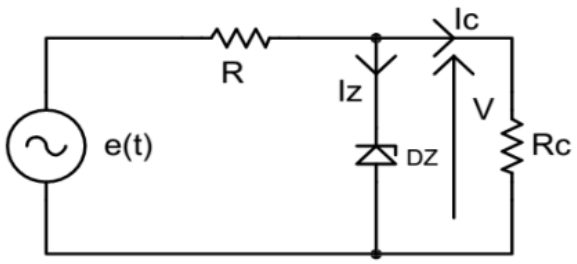
In electrical circuits, voltage regulation refers to the ability to maintain an output voltage as constant as possible despite variations in other circuit parameters.

From the characteristic shown in Figure III.21, it is clear that as long as the current through the Zener diode satisfies

$$I_{Z\min} < I_Z < I_{Z\max},$$

the diode can serve as a voltage regulator. However, for loads requiring high current levels, the Zener diode alone becomes an inefficient regulator due to its limited current-handling capability.

Consider the circuit illustrated in Figure III.22, where the Zener diode D_Z is employed to regulate the voltage across the load resistance R_C.



The load R_C is supplied by the voltage V, which is equal to the voltage across the Zener diode. Provided that

$$I_{Z\min} < I_Z < I_{Z\max},$$

the output voltage remains constant and equal to V_Z.

For proper regulation, the following conditions must be satisfied:

1. The input voltage $e(t)$ must always exceed the Zener voltage:

$$e(t) > V_Z \quad \text{for all } t.$$

2. The Zener current must remain above the minimum specified value:

$$I_Z > I_{Z\min}.$$

3. The maximum allowable current must not be exceeded:

$$I_Z < I_{Z\max}.$$

Two operating situations are typically examined:

Case 1: Variation of the Input Voltage $e(t)$ with Fixed Load R_C

Assume that the input voltage fluctuates between two limits, $E_{\min}$ and $E_{\max}$, while the load resistance remains constant.

For regulation to hold ($V = V_Z$), the series resistance R must be selected so that the Zener current remains within its permissible range.

From the circuit relations:

$$I_Z = I - I_C$$

where

$$I_C = \frac{V_Z}{R_C}$$

and

$$I = \frac{e(t) - V_Z}{R}.$$

Thus,

$$I_Z = \frac{e(t) - V_Z}{R} - \frac{V_Z}{R_C}.$$

- **Lower limit condition:**
  The most critical situation occurs when $e(t) = E_{\min}$. To ensure

  $$I_Z > I_{Z\min},$$

  the resistance R must satisfy an upper bound determined by this inequality.

- **Upper limit condition:**
  The most severe condition for excessive current occurs when $e(t) = E_{\max}$. To ensure

$$I_Z < I_{Z\max},$$

  the value of R must also satisfy the corresponding inequality.

The appropriate value of R is therefore chosen so that both conditions are simultaneously fulfilled.

Case 2: Variation of the Load Resistance R_C with Constant Input Voltage $e(t) = E$

In this scenario, the input voltage is fixed, while the load resistance varies between $R_{C\min}$ and $R_{C\max}$.

When regulation is achieved, the total current

$$I = \frac{E - V_Z}{R}$$

remains constant. However, the load current

$$I_C = \frac{V_Z}{R_C}$$

changes as R_C varies.

- If R_C decreases, the load current I_C increases, causing the Zener current I_Z to decrease.
- If R_C increases, I_C decreases and I_Z correspondingly increases.

Therefore, the series resistance R must be determined so that, under all load conditions, the Zener current always remains within the allowable interval:

$$I_{Z\min} < I_Z < I_{Z\max}.$$

This ensures proper voltage regulation without exceeding the diode's rated limits.

$$\frac{E - V_Z}{R} - \frac{V_Z}{R_{cmin}} > I_{zmin} \Rightarrow R < \frac{E - V_Z}{V_Z + R_{cmin} I_{zmin}} R_{cmin}$$

$$\frac{E - V_Z}{R} - \frac{V_Z}{R_{cmax}} < I_{zmax} \Rightarrow R > \frac{E - V_Z}{V_Z + R_{cmax} I_{zmax}} R_{cmax}$$