9.6: Band Theory of Solids

Learning Objectives

By the end of this section, you will be able to:

- Describe two main approaches to determining the energy levels of an electron in a crystal
- Explain the presence of energy bands and gaps in the energy structure of a crystal
- Explain why some materials are good conductors and others are good insulators
- Differentiate between an insulator and a semiconductor

The free electron model explains many important properties of conductors but is weak in at least two areas. First, it assumes a constant potential energy within the solid. (Recall that a constant potential energy is associated with no forces.) Figure $9.6.1$ compares the assumption of a constant potential energy (dotted line) with the periodic Coulomb potential, which drops as $-1/r$ at each lattice point, where **r** is the distance from the ion core (solid line). Second, the free electron model assumes an impenetrable barrier at the surface. This assumption is not valid, because under certain conditions, electrons can escape the surface—such as in the photoelectric effect. In addition to these assumptions, the free electron model does not explain the dramatic differences in electronic properties of conductors, semiconductors, and insulators. Therefore, a more complete model is needed.

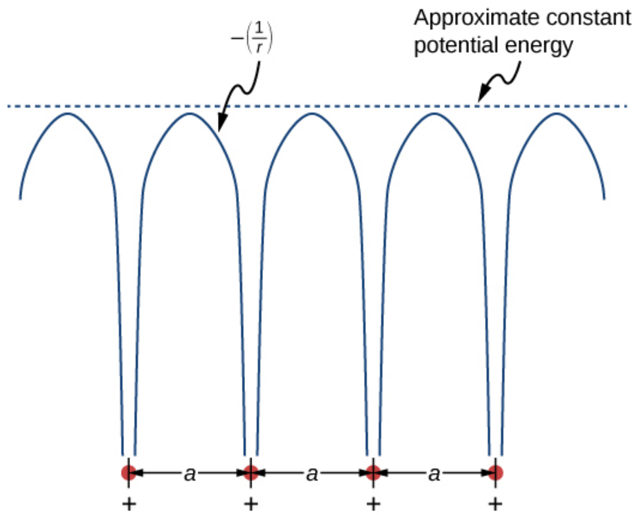

Figure $9.6.1$: The periodic potential used to model electrons in a conductor. Each ion in the solid is the source of a Coulomb potential. Notice that the free electron model is productive because the average of this field is approximately constant.

We can produce an improved model by solving Schrödinger's equation for the periodic potential shown in Figure $9.6.1$. However, the solution requires technical mathematics far beyond our scope. We again seek a qualitative argument based on quantum mechanics to find a way forward.

We first review the argument used to explain the energy structure of a covalent bond. Consider two identical hydrogen atoms so far apart that there is no interaction whatsoever between them. Further suppose that the electron in each atom is in the same ground state: a 1**s** electron with an energy of $-13.6\,eV$ (ignore spin). When the hydrogen atoms are brought closer together, the individual wave functions of the electrons overlap and, by the exclusion principle, can no longer be in the same quantum state, which splits the original equivalent energy levels into two different energy levels. The energies of these levels depend on the interatomic distance, a (Figure $9.6.2a$).

If four hydrogen atoms are brought together, four levels are formed from the four possible symmetries—a single sine wave "hump" in each well, alternating up and down, and so on. In the limit of a very large number **N** of atoms, we expect a spread of nearly continuous bands of electronic energy levels in a solid (Figure $9.6.2c$). Each of these bands is known as an **energy band**. (The allowed states of energy and wave number are still technically quantized, but for large numbers of atoms, these states are so close together that they are consider to be continuous or "in the continuum.")

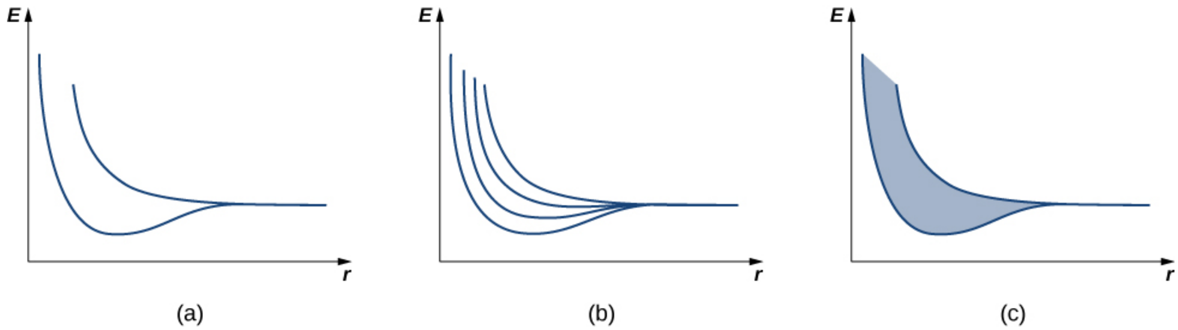

Figure 9.6.2: The dependence of energy-level splitting on the average distance between (a) two atoms, (b) four atoms, and (c) a large number of atoms. For a large number of electrons, a continuous band of energies is produced

Energy bands differ in the number of electrons they hold. In the 1**s** and 2**s** energy bands, each energy level holds up to two electrons (spin up and spin down), so this band has a maximum occupancy of 2**N** electrons. In the 2**p** energy band, each energy level holds up to six electrons, so this band has a maximum occupancy of 6**N** electrons (Figure 9.6.3).

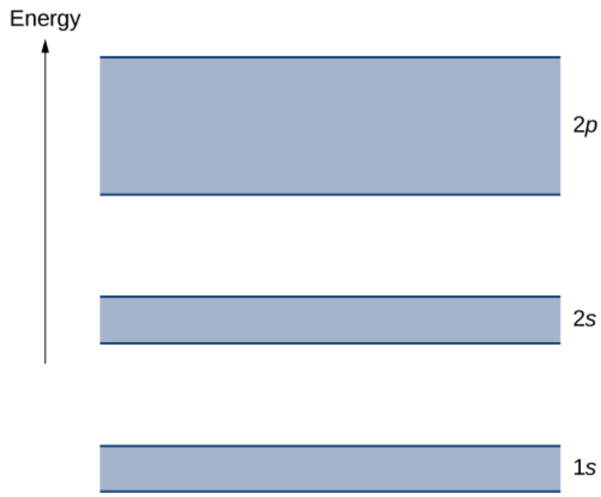

Figure 9.6.3: A simple representation of the energy structure of a solid. Electrons belong to energy bands separated by energy gaps.

Each energy band is separated from the other by an **energy gap**. The electrical properties of conductors and insulators can be understood in terms of energy bands and gaps. The highest energy band that is filled is known as a **valence band**. The next available band in the energy structure is known as a **conduction band**. In a conductor, the highest energy band that contains electrons is partially filled, whereas in an insulator, the highest energy band containing electrons is completely filled. The difference between a conductor and insulator is illustrated in Figure 9.6.4.

A conductor differs from an insulator in how its electrons respond to an applied electric field. If a significant number of electrons are set into motion by the field, the material is a conductor. In terms of the band model, electrons in the partially filled conduction band gain kinetic energy from the electric field by filling higher energy states in the conduction band. By contrast, in an insulator, electrons belong to completely filled bands. When the field is applied, the electrons cannot make such transitions (acquire kinetic energy from the electric field) due to the exclusion principle. As a result, the material does not conduct electricity.

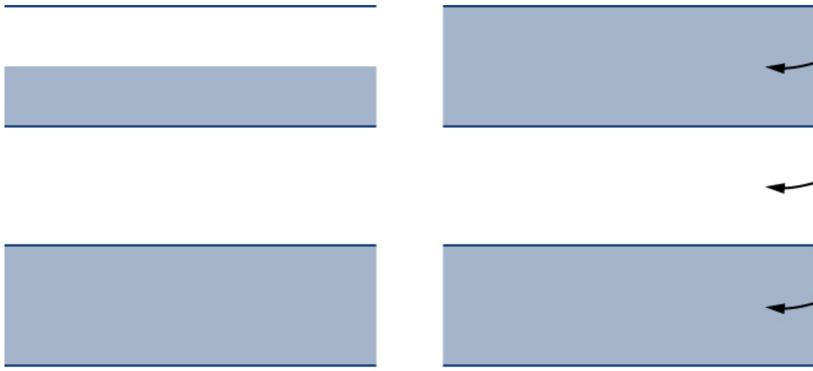

Figure 9.6.4: Comparison of a conductor and insulator. The highest energy band is partially filled in a conductor but completely filled in an insulator.

Simulation

Visit this simulation to learn about the origin of energy bands in crystals of atoms and how the structure of bands determines how a material conducts electricity. Explore how band structure creates a lattice of many wells.

A semiconductor has a similar energy structure to an insulator except it has a relatively small energy gap between the lowest completely filled band and the next available unfilled band. This type of material forms the basis of modern electronics. At $T = 0\,K$, the semiconductor and insulator both have completely filled bands. The only difference is in the size of the energy gap (or **band gap**) $\mathbf{E_g}$ between the highest energy band that is filled (the valence band) and the next-higher empty band (the conduction band). In a semiconductor, this gap is small enough that a substantial number of electrons from the valence band are thermally excited into the conduction band at room temperature. These electrons are then in a nearly empty band and can respond to an applied field. As a general rule of thumb, the band gap of a semiconductor is about 1 eV. (Table 9.6.1 for silicon.) A band gap of greater than approximately 1 eV is considered an insulator. For comparison, the energy gap of diamond (an insulator) is several electron-volts.

Table 9.6.1: Energy Gap for Various Materials at 300 KNote: Except for diamond, the materials listed are all semiconductors.

Material	Energy Gap $E_g(eV)$
Si	1.14
Ge	0.67
GaAs	1.43
GaP	2.26
GaSb	0.69
InAs	0.35
InP	1.35
InSb	0.16
C(diamond)	5.48

This page titled 9.6: Band Theory of Solids is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform; a detailed edit history is available upon request.

9.7: Semiconductors and Doping

Learning Objectives

By the end of this section, you will be able to:

- Describe changes to the energy structure of a semiconductor due to doping
- Distinguish between an n-type and p-type semiconductor
- Describe the Hall effect and explain its significance
- Calculate the charge, drift velocity, and charge carrier number density of a semiconductor using information from a Hall effect experiment

In the preceding section, we considered only the contribution to the electric current due to electrons occupying states in the conduction band. However, moving an electron from the valence band to the conduction band leaves an unoccupied state or **hole** in the energy structure of the valence band, which a nearby electron can move into. As these holes are filled by other electrons, new holes are created. The electric current associated with this filling can be viewed as the collective motion of many negatively charged electrons or the motion of the positively charged electron holes.

To illustrate, consider the one-dimensional lattice in Figure $9.7.1$. Assume that each lattice atom contributes one valence electron to the current. As the hole on the right is filled, this hole moves to the left. The current can be interpreted as the flow of positive charge to the left. The density of holes, or the number of holes per unit volume, is represented by **p**. Each electron that transitions into the conduction band leaves behind a hole. If the conduction band is originally empty, the conduction electron density **p** is equal to the hole density, that is, $n = p$.

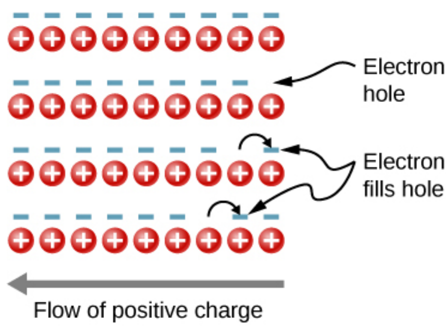

Figure $9.7.1$: The motion of holes in a crystal lattice. As electrons shift to the right, an electron hole moves to the left.

As mentioned, a semiconductor is a material with a filled valence band, an unfilled conduction band, and a relatively small energy gap between the bands. Excess electrons or holes can be introduced into the material by the substitution into the crystal lattice of an impurity atom, which is an atom of a slightly different valence number. This process is known as doping. For example, suppose we add an arsenic atom to a crystal of silicon (Figure $9.7.2a$).

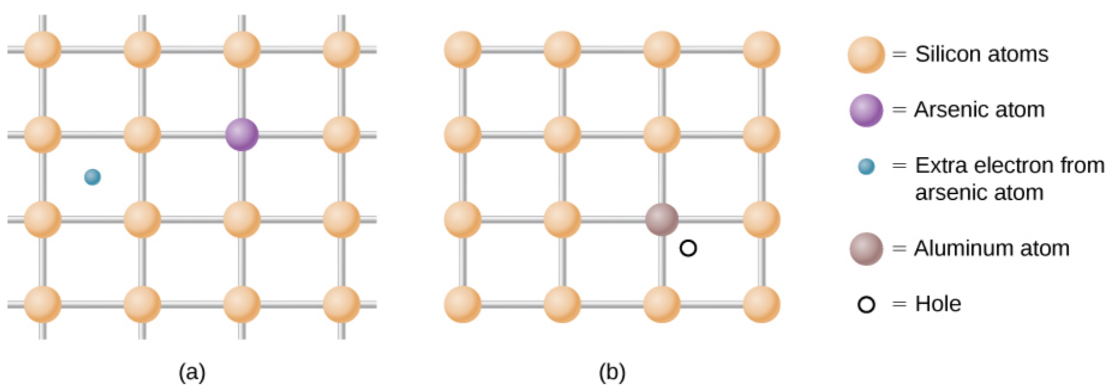

Figure $9.7.2$: (a) A donor impurity and (b) an acceptor impurity. The introduction to impurities and acceptors into a semiconductor significantly changes the electronic properties of this material.

Arsenic has five valence electrons, whereas silicon has only four. This extra electron must therefore go into the conduction band, since there is no room in the valence band. The arsenic ion left behind has a net positive charge that weakly binds the delocalized electron. The binding is weak because the surrounding atomic lattice shields the ion's electric field. As a result, the binding energy of the extra electron is only about 0.02 eV. In other words, the energy level of the impurity electron is in the band gap below the conduction band by 0.02 eV, a much smaller value than the energy of the gap, 1.14 eV. At room temperature, this impurity electron is easily excited into the conduction band and therefore contributes to the conductivity (Figure $9.7.3a$). An impurity with an extra electron is known as a **donor impurity**, and the doped semiconductor is called an **n-type semiconductor** because the primary carriers of charge (electrons) are negative.

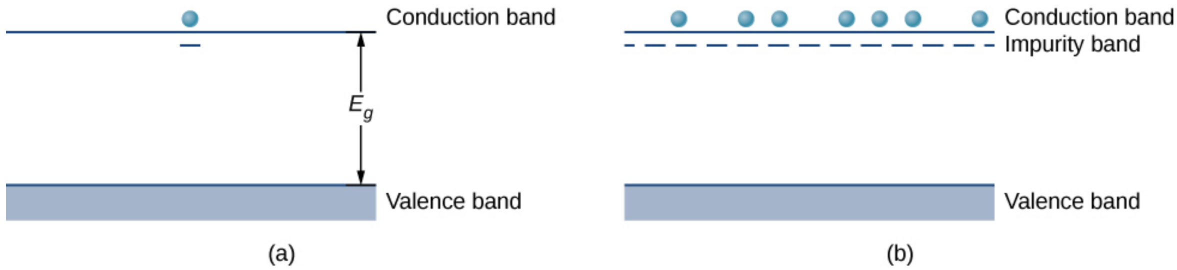

Figure $9.7.3$: The extra electron from a donor impurity is excited into the conduction band; (b) formation of an impurity band in an n-type semiconductor.

By adding more donor impurities, we can create an **impurity band**, a new energy band created by semiconductor doping, as shown in Figure $9.7.3b$. The Fermi level is now between this band and the conduction band. At room temperature, many impurity electrons are thermally excited into the conduction band and contribute to the conductivity. Conduction can then also occur in the impurity band as vacancies are created there. Note that changes in the energy of an electron correspond to a change in the motion (velocities or kinetic energy) of these charge carriers with the semiconductor, but not the bulk motion of the semiconductor itself.

Doping can also be accomplished using impurity atoms that typically have one **fewer** valence electron than the semiconductor atoms. For example, Al, which has three valence electrons, can be substituted for Si, as shown in Figure $9.7.2b$. Such an impurity is known as an **acceptor impurity**, and the doped semiconductor is called a **p-type semiconductor**, because the primary carriers of charge (holes) are positive. If a hole is treated as a positive particle weakly bound to the impurity site, then an empty electron state is created in the band gap just above the valence band. When this state is filled by an electron thermally excited from the valence band (Figure $9.7.1a$), a mobile hole is created in the valence band. By adding more acceptor impurities, we can create an impurity band, as shown in Figure $9.7.1b$.

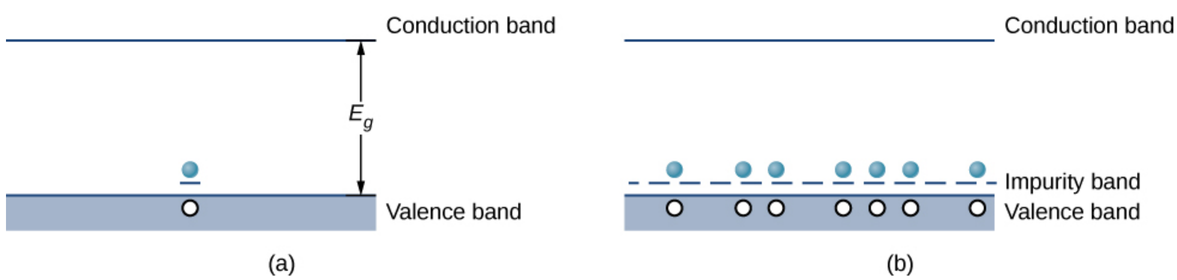

Figure $9.7.4$: (a) An electron from the conduction band is excited into the empty state resulting from the acceptor impurity; (b) formation of an impurity band in a p-type semiconductor.

The electric current of a doped semiconductor can be due to the motion of a **majority carrier**, in which holes are contributed by an impurity atom, or due to a **minority carrier**, in which holes are contributed purely by thermal excitations of electrons across the energy gap. In an **n**-type semiconductor, majority carriers are free electrons contributed by impurity atoms, and minority carriers are free electrons produced by thermal excitations from the valence to the conduction band. In a **p**-type semiconductor, the majority carriers are free holes contributed by impurity atoms, and minority carriers are free holes left by the filling of states due to thermal excitation of electrons across the gap. In general, the number of majority carriers far exceeds the minority carriers. The concept of a majority and minority carriers will be used in the next section to explain the operation of diodes and transistors.

Hall Effect

In studying **p**- and **n**-type doping, it is natural to ask: Do "electron holes" really act like particles? The existence of holes in a doped **p**-type semiconductor is demonstrated by the Hall effect. The Hall effect is the production of a potential difference due to the motion of a conductor through an external magnetic field. A schematic of the Hall effect is shown in Figure $9.7.5a$.

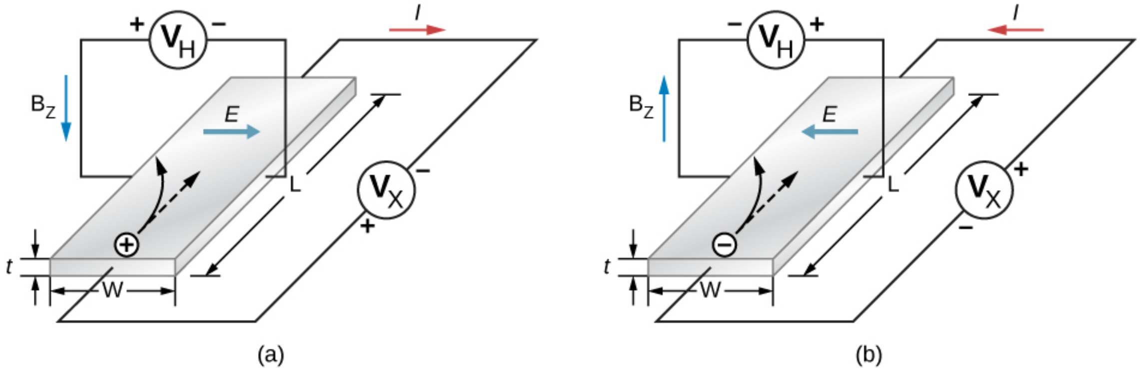

Figure 9.7.5: The Hall effect. (a) Positively charged electron holes are drawn to the left by a uniform magnetic field that points downward. An electric field is generated to the right. (b) Negative charged electrons are drawn to the left by a magnetic field that points up. An electric field is generated to the left.

A semiconductor strip is bathed in a uniform magnetic field (which points into the paper). As the electron holes move from left to right through the semiconductor, a **Lorentz force** drives these charges toward the upper end of the strip. (Recall that the motion of the positively charged carriers is determined by the right-hand rule.) Positive charge continues to collect on the upper edge of the strip until the force associated with the downward electric field between the upper and lower edges of the strip ($F_E = E_q$) just balances the upward magnetic force ($F_B = qvB$). Setting these forces equal to each other, we have $E = vB$. The voltage that develops across the strip is therefore

$$V_H = vBw,$$

where V_H is the Hall voltage; v is the hole's **drift velocity**, or average velocity of a particle that moves in a partially random fashion; **B** is the magnetic field strength; and **w** is the width of the strip. Note that the Hall voltage is transverse to the voltage that initially produces current through the material. A measurement of the sign of this voltage (or potential difference) confirms the collection of holes on the top side of the strip. The magnitude of the Hall voltage yields the drift velocity (**v**) of the majority carriers.

Additional information can also be extracted from the Hall voltage. Note that the electron current density (the amount of current per unit cross-sectional area of the semiconductor strip) is

$$j = nqv, \tag{9.7.1}$$

where **q** is the magnitude of the charge, **n** is the number of charge carriers per unit volume, and **v** is the drift velocity. The current density is easily determined by dividing the total current by the cross-sectional area of the strip, **q** is charge of the hole (the magnitude of the charge of a single electron), and **u** is determined by Equation 9.7.1. Hence, the above expression for the electron current density gives the number of charge carriers per unit volume, **n**. A similar analysis can be conducted for negatively charged carriers in an **n**-type material (see Figure 9.7.5).

This page titled 9.7: Semiconductors and Doping is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform; a detailed edit history is available upon request.

9.8: Semiconductor Devices

Learning Objectives

By the end of this section, you will be able to:

- Describe what occurs when n- and p-type materials are joined together using the concept of diffusion and drift current (zero applied voltage)
- Explain the response of a p-n junction to a forward and reverse bias voltage
- Describe the function of a transistor in an electric circuit
- Use the concept of a p-n junction to explain its applications in audio amplifiers and computers

Semiconductors have many applications in modern electronics. We describe some basic semiconductor devices in this section. A great advantage of using semiconductors for circuit elements is the fact that many thousands or millions of semiconductor devices can be combined on the same tiny piece of silicon and connected by conducting paths. The resulting structure is called an integrated circuit (ic), and ic chips are the basis of many modern devices, from computers and smartphones to the internet and global communications networks.

Diodes

Perhaps the simplest device that can be created with a semiconductor is a diode. A diode is a circuit element that allows electric current to flow in only one direction, like a one-way valve (see Model of Conduction in Metals). A diode is created by joining a **p**-type semiconductor to an **n**-type semiconductor (Figure $9.8.1$). The junction between these materials is called a **p-n junction**. A comparison of the energy bands of a silicon-based diode is shown in Figure $9.8.1b$. The positions of the valence and conduction bands are the same, but the impurity levels are quite different. When a **p-n** junction is formed, electrons from the conduction band of the **n**-type material diffuse to the **p**-side, where they combine with holes in the valence band. This migration of charge leaves positive ionized donor ions on the **n**-side and negative ionized acceptor ions on the **p**-side, producing a narrow double layer of charge at the **p-n** junction called the **depletion layer**. The electric field associated with the depletion layer prevents further diffusion. The potential energy for electrons across the **p-n** junction is given by Figure $9.8.2$.

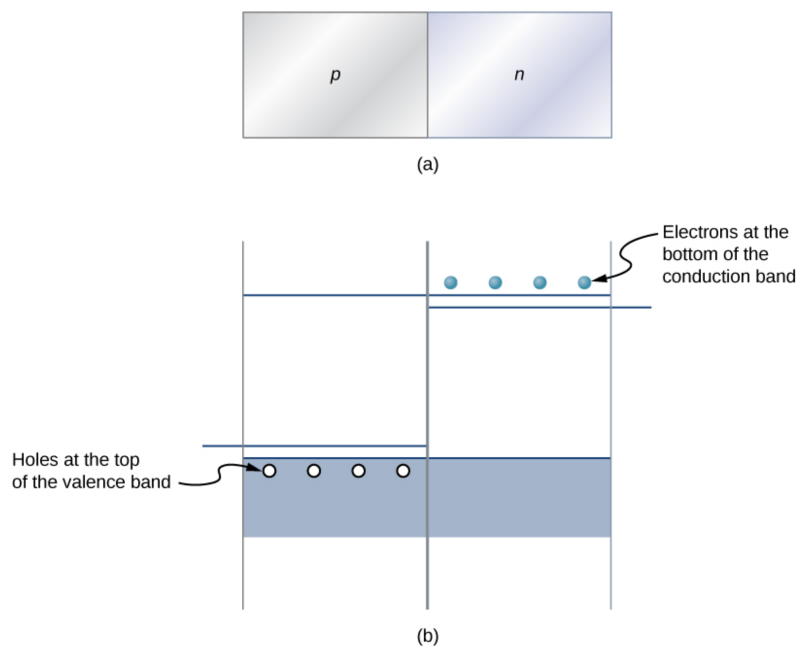

Figure $9.8.1$: (a) Representation of a **p-n** junction. (b) A comparison of the energy bands of **p**-type and **n**-type silicon prior to equilibrium.

The behavior of a semiconductor diode can now be understood. If the positive side of the battery is connected to the **n**-type material, the depletion layer is widened, and the potential energy difference across the **p-n** junction is increased. Few or none of the electrons (holes) have enough energy to climb the potential barrier, and current is significantly reduced. This is called the **reverse** bias configuration**. On the other hand, if the positive side of a battery is connected to the **p**-type material, the depletion layer is narrowed, the potential energy difference across the **p-n** junction is reduced, and electrons (holes) flow easily. This is called the **forward bias configuration** of the diode. In sum, the diode allows current to flow freely in one direction but prevents current flow in the opposite direction. In this sense, the semiconductor diode is a one-way valve.

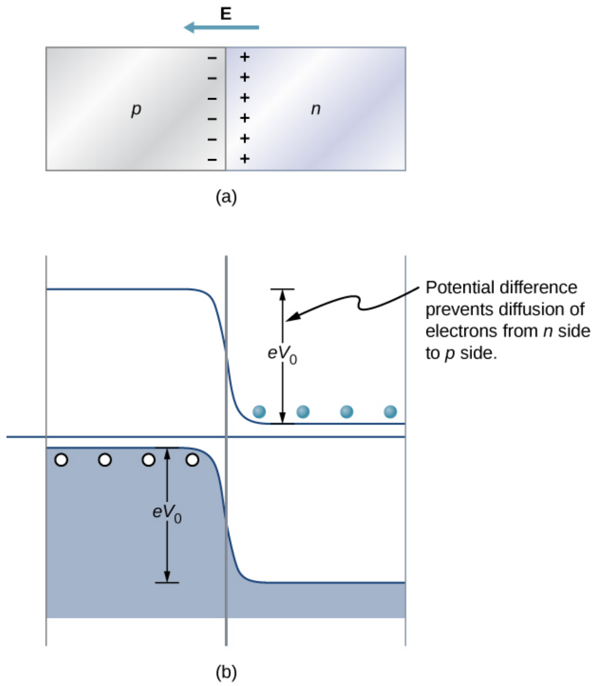

Figure 9.8.2: At equilibrium, (a) excess charge resides near the interface and the net current is zero, and (b) the potential energy difference for electrons (in light blue) prevents further diffusion of electrons into the **p**-side.

We can estimate the mathematical relationship between the current and voltage for a diode using the electric potential concept. Consider **N** negatively charged majority carriers (electrons donated by impurity atoms) in the **n**-type material and a potential barrier **V** across the **p-n** junction. According to the Maxwell-Boltzmann distribution, the fraction of electrons that have enough energy to diffuse across the potential barrier is Ne^{-eV/k_BT}. However, if a battery of voltage V_b is applied in the forward-bias configuration, this fraction improves to $Ne^{-e(V-V_b)/k_BT}$. The electric current due to the majority carriers from the **n**-side to the **p**-side is therefore

$$I = Ne^{-eV/k_BT} e^{eV_b/k_BT} = I_0 e^{eV_b/k_BT},$$

where I_0 is the current with no applied voltage and **T** is the temperature. Current due to the minority carriers (thermal excitation of electrons from the valence band to the conduction band on the **p**-side and subsequent attraction to the **n**-side) is $-I_0$, independent of the bias voltage. The net current is therefore

$$I_{net} = I_0 \left(e^{eV_b/k_BT} - 1 \right).$$

A sample graph of the current versus bias voltage is given in Figure 9.8.3. In the forward bias configuration, small changes in the bias voltage lead to large changes in the current. In the reverse bias configuration, the current is $I_{net} \approx -I_0$. For extreme values of reverse bias, the atoms in the material are ionized which triggers an avalanche of current. This case occurs at the **breakdown voltage**.

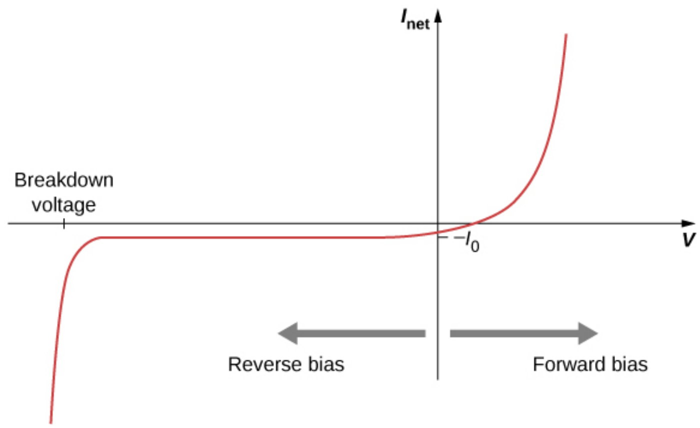

Figure 9.8.3: Current versus voltage across a **p-n** junction (diode). In the forward bias configuration, electric current flows easily. However, in the reverse bias configuration, electric current flow very little.

Example 9.8.1: Diode Current

Attaching the positive end of a battery to the **p**-side and the negative end to the **n**-side of a semiconductor diode produces a current of $4.5 \times 10^{-1} A$. The reverse saturation current is $2.2 \times 10^{-8} A$. (The reverse saturation current is the current of a diode in a reverse bias configuration such as this.) The battery voltage is 0.12 V. What is the diode temperature?

Strategy

The first arrangement is a forward bias configuration, and the second is the reverse bias configuration.

Solution

The current in the forward and reverse bias configurations is given by

$$I_{net} = I_0 \left(e^{eV_b/k_BT} - 1\right).$$

The current with no bias is related to the reverse saturation current by

$$I_0 \approx -I_{sat} = 2.2 \times 10^{-8}.$$

Therefore

$$\frac{I_{net}}{I_0} = \frac{4.5 \times 10^{-1} A}{2.2 \times 10^{-8} A} = 2.0 \times 10^8.$$

this can be written as

$$\frac{I_{net}}{I_0} + 1 = e^{eV_b/k_BT}.$$

This ratio is much greater than one, so the second term on the left-hand side of the equation vanishes. Taking the natural log of both sides gives

$$\frac{eV_b}{k_BT} = 19.$$

The temperature is therefore

$$T = \frac{eV_b}{k_B}\left(\frac{1}{19}\right) = \frac{e(0.12\,V)}{8.617 \times 10^{-5} eV/K}\left(\frac{1}{19}\right) = 73\,K.$$

Significance

The current moving through a diode in the forward and reverse bias configuration is sensitive to the temperature of the diode. If the potential energy supplied by the battery is large compared to the thermal energy of the diode's surroundings, k_BT, then the forward bias current is very large compared to the reverse saturation current.

? Exercise 9.8.1

How does the magnitude of the forward bias current compare with the reverse bias current?

Solution

The forward bias current is much larger. To a good approximation, diodes permit current flow in only one direction.

Create a **p-n** junction and observe the behavior of a simple circuit for forward and reverse bias voltages. Visit this site to learn more about semiconductor diodes.

Junction Transistor

If diodes are one-way valves, transistors are one-way valves that can be carefully opened and closed to control current. A special kind of transistor is a junction transistor. A **junction transistor** has three parts, including an **n**-type semiconductor, also called the emitter; a thin **p**-type semiconductor, which is the base; and another **n**-type semiconductor, called the collector (Figure 9.8.4). When a positive terminal is connected to the **p**-type layer (the base), a small current of electrons, called the **base current** I_B, flows to the terminal. This causes a large **collector current** I_C to flow through the collector. The base current can be adjusted to control the large collector current. The current gain is therefore

$$I_c = \beta I_B.$$

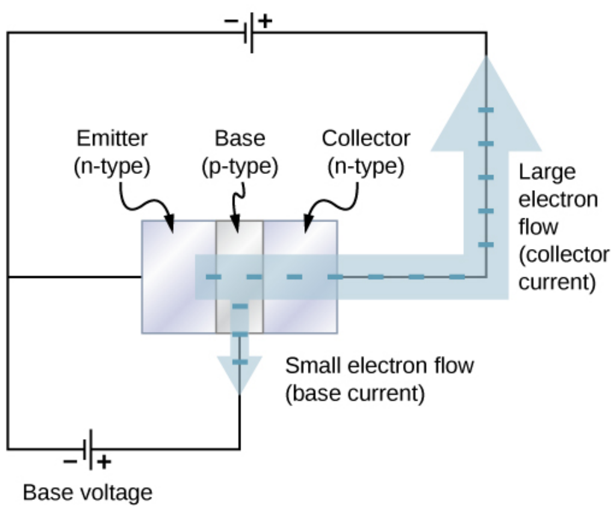

Figure 9.8.4: A junction transistor has three parts: emitter, base, and collector. Voltage applied to the base acts as a valve to control electric current from the emitter to the collector.

A junction transistor can be used to amplify the voltage from a microphone to drive a loudspeaker. In this application, sound waves cause a diaphragm inside the microphone to move in and out rapidly (Figure 9.8.5). When the diaphragm is in the “in” position, a tiny positive voltage is applied to the base of the transistor. This opens the transistor “valve” and allows a large electrical current flow to the loudspeaker. When the diaphragm is in the “out” position, a tiny negative voltage is applied to the base of the transistor, which shuts off the transistor valve so that no current flows to the loudspeaker. This shuts the transistor “valve” off so no current flows to the loudspeaker. In this way, current to the speaker is controlled by the sound waves, and the sound is amplified. Any electric device that amplifies a signal is called an **amplifier**.

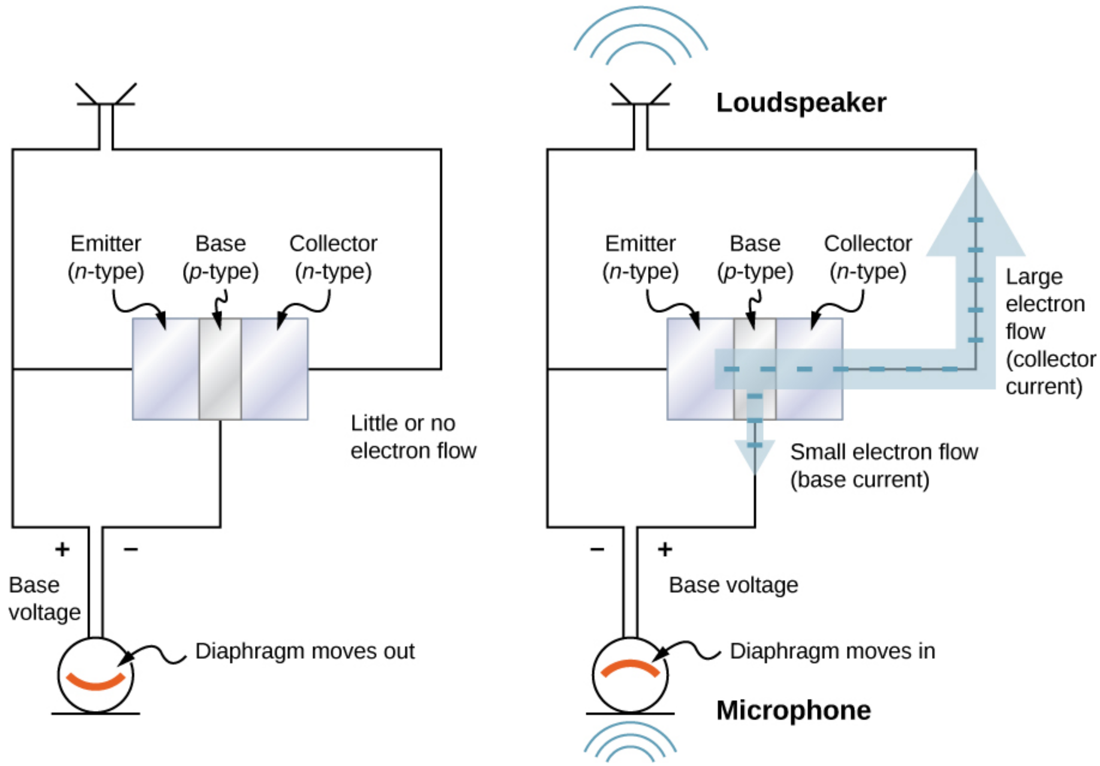

Figure 9.8.5: An audio amplifier based on a junction transistor. Voltage applied to the base by a microphone acts as a valve to control a larger electric current that passes through a loudspeaker.

In modern electronic devices, digital signals are used with diodes and transistors to perform tasks such as data manipulation. Electric circuits carry two types of electrical signals: analog and digital (Figure 9.8.6). An analog signal varies continuously, whereas a digital signal switches between two fixed voltage values, such as plus 1 volt and zero volts. In digital circuits like those found in computers, a transistor behaves like an on-off switch. The transistor is either on, meaning the valve is completely open, or it is off, meaning the valve is completely closed. Integrated circuits contain vast collections of transistors on a single piece of silicon. They are designed to handle digital signals that represent ones and zeroes, which is also known as binary code. The invention of the ic helped to launch the modern computer revolution.

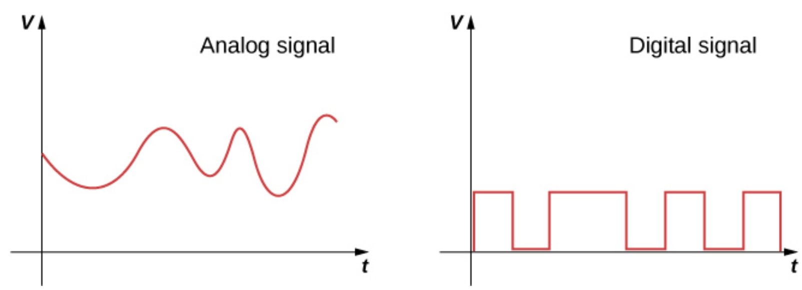

Figure 9.8.6: Real-world data are often analog, meaning data can vary continuously. Intensity values of sound or visual images are usually analog. These data are converted into digital signals for electronic processing in recording devices or computers. The digital signal is generated from the analog signal by requiring certain voltage cut-off value.

This page titled 9.8: Semiconductor Devices is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform; a detailed edit history is available upon request.